tmt30

expertise in audio media

30. Tonmeistertagung
VDT International Convention

14.11.–17.11.2018 Köln | Cologne

Tagungsbericht
Conference Proceedings

ISBN 978-3-9812830-9-9

tmt30

expertise in audio media

30. Tonmeistertagung
VDT International Convention

14.11.–17.11.2018 Köln | Cologne

Hinweis:
Der Tagungsbericht enthält die von den Referenten eingereichten schriftlichen Arbeiten. Weitere Informationen sind auf der Website https://tonmeistertagung.com verfügbar.

Note:
The Proceedings contain written Papers submitted by the presenters. More information is available at our website https://tonmeistertagung.com.

ISBN 978-3-9812830-9-9

tmt30

expertise in audio media

30. Tonmeistertagung
VDT International Convention

Verband Deutscher Tonmeister e.V.
Am Zaarshäuschen 9 · DE-51427 Bergisch Gladbach
Telefon: +49 2204 2 35 95 · Telefax: +49 22 04 20 15 43
E-Mail: vdt@tonmeister.de

Für die in dieser Publikation enthaltenen Angaben wird keine Gewähr hinsichtlich der Freiheit von gewerblichen Schutzrechten (Patente, Gebrauchsmuster, Warenzeichen) übernommen. Auch die wiedergegebenen Gebrauchsnamen, Handelsnamen und Warenbezeichnungen dürfen nicht als frei zur allgemeinen Benutzung im Sinne der Warenzeichen- und Markenschutzgesetzgebung betrachtet werden. Die Verletzung dieser Rechte ist im Rahmen der geltenden Gesetze strafbar.

No guarantee is made that the information contained in this publication does not infringe industrial property rights (patents, utility models, trademarks).The utility names, trade names and product designations may not be deemed for general usage in the meaning of trade name and trademark protection laws. An infringement of these rights is liable for prosecution under applicable law.

ISBN 978-3-9812830-9-9

tmt30

expertise in audio media

Themenübersicht | Topic Overview

Vorspann | Intro 2

Beiträge | Papers

3D	3D Audio	5
3DP	3D Production	18
AC	Acoustics	26
AP	The Arts of Production	59
LN	Loudness in Radio	66
MIC	Microphone Setups	73
MU	Musical Aspects in Production	80
NA	Networked Audio & Video	83
OB	Object Based Production	88
PE	Perception & Aesthetics	111
PP	Poster Presentation	136
RA	Restoration & Archiving	166
RD	Research & Development	179
SI	Speech Intelligibility in Broadcast and Cinema	212
SP	Spatial Audio	229
SR	Sound Reinforcement	245
TA	TMT Academy	269
TB	Talk Back	279
VR	VR and 360°	283

Impressum 298

ISBN 978-3-9812830-9-9

Amplitude Panning and the Interior Pan

M. R. P. Thomas, C. Q. Robinson
[1] *Dolby Laboratories, San Francisco, USA, Email: {mrthom, cqr}@dolby.com*

Abstract

The perception of source location using multi-loudspeaker amplitude panning is considered. While there have been many studies on the localization of pairwise panned sound sources, relatively few studies investigate the multi-loudspeaker case. This paper evaluates panning scenarios in which a source is panned on the boundary or within the area bounded by discrete loudspeakers, referred to as boundary and interior pans respectively. Subjective testing of a variety of pan locations reveal the following: (1) as expected, pans to a single loudspeaker yield lowest localization error, (2) pairwise pans tend to be consistently localized closer to the listener than single loudspeaker pans, (3) largest errors occur when the virtual source is panned close to the listener, and (4) interior pans are accurately perceived and, surprisingly, in some cases more accurately than pairwise boundary pans.

1. Introduction

The panning of a virtual source has been of interest to the audio community since the invention of stereophonic reproduction [1], whose purpose is to give the listener the perception of continuous source location or movement when reproduced with a discrete set of loudspeakers. Amplitude panning is commonly used for producing a virtual source at an arbitrary location in between a pair of loudspeakers [1–6]. Several multi-loudspeaker amplitude panning extensions have appeared in the academic literature, including Vector Base Amplitude Panning (VBAP) [7], Multi-Direction Amplitude Panning (MDAP) [8], and Ambisonics [9, 10], that consider pans on a convex hull of loudspeakers surrounding the user, with a view to creating an immersive listening environment. While these extensions use loudspeakers arranged within a 3D space, the object's position has degrees of freedom in azimuth and elevation, but not range, and are therefore considered to be 2D panners.

The concept of interior panning is to render acoustic sound sources inside the listening area, as opposed to on its boundaries. Several approaches achieve this by controlling both the amplitude and phase of a loudspeaker channel gain [11], notably in the field of acoustic wavefield synthesis [10]. While such approaches can provide a compelling listening experience, they may suffer from small sweet spots, outside which the artefacts can become objectionable. Amplitude panning algorithms that are capable of interior panning include Distance-Based Amplitude Panning (DBAP) [12], Auditory Distance Rendering [13], and Optimal 3D Speaking Panning [14]. Dual-Balance Panning [15] has found extensive use in content creation tools for cinema and home theater systems such as Dolby Atmos. An overview of the perceived size, timbre, and location of a virtual source with various amplitude panning algorithms can be found in [16]. Investigations into the subjective preference of interior panning algorithms on the horizontal plane is given in [15].

Given the prevalence of interior panning tools in cinematography, it is remarkable that no known study has been undertaken to determine whether or not listeners can consistently perceive a static virtual source inside the listening volume, and if so to what degree. This paper describes a methodology for presenting a listener with a set of random static interior pans, capturing the perceived location with a graphical panner as well as informal comments on the timbre, spatial extent, and perceived height. Dual-Balance panning is used as an example case. The results reveal that listeners can consistently locate a interior virtual source, in some cases to at least the accuracy of a pairwise pan on the boundary. The perceived size and height are reported to increase as the source moves closer to the listener.

The remainder of this paper is organized as follows. Several amplitude panning algorithms are reviewed in Sec. 2. Experimental results are presented and discussed in Sec. 3 and Sec. 4. A summary is given in Sec. 5.

2. Amplitude Panning Algorithms

The following is a review of common amplitude panning algorithms for sources on the horizontal plane, progressing from pairwise pans, through boundary pans, and interior pans.

Let g_i be non-negative amplitude panning gains assigned to loudspeaker $i \in \{1, \ldots, M\}$ such that loudspeaker signals $y_i(t)$ are derived from a mono source signal $x(t)$ at time t by

$$y_i(t) = g_i x(t). \tag{1}$$

Typically, including in cinemas, the playback environment is a rectilinear geometry with several loudspeakers on the front and back wall and a somewhat regular spacing along the sidewalls. In these cases is convenient to use a Cartesian indexing system with gains $g_{j,k}$ associated with row $j \in \{1, \ldots, J\}$ and column $k \in \{1, \ldots K\}$, yielding corresponding loudspeaker signals $y_{j,k}(t)$. It is always assumed that amplitude gains are normalized to preserve unit norm:

$$\sum_{i=1}^{M} g_i^p = 1. \tag{2}$$

ISBN
978-3-9812830-9-9

The authors hold the copyrights of all of content in this manuscript. They have granted the VDT the right to publish it in print, online and as PDF download on https://tonmeistertagung.com

For all amplitude panning algorithms in this paper, with the exception of Ambisonics, typically $p = 2$.

2.1. Pairwise Panning

The pairwise pan [4, 17] sets $M = 2$, with loudspeakers placed equidistant from a listening position at azimuth angles $\phi_{\mathrm{s},1} = -\phi_{\mathrm{s}}$ and $\phi_{\mathrm{s},2} = \phi_{\mathrm{s}}$, $0 \leq \phi_{\mathrm{s}} \leq 90°$, so that the subtended aperture is $2\phi_{\mathrm{s}}$. The forward direction is $0°$. The angle of the virtual source, sometimes referred to as the predicted phantom source, is ϕ_{p}.

2.1.1. Trigonometric Laws

The law of sines [1–3] relates loudspeaker angle, virtual source angle and gain by

$$\sin\phi_{\mathrm{p}} = \sin\phi_{\mathrm{s}} \frac{g_2 - g_1}{g_2 + g_1}. \tag{3}$$

It has been observed that the sine law is valid if the listener's head is pointing directly forward. The tangent law [18,19] has been shown to be more correct if the listener turns his or head to track the virtual source:

$$\tan\phi_{\mathrm{p}} = \tan\phi_{\mathrm{s}} \frac{g_2 - g_1}{g_2 + g_1}. \tag{4}$$

Some studies have however shown that the effect of head rotation is negligible [18].

The sine-cosine law has proved popular in the entertainment industry [20]

$$g_1 = \cos\alpha \tag{5}$$

$$g_2 = \sin\alpha, \tag{6}$$

with α a panning trajectory parameter, $0 \leq \alpha \leq 90°$. Unlike sine and tangent laws, the sine-cosine law does not have listener position dependence. In the case where $\phi_{\mathrm{s}} = 45°$, the sine-cosine law becomes equivalent to the tangent law.

For small angles ϕ such that $\phi \simeq \sin\phi \simeq \tan\phi$, the angular panning law may also be used:

$$\phi_{\mathrm{p}} = \phi_{\mathrm{s}} \frac{g_2 - g_1}{g_2 + g_1}. \tag{7}$$

2.2. Multi-Loudspeaker Panning – Boundary

The following summarizes some popular methods for amplitude panning a virtual source on the loudspeaker boundaries, giving freedom in azimuth and elevation with range fixed by the loudspeaker radius relative to the listening position.

2.2.1. VBAP

Vector Base Amplitude Panning (VBAP) [7] is a technique for calculating amplitude gains by decomposing a source location on a basis defined by loudspeaker geometry. Define Cartesian unit vectors $\mathbf{l}_1 = [l_{11}\ l_{12}]^T$ and $\mathbf{l}_2 = [l_{21}\ l_{22}]^T$ pointing from the listener position towards loudspeakers 1 and 2 such that $\mathbf{L}_{12} = [\mathbf{l}_1\ \mathbf{l}_2]^T$. Unit vector $\mathbf{p} = [p_1\ p_2]^T$ points towards the virtual source. VBAP decomposes $\mathbf{p}$ on a linear combination of loudspeaker vectors

$$\mathbf{p}^T = \mathbf{g}\mathbf{L}_{12}, \tag{8}$$

where $\mathbf{g} = [g_1\ g_2]$ are loudspeaker gains that can be solved by

$$\mathbf{g} = \mathbf{p}^T \mathbf{L}_{12}^{-1}. \tag{9}$$

Gains $\mathbf{g}$ are then normalized so that (2) holds. The matrix inverse $\mathbf{L}_{12}^{-1}$ exists provided $\phi_{\mathrm{p}} \neq 0°$ and $\phi_{\mathrm{p}} \neq 90°$. When the $\phi_{\mathrm{s}} = 45°$, the panning law is equivalent to the magnitude of the Ambisonics panning function in (10).

2.2.2. MDAP

It was noted that the perceived spreading of the VBAP source varied depending upon panning directions, due to certain directions yielding small or zero panning gains to the subset of active loudspeakers. It was suggested to spawn additional panning angles such that at least two loudspeakers are active at any one time, termed Multi-Direction Amplitude Panning (MDAP) [8].

2.2.3. Ambisonics Panning Functions

Ambisonics panning functions are derived assuming a continuous circular distribution of monopole loudspeakers, often referred to as *secondary sources*, that are driven so as to produce a virtual monopole located on the secondary source distribution. Assuming the listener is at the center of the circle, the corresponding driving function (gains) for Ambisonics order N is [9, 10]

$$g(\phi_{\mathrm{s}}) = \frac{1}{2N+1} \sum_{m=-N}^{N} e^{im(\phi_{\mathrm{s}} - \phi_{\mathrm{p}})}. \tag{10}$$

Ambisonics gains differ from the other listed algorithms in that they are not constrained to be non-negative.

2.3. Multi-Loudspeaker Panning – Interior

Interior amplitude panning adds a degree of freedom to the virtual source location by allowing it to pass through the volume enclosed by the convex hull of loudspeakers.

2.3.1. DBAP

Distance-Based Amplitude Panning (DBAP) [12] is a technique that weights loudspeaker gains based upon the distance d_i from the virtual source to loudspeaker i:

$$d_i = \sqrt{(x_i - x_{\mathrm{p}})^2 + (y_i - y_{\mathrm{p}})^2 + r_{\mathrm{s}}^2}, \tag{11}$$

where r_{s} is an optional spatial blur that avoids the firing of a single loudspeaker when the source position and loudspeaker position are equal. DBAP can also be extended to support virtual source height. Loudspeaker gains are calculated by

$$g_i = \frac{k}{d_i^a}, \tag{12}$$

where a is a distance attenuation factor

$$a = \frac{R}{20\log_{10} 2}, \tag{13}$$

and R varies between 3 dB (power summing) and 6 dB (amplitude summing), and coefficient k enforces a unit power sum

$$k = \frac{1}{\sqrt{\sum_{i=1}^{M} \frac{1}{d_i^{2a}}}}. \tag{14}$$

For $r_\mathrm{s} > 0$, all loudspeakers are active for any virtual source position. For sources outside the convex hull of loudspeakers, their position is projected onto the closest point on the boundary.

2.3.2. Auditory Distance Rendering

Auditory Distance Rendering [13] aims to achieve an interior pan by exploiting the observation that interaural coherence and phase coherence increase as a source gets closer to the listener. A model for controlling the distance was proposed that coherently adds energy to the loudspeakers flanking the active loudspeaker. For example, a source panned hard to the center channel is played at a lower level through the left and right channel according to a distance parameter $0 \leq d_2 \leq 1$, where a value of 1 denotes halving the distance to the listener. The resulting gains for the center, left, and right channels (denoted by subscripts c, l, r respectively) are

$$g_\mathrm{c} = g_\mathrm{dry} = 1 \tag{15}$$

$$g_\mathrm{l}(d_2) = g_\mathrm{r}(d_2) = g_\mathrm{coh}(d_2) = (2/(2-d_2)-1)\cdot 0.6. \tag{16}$$

Gains are also normalized such that the presented signals are preserve unit power:

$$\begin{aligned} g_\mathrm{norm} &= \sqrt{(g_\mathrm{dry})^2 + 2\cdot(g_\mathrm{coh}(d_2))^2} \\ \hat{g}_\mathrm{dry}(d_2) &= \frac{g_\mathrm{dry}(d_2)}{g_\mathrm{norm}} \\ \hat{g}_\mathrm{coh}(d_2) &= \frac{g_\mathrm{coh}(d_2)}{g_\mathrm{norm}}. \end{aligned} \tag{17}$$

The authors also consider pans outside the boundary by use of decorrelation in addition to amplitude gains, however they are outside the scope of this paper.

2.3.3. Dual Balance

Consider a loudspeaker layout on the periphery of a rectangular listening area, where each loudspeaker is assigned a row and column index. In a Dolby Atmos cinema, typically the row behind the screen consists of at least three loudspeakers (L, C, R), sidewall rows contain two (e.g. Lss1, Rss2), and the rear wall contains at least two (e.g. Lrs1, Crs, Rrs1). Let $(x_{j,k}\ y_{j,k})$ be the Cartesian coordinates of the loudspeaker in row $j \in \{1,\ldots,J\}$ and column $k \in \{1,\ldots,K\}$. Let $(x_\mathrm{p}, y_\mathrm{p})$ be the virtual object's intended position on the horizontal plane. The dual-balance algorithm treats the x and y components as independent panning gains that are later combined.

A row-wise pass first calculates row-wise gains as according to the sine-cosine panning law,

$$g_{j,k}^\mathrm{rows} = \begin{cases} \cos\left(\frac{\pi}{2}\frac{x_\mathrm{p}-x_{j,k}}{x_{j,k+1}-x_{j,k}}\right) & (x_{j,k} < x_\mathrm{p})\ \&\ (x_{j,k+1} > x_\mathrm{p}) \\ \sin\left(\frac{\pi}{2}\frac{x_\mathrm{p}-x_{j,k-1}}{x_{j,k}-x_{j,k-1}}\right) & (x_{j,k} > x_\mathrm{p})\ \&\ (x_{j,k-1} < x_\mathrm{p}) \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

A column-wise pass then calculates column-wise gains in a similar way,

$$g_{j,k}^\mathrm{cols} = \begin{cases} \cos\left(\frac{\pi}{2}\frac{y_\mathrm{p}-y_{j,k}}{y_{j+1,k}-y_{j,k}}\right) & (y_{j,k} < y_\mathrm{p})\ \&\ (y_{j,k+1} > y_\mathrm{p}) \\ \sin\left(\frac{\pi}{2}\frac{y_\mathrm{p}-y_{j-1,k}}{y_{j,k}-y_{j-1,k}}\right) & (y_{j,k} > y_\mathrm{p})\ \&\ (y_{j-1,k} < y_\mathrm{p}) \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

The final loudspeaker gains are the product of the row and column gains,

$$g_{j,k} = g_{j,k}^\mathrm{rows} g_{j,k}^\mathrm{cols}. \tag{20}$$

Up to four loudspeakers enclosing the object's position are active at any one time. An object lying on the line connecting pair of loudspeakers receives a pair-wise sine-cosine pan. The extension to include a z component, the so-called *triple balance* algorithm, is straightforward.

3. Experimental Setup

The following describes an experimental setup for playback of an arbitrary amplitude pan, both boundary and interior, and recording its perceived location along with informal comments on perceived size, height and timbre. The Dual-Balance panner was used to render the pans, all of which lay on the horizonal plane in the front-left quadrant.

3.1. Listening Environment

Five Revel Ultima Salon 2 loudspeakers were configured in a rectangular arrangement measuring 2 × 4 m as shown in Fig. 1, of which at most four were active at any given time. The corner loudspeakers, that lay $2\sqrt{2}$ m from the subject, were compensated for amplitude and time of arrival.[1] The testing environment was a dedicated listening room measuring 6.9 × 6.4 × 3 m with T_{60} = 0.28 s and a measured noise floor below the ISO threshold of human hearing [21]. The loudspeakers were characterized in-situ by averaging magnitude responses at four locations in a ∼60×60 cm square centered around the listening area. Four second order parametric equalization filters were applied 50–300 Hz with a Nion N6 MediaMatrix. The final chain consisted of a Lynx Aurora 16 D/A converter driving ATI AT6005 power amplifiers.

3.2. Listening Method

The listener was located equidistant from the three edge loudspeakers, as marked by spots on the floor and ceiling. Acoustically transparent optically opaque cloth was hung between the listener and the loudspeakers and the room was darkened so that only a 24" screen, keyboard and mouse were visible. The listener was then presented with an array of 16 buttons that each played a static pan randomly drawn from a uniform distribution in the front left quadrant. Listeners were made aware that pans would be produced only in the front left quadrant. Three additional buttons were provided to drive only the left, center or left side surround to provide acoustic

[1] This layout was chosen to provide a mapping between a rectangular panner graphic used to record the perceived phantom source; alternatively the loudspeakers could have been placed in a circle but this would have led to a discrepancy between the graphic and the room in the corners.

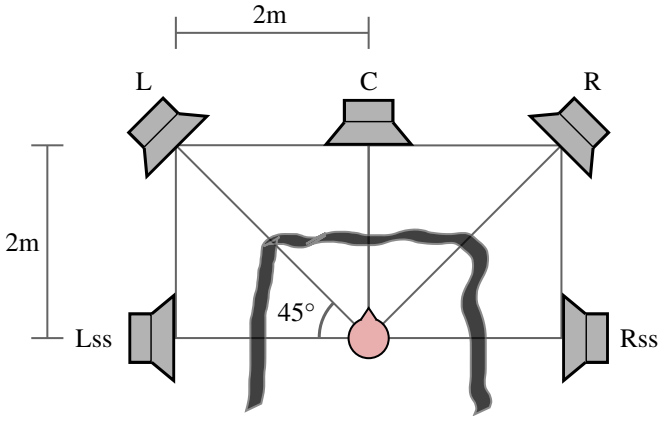


Fig. 1: Loudspeaker Layout. Corner loudspeakers were placed at 45° relative to the listener and were time- and amplitude-compensated.

anchors for the boundaries of the panning area. Localization accuracy is known to be sensitive to source type [22], however a single 10 s excerpt of male speech was used for these experiments. Listeners were also given the option of changing the loop points with 4.3 ms crossfades to suppress audible discontinuities during loop and switching between pans.

The task was to mark the perceived source location of each of the 16 stimuli by means of a perspective panner graphic that could be manipulated with a mouse. The buttons could be traversed in any order and be replayed at will. Listeners were encouraged to return to the acoustic anchors throughout the session. In the case of a source with a perceived width or height offset, they were asked to provide the centroid of the horizontal component. The listeners were unaware of the underlying set of source locations and were able to place the graphical panner anywhere on the horizontal plane within the panning area. They were asked to remain in the listening position throughout the test but they were permitted to turn their heads if they wished.

A total of 21 listeners took part in the study. All were audio professionals and had experience with critical listening and subjective tests but had not previously conducted such an experiment.

4. Results

The results are shown as scatter plots in Fig. 2 for each of the 16 uniformly-spaced pans in the front left quadrant, in which the intended auditory source location is a red '$\times$' and the responses are a white '$\circ$' over a 2D Gaussian fit to the results. The listener was seated at the $\times$ in the bottom right subplot. Mean distance errors are recorded in Table 1 and standard deviation in Table 2. Pans lying on the center and left side surround loudspeakers were identified most accurately with a mean error of 8 and 5 cm respectively. Left was less accurate at 16 cm; this may be due, in part, to the increased distance and lower direct-to-reverberant ratio leading to increased ambiguity. Three outliers, included in all calculations, may also have skewed the statistics. Fig. 3 shows the mean responses and intended source position. In all cases the rank order is correct across any given row or column.

4.1. Left-Center Pairwise Pans

Left-center pairwise pans are the closest approximation to a typical stereo configuration, with the exception that the listener is offset with respect to the loudspeakers so that they lay at 45° and 0° relative to the listener respectively. Such offsets have been considered in the literature [23], however no explicit compensation was applied in this paper. Surprisingly, both pans lying between left and center produced large errors, with means of 52 and 62 cm, both of which were biased on a line pointing somewhat towards the listener. Left-center pairwise pans were therefore interpreted as interior pans with error mostly in the y-direction; this is consistent with the phase and amplitude coherence argument for interior panning in [13], albeit in this case with two and not three loudspeakers. Given that the listeners were able to set the panner graphic up to the edge of the panning area but not beyond, causing all errors to lie on or inside the room.

4.2. Left Wall Pans

Pairwise pans between left and left side surround also differ from a typical stereo configuration, with loudspeakers at 45° and 90° respectively. Surprisingly, they produced smaller mean errors than the left-center pairwise pans at 33 and 40 cm in both x and y directions. Like the left-center pairwise pans, all errors occurred inside the panning area and not outside due to the constraints of the panner graphic.

4.3. Pans on the Interaural Line

The interaural line is defined as the line passing through the listener's ears. Dual Balance acheives an interaural pan by amplitude panning between left side and right side surround loudspeakers, which lie at $\pm 90°$ azimuth relative to the listener. Such pans are known to produce a small sweet spot [19]. The results show that pans along the interaural line tend to produce large errors, in particular the pan immediately to the left of the listening position. Several listeners also noted that the auditory source sounded elevated relative to the reference pans, and the greatest apparent source size. The *phantom image elevation effect* is a known phenomenon when ϕ_s is large [24, 25]. In the case where the target auditory source source lay precisely half way between the two loudspeakers, there is an apparent leftward bias. It is expected that this was due to a bias in the listeners who had prior knowledge that all pans lay in the front left quadrant. A similar leftward bias is seen in the pans on the median plane.

4.4. Pans on the Median Plane

Auditory sources lying on the line passing through the median plane do not produce interaural time delay and are known to be difficult to localize even with real loudspeakers [26]. However, the mean estimates are relatively close to the ground truth with errors 13 and 36 cm. With the exception of a small leftward bias, most estimates lay on the median plane. Like pans on the interaural line, it was reported that apparent elevation increased as the auditory source came closer to the listener.

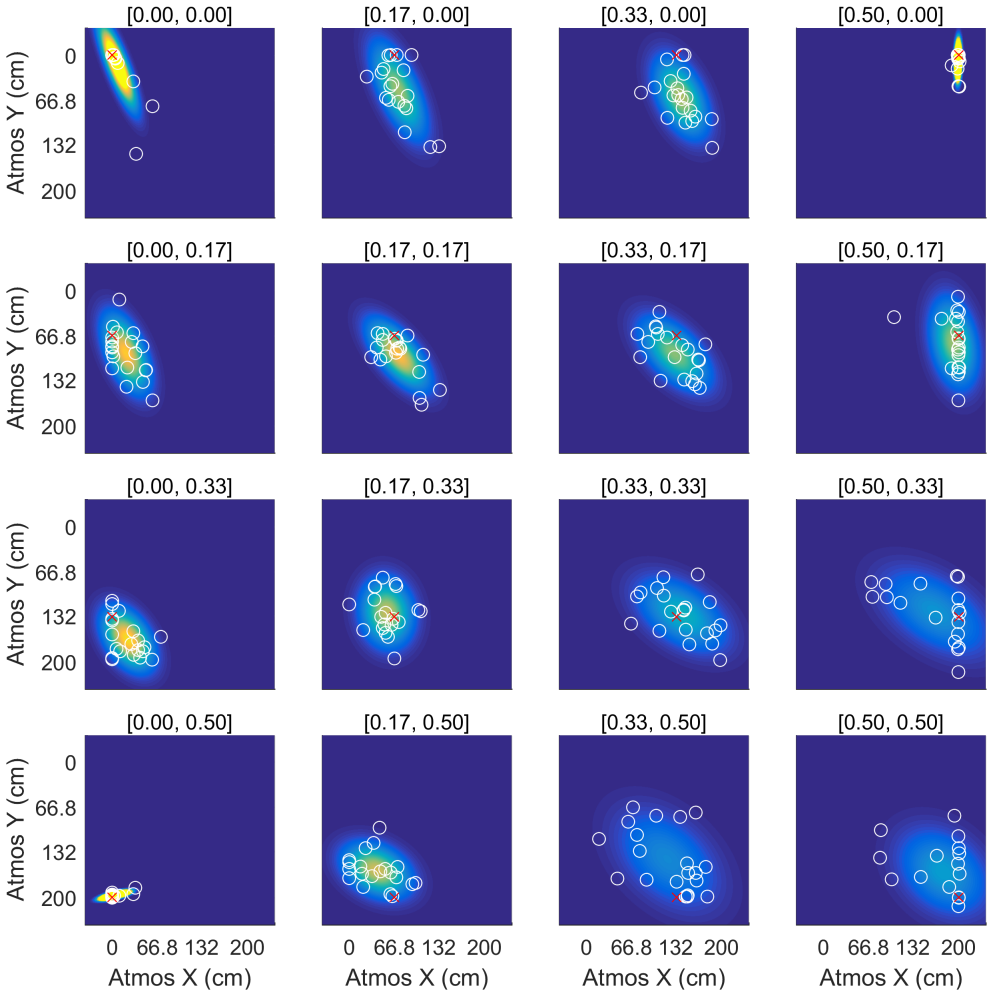


Fig. 2: Intended pan (red ×) and responses (white ○) overlaid on a 2D Gaussian fit in the front-left quadrant on the horizontal plane. Pans are in the left-handed Dolby Atmos coordinate system.

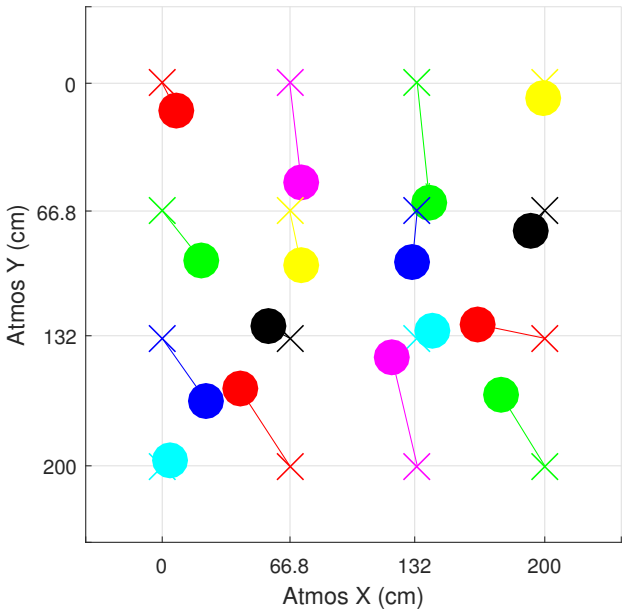


Fig. 3: Intended pan (×) and mean response (○). Lowest errors are for single loudspeaker pans. The mid-interior pan [0.33, 0.17] and [0.33, 0.33] errors are surprisingly low.

	0	1/6	1/3	1/2
0	16	52	62	8
1/6	33	29	27	13
1/3	40	13	8	36
1/2	5	48	58	44

Tab. 1: Distance between sample mean and intended location (cm).

	0	1/6	1/3	1/2
0	39	43	36	14
1/6	29	32	20	26
1/3	25	19	23	43
1/2	10	27	46	50

Tab. 2: Standard deviation of error distance (cm).

4.5. Other Interior Pans

There is some disagreement within the audio community regarding the efficacy of interior pans using amplitude-based panners. The results of this study suggest that, for listeners placed at the sweet spot, some interior pans can be identified no less accurately than boundary pans. Furthermore, the mean errors between 8 and 58 cm suggest that some are identified more accurately than some boundary pans. In most cases the Gaussian fit tends to be orientated with its major axis towards listener, implying that azimuth angle is more easily identifiable than distance. Many earlier studies have shown

this higher distance error to be the case for real sources [26] so it is to be expected with virtual sources. The listeners relied upon the reference buttons to help identify different distances; some reported recalling the pans associated with different buttons and switching between them to produce a step change in the object's position. It is expected that the trajectory of moving objects would have provided additional cues to help identify the starting and finishing locations for the same reason.

Dual-Balance interior pans fire four loudspeakers, producing different timbre to pairwise (2-loudspeaker) or single loudspeaker pans [8]. However, most listeners did not state that timbre differences were objectionable.

5. Summary

A methodology for comparing intended boundary and interior pans to the perceived location was presented. Using the Dual-Balance algorithm as an example case, the results from 21 listeners showed close agreement between the perceived and intended pans. Lowest errors were seen for cases that fired a single loudspeaker and highest were for pans intended to be close to the listener. Surprisingly, errors for several interior pans were comparable to—and in some cases lower than—pairwise pans. The phantom image elevation effect for virtual sources panned close to the listener was observed as previously recognized in the literature. The results presented in this paper suggests that static interior amplitude pans can be consistently perceived by a listener, warranting further investigation into the perception of source height and size and sensitivity to other variables such as source type and listener position.

6. References

[1] A. D. Blumlein, "British patent specification 394,325 (improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems)," *Journal Acoust. Eng. Soc.*, vol. 6, no. 2, pp. 91–98, 1958.

[2] H. A. M. Clark, G. F. Dutton, and P. B. Vanderlyn, "The 'Stereosonic' recording and reproducing system," *Journal Acoust. Eng. Soc.*, vol. 6, pp. 102–117, Apr. 1958.

[3] B. B. Bauer, "Phasor analysis of some sterophonic phenomena," *Journal Acoust. Soc. America*, vol. 33, pp. 1536–1539, Nov. 1961.

[4] N. V. Franssen, "Sterophony," Tech. Rep., Philips Technical Library, Eindhoven, The Netherlands, 1964.

[5] C. Ceoen, "Comparative sterophonic listening tests," *Journal Acoust. Eng. Soc.*, vol. 20, no. 1, pp. 19–27, 1972.

[6] B. Bernfeld, "Attemps for better understanding of the directional stereophonic listening mechanism," in *Proc. Audio Engineering Soc. Conv.*, Mar. 1973.

[7] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal Acoust. Eng. Soc.*, vol. 45, no. 6, pp. 345–466, June 1997.

[8] V. Pulkki, "Uniform spreading of amplitude panned virtual sources," in *Proc. IEEE Workshop on Applications of Signal Proc. to Audio and Acoustics*, New Paltz, New York, Oct. 1999.

[9] J. Daniel, "Further study of sound field coding with higher order Ambisonics," in *Proc. Audio Engineering Soc. Conv.*, Berlin, Germany, May 2004.

[10] J. Ahrens, *Analytic Methods of Wavefield Synthesis*, Springer-Verlag Berlin Heidelberg, 2012.

[11] H. Lee and F. Rumsey, "Level and time panning of phantom images for musical sources," *Journal Acoust. Eng. Soc.*, vol. 61, no. 12, pp. 978–988, 2013.

[12] T. Lossius, P. Balatazar, and T. de la Hogue, "Distance-based amplitude panning," in *Proc. International Computer Music Conf.*, Montreal, Canada, Aug. 2009.

[13] M-V Laitinen, A. Walther, J. Plogsties, and V. Pulkki, "Auditory distance rendering using a standard 5.1 loudspeaker layout," in *Audio Engineering Soc.*, New York, USA, Oct. 2015.

[14] G. Dickens, M. Flax, A. McKeag, and D. McGrath, "Optimal 3D speaker panning," in *Audio Engineering Soc.*, Rovaniemi, Finland, Apr. 1999, pp. 421–426.

[15] N. Tsingos, C. Q. Robinson, D. P. Darcy, and P. A. Crum, "Evaluation of panning algorithms for theatrical applications," in *Proc. International Conf. on Spatial Audio*, Erlangen, Germany, Feb. 2014.

[16] M. Frank, *Phantom Sources using Multiple Loudspeakers in the Horizontal Plane*, Ph.D. thesis, University of Music and Performing Arts, Graz, Austria, June 2013.

[17] M. Williams, "Unified theory of microphone systems for sterophonic sound recording," in *Proc. Audio Engineering Soc. Conv.*, London, Mar. 1987.

[18] D. M. Leakey, "Some measurements on the effect of interchannel intensity and time differences in two channel sound systems," *Journal Acoust. Soc. America*, vol. 31, pp. 977–986, July 1959.

[19] J. C. Bennett, K. Barker, and F. P. Edeko, "A new approach to the assessment of stereophonic sound system performance," *Journal Acoust. Eng. Soc.*, vol. 33, pp. 314–321, May 1985.

[20] D. Griesinger, "Stereo and surround panning in practice," in *Proc. Audio Engineering Soc. Conv.*, Munich, Germany, May 2002.

[21] "Acoustics – normal equal-loudness-level contours," Tech. Rep. ISO 226:2003, International Organization for Standardization, New York, USA, Oct. 2003.

[22] J. Corey and W. Woszczyk, "Localization of lateral phantom images in a 5-channel system with and without simulated early reflections," in *Proc. Audio*

Engineering Soc. Conv., Los Angeles, CA, Oct. 2002.

[23] V. Pulkki, "Compensating displacement of amplitude-panned virtual sources," in *AES International Conf. on Virtual, Synthetic and Entertainment Audio*, 2002.

[24] K. de Boer, "A remarkable phenomenon with stereophonic sound reproduction," Tech. Rep. 9, Philips, 8–13, 1947.

[25] H. Lee, "Investigation on the phantom image elevation effect," in *Proc. Audio Engineering Soc. Conv.*, 2015.

[26] B. Moore, *Hearing*, New York: Academic, 1995.

HOA convolution reverb: simulated multichannel impulse responses for a maximum of flexibility in ambisonics

Leon Merkel [1, 2]
[1] *University of Music and Performing Arts Graz*
[2] *Graz University of Technology*

Correspondence should be addressed to Leon Merkel (leon.merkel@tonmeister.de)

Abstract

This work presents the development of a convolution reverb software for 3D audio productions by using object based programming and acoustically simulated hoa impulse responses of an existing concert hall. An innovative graphical user interface enables the user to change the position of the listener and different sound sources in the spatial environment with ease. An individually simulated ambisonics impulse response exists for every possible position in order to create a feeling of realistic depth and accurate localization. The resulting 3D sound field can be played back on hoa systems, in binaural using head tracking headphones, or on every common loudspeaker setup up to Auro 3D.

1. Introduction

Since the beginning of classical music production, one of the main goals for audio engineers has been the realistic sound field reproduction of the worlds greatest concert halls [4]. It is well known that besides the placement of the microphones during the recording process, the added reverberation in postproduction plays a crucial role for a successful music production [4, 12].

Nowadays, the relevance of tools for 3D-Audio applications is growing. Additionally software solutions are able to simulate complex room acoustics in high quality and generate excellent ambisonics impulse responses with an ultra-low S/N ratio based on perfect omnidirectional sound sources [9, 22].

This work presents the development of a flexible ambisonics convolution reverb of an existing room for 3D-Audio music productions. Object programming and simulated multichannel impulse responses enable the user to change the position of the listener and the sound source in a realistic sound field.

Fig. 1: Concert hall of the University of Music and Performing Arts Frankfurt during a string quartett recording

2. Acoustical environment

The chosen acoustic environment for the simulation was the 3000m³ concert hall at the University of Music and Performing Arts „HfMdK“ in Frankfurt am Main (**Fig. 1**). Willi Cahn built the room in 1929 as the large broadcasting hall for the southwest german broadcast company (SÜWRAG) [18]. The „HfMdK“ has used the 350 seat room as a concert hall for student productions since 1954. With an average reverberation time of 1.6 seconds, Andreas Odenkirchen denoted it as „the hall with the best acoustics for chamber music in Frankfurt“ [18].

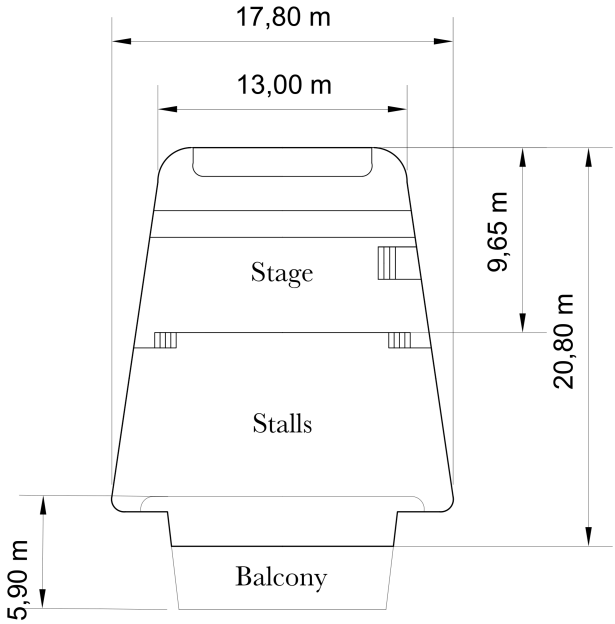


Fig. 2: Sketch of the concert Hall of the University of Music and Performing Arts Frankfurt

ISBN
978-3-9812830-9-9

The authors hold the copyrights of all of content in this manuscript. They have granted the VDT the right to publish it in print, online and as PDF download on https://tonmeistertagung.com

The exact geometric and acoustical surveying of the room constitutes the groundwork for the accurate acoustical simulation. Due to the mathematical restrictions of geometrical room acoustics, details in magnitude of 0.1 meters were not considered [21]. The acoustical characteristics are ascertained from the evaluation of sweep measurements, made according to the german industry norm „DIN 3382-1“ [6]. **Fig. 3** shows the average reverberation time of the measurements, recorded at the positions, illustrated in **Fig. 4** in limits of the „DIN 18041“ [5].

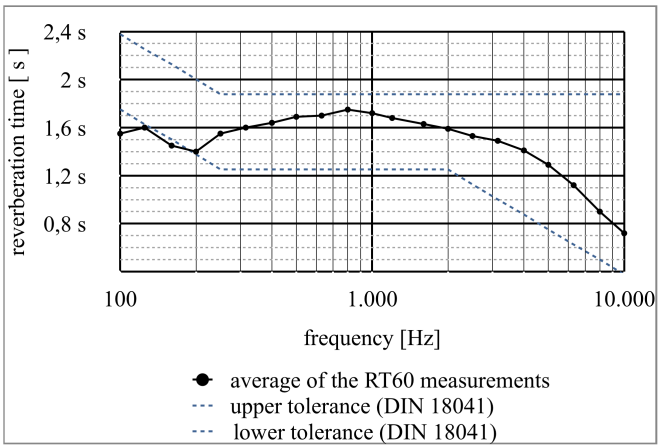


Fig. 3: Average reverberation time of the measurements

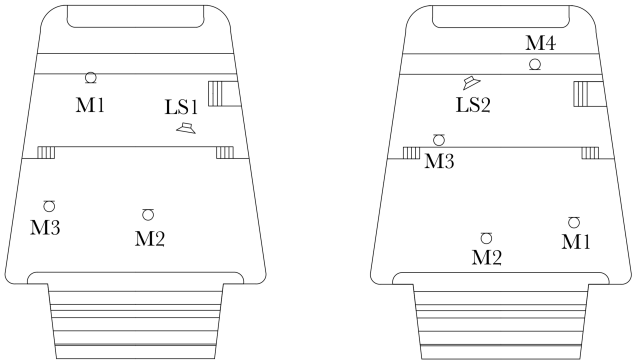


Fig. 4: Mircophone (M) and loudspeaker (LS) placement during the measurements

3. Acoustic verification and simulation

The software „EASE“ (enhanced acoustic simulator for engineers) was used for the room simulation and the „AURA 3D-Response“-modul for the generation of multichannel room impulse responses (RIR) at specific positions [1]. In order to simulate the RIR’s, a geometric model including boundary surfaces and their absorption and scattering coefficients was created.

The hybrid simulation procedure AURA-3D-Response, which includes material scattering, of the RWTH Aachen, based on the CEASAR-algorithm [20] was used for the best possible time resolution in the resulting RIR. The first reflections are hybrid simulated. Every reflection path is ascertained by ray tracing and based on these paths, mirror sources are placed [20]. From a defined order of reflection, the diffuse part is simulated with ray tracing. The high time resolution isn’t required. The scattering coefficients, which are playing a crucial role in the diffuse part of the impulse response, are considered [16].
To verify the accuracy of this model, RIR’s were recorded [6] using the shown Equipment (**Tab. 1**) and were then compared with the simulated RIR’s.

Software	SATlive - Version 1.70.26
Soundcard	RME Fireface UFX
Loudspeaker	Genelec 8030A
Subwoofer	Genelec 7050B
Mic	iSEMcon EMX 7150

Tab. 1: RIR-easurement equipment

For accurate comparability, the same positions were selected for loudspeakers and microphones as in the RIR recordings. The average of the calculated RT60 values of the simulated RIR’s has been compared with the average of the RT60 values of the recorded RIR’s .
In addition, the reflection pattern of the first 100ms of the simulated RIR’s are also compared with the recorded RIR’s (**Fig. 5**).

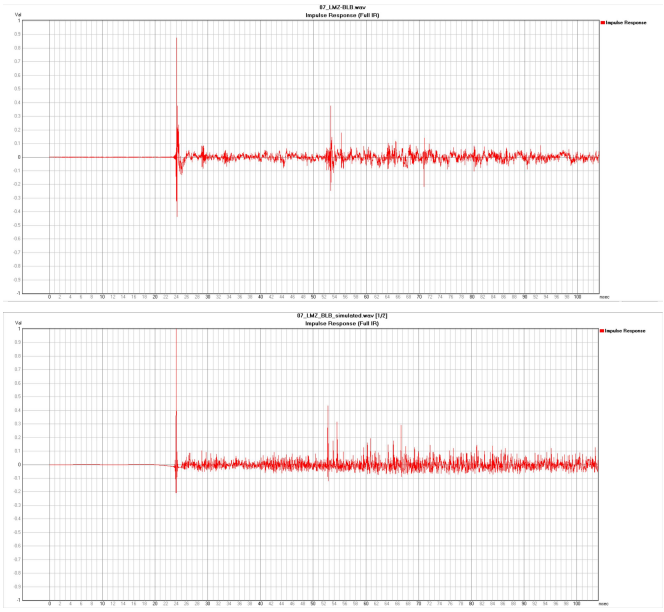


Fig. 5: RIR-Comparasion (Top: Simulation, Bottom: Measurement)

After the first comparison, some absorption coefficients of walls and surfaces were adapted. Having changed the values for walls and surfaces, a new comparison of both averaged RT60 values presented a deviation of 5% (**Fig. 6**), which is within the measuring tolerance. This confirms the validity of the model.

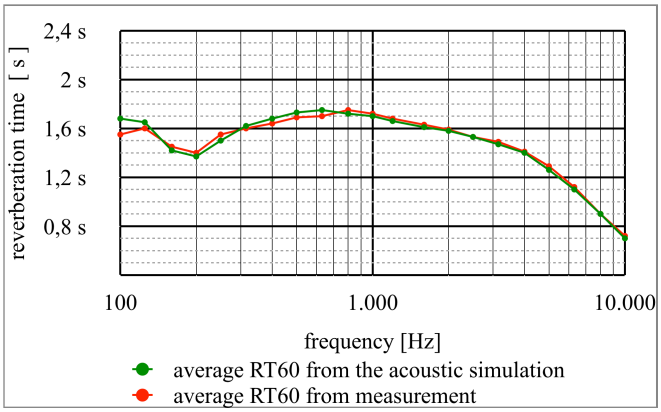


Fig. 6: Comparison of the simulated and measured RT60-Values

In order to simulate the RIR's for a convolution reverb, the parameters were selected to obtain the highest time resolution of the IR by using a high number of particles and a high resolution in the time domain (**Tab. 2**).

Parameter	Value
Particels	3486000
Length	1830
Cutoff Order	30
Density Factor	30
Tail Resolution	1 ms

Tab. 2: Important simulation parameters

Using the modeling software, sixteen virtual omnidirectional sound sources were placed on a stage in a half circle, recreating the placement of musicians in a classical orchestra. The eleven selectable virtual positions of the listener are located in the stalls and on the balcony. The virtual second order ambisonics receiver was placed at ear height on the different listener positions. The exact simulated locations of sound sources and listeners are delineated in **Fig. 7**.

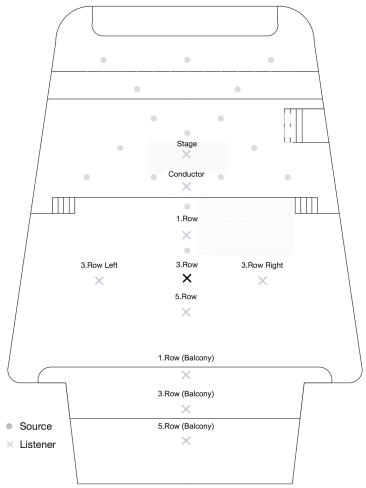


Fig. 7: Sketch of the source and listener positions

Due to technical restrictions of geometric room acoustics, which are used for the simulation, the resulting RIR's exclude low frequency content. In the acoustic theory low frequencies are considered as a wave, because of their relation between the wavelength and room dimensions. In the case of acoustical simulation, ray tracing and the mirror source method are not able to simulate wave propagation. The software package AURA 3D-Response simulates in the frequency range of 100 Hz to 20000 Hz. Artificial impulse responses for those low frequencies were generated with pink noise and time-automated low-pass-filters, as the solution for these missing low frequencies. This artificial IR will be added later to the sound pressure component of the ambisonics (W) RIR.

The estimated reverberation time values are based on the „bass ratio" at concert halls, introduced by F. Alton Everest and Ken C. Pohlmann [7]. This low frequency trend function is typical for concert halls and responsible for a natural warm sound and sonority, according to Everest and Pohlmann. The „BR" describes the mathematic ratio of the reverberation time between the frequency ranges 125-250 Hz and 500-1000 Hz.

$$BR = \frac{RT60_{125Hz} + RT60_{250Hz}}{RT60_{500Hz} + RT60_{1000Hz}} \qquad (1)$$

In order to calculate the BR of the existing concert hall modified RT60 values were used to consider audience absorption. The resulting BR corresponds to *1.27*. The calculated lower octave interval of the reverberation time allows for the creation of a fitting curve for the third band time values in the low frequencies. The estimated values are shown as a fitting curve in **Fig. 8**.

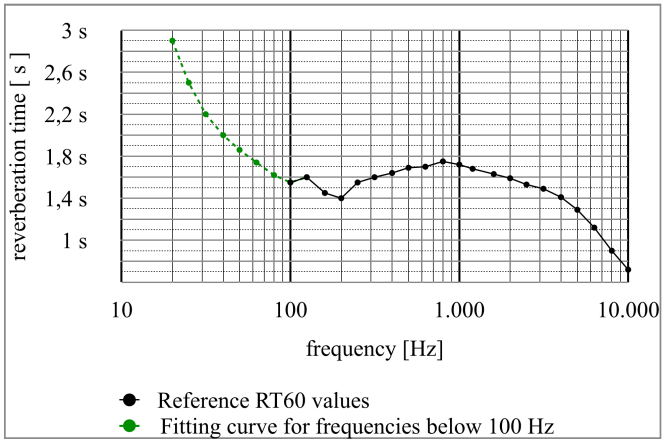


Fig. 8: Interpolation of RT60 values for low frequencies

4. SDM upscaling

To improve the spatial impression and three-dimensional resolution, the simulated second order ambisonics impulse responses are upscaled to third order ambisonics, using the spatial decomposition method (SDM) invented by Sakari Tervo, Jukka Pätynen and Anitti Kuusinen [19]. Based on the assumption, that the sound propagation direction is characterized by the average of all arriving waves, SDM analyses the RIR and encodes these values to a sound pressure component and arrival directions coordinates [19].

At the beginning SDM locates the sound position from the RIR, estimating the average location of the sound for every sample by using a localization function. Secondly every sample of the sound pressure component is assigned to a determined location [19]. The resulting upscaled RIR contains the same frequency response as the simulated sound pressure component of the RIR.

Increasing the ambisonics order of the RIR's using SDM results in enhanced localization, a larger sweet spot and a more enveloped and deep reverberation feeling [8].

5. Software structure

The visual object based programming software Max/MSP 8 was used for the development of the convolution reverb [3]. In the implementation process of the software the „HoaLibrary" by Julien Colafrancesco, Pierre Guillot and Eliott Paris [17] and the ICST ambisonics spatialization tools by Jan C. Schacher and Philippe Kocher [15] were used for monitoring and spatialization tasks. Additionally, the „mcfx_convolver16" -vst-plugin from the „Plugin-Suite for Ambisonics" by Matthias Kronlachner calculated the convolutions, since the latency and CPU usage of this plugin are considerably low [9]. The schematic structure of the software is visualized in **Fig. 9**.

to arrive at the listener. After the spatialization the signal is attenuated in level in order to adjust the sound pressure loss caused by distance. The second component of the signal is split at 100Hz to implement two different convolutions. Since the artificially generated impulse response for the lower frequencies isn't compensated in time, every signal is delayed before convolution by the same amount of milliseconds as the dry signal. The convolution of every position is accomplished by only one instance of the „mcfx_convolver16" using sixteen separate inputs and outputs. After the convolution of the frequencies below 100Hz the signal is also attenuated in level to adjust the described sound pressure loss. Every input channel of the upper frequency component is immediately convolved with the upscaled third order RIR's by using the matrix convolution „mcfx_convolver16". All of the sixteen inputs are convolved with one set of the ambisonics RIR's which represent one position in the room. One input channel, allocated to one specific position, feeds the affiliated stream of the matrix convolution. Owed to the fact, that RIR's are not distance compensated in level, they are also separately attenuated. After the convolution and level adjustment, the signal component below 100Hz is merged with the sound pressure component (W) of the ambisonics signal. To adjust levels between the signal convolved with simulated and artificial RIR's, the level of the component below 100Hz can be customized by using the „bass ratio"-knob on the graphical user interface.

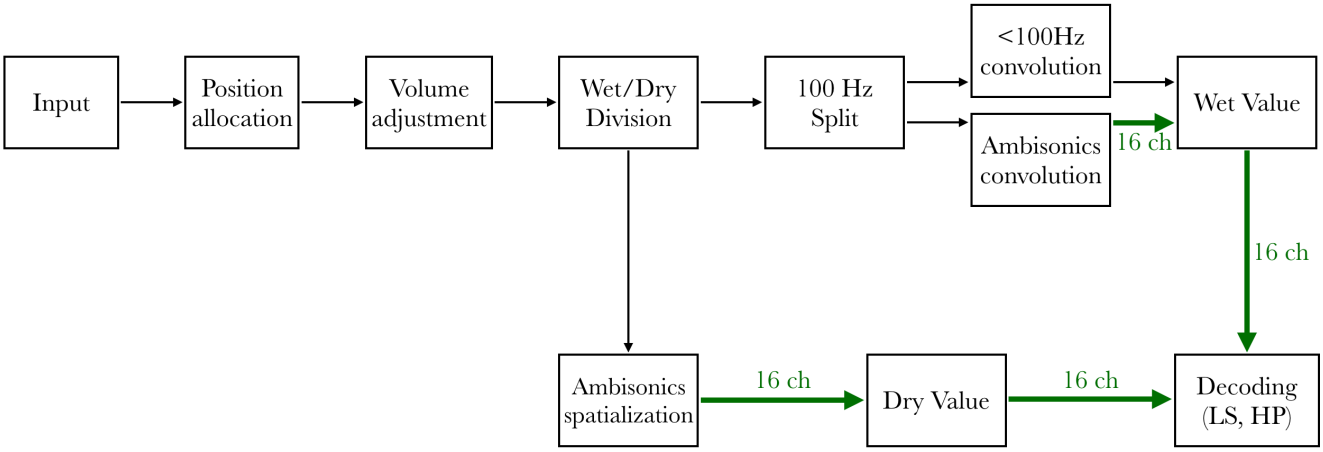


Fig. 9: Schematic structure of the software

Using any possible input e.g. sound card mic preamp or internal digital audio workstation (DAW) multichannel stream, every input can be separately selected and allocated to one of sixteen source positions. Multiple positioning of sound sources allows the user to deploy the same signal all over the hall. The volume of all sixteen possible input channels are separately adjustable before convolution. In order to mix the dry and reverberated component of the signal, every input channel is divided in two signal chains. The dry component is spatialized by using the third order ambisonics version of the „ambiencode~"-object of the ICST ambisonics spatialization tools. For accurate localization of the source in the room, the exact coordinates of both listener and sound source, as determined in the acoustical simulation, are feeding the spatialization tool. To generate a feeling of distance in the dry component every input is delayed before spatialization by the amount of milliseconds the sound needs

The ratio between the dry and reverberated third order ambisonics streams, can be separately manipulated infinitely variable from completely dry to fully reverberated. Optionally the merged signal can be decoded for binaural headphone listening or any common loudspeaker setup.
Binaural playback is implemented by decoding the ambisonic stream using the „ambidecode" -object from the ICST ambisonics spatialization tool to feed 25 virtual loudspeakers. These are convolved to a two channel binaural audio stream using 25 stereo binaural room impulse responses, from the „Listen HRTF Database" recorded by G. Vandernoot at the IRCAM Institute in 2002 [10]. For a better spatial impression, headphone-tracking can be implemented utilizing the „MrHeadTracker" -device by Michael Romanov [13]. Depending on the head rotation, the ambisonic sound field will then rotate before decoding to binaural. The user gets the impression of fixed sound sources

positions in the concert hall during head movement and rotation. The „ambidecode“ -object converts the ambisonic stream to different common loudspeaker setups. Alongside presets for small one-dimensional setups like mono, stereo and LCR, presets for two-dimensional setups (5.1, 7.1), as well as three-dimensional setups (5.4.1, 7.4.1) are available and pre-configured. Additionally a preset for the 24 loudspeaker ambisonic setup of the „Cube“ in the IEM (Institute for Electronic Music and Acoustics) of the University of Music and Performing Arts Graz can also be selected.

6. Graphical user interface

The graphical user interface is unique compared to conventional reverb software. A great advantage is the allocation of sound sources to specific three-dimensional positions, without having to change different parameters to create a feeling of depth and localization in the room. **Fig. 10** shows the GUI of the software.

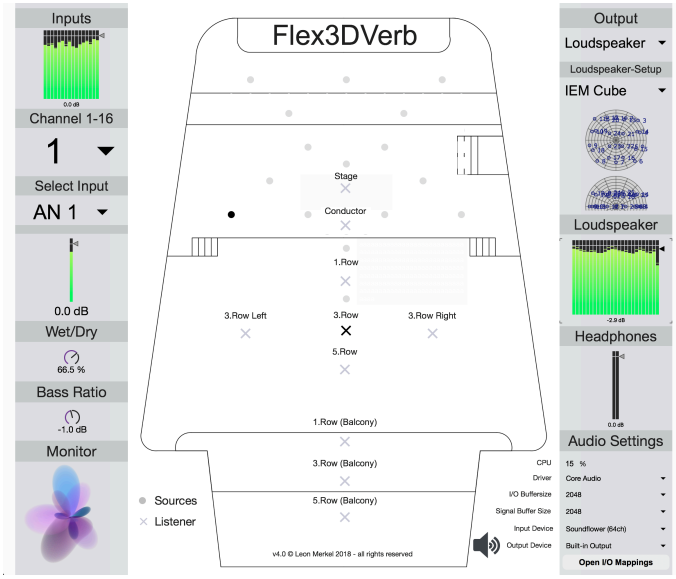


Fig. 10: Graphical user interface of the software

In the audio settings, the audio driver, I/O buffer size, signal buffer size, as well as the in- and output device can be selected. The current cpu performance is displayed. The I/O mappings enable the user to create complex routing setups.
The input section meters the levels of the sixteen inputs of the selected audio device. Below the metering the user can choose between sixteen track channels, which are fed by one of the sixteen different inputs. A separate meter shows the level of the selected track and allows for volume adjustments. The selected track channel can then be allocated, to the specific source positions. Multiple positioning of one source is also possible. The positions in the hall and the selected input are stored. The „Wet/Dry“ -knob is used to mix the dry and the reverberated signal. The „Bass ratio“ -knob makes it possible to manipulate the amount of reverb in the lower frequency range. In the bottom left corner of the GUI, a visualization of the ambisonics stream allows for monitoring of the signal in spherical harmonics. The reverberated signal component is colored in purple, the dry signal in blue. The main section of the GUI displays a two-dimension schematic of the concert hall with the selectable source and listener positions. All of the sixteen input signals allocated to the different source positions can only be monitored to from one of the eleven listener positions. The source positions are static while the listener position can be changed in real time.
In the output-section of the GUI, the user can choose between headphone- and loudspeaker decoding. When selecting headphone playback the resulting sound field will be decoded to binaural as described in „5. Software structure“. A master headphone meter shows the output level, which can also be adjusted. When using loudspeaker playback the specific speaker setup must be selected. An up to 25-channel meter for the loudspeakers displays the different levels. The master fader adjusts all speaker levels simultaneously .

7. Conclusion

The resulting convolution reverb software enables the user to choose an audio signal and place it on one of the sixteen positions on the stage of the concert hall and plays it back on every common loudspeaker setup on up to 25 channels. Binaural sound with head tracking headphones is also available. It is possible to switch between eleven listeners positions in the stalls and on the balcony of the concert hall.

For a subjective comparison of the simulated and the original room reverberation, a 3D recording of a musical performance was first recorded in the concert hall of the HfMdK and then in the anechoic chamber of the University of Applied Sciences Darmstadt. The audio signal of the latter recording is convolved with both the simulated and recorded IR’s, and later compared using a 9.0 Auro 3D loudspeaker setup and professional headphones.

For further subjective evaluation, the anechoic symphony orchestra recordings by Tapio Lokki and Jukka Pätynen [11] were used to reproduce a realistic orchestra performance in the virtual acoustic environment. Played back on a 7.4.1 loudspeaker system in the production studio of the IEM Graz, a realistic feeling of depth and accurate localization was created. This gives the impression of actually being in a concert hall.

8. Acknowledgments

Firstly I would like to thank Jorge Medina for his outstanding supervision on this project. I’m also thankful for the technical possibilities offered to me by the University of Applied Sciences Darmstadt in the early stages of the development process. I would like to thank Matthias Frank and Daniel Rudrich form the Institute for Electronic Music and Acoustics for their secondary supervision and their ideas for improving the software, in regards to spatial resolution and cpu-efficiency.

References

[1] Ahnert, W.: „Tutorial für EASE 4.0“, URL http://www.afmg-support.de/SoftwareDownloadBase/AFMG/EASE/ (2000)

[2] Bork, I.: „A Comparison of Room Simulation Software - The 2nd Round Robin on Acoustical Computer Simulation“ In Acoustica, Vol. 86 S. 943-956. (2000)

[3] Cycling’ 74. „Max/MSP“ URL https://cycling74.com

[4] De Man, B.; McNally, K.; Reiss, J.: „Perceptual Evaluation and Analysis of Reverberation in Multitrack Music Production“. In Journal of the AES, Vol. 65, No. 1/2. (2017)

[5] DIN 18041: „Hörsamkeit in kleinen bis mittelgroßen Räumen“ (2004)

[6] DIN EN ISO 3382-1: „Akustik - Messung von Parametern der Raumakustik - Teil 1: Aufführungsräume“ (2009)

[7] Everst, F. A.; Pohlmann, K.C.: „Master Handbook of Acoustics“, Fifth Edition, Mc Graw Hill. (2009)

[8] Frank, M.; Zotter F.: „Spatial impression and directional resolution in the reproduction of reverberation“ Fortschritte der Akustik - DAGA 2016, Aachen (DAGA 2016) (2016)

[9] Kronlachner, M.: „Plug-in Suite for Mastering the Production and Playback in Surround Sound and Ambisonics“ AES Students Design Competition (2014)

[10] Listen HRTF Database: http://recherche.ircam.fr/equipes/salles/listen/index.html File-ID: IRC52 measured by G. Vandernoot (2002)

[11] Pätynen, J.; Pulkki, V.; Lokki, T.: Anechoic recording system for symphony orchestra“, Acta Acustica united with Acustica, 94, 856-865. (2008)

[12] Pujahari, A.: „Towards Automatic Reverb Addition for Production Oriented Multi-Track Audio Mixing“. Masters Thesis. TU Delft. (2017)

[13] Romanov, M.; Berghold, P.; Rudrich, D., Zaunschirm, M.; Frank, M., Zotter, F.: „Implementation and Evaluation of a Low-cost Head-tracker for Binaural Synthesis“ AES Convention Paper (2017)

[14] Rumsey, F.: „Spatial Audio“. Preface. Oxford: Focal Press. (2005)

[15] Schacher, J. C.; Kocher, P.: „Ambisonics Spatialzation Tools for Max/Msp“ (2016)

[16] Schmitz, O.; Feistel, S.; Ahnert, W.; Vorländer, M.: "Software for sound reinforcement system and for room acoustics“, AES Conference Paper (2001)

[17] Sèdes A.; Guillot P.; Paris, E.: „The HOA library, review and prospects.“ International Computer Music Conference | Sound and Music Computing, Sep 2014, Athènes, Greece. Proceedings ICMC|SMC, pp.855 -860, 2014, Proceedings ICMC|SMC. (2014)

[18] Stecker, A.: „Schwimmen im Klang“, Frankfurter Rundschau, URL http://www.fr.de/rhein-main/wie-man-frankfurt-lieben-lernt-schwimmen-im-klang-a-1179979 (2008)

[19] Tervo, S.; Pätynen, J.; Kuusinen A.; Lokki T.: „Spatial Decomposition Method for Room Impulse Responses“ In Journal of the AES, Vol. 61, No. 1/2. (2013)

[20] Vorländer, M.: „Simulation of the transient and steady-state sound propagation in rooms using a new combined ray-tracing/image-source algorithm“, In J. Acoust. Sol. Am. 86, S. 172. (1989)

[21] Weinzierl, S.: „Handbuch der Audiotechnik“, Berlin: Springer. (2008)

[22] Wiggins, B.; Dring, M.: „AmbiFreeverb 2 - Development of a 3D Ambisonis Reverb with Spatial Warping and Variable Scattering“ AES Conference Paper (2016)

Die Bedeutung von 3D bei Immersive Audio

L. Nipkow[1], G. Zielinsky[2], T. Ammermann[3]

[1] *Silent Work GmbH, Zürich, Switzerland, Email: lasse.nipkow@silentwork.com*

[2] *Dipl. Tonmeister, Hannover, Germany, Email: g.zielinsky@me.com*

[3] *New Audio Technology, Hamburg, Germany, Email: tom@newaudiotechnology.com*

Abstract

Immersive Audio steht für eine Reihe verschiedener Audioformate und Betrachtungsweisen, darunter Auro 3D, Dolby Atmos, Ambisonics und VR / AR. 3D-Audio ist aus geometrischer Sicht viel mehr als nur Surround + Height oder das Verschieben von Audioobjekten im 3D-Raum. Es ist die konsequente Erweiterung der Dimensionen: von Mono (Punkt) und Stereo (Linie) über Surround Sound (Ebene) bis 3D (Volumen). Das Ergebnis ist eine atemberaubende Räumlichkeit sowie eine Natürlichkeit des Instrumentenklanges, wie sie in Stereo nicht möglich ist.

Lasse Nipkow, Gregor Zielinsky und Tom Ammermann präsentieren in diesem Beitrag ihre über viele Jahre erarbeiteten Erkenntnisse und Ergebnisse mit 3D-Audio. Sie erklären, wie das Potential mittels Lautsprecher- und Kopfhörerwiedergabe weitgehend ausgeschöpft werden kann.

1. Einführung

Was ist 3D-Audio? Es gibt verschiedene Bezeichnungen wie 2+2+2, 13.1, 7.1.4, Headphone 3D etc. In den meisten Fällen werden damit die Anzahl Lautsprecher angegeben und deren Anordnung im Wiedergaberaum.

Aufgrund von bestehenden Lautsprecheranordnungen mit teils weit über 10 Lautsprechern und binauralen Systemen zur Virtualisierung von 3D-Audio sind in den vergangenen Jahren eine Vielzahl an objektorientierten Richtungsmischer auf den Markt gekommen. Die Objekte werden oftmals als Mono-Schallquellen behandelt. Eine Ansammlung von Mono-Punkten im 3D-Raum entspricht alleine allerdings nicht 3D-Audio.

Eine der grössten Errungenschaften von 3D-Audio ist eine eindrückliche Umhüllung mit einem realistisch wirkenden Raumklang. Darüber hinaus können Schallquellen viel natürlicher klingen als bei Stereo oder Surround Sound. 3D-Audio ist deshalb wesentlich mehr als Surround + Height, wie es manchmal auch genannt wird.

2. Abbildung

Wie bei Stereo gibt es bei 3D-Audio generell zwei sich gegenüberstehende Komponenten: Abbildung und Räumlichkeit. Bei der Abbildung in der Horizontalen wird zwischen Real- und Phantomschallquellen unterschieden, wobei Phantomschallquellen eine wesentlich grössere Ausdehnung aufweisen als Realschallquellen [1]. Diese klingen im Vergleich zu Phantomschallquellen wesentlich klarer und haben daher eine markant höhere Durchdringungskraft. Darüber hinaus sind sie immer aus der Richtung des betreffenden Lautsprechers zu hören. Realschallquellen führen somit zum maximal möglichen Sweet Spot.

Phantomschallquellen können bewegt werden, Realschallquellen aufgrund ihrer fixen Position nicht. Eine Bewegung führt zu erhöhter Aufmerksamkeit des Hörers [2]. Bewegungen von Phantomschallquellen können über mehrere Lautsprecher im Raum erfolgen wie z.B. eine Kreisbewegung oder aber zwischen zwei Lautsprechern wie z.B. eine Pegelmodulation, spektrale Bewegungen etc. [3].

Phantomschallquellen können sowohl in der Horizontalen als auch in der Vertikalen erzeugt werden, wobei die Lokalisationsschärfe des menschlichen Gehörs in der Vertikalen lediglich ein Drittel im Vergleich zur Horizontalen ist [4]. Dafür ist die Ausdehnung der vertikalen Phantomschallquelle in der Horizontalen gleich wie bei der Realschallquelle und somit sehr gering [5].

2.1. Panning von Direktschall

Die einfachste Form von Panning eines Mono-Signals ist zwischen zwei Lautsprechern wie L-R. Dies kann mit dem Panner der Mischeinheit in der benutzten Digital Audio Workstation (DAW) umgesetzt werden. Dasselbe gilt für Panning bei einem 9.1 Lautsprecher-Setup zwischen LS-RS,

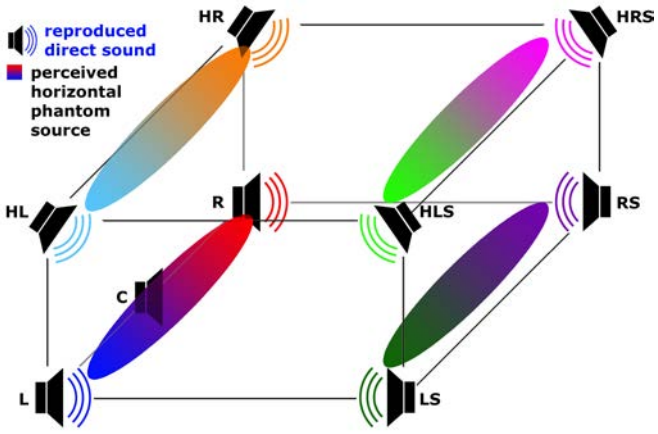


Abb. 1: Schematische Darstellung eines 9.1 Lautsprecher-Setups. Statische, horizontale Phantomschallquellen werden sinnvollerweise lediglich zwischen den angegebenen Lautsprechern erzeugt. Statische Phantomschallquellen zwischen vorne und hinten sind mit nach vorne gerichtetem Kopf nicht stabil lokalisierbar.

ISBN
978-3-9812830-9-9

Der Autor hält die Urheberrechte für den gesamten Inhalt seines Manuskripts. Er überlässt dem VDT das Recht auf Veröffentlichung in gedruckter Form, Online oder als PDF-Download von https://tonmeistertagung.com

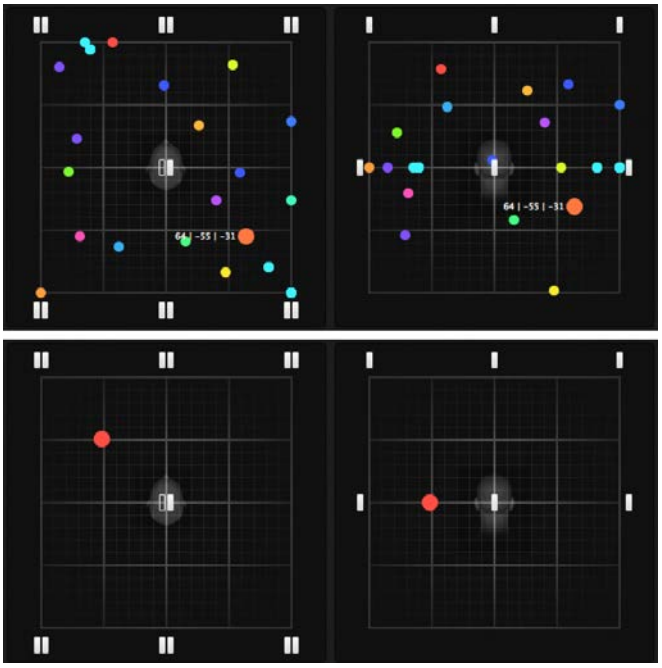


Abb. 2a + 2b: Screenshots des SAD Panning-Tools von New Audio Technologies. Es lassen sich an beliebigen Positionen im Raum Schallquellen positionieren, siehe oben. Dies entspricht aber nicht dem, was als Hörereignis resultiert. Die untere Grafik zeigt ein Beispiel, bei dem vier Lautsprecher für die Phantomschallquelle involviert sind: L/C/LS/CS. Resultat ist eine Phantomschallquelle mit einer sehr grossen räumlichen Ausdehnung.

HL-HR und HLS-HRS, siehe **Abb. 1**.

Für andere Positionen und insbesondere Bewegungen ist es ratsam, mit einem Panning-Tool zu arbeiten wie z.B. dem SAD von New Audio Technologies. Dieses Tool zeigt mittels grafischem Userinterface Schallquellen als farbige Punkte im 3D-Raum an und kreiert Lautsprechersignale, die am ehesten den dargestellten Punkten entsprechen, siehe **Abb. 2a**. Mittels Automatisation können programmierte Bewegungen der Schallquellen realisiert werden, was die Arbeit markant erleichtert. Stereo-Signale werden durch zwei Punkte repräsentiert, welche für Bewegungen miteinander gekoppelt werden.

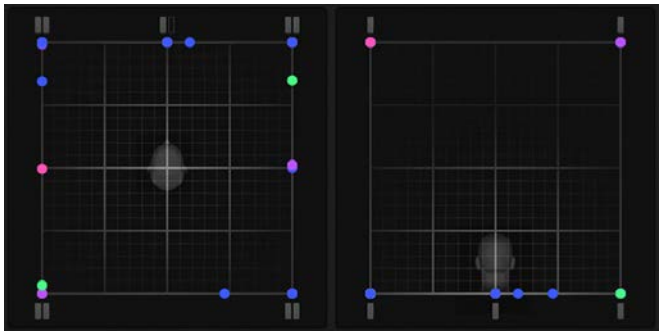


Abb. 3: Screenshot des SAD Panning-Tools von New Audio Technologies. Werden Schallquellen an den Rändern des 3D-Raums angeordnet, sind weniger Lautsprecher für die Abbildung involviert. Dies führt zu einer geringeren Ausdehnung der Phantomschallquelle. Die Position in einem Lautsprecher bedeutet die Kreation einer Realschallquelle; diese hat die kleinstmögliche Ausdehnung zur Folge.

Die freie Positionierung von Schallquellen im Raum ist, wie oben erwähnt, aus grafischer Sicht sehr einfach in einem Tool zu realisieren. Dies entspricht allerdings nicht dem, was als Hörereignis resultiert. **Abb. 2b** zeigt ein Beispiel, bei dem 4 Lautsprecher für eine statische Phantomschallquelle involviert sind: L/C/LS/CS. Weil 4 anstatt üblicherweise 2 Lautsprecher für die Erzeugung der Phantomschallquelle beteiligt sind, ergibt sich eine wesentlich grössere räumliche Ausdehnung. Darüber hinaus wird die Phantomschallquelle nicht genau an jenem Punkt wahrgenommen, wo sie zu liegen scheint: Statische, gut lokalisierbare Phantomschallquellen können mittels einem 9.1 Lautsprechersetup nicht innerhalb des 3D-Raumes positioniert werden sondern lediglich zwischen Lautsprechern auf den Aussenflächen des 3D-Raumes. Daher kann eine statische, gut lokalisierbare Phantomschallquelle als Projektion auf der Aussenfläche aufgefasst werden, Vergleich **Abb. 3**. Statische Phantomschallquellen, welche den Hörraum durchkreuzen, führen zu widersprüchlichen Hörereignissen; dies sollte deshalb vermieden werden [5]. Wird mittels des Tools ein farbiger Punkt an Ort eines Lautsprechers positioniert, entsteht eine Realschallquelle d.h. eine Schallquelle mit maximal möglicher Lokalisationsschärfe. Es ist zu beachten, dass seitlich positionierte Phantomschallquellen für einen Hörer mit nach vorne gerichtetem Kopf nicht stabil lokalisierbar sind [6].

2.2. Direktschall aus verschiedenen Richtungen

Bei 3D-Audio Aufnahmen klassischer Musik werden in der Regel Instrumente ausschliesslich in der Front abgebildet, wie dies bei Konzerten aufgrund der konventionellen Aufstellung der Musiker auf der Bühne der Fall ist. Ausnahmen bilden Werke mit Instrumenten, die ausserhalb des Raumes oder z.B. hinten im Raum gespielt werden. Orgeln sind teilweise so aufgebaut, dass ihre Register im Raum verteilt sind. Ein ausgeprägtes Beispiel stellt die Grosse Hoforgel der Hofkirche Luzern, CH dar. Sie hat drei Orgelteile: grosses Gehäuse (inkl. Rückpositiv), Fernwerk und Echowerk, siehe **Abb. 4**.

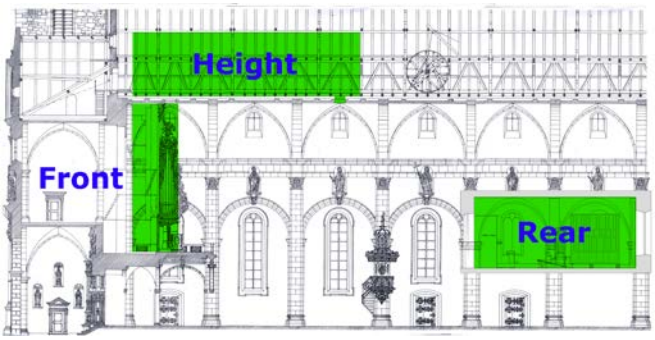


Abb. 4: Plan der Hofkirche Luzern (Seitenansicht). Die drei Teile grosses Gehäuse inkl. Rückpositiv (links), Fernwerk (oben) und seit 2015 das Echowerk (rechts) bilden zusammen die Orgel und sind grün eingezeichnet. Bei einer 3D-Audio Aufnahme wird sinngemäss der Direktschall aus der Front, von oben und hinten abgestrahlt.

Dieses Instrument kann als eine Schallquelle betrachtet werden, die um den Zuhörer gebaut ist. Wolfgang Sieber,

der Organist der Hofkirche Luzern, komponiert Werke für diese Orgel und spielt ganz bewusst mit der räumlichen Verteilung des Instruments. Durch seine jahrelange Erfahrung mit dem Orgelspiel in diesem Raum beherrscht er wie kein anderer, die Registrierung so zu wählen, dass für die Zuhörer in der Kirche eine imposante klangliche Wirkung entsteht.

Abb. 5: Mikrofonierung des Grossen Gehäuses in der Hofkirche Luzern, CH (Prospektpfeifen Mitte). Aufgrund der kleinen Distanz zum Instrument wird fast ausschliesslich Direktschall eingefangen (Klein-AB Phantomschallquellen für die Lautsprechersignale L-R).

In der Hofkirche Luzern sind für 3D-Audio Orgel-Aufnahmen eine Vielzahl der Register mit Klein-AB Stereo-Stützmikrofonen versehen, siehe **Abb. 5**. Dies führt aufgrund der geringen Stereobasis zu einer schwach ausgeprägten Abbildung in der Mitte; der Klang der Register ist aber stereofon, vergleich Stereo Synthi-Sound [3]. Mittels Pegeldifferenzen kann die Phantomschallquelle nach links und rechts verschoben werden.

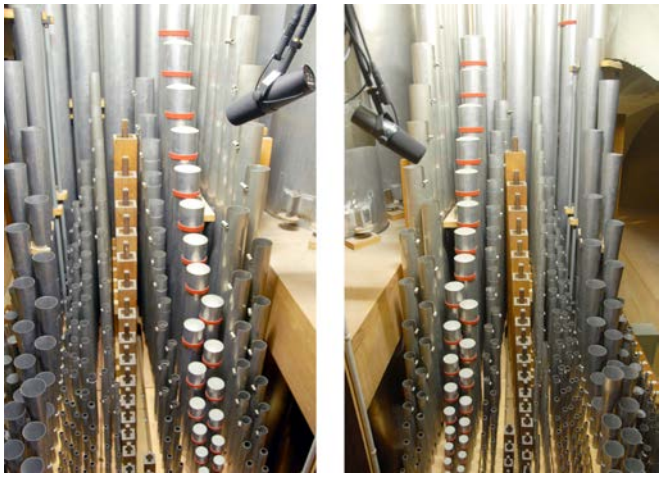


Abb. 6: Das Register ‚Oberwerk' des Grossen Gehäuses in der Hofkirche Luzern, CH. Dieses Register ist in zwei Teile aufgeteilt. Diese befinden sich in zwei Kammern, die akustisch voneinander getrennt sind und die eine links und die andere rechts in der Orgel steht (Ping-Pong-Stereofonie für die Lautsprechersignale HL-HR).

Die Signale der Stützmikrofone werden entweder für L-R, LS-RS, HL-HR oder HLS-HRS Lautsprecher-Paare verwendet, vergleich **Abb. 1**. Aufgrund der geringen Distanz zwischen den Orgelpfeifen und den Mikrofonen ergibt sich für die Stützmikrofone sehr geringes Übersprechen. Dies hat wiederum zur Folge, dass sich keine unerwünschten Phantomschallquellen zwischen vorne und hinten sowie zwischen unten und oben ausbilden. Weiter ist der Anteil früher Reflexionen dadurch gering, was zu einer hohen Klarheit der mikrofonierten Schallquelle führt.

Einige Register wie das Récit und das Oberwerk sind in zwei Hälften aufgeteilt und räumlich verteilt d.h. die eine Hälfte auf der linken Seite des Grossen Gehäuses und die andere Hälfte auf der rechten Seite, siehe **Abb. 6**. Die Töne hüpfen bei einer chromatischen Tonleiter alternierend zwischen links und rechts. In jenen Registern kommen Mono-Stützen zum Einsatz, welche als Ping-Pong-Stereo direkt den Lautsprechern L-R bzw. HL-HR zugeordnet werden. Somit werden die Register aus Richtung der Lautsprecher lokalisiert, wie es in etwa dem natürlichen Hören in der Hofkirche entspricht. Die Wahl von Ping-Pong-Stereo jener Register führt zu Realschallquellen, welche den hohen Orgelregistern einen äusserst transparenten Klang verleihen.

Der Hörer wird aufgrund der gewählten Mikrofonierung und der sinngemässen Zuordnung der Signale zu den Lautsprechern im Wiedergaberaum aus verschiedenen Richtungen mit Direktschall versorgt; dies führt zu einer intensiv wirkenden Umhüllung. Aufgrund der Klangeigenschaften der Orgel und der Komposition von Wolfgang Sieber fällt der Direktschall von hinten jedoch nicht negativ auf, vergleich [5].

2.3. Seitliche Lautsprecher

Bei manchen Lautsprecher-Setups werden seitliche Lautsprecher (±90°) eingesetzt. Bei 7.1 werden diese Surround L bzw. Surround R genannt. Einseitig seitliche Beschallung eines Hörers mittels Direktschall führt zu hohen Pegeldifferenzen bei hohen und mittleren Frequenzen. Dies ist vergleichbar mit einseitiger Kopfhörerwiedergabe, was je nach Lautstärke unangenehm bis schmerzhaft ist.

Wird Direktschall wie Stimmen oder statische Klänge von beiden seitlichen Lautsprechern wiedergegeben, ist der Sweet Spot auf die Symmetrieachse reduziert. Verschiebt sich der Hörer leicht nach links, springt das Hörereignis zum linken Seitenlautsprecher, da es lauter und gleichzeitig früher zu hören ist als auf der rechten Seite und umgekehrt. Dieser Fall sollte demzufolge möglichst vermieden werden.

Seitliche Lautsprecher eignen sich aber für die Reproduktion von kurzen Direktschallsignalen mit geringer bis mittlerer Lautstärke wie z.B. Atmo-Sounds oder perkussiven Sounds, die nicht aufdringlich klingen.

3. Räumlichkeit

Räumlichkeit bildet sich aus, wenn zwei oder mehrere Lautsprechersignale mit ähnlichem Inhalt wie Raumschall bei sehr tiefer Korrelation abstrahlen. Da diese Signale das Gegenstück zu Phantomschallquellen (Korrelation zwischen den Kanälen = 1) darstellen, ergibt sich durch Räumlichkeit keine Abbildung.

3.1. Sound Content Correlation

Nicht in jedem Fall führen Signale mit einer Korrelation um Null herum, die von mehreren Lautsprechern wiedergegeben werden, zu einem Gefühl von räumlichem Eindruck. Wenn z.B. mehrere Stimmen verschiedener Sprecher als Lautsprechersignale verwendet werden, wird jede von ihnen als eine einzelne lokalisierte Klangquelle aus dem jeweiligen Lautsprecher wahrgenommen. Dies bedeutet, dass unkorrelierte Signale, die keinen ähnlichen Klang haben, nicht zu einem räumlichen Eindruck führen.

Raumschall hat seinen Ursprung in Schallquellen, die sich im Aufnahmeraum befinden. Aufgrund der gleichmässigen Verteilung des Raumschalls klingt mikrofonierter Raumschall von Mikrofonen an verschiedenen Positionen im Aufnahmeraum sehr ähnlich und ist bei sequenzieller Wiedergabe praktisch nicht voneinander zu unterscheiden. Diese Eigenschaft von Signalen wird als Sound Content Correlation bezeichnet. Es ist zu beachten, dass die klangliche Ähnlichkeit von Signalen alleine nicht ausreicht, um Räumlichkeit zu erzeugen: Je impulsartiger die Signale sind, desto genauer muss die Zeitstruktur übereinstimmen; ansonsten zerfällt das Stereo-Signal in zwei unabhängige Mono-Signale [7].

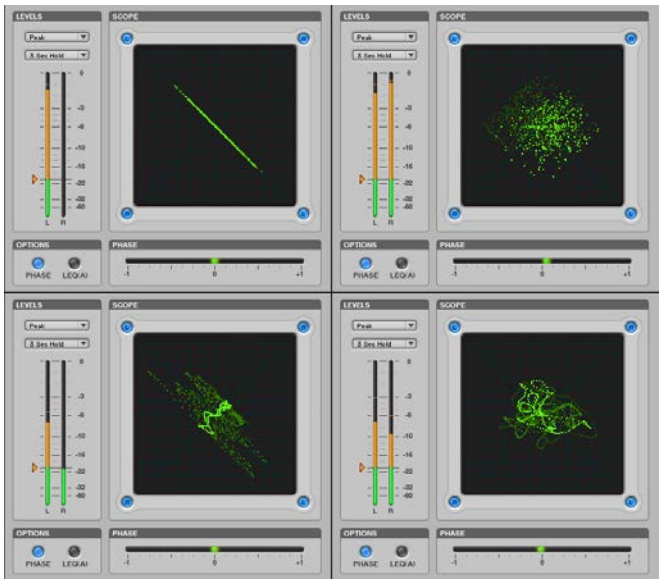


Abb. 7a..d: Screenshots des Phasescopes von Digidesign. Die Korrelation aller dargestellten Signale ist um Null herum (v.l.n.r.): rosa Rauschen lediglich auf dem linken Kanal, zwei unabhängige Instanzen rosa Rauschens, zwei komplett unabhängige Signale und Raumsignale von Mikrofonen mit ca. 10..20m Distanz zueinander.

Signale, welche eine Korrelation um Null herum haben, können, wie oben beschrieben, unterschiedliche Eigenschaften haben. Die Korrelation kann keine eindeutigen Aufschlüsse darüber geben, ob die beiden Signale ähnlich sind oder komplett unabhängige Signale darstellen. Mittels eines Phase Scopes werden Unterschiede der Signale sichtbar. Bei Signalen mit Sound Content Correlation ist die Verteilung über die Fläche des Scopes gleichmässig bzw. zufällig; es ist keine Ausprägung der beiden Kanäle sichtbar, siehe Raumschall in **Abb. 7d.** Das Bild gleicht dem Spezialfall von stereofonen rosa Rauschen, das mittels zwei unabhängigen Rauschgeneratoren erzeugt wird, siehe **Abb. 7b**. Bei zwei vollkommen unabhängigen Signalen ist im Phase Scope eine Überlagerung des einen durch den anderen Kanals erkennbar, siehe **Abb. 7c**.

3.2. Umhüllung

Wird aus allen Lautsprechern eines 9.1 Lautsprecher-Setups Nachhall mit einer Korrelation um Null herum bei etwa gleich hohem Pegel wiedergegeben, führt dies zu einer Empfindung von Umhüllung. Der Hörer hat innerhalb des Hörvolumens den Eindruck, sich im Aufnahmeraum zu befinden, siehe **Abb. 8**.

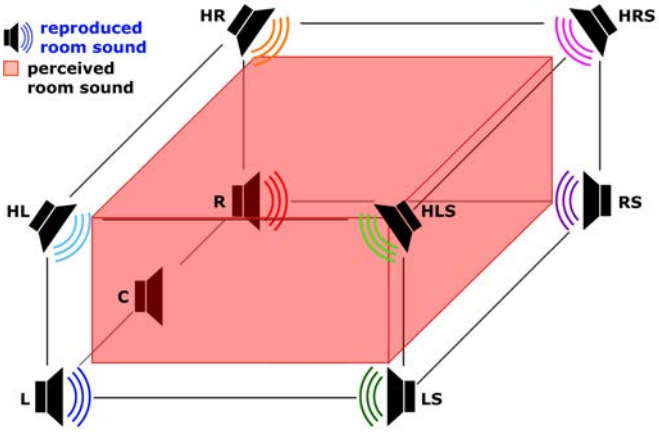


Abb. 8: Der wiedergegebene Raumschall zwischen den vorderen und hinteren Lautsprechern sollte bezüglich Pegel ausbalanciert sein, um die hörbare Verbindung zwischen beiden Seiten sicherzustellen. In diesem Fall kann ein Hörer fast überall im Setup platziert werden und hat den Eindruck, sich im Aufnahmeraum zu befinden.

Das empfundene Mass von Umhüllung ist stark von der Korrelation der Signale abhängig: Eine Korrelation um 0 führt bei sinnvoller Signalwahl zu einem sehr eindrücklichen Hörerlebnis. Anders beschrieben, entsteht eine akustische Verbindung zwischen den jeweils benachbarten Lautsprechern, so dass die einzelnen Lautsprecher als Schallquellen selber nicht mehr wahrgenommen werden [5]. Eine Korrelation von 1 d.h. aus allen Lautsprechern identische Signale führt dagegen in den meisten Fällen, insbesondere mit statischen Signalen wie Rauschen, zu einem unangenehmen, unnatürlich wirkenden Hörereignis.

3.3. Seitliche Hörereignisse

Wie im Kapitel 2.1. beschrieben, können entlang einer Seitenkante keine stabil lokalisierbare stationäre Phantomschallquellen erzeugt werden. Dagegen kann mittels stereofonen Signalen mit Sound Content Correlation eine akustische Verbindung entlang einer Seitenkante erzeugt werden z.B. zwischen L und LS, siehe **Abb. 9**. Die Verbindung kann dadurch verstärkt werden, indem auch die oberen Kanäle HL und HLS herbeigezogen werden; dadurch entsteht eine akustische Verbindung entlang der linken Seitenfläche.

Somit können Hörereignisse auch von der Seite her wahrgenommen werden, wenn auch sie nicht als Schallquellen lokalisierbar sind. Vielmehr werden sie im Vergleich zu

Phantomschallquellen als feine Bewegungen in Form von Schwebung entlang der Seitenkante bzw. Seitenfläche wahrgenommen.

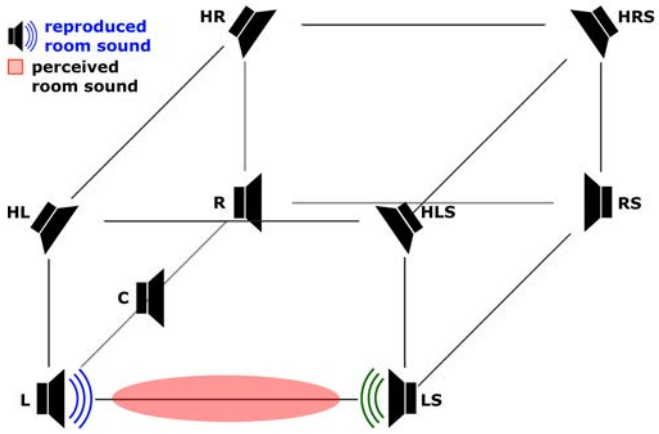


Abb. 9: Mittels eines stereofonen Signals mit einer Sound Content Correlation kann eine akustische Verbindung zwischen vorne und hinten erzeugt werden. Somit ist es möglich, stabile Hörereignisse entlang einer Seitenkante zu erzeugen; es handelt sich allerdings nicht um Phantomschallquellen.

3.4. Bewegung und Räumlichkeit

Vielfalt spielt eine wichtige Rolle für die Räumlichkeit. Ein einfaches Hallgerät erreicht nur eine begrenzte Räumlichkeit. Sehr komplexe Nachhallgeräte wie z.B. ein Faltungshall mit langen Impulsantworten erzielen eine bessere Leistung. Bewegungen können auf verschiedene Arten umgesetzt werden: Pegeländerungen oder Pegelmodulation, Spektraländerungen, Überlagerung verschiedener Sounds etc.

Wird Räumlichkeit in Bewegung gebracht, kann das unter Umständen zu einer Mischform von Räumlichkeit und Abbildung führen, da Bewegungen von Schallflächen d.h. abgrenzbare Bereiche von Raumschall vom Hörer als Objekte interpretiert werden.

4. Vertikale Stereofonie

Bei Stereo steht lediglich die 1. Dimension zur Verfügung: eine Verbindunglinie zwischen links und rechts. Betrachtet man jene Linie auf einer Fläche zwischen unten L-R und oben HL-HR, ist Stereo in Bezug auf die vertikale Ausrichtung Mono. Wie in Kapitel 2 beschrieben, ist die Lokalisation senkrechter Phantomschallquellen durch den Hörer stark eingeschränkt. Die vertikale Räumlichkeit wird dagegen sehr deutlich vom Hörer wahrgenommen [8].

4.1. 5-CH Stereo aus der Front

Frühe Reflexionen können als eine Sonderform von Räumlichkeit angesehen werden. Im Zusammenhang mit Direktschall kann ein Entfernungseindruck von Schallquellen erzeugt werden [9]. Dies erfolgt, wenn Direktschall und frühe Reflexionen aus derselben Richtung abgestrahlt werden. Wird Direktschall vorwiegend von den unteren Lautsprechern in der Front L-C-R und frühe Reflexionen von den oberen Lautsprechern in der Front HL-HR wiedergegeben, ergibt sich eine hörbare Verbindung der unteren und oberen vorderen Ebene; dies kann als Projektion des Klangkörpers in der Front beschrieben werden, siehe **Abb. 10**.

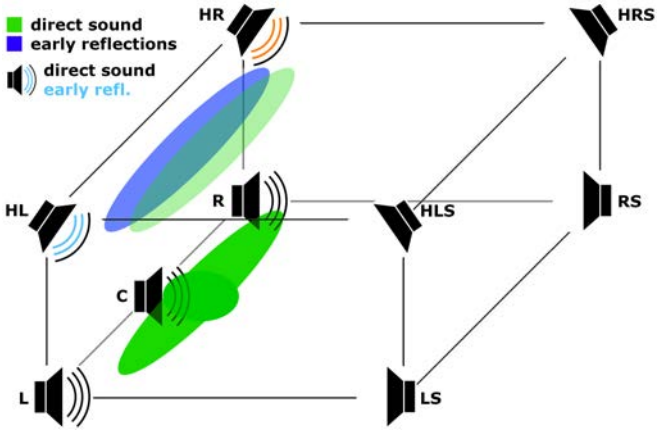


Abb. 10: Wird Direktschall vorwiegend von den unteren Lautsprechern in der Front L-C-R und frühe Reflexionen von den oberen Lautsprechern in der Front HL-HR wiedergegeben, ergibt sich eine Projektion des Klangkörpers in der Front.

Die Instrumente werden dabei vorwiegend unten lokalisiert, währenddem sie natürlicher klingen als ohne den frühen Reflexionen; es entspricht in etwa den Gegebenheiten in einem Konzertsaal in Bühnennähe.

4.2. Vertikal abgestrahlter Direktschall

Wenn stereofoner Direktschall mit Sound Content Correlation von einem der unteren und gleichzeitig oberen Lautsprechern abgestrahlt wird z.B. L-HL, führt dies zu einem ähnlichen Effekt der Natürlichkeit, siehe **Abb. 11**.

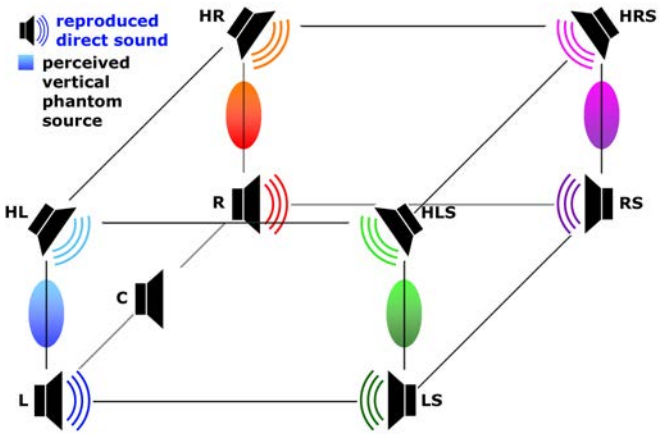


Abb. 11: Wenn Stereo-Paare der vier senkrechten Kanten des 9.1 Lautsprecher-Setups gebildet und Instrumente / Sounds auf diese Weise zugeordnet werden, wird eine hohe Durchhörbarkeit erreicht. Diese ist gleich hoch, wie wenn Realschallquellen bei einer Surround Sound Wiedergabe verwendet würden.

Die Abbildung ist in diesem Fall nicht eindeutig unten sondern befindet sich zwischen den beiden beteiligten Lautsprechern. Bemerkenswert ist bei diesem Fall die Lokalisationsschärfe in der Horizontalen, die derjenigen

einer Realschallquelle entspricht [5]. Die Kombination von sehr scharfer Lokalisation sowie gleichzeitig stark wirkenden Natürlichkeit des Instruments bzw. eindrücklichen Klang eines synthetischen Sounds führt dazu, dass die Schallquelle als sehr unmittelbar-direkt empfunden wird und somit die Aufmerksamkeit auf sich zieht. Dies gilt insbesondere für Schallquellen mit hohem Anteil hoher Frequenzen.

4.3. Kombination von Abbildung und Räumlichkeit

Die Abbildung einer Schallquelle erfolgt am präzisesten, wenn lediglich Direktschall beteiligt ist und dieser von einer Realschallquelle abgestrahlt wird. Werden mehrere Realschallquellen gleichzeitig eingesetzt, liegen diese als Punktschallquellen im 3D-Raum, ohne eine hörbare Verbindung zwischen einander zu haben. Es entsteht für den Hörer den Eindruck, Löcher zwischen den Schallquellen zu empfinden. Noch ausgeprägter wird diese Wirkung, wenn aus den Realschallquellen zusätzlich stark färbende frühe Reflexionen abgestrahlt werden. Senkrechte Phantomschallquellen verhalten sich bezüglich Abgrenzung im 3D-Raum aufgrund ihrer geringen horizontalen Ausdehnung sehr ähnlich wie Realschallquellen.

Räumlichkeit erzeugt dagegen eine hörbare Verbindung zwischen den beteiligten Lautsprechern. Je mehr Dimensionen beteiligt sind, desto plastischer wirkt die Räumlichkeit. Bei einer 3D-Audio Wiedergabe entsteht für den Hörer der Eindruck, sich im Aufnahmeraum zu befinden. Horizontale Phantomschallquellen füllen aufgrund ihrer ausgeprägten Ausdehnung die oben beschriebenen wahrnehmbaren Löcher aus. Somit ist eine geringe Anzahl überlagerter horizontaler Phantomschallquellen für eine 3D-Audio Mischung förderlich.

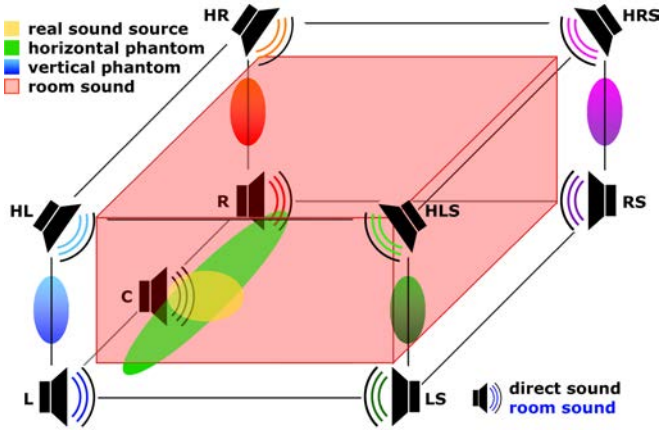


Abb. 12: Bei sehr eindrücklich wirkenden Aufnahmen werden zwei Komponenten miteinander kombiniert: sehr scharf lokalisierbare Schallquellen und verbindende Elemente wie Räumlichkeit und horizontale Phantomschallquellen. In der Grafik sind fünf scharf lokalisierbare Richtungen dargestellt: Center (Realschallquelle) und vier senkrechte Kanten (AB-Phantomschallquellen). Verbindende Elemente sind Raumschall, der zum Eindruck von Umhüllung führt, sowie horizontale Phantomschallquellen in der Front und hinten (in dieser Grafik lediglich zwischen L-R dargestellt).

Die Kombination von Abbildung und Räumlichkeit führt zum gleichzeitig vorliegenden Eindruck von scharfer Abbildung und einer hörbaren Verbindung zwischen den Schallquellen, siehe **Abb. 12**. Bei günstig gewählten Signalen führt dies zu einem sehr eindrücklichen Klangerlebnis. Ein Beispiel dafür sind Aufnahmen der Grossen Hoforgel in der Hofkirche Luzern wie die Produktion ‚Sieberspace IV'. Die verschiedenen Register werden als Kombination von Realschallquellen aus den Richtungen der vorliegenden Lautsprechern eines 9.1 Lautsprecher-Setups und stereofonen Phantomschallquellen wiedergegeben. Die Phantomschallquellen unterstützen dabei die Wirkung, keine Löcher in der Front und hinten zu empfinden, obwohl keine Mikrofonsignale für den Centerkanal vorliegen. Der Raumklang wird von Mikrofonen eingefangen, die sehr grosse Abstände zueinander haben und somit bis zu tiefsten Frequenzen eine Korrelation um Null herum aufweisen, siehe **Abb. 13**.

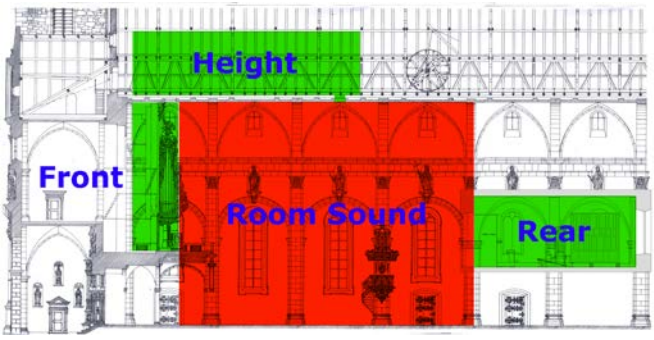


Abb. 13: Plan der Hofkirche Luzern, CH (Seitenansicht). Die drei Teile grosses Gehäuse inkl. Rückpositiv (links), Fernwerk (oben) und seit 2015 das Echowerk (rechts) bilden zusammen die Grosse Hoforgel und sind grün eingezeichnet und geben Direktschall wieder. Der aus dem Direktschall resultierende Raumklang (rot) wird mittels Mikrofonen in den seitlichen Kirchenschiffen sowie von der Decke eingefangen.

5. Upmix

5.1. L-R → L-C-R Konversion für den Center

Der Center-Kanal hat bei Surround Sound und 3D-Audio Produktionen eine wichtige Bedeutung: Er unterstützt die Stabilität der Abbildung und führt zu einem grösseren Sweet Spot. Prominente Schallquellen wie Stimmen und Solo-Instrumente wie auch Kick und Snare werden bei 2-Kanal Stereo in der Regel als Phantomschallquellen in der Mitte zwischen L-R positioniert. Mitten-Signale einer Phantomschallquelle sind allerdings bezüglich Abbildung nicht stabil; sie befindet sich in etwa vor dem Hörer auf der Verbindungslinie der beiden Lautsprecher L-R [5]. Befindet sich ein Hörer nicht auf der Symmetrieachse der beiden Lautsprecher L-R, wird die Abbildung verzerrt, siehe **Abb. 14**. Somit wird der Sweet Spot auf die Symmetrieachse reduziert.

Bei Aufnahmen akustischer Musik ist es nicht immer möglich oder sinnvoll, ein Center-Signal mittels Mikrofonen zu gewinnen. Auf der einen Seite erzeugen, abgesehen vom OCT, die meisten vorgeschlagenen Surround Sound Hauptmikrofonsysteme zu hohes Übersprechen zwischen den Front-Kanälen. Auf der anderen Seite können Stereo-Stützsignale nicht ohne weiteres in den Center-Kanal eingefügt

werden; das wäre eine Reduktion von Stereo (1. Dimension) auf Mono (0. Dimension). Es ist allerdings keine sinnvolle Lösung, deswegen den Centerkanal wegzulassen, oder gar unsinnige Dinge vorzunehmen, wie dem Center-Kanal eine Monomischung aus L-R zuzuführen.

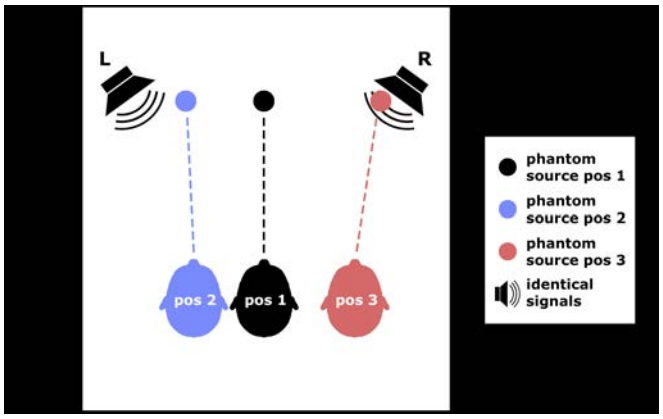


Abb. 14: Im Fall von zwei Lautsprechern und identischen Signalen wird der Hörer die Phantomschallquelle unabhängig von seiner Position in etwa vor ihm wahrnehmen. Sobald sich der Hörer näher an einem Lautsprecher befindet, zieht es die Phantomschallquelle zum näheren Lautsprecher (Präzedenz-Effekt). In dieser Abbildung sind drei verschiedene Positionen von demselben Hörer dargestellt. Die Lautsprechersignale sind in jedem Falle dieselben.

Eine Möglichkeit besteht darin, aus einem L-R Signal mittels Signalprocessing ein L-C-R Signal zu erzeugen, wie dies der Illusonic Processor der IAP-Serie macht. Dabei werden, grob beschrieben, Mitten-Anteile des Stereo-Signals aus dem L-R-Signal gewonnen, klanglich optimiert und als Center-Signal ausgegeben. Die Mitten-Signale werden aus dem L-R-Signal entfernt. Das heisst, durch Processing wird von der Stereo-Mischung (Hauptmikrofon plus Stützen und allfällige Stereo-Effekte wie zusätzlicher Hall) ein optimales L-C-R Signal erzeugt. Die Stabilität des Resultats ist verblüffend, denn anstelle der Phantomschallquelle tritt an Ort der Mitten-Signale bzw. des Center-Lautsprechers eine Realschallquelle auf.

5.2. Raumschall-Upmix

Ähnlich wie beim Center-Signal kann es bei der Mikrofonierung von Raumschall zu Problemen kommen. Die möglichen Gründe sind vielfältig: zu wenige Raum-Mikrofone vorhanden, limitierte Kanalzahl, fehlender Platz, ungenügende Raumakustik, Aussenlärm etc. Die Herausforderung beim Processing ist der Umstand, dass der bestehende Raum auf der Aufnahme nicht durch einen anderen, stark abweichenden Raumklang überlagert wird, wie dies bei einem Raumsimulator mittels Impulsantworten oder auch herkömmlichen Hallgeräten der Fall sein kann.

Eine clevere Lösung besteht darin, den bestehenden Raumklang einer Aufnahme zu unterdrücken bzw. bei der Aufnahme zu vermeiden, und durch einen eigenen, optimalen Raumklang zu ersetzen.

Raumklang wird am besten mittels aller vorhandenen Lautsprecher reproduziert. Wie bei echtem Raumklang sollten die Signale aus den verschiedenen Richtungen bzw. Lautsprechern eine sehr tiefe Korrelation aufweisen. Es wird ein hochwertiger Multikanal Hall oder Akustik-Raum-Modell benötigt. Frühe Reflexionen hängen richtungsmässig von der Richtung der Ursprungssignale ab. Später Hall sollte diffus sein, d.h. unabhängige Signale aus allen Richtungen, die wenn einzeln abgehört, gleich klingen, vergleich Sound Content Correlation. Mittels Signalverarbeitung, wie dies der Illusonic Processor vornimmt, wird die Stereo-Aufnahme leicht enthallt (optional) bevor aus dem trockenen Stereo-Signal direktschallarme, schwach korrelierende und schwach verfärbende Raumschallsignale für die vorhandene Anzahl Kanäle erzeugt werden.

6. Binaurale Lautsprecher-Virtualisierung

Kopfhörerwiedergabe hat bei den Konsumenten mit dem Aufkommen von Smartphones und Tablets markant zugenommen. Daher ist es naheliegend, 3D-Audio Produktionen auch für Kopfhörer zur Verfügung zu stellen. Die Umsetzung ist allerdings nicht trivial: Bei Kopfhörern stehen lediglich zwei Kanäle zur Verfügung.

Für eine derartige Transformation werden die Signale eines 3D-Audio Lautsprecher-Setups binauralisiert. Dies kann z.B. mittels Kunstkopf bzw. HRTFs (Head-Related Transfer Function) erfolgen. Eine 3D-Audio Lautsprecherwiedergabe lässt sich auf diese Weise erstaunlich präzise für die Kopfhörerwiedergabe binauralisieren. Ein Beispiel hierfür sind die Realiser A8 und A16 von Smyth Research mit Head-Tracking.

Bei objektorientierten Audio-Formaten kann jedes Objekt mit spezifischen HRTFs für die präzise Objektposition gerendert werden. Somit werden die Objekte als Schallquellen und nicht als Phantomschallquellen reproduziert.

6.1. Multimodale Wahrnehmung

Die Binauralisierung basiert auf einem Lautsprechersystem in einem bestimmten Wiedergaberaum, in dem personalisierte HRTFs gemessen werden. Solange sich der Hörer an der Position der Messung befindet, sind Lautsprecher- und Kopfhöherwiedergabe mit Head-Tracking auch bei Kopfdrehungen praktisch nicht voneinander zu unterscheiden. Binauralisierte Realschallquellen werden an Ort der im Raum platzierten Lautsprecher lokalisiert. Somit stimmen die Positionsinformation des auditiven und visuellen Reizes für den Hörer überein.

Es ist allerdings zu beachten, dass mehrere Modalitäten zum Hörereignis integriert werden. Werden nämlich binauralisierte Realschallquellen ausserhalb des zuvor beschriebenen Wiedergaberaums reproduziert, sind am Hörereignisort keine Lautsprecher sichtbar, und der Schall wird scheinbar aus dem Nichts wiedergegeben. Das führt beim Hörer zu einem Widerspruch in der multimodalen Wahrnehmung. Bei Kopfhörerwiedergabe von Standard 2-CH Stereo ist dagegen die Lokalisation von Schallquellen unabhängig vom Wiedergaberaum und abhängig von der Ausrichtung des Kopfes; dreht der Hörer seinen Kopf, bewegen sich die Hörereignisse starr gekoppelt mit den Kopf mit. Dieser fix gekoppelte

Mechanismus führt zu Binding [10], d.h. sobald eine Hörperson einen Kopfhörer aufsetzt, erwartet sie unbewusst dieses Verhalten unabhängig davon, ob Lautsprecher an bestimmten Positionen im Raum stehen oder nicht.

Ein anderer Aspekt bezieht sich auf den Wiedergaberaum mit seiner individuellen Akustik. Jede Hörperson mit gesundem Gehör hat in einem Raum eine bestimmte Hörerwartung. Es kommt selten vor, dass die meist unbewusste Hörerwartung nicht mit der tatsächlich vorhandenen Akustik übereinstimmt. Die Raumakustik wird während der Rezeption durch den Hörer auf die reproduzierte Aufnahme aufgeprägt. Diese Aufprägung wird bis zu einem gewissen Grad unbewusst erwartet und Verfärbungen und Verschleifungen von Transienten der Aufnahme nicht als störend wahrgenommen. Würden diese Verfremdungen fehlen, würde dies einem Widerspruch zur Hörerwartung entsprechen.

Bei einer Reproduktion von Standard 2-CH Stereo mittels Kopfhörer haben Hörpersonen eine andere unbewusste Erwartungshaltung: Es existiert kein Wiedergaberaum, welcher auf die Aufnahme aufgeprägt wird, und somit sind keine störenden Verfärbungen und Verschleifungen der Transienten hörbar. Daher sollte auch eine 3D-Audio Reproduktion über Kopfhörer weitgehend frei von störenden und verfremdenden Komponenten sein.

6.2. Binauralisierung und Umhüllung

Ein Schwachpunkt bei den meisten Produkten binauraler Lautsprecher-Virtualisierungen ist der Meinung der Autoren nach die wahrnehmbare Umhüllung. Wie bei der Lautsprecherwiedergabe mittels 9.1 müsste die Räumlichkeit zwischen vorne und hinten stark ausgeprägt sein und das Gefühl von Umhüllung erzeugen. Stattdessen scheint die Räumlichkeit auf einer senkrecht aufgestellten Scheibe an Ort zwischen den beiden Ohren ausgebildet zu sein. Somit fehlt bei der Kopfhörerwiedergabe eine der wichtigsten Komponenten, die 3D-Audio bietet. Eines der wenigen Produkte, welches derzeit Umhüllung überzeugend reproduzieren kann und auf nicht personalisierten HRTFs beruht, ist das EMT Phönix 5.1 BRS System.

7. Schlussfolgerungen

3D-Audio ist die konsequente Fortsetzung der Evolution von Mono, Stereo und Surround Sound. Daher ist es mehr als bloss Effekthascherei oder ein Marketing-Gag. Der sinnvolle Umgang mit 3D-Audio ist allerdings weitaus anspruchsvoller als bei den vorangehenden Formaten und setzt viel Wissen und Hörerfahrung voraus.

Bei 3D-Audio treten im Vergleich zu den vorangehenden Formaten neue psychoakustische Effekte auf, die zuvor nicht beobachtet worden sind. Sie wurden hier im Kapitel vertikale Stereofonie beschrieben. Kernpunkt ist die Kombination von sehr scharf lokalisierbaren Schallquellen und verbindenden Komponenten wie Raumschall und horizontale Phantomschallquellen. Es ist naheliegend, in diesem Bereich weitere Untersuchungen vorzunehmen, um allenfalls weitere psychoakustische Phänomene zu entdecken, die zum Wohlklang einer Aufnahme beitragen!

3D-Audio mittels Kopfhörer ist ein vielversprechendes Feld. Die Umsetzung ist derzeit aber noch nicht so weit vorangeschritten, dass es unabhängig vom Standort des Hörers vergleichbar eindrückliche Hörergebnisse erzielen kann wie ein 9.1 Lautsprecher-Setup. Nichtsdestotrotz sind die heute erhältlichen Systeme wichtig für die Auseinandersetzung, Hörerfahrung und Verbreitung von 3D-Audio.

Um 3D-Audio erfolgreich im Markt zu etablieren, benötigt es noch viel Aufklärungsarbeit. Insbesondere die Vermittlung des notwendigen Wissens an Tonschaffende wie Produzenten, Musiker und Sound Engineers ist wichtig. Denn psychoakustische Phänomene alleine führen nicht zu neuen kreativen musikalischen Ideen; es ist vielmehr der Austausch zwischen den beiden Disziplinen Wissenschaft und Kunst.

8. References

[1] Wittek, H.: Untersuchungen zur Richtungsabbildung mit L-C-R Hauptmikrofonen. Diploma thesis (2000), Institut für Rundfunktechnik, Fachhochschule Düsseldorf Studiengang Ton- und Bildtechnik

[2] Lee, H.: Effects of Interchannel Crosstalk in Multichannel Microphone Technique. Doctoral thesis (2006), Institute of Sound Recording, University of Surrey.

[3] Nipkow, L.: The transparency of sound in Stereo- and 3D 9.1 recordings. 3. ICSA, Graz (2015), Proceedings

[4] Steinke, G.: Mit den Ohren sehen – mit den Augen hören; Ansprüche und Wege zum optimalen Hörerlebnis, Verlag Kopie & Druck Adlershof, Berlin, 2009, ISBN 978-3- 9811396-5-5

[5] Nipkow, L.: Sound Design in 3D for Dance Music. 4. ICSA, Graz (2017), Proceedings

[6] W. Zieglmeier, G. Theile: Darstellung seitlicher Schallquellen bei Anwendung des 3/2-Stereo-Formats, 19. Tonmeistertagung Karlsruhe, Tagungsband pp. 159–169, 1996

[7] Nipkow, L.: Room signals – properties and influence on Auro 3D recordings. Proceedings of ICSA 2014 in Erlangen, ISBN 978-3-98 12830-4-4, 96-101

[8] Zielinsky, G.: More reality mit Auro 3D. VDT-Magazin, 2011, Heft 2, 24-27

[9] Potratz, U.: Untersuchung der Gestaltungsmöglichkeiten früher Reflexionen mit Hilfe eines raumakustischen Modells. Diplomarbeit, Erich-Thienhaus-Institut der Hochschule für Musik Detmold (2005), http://www.eti.hfm-detmold.de/lehraktiv/diplomarbeiten/diplomarbeitenordner/da-potratz.pdf

[10] U. Ansorge, H. Leder: Wahrnehmung und Aufmerksamkeit. VS Verlag für Sozialwissenschaften, Springer Fachmedien Wiesbaden GmbH (2011), ISBN 978-3-531-16704-6

Klang ist „kING“ – Variable Konzertsaal-Akustik in Ingelheim am Rhein

Dominik Schenke, Manuel Marx
Graner + Partner Ingenieure GmbH, Bergisch Gladbach, Germany
E-Mail: dominik.schenke@graner-ingenieure.de

Abstract

Die Kultur- und Veranstaltungshalle „kING“ in Ingelheim am Rhein wurde im August 2017 eröffnet. Das Herzstück ist der ca. 800 Plätze fassende Konzertsaal. Die Bandbreite der Veranstaltungen erstreckt sich vom Kongress über Theater-Produktionen, Pop-, Rock- und Jazz-Konzerten bis hin zu hochkarätigen Klassik-Events. Dabei hebt sich die „kING“ vom Konzept einer herkömmlichen Mehrzweckhalle durch die variable Raumakustik auf allerhöchstem klanglichem Niveau deutlich ab und besitzt damit in Deutschland ein Alleinstellungsmerkmal. Die Variabilität der Raumakustik wird dabei durch ein elektroakustisches Nachhallzeitverlängerungssystem erreicht, welches eine besonders hohe Flexibilität unterschiedlichster akustischer Situationen bietet. Die Nachhallzeit kann auf Kopfdruck verändert werden. Das Projekt wurde im Februar 2018 auf der Fachmesse Integrated Systems Europe (ISE) in Amsterdam mit dem „InAVation-Award“ als innovativstes Projekt in der Kategorie Kulturbauten ausgezeichnet. Graner+Partner Ingenieure GmbH zeichnete für die Fachplanung und Bauüberwachung in allen Leistungsphasen in den Fachbereichen Raum- u. Bauakustik und Elektroakustik verantwortlich. Der Vortrag verschafft sowohl einen Einblick in innovative akustische Lösungsprinzipien am Beispiel des genannten Projekts, als auch einen generellen Ausblick auf die sich stetig verändernden Anforderungen im Bereich der Kulturbauten.

1. Anforderungen

Bei der „kING“ handelte es sich um ein Neubau-Projekt, so dass der Konzertsaal von vornherein optimal auf die Anforderungen ausgelegt werden konnte. Graner+Partner waren von Beginn an in die Planungen eingebunden und konnten so durch frühzeitige Beratung den Bauherrn bei wichtigen Entscheidungen unterstützen.

Das Nutzungsspektrum war seitens des Bauherrn und des zukünftigen Nutzers derart definiert, dass eine große Bandbreite unterschiedlicher Veranstaltungen die „kING“ mit Leben füllen sollten. Ganz oben auf der Liste stand eine hochwertige konzertante Nutzung des Saals, aber auch Rockkonzerte, Partys, Tagungen, Kongresse und sogar Filmvorführungen sollten im kING beste Bedingungen vorfinden.

Die Anforderungen an die Auslegung der Raumakustik, die aus diesen Nutzungen resultieren, sind jedoch sehr unterschiedlich. So ist für Sprache und Pop-/Rock-Konzerte eine kurze Nachhallzeit (ca. 0,8 – 1,0 Sekunden) optimal, während für klassische Konzerte eine lange Nachhallzeit (1,8 – 2,0 Sekunden) erforderlich ist. Um diese Anforderungen optimal umzusetzen, war es also notwendig, die Raumakustik des Saals variabel zu gestalten und die Nachhallzeit je nach Nutzung verändern zu können.

Nach Gegenüberstellung unterschiedlicher Varianten wurde gemeinsam die Entscheidung getroffen, im Konzertsaal ein elektroakustisches Nachhallzeitverlängerungssystem zu installieren, um die gewünschte Variabilität und Flexibilität zu erreichen.

Abb. 1: Großer Saal der Kultur- und Kongresshalle „kING“ in Ingelheim am Rhein, Kapazität: 800 Sitzplätze.

2. Elektronisches Akustiksystem

Der Ansatz einer elektroakustischen Nachhallverlängerung ist im Grundsatz nicht neu. Erste Funktionsprinzipien und Systeme wurden bereits in den 1970er Jahren entwickelt. Allerdings hat die mit solchen Raumsimulationen erreichbare akustische Qualität in den letzten 10 Jahren u.a. durch die rasante Entwicklung von DSP-Leistung, aber nicht zuletzt auch durch die wachsende Erfahrung mit solchen Systemen einen enormen Sprung gemacht. Wurden vor einigen Jahren solche Systeme häufig noch mit einem unnatürlichen, sterilen und metallischen Klang in Verbindung gebracht, lassen sich heute akustische Situationen schaffen, die von einer natürlichen Konzertakustik nicht mehr zu unterscheiden sind.

ISBN
978-3-9812830-9-9

Der Autor hält die Urheberrechte für den gesamten Inhalt seines Manuskripts. Er überlässt dem VDT das Recht auf Veröffentlichung in gedruckter Form, Online oder als PDF-Download von https://tonmeistertagung.com

Das Prinzip des elektronischen Akustiksystems basiert darauf, dass der Raum, der eine variable Akustik erhalten soll, zunächst mit schallabsorbierenden Oberflächen so ausgestattet wird, dass die innerhalb des Nutzungsspektrums erforderliche kürzeste Soll-Nachhallzeit erreicht wird. Diese Nutzungen finden dann auch ohne das elektronische Akustiksystem optimale Bedingungen vor. Bei allen anderen Nutzungen, die eine längere Nachhallzeit erfordern, wird dann das System zur Erzeugung des Raumklangs eingesetzt.

Hierfür kommen eine Vielzahl an Mikrofonen über der Bühne, aber auch über den Publikumsflächen zum Einsatz, welche die akustischen Ereignisse im Raum aufnehmen. Die Mikrofonsignale werden einer Matrizierung und einem Signal-Processing zugeführt und anschließend über im gesamten Raum installierte Lautsprecher wiedergegeben. Vereinfacht gesprochen werden auf diese Weise Reflexionen, die normalerweise durch Raumoberflächen erzeugt würden, durch die vielen verschiedenen Mikrofon-Lautsprecher-Kombinationen reproduziert und werden dadurch steuer- und kontrollierbar. Die Akustik lässt sich nun auf Knopfdruck verändern.

Im Projekt in Ingelheim wurden insgesamt 32 Kondensator-Mikrofone (Niere) und 180 Lautsprecher installiert. Die Lautsprecher sind auf zwei umlaufende „Bänder“ im Wandbereich, sowie in einem ca. 4 m messenden Raster in der Decke aufgeteilt.

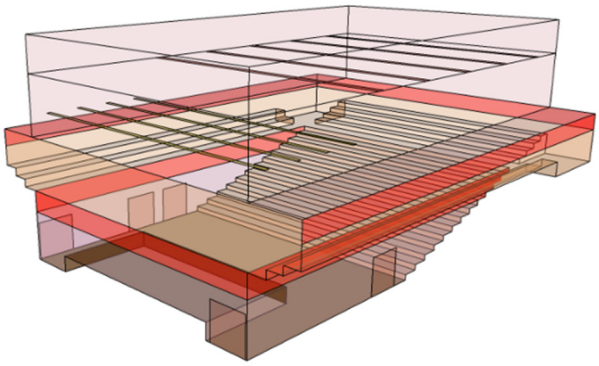


Abb. 2: 3D-Modell des großen Saals mit Lautsprecher-Bändern an den Wänden (rot) und an der Decke (grau)

3. Integration

Sämtliche Lautsprecher des elektronischen Akustiksystems sollten möglichst unsichtbar oder zumindest unauffällig in den Baukörper integriert werden, um auch architektonisch einen natürlichen Raumeindruck zu bewahren und die technische Ausstattung nicht zu sehr in den Vordergrund zu rücken. Genauso wie es akustisch das Ziel ist, einzelne Lautsprecher nicht wahrnehmen zu können, sollten sie auch optisch möglichst nicht in Erscheinung treten.

Im Wandbereich wurden die Lautsprecher daher in die Wandverkleidung integriert, die aus raumakustischen Gründen ohnehin eine Tiefe von bis zu 50 cm aufweisen musste. Die Lautsprecher sind dabei hinter einer akustisch transparenten Textilbespannung versteckt.

Abb. 3: Offene Wandverkleidung mit dahinter liegenden Installationen und Lautsprechern

Abb. 4: Wandpaneel mit akustisch transparenter Textilbespannung zur Kaschierung der Lautsprecher

Für die Textilverkleidung wurden verschiedene Stoffe messtechnisch untersucht, um das optimale Material zu finden, und die Abstrahlung der Lautsprecher nicht zu beeinträchtigen.

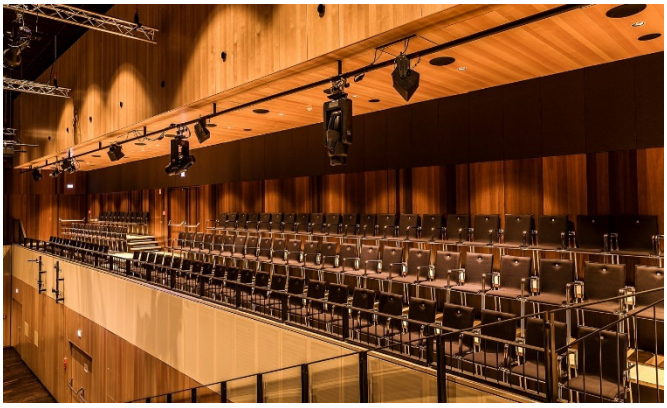


Abb. 5: Ansicht der Seitenwand mit fertig gestellten Lautsprecher-bändern hinter der Textilverkleidung (unten beige, oben schwarz)

Die Lautsprecher an der Decke konnten auf einfache Weise innerhalb der als Gitterrost-Ebene ausgeführten Saaldecke integriert werden.

Für den Bühnenbereich war es erforderlich, die Decken-Lautsprecher in ca. 7 Meter Höhe über der Bühne zu positionieren, um die für das gegenseitige Hören der Musiker so wichtigen frühen Reflexionen nicht zu stark zu verzögern. Die Lautsprecher ersetzen an dieser Stelle gewissermaßen den Konzertsaal-typischen Deckenreflektor. Dazu wurden sowohl die Lautsprecher als auch Mikrofone an verfahrbaren Bühnenzügen installiert. Auf diese Weise kann die Technik bei Veranstaltungsformen, für die die Nachhallzeit-verlängerung nicht benötigt wird, gänzlich aus dem Sichtfeld heraus gefahren werden.

Abb. 6: „Virtuelles Deckensegel“: Bühnenzug mit Lautsprechern über dem Orchester

4. Steuerung

Im Rahmen der Inbetriebnahme und Einmessung, sowie im Probenbetrieb für die ersten Veranstaltungen, wurden mehrere Presets mit unterschiedlichen akustischen Situationen und Nachhallzeiten für die jeweiligen Veranstaltungsformen eingerichtet. Diese Presets sind über fest installierte und drahtlose Touchpanels abrufbar. Innerhalb eines jeden Presets sind darüber hinaus geschmackliche Fein-Anpassungen, z.B. der Stärke oder der Klangfarbe des Nachhalls, möglich.

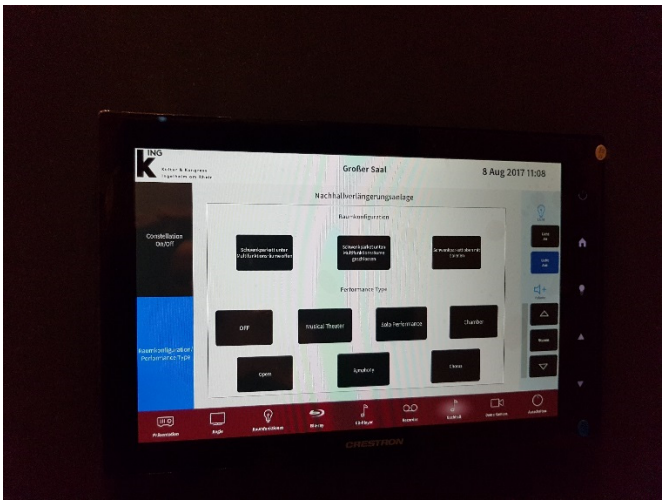


Abb. 7: Touchpanel mit Presets für das Akustik-System

5. Raumakustik-Design

Als Ziel-Nachhallzeit für den Saal wurden 0,80 Sekunden mit einem leichten Anstieg zu tiefen Frequenzen definiert, um für die Nutzung der Beschallungsanlage bei Sprache oder Pop- und Rock-Konzerten eine trockene Akustik und einen definierten Sound zu erreichen und für die Nutzung des elektronischen Akustiksystems eine gute Grundlage zu bilden.

Zur Bedämpfung des Saals kamen an den Wänden mikroperforierte Holzpaneele zum Einsatz, welche in Kombination mit dem großen Installationsraum und einer Mineralfaserdämmung dahinter eine sehr gleichmäßige, breitbandige Absorption liefern. Es sollten jedoch nicht alle Wandflächen absorptiv gestaltet werden, um auch auf natürliche Weise noch frühe Reflexionen zu erzeugen und einen hörbaren Raumeindruck zu erhalten. Der Raum sollte keinesfalls überdämpft werden. Daher wurden die mikroperforierten, absorbierenden Paneele jeweils im Wechsel mit glatten, reflektierenden Paneelen eingesetzt. Auf diese Weise wird gleichzeitig eine hohe Diffusität erreicht und störende Reflexionen wie z.B. Flatterechos vermieden. Die Mikroperforation mit einem Loch-Durchmesser von 0,3 mm ist optisch kaum wahrnehmbar, so dass auch in Kombination mit den glatten Paneelen eine homogene Wandoberfläche hergestellt werden konnte.

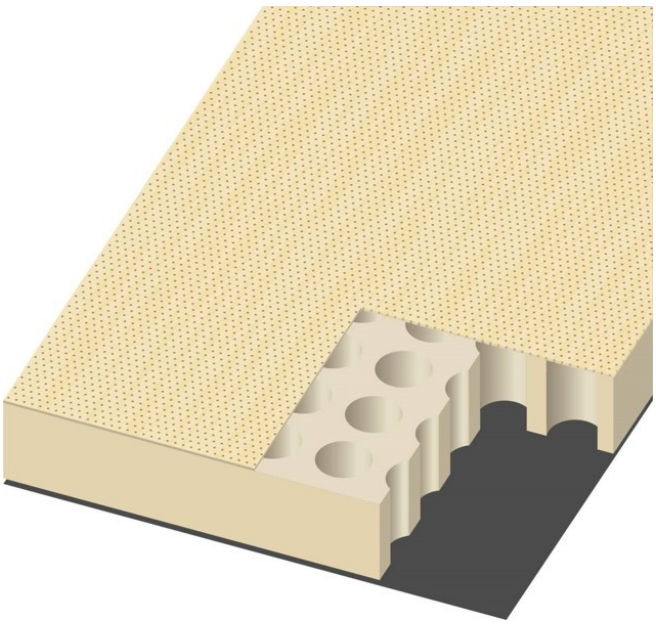


Abb. 8: Aufbau der absorbierenden Holzpaneele: Frontseitiges Furnier mit Mikroperforation auf gelochter MDF-Trägerplatte, rückseitiges Akustikvlies

An der Roh-Decke oberhalb der Gitterrostebene wurden vlieskaschierte Mineralwollplatten mit 140 mm Dicke eingesetzt, um auch hier breitbandige Absorption zu erreichen.

Alle geplanten raumakustischen Maßnahmen wurden in einem 3D-Modell implementiert, um die Erreichung der Planungsziele mittels Computersimulationen verifizieren zu können.

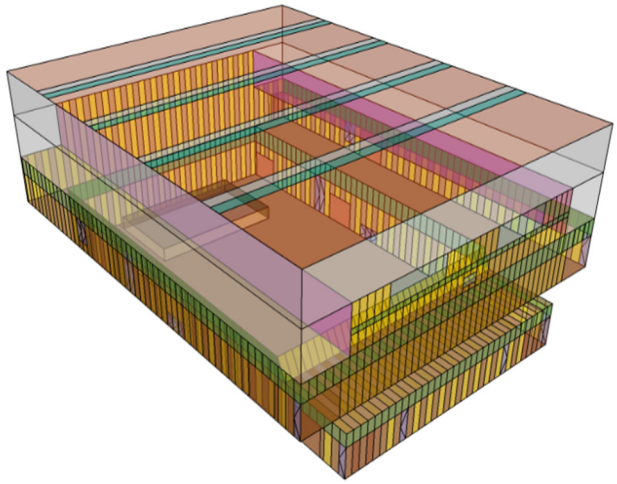
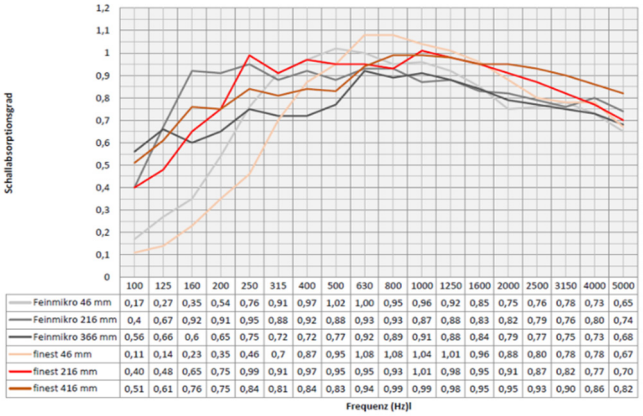


Abb. 9: 3D-Simulationsmodell mit wechselseitiger Anordnung von absorbierenden (braun) und reflektierenden (gelb) Wandpaneelen

6. Akustische Messungen

Für die mikroperforierten Holzpaneele wurden im Rahmen der Ausschreibungen unterschiedliche Fabrikate angeboten. Nicht für alle dieser Produkte lagen jedoch Prüfzeugnisse der Schallabsorption vor. Um die Wirksamkeit der Produkte und die Angebote vergleichen zu können und Planungssicherheit zu erhalten, wurden die Paneele im hauseigenen Hallraum messtechnisch untersucht und die frequenzabhängige Schallabsorption ermittelt.



	100	125	160	200	250	315	400	500	630	800	1000	1250	1600	2000	2500	3150	4000	5000
Feinmikro 46 mm	0,17	0,27	0,35	0,54	0,76	0,91	0,97	1,02	1,00	0,95	0,96	0,92	0,85	0,75	0,76	0,78	0,73	0,65
Feinmikro 216 mm	0,4	0,67	0,92	0,91	0,95	0,88	0,92	0,88	0,93	0,93	0,87	0,88	0,83	0,82	0,79	0,76	0,80	0,74
Feinmikro 366 mm	0,56	0,66	0,6	0,65	0,75	0,72	0,72	0,77	0,92	0,89	0,91	0,88	0,84	0,79	0,77	0,75	0,73	0,68
finest 46 mm	0,11	0,14	0,23	0,35	0,46	0,7	0,87	0,95	1,08	1,08	1,04	1,01	0,96	0,88	0,80	0,78	0,78	0,67
finest 216 mm	0,40	0,48	0,65	0,75	0,99	0,91	0,97	0,95	0,95	0,93	1,01	0,98	0,95	0,91	0,87	0,82	0,77	0,70
finest 416 mm	0,51	0,61	0,76	0,75	0,84	0,81	0,84	0,83	0,94	0,99	0,99	0,98	0,95	0,95	0,93	0,90	0,86	0,82

Abb. 10: Vergleich der Messergebnisse der Schallabsorption unterschiedlicher Paneele und Aufbauten

Nach Fertigstellung aller Oberflächen sowie des elektronischen Akustiksystems wurden die im Großen Saal erreichten Nachhallzeiten ebenfalls messtechnisch gemäß DIN EN ISO 3382-2 überprüft. Die Ergebnisse der rein natürlichen Akustik, ohne den Einfluss des Nachhallzeitverlängerungssystems entsprachen vollständig den Planungszielen. Die mittlere Nachhallzeit beträgt rund 0,80 Sekunden und zeigt einen gleichmäßigen Frequenzverlauf mit einem leichten Anstieg zu tiefen Frequenzen

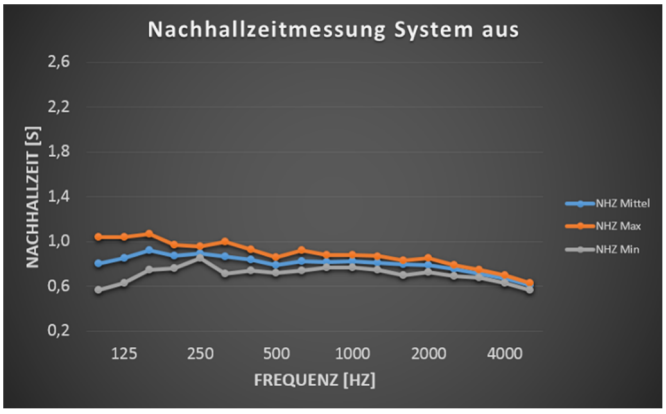


Abb. 11: Nachhallzeitmessung mit ausgeschaltetem Akustiksystem

Darüber hinaus wurde die Nachhallzeit mit eingeschaltetem Akustiksystem mit dem Preset „Symphony“ für größere Orchesterbesetzungen gemessen. Die mittlere Nachhallzeit beträgt rund 1,80 Sekunden, ebenfalls mit einem Anstieg zu tiefen Frequenzen. Die Messergebnisse zeigen, dass die Nachhallzeitverlängerung nicht nur eine rein subjektive Aufwertung des Raumklangs zur Folge hat, sondern tatsächlich ein ausgeprägtes Frühreflexionsmuster und eine dichte Hallfahne erzeugt wird, und zwar mit hoher Diffusität und sehr gleichmäßiger Schallverteilung im gesamten Saal, und nicht etwa ausschließlich auf den Bühnenbereich beschränkt. Das System verhält sich schließlich wie ein natürlicher Konzertsaal.

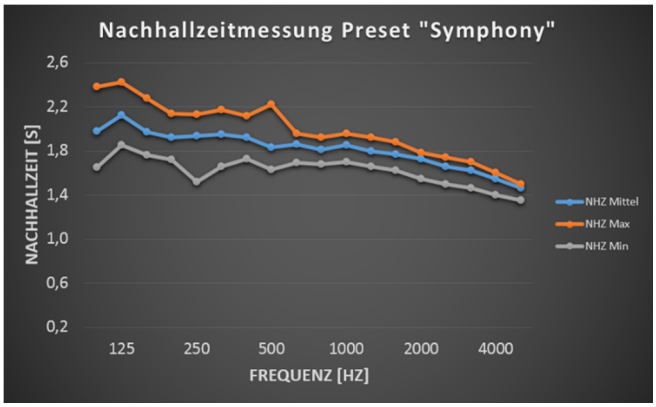


Abb. 12: Nachhallzeitmessung mit eingeschaltetem Akustiksystem, Preset „Symphony“

7. Zusammenfassung und Ausblick

Im Großen Saal der Kultur- und Kongresshalle „kING“ in Ingelheim am Rhein wurde ein elektroakustisches Nachhallzeitverlängerungssystem installiert. Das System sowie die natürliche Raumakustik wurden vom Beginn der gesamten Neubaumaßnahme an aufeinander abgestimmt geplant und optisch absolut unauffällig in den Baukörper integriert.

Nach ersten rein technischen Einmessungen wurde im Rahmen des Probetriebs für die Eröffnungsveranstaltung das Feintuning des Systems nach Gehör vorgenommen. Dies geschah in enger Zusammenarbeit mit dem Mainzer Staatsorchester, welches dankeswerterweise sehr offen und unvoreingenommen an die Thematik heranging. Die Rückmeldungen des Dirigenten und der Musiker waren durchgehend überaus positiv. Unter anderem Aussagen wie: „Wenn ich es nicht wüsste, hätte ich nie vermutet, dass bei diesem tollen Raumklang hier Mikrofone und Lautsprecher zum Einsatz kommen…“ unterstreichen die subtile und absolut natürliche Wirkung des Systems. Auch wurde die Möglichkeit, während der Proben noch individuell Einfluss auf den Raumklang nehmen zu können, von den Musikern sehr positiv bewertet.

Es lässt sich demnach festhalten, dass bei sorgfältiger Planung und Ausführung die mit einem elektronischen Akustiksystem erreichbare Klangqualität einer natürlichen Konzertakustik in nichts nachsteht. Perspektivisch ist auch davon auszugehen, dass mit einer steigenden Anzahl erfolgreicher Projekte auch die bei Profimusikern z.T. noch vorhandenen Vorbehalte gegenüber solchen Systemen nach und nach verringert und schließlich ganz ausgeräumt werden können.

Für Bauherren und Betreiber von Konzertsälen, Veranstaltungshallen und Kulturbauten allgemein stellt sich immer öfter auch die Frage nach einer multifunktionalen Nutzung von Aufführungsräumen. Auch neue, interdisziplinäre, oder Genre-übergreifende Formate stellen deutlich vielfältigere Anforderungen an Akustik und technische Ausstattung, als dies mit einem z.B. rein auf klassische Musik ausgelegten Konzertsaal abzudecken wäre. Eine variable Akustik auf höchstem klanglichem Niveau dagegen erschließt eine große Bandbreite unterschiedlicher Veranstaltungsformen. Die kurzen, bzw. quasi nicht vorhandenen Umrüstzeiten bei der Nutzung eines elektronischen Akustiksystems tragen zu einem effizienten Betrieb bei. Aufführungsräume können vielfältiger genutzt und stärker ausgelastet werden. Dies alles stellt für die Betreiber von Kulturbauten auch einen deutlichen wirtschaftlichen Vorteil dar.

Abb. 13: Eröffnungsveranstaltung mit dem Mainzer Staatsorchester

Der Partika-Saal der Robert Schumann-Hochschule Düsseldorf – Raumakustische Messungen und Verbesserungsmöglichkeiten

K.-H. Lorenz-Kierakiewitz[1], B. Pfändner[2], M. Oehler[3]

[1] *Peutz Consult GmbH, Germany, Email: khl@peutz.de*

[2] *Peutz Consult GmbH, Germany, Email: bp@peutz.de*

[3] *Universität Osnabrück, Institut für Musikwissenschaft und Musikpädagogik, Email: michael.oehler@uos.de*

Abstract

Die Robert Schumann-Hochschule Düsseldorf besitzt für Proben, Aufführungen und Aufnahmen einen eigenen Saal, den Partika-Saal. Für den Studiengang Ton und Bild des Instituts für Musik und Medien (IMM) der Musikhochschule ist es zugleich der größte innerhalb der Hochschule zur Verfügung stehende Aufnahmeraum. Dozenten und Studierenden seit Jahrzehnten wohlbekannt, verfügt dieser Saal über eine besondere Architektur und raumakustische Eigenschaften, z. B. dass es innerhalb des annähernd sechseckigen Raums eine Vorzugsrichtung für Musizieren und Aufnahmen zu geben scheint. Im Rahmen einer Bachelorarbeit des Studiengangs Ton und Bild des IMM sollte dem wissenschaftlich auf den Grund gegangen werden, wofür zunächst im unbestuhlten, im bestuhlten, aber unbesetzten, sowie im mit Besetzungssimulation versehenen Zustand des Raums ausführliche raumakustische Messungen inklusive Raumimpulsantworten mit einem Maximalfolgenmesssystem konform DIN EN ISO 3382-1 und gemäß DIN EN ISO 14257 durchgeführt und ausgewertet wurden. Dabei konnte gezeigt werden, dass es im Raum tatsächlich eine Vorzugsrichtung gibt sowie, dass eine Verbesserung der Hörsamkeit im Partika-Saal für die derzeitigen Nutzungen erreicht werden kann. Mit Hilfe eines 3D-Computer-Simulationsmodells wurden nach sorgfältiger Kalibrierung des Bestandsmodells an der messtechnisch erfassten Realität beispielhafte Maßnahmenprinzipien erarbeitet, die Grundlage für optimierende raumakustische Maßnahmen bilden können. Die sich hier bietenden Optionen sollen in diesem Beitrag vorgestellt und durch Auralisationen hörbar gemacht werden.

1. Einführung

Die Robert Schumann-Hochschule Düsseldorf besitzt für Proben, Aufführungen und Aufnahmen der Studierenden auf dem Hochschulcampus einen eigenen Saal, den Partika-Saal. Für die Studierenden des Studiengangs Ton- und Bild des IMM der Musikhochschule ist es der größte, innerhalb der Hochschule zur Verfügung stehende Aufnahmeraum. Durch studentische Projekte während des Studiums wurde in der Vergangenheit festgestellt, dass die Raumakustik dieses Saals durch einige Dozenten und Studierende der Hochschule im Bestand als nicht zufriedenstellend beurteilt wurde und wird. Daher lag es nahe, die raumakustischen Eigenschaften des Raums im Rahmen einer Bachelor-Arbeit [1] am IMM mit ausführlichen raumakustischen Messungen des Bestands näher zu untersuchen mit dem Ziel, mögliche Prinzipien für Maßnahmen zur Verbesserung der Akustik dieses Raums zu erarbeiten. Dieser Beitrag stellt di Vorgehensweise, Methoden und Ergebnisse dieser Arbeit in komprimierter Form vor. Die wesentlichen Passagen wurden dabei aus [1] unverändert übernommen.

2. Vorgehensweise der Untersuchung

Zunächst wurden zur Vorbereitung der Bachelorarbeit die derzeitigen Nutzungen des Saals erfasst und dafür auf Basis der Empfehlungen gängiger Literatur Anforderungen an die Werte raumakustischer Parameter für diese Nutzungen formuliert. Darauf aufbauend wurden die im Raum im Bestand vorliegende Bewertung der akustischen Eigenschaften mit einer Umfrage zur subjektiven Beurteilung durch die Nutzer präzisiert. Nach der Definition raumakustischer Anforderungen für die dort stattfindenen Nutzungen wurde ein Messplan zur Durchführung der erforderlichen raumakustischen Messungen im Partika-Saal erstellt. Zur Erfassung der vorhandenen raumakustischen Eigenschaften des Partika-Saals wurden ausführliche raumakustische Messungen durchgeführt und anschließend ausgewertet, beurteilt und anhand der zuvor definierten Anforderungswerte bewertet.

Abschließend wurden mit Hilfe eines 3D-Computer-simulationsmodells Prinzipien zur Verbesserung der Raumakustik erarbeitet und beispielhafte Konzepte für optimierende raumakustische Maßnahmen entwickelt.

3. Untersuchungsobjekt Partika-Saal

3.1. Geschichte

Der Partika-Saal der RSH, siehe Abb. 1 und 2, wurde 1993 erbaut. Er wurde von Architekt Klaus Reese als Probensaal entworfen und entstand unter der Leitung der Musikhochschule von Prof. Dr. Helmut Kirchmeyer. Seinen Namen verdanken Gebäude und Raum seinem Stifter, dem Oberstudiendirektor Josef Partika.

3.2. Architektur

Der Saal ist in einem eigenständigen Gebäude auf dem Hochschulcampus untergebracht. Den Saal umgibt ein Foyer, im Obergeschoss ist eine Tonregie angegliedert.

ISBN 978-3-9812830-9-9

Der Autor hält die Urheberrechte für den gesamten Inhalt seines Manuskripts. Er überlässt dem VDT das Recht auf Veröffentlichung in gedruckter Form, online oder als PDF-Download von https://tonmeistertagung.com

Der Saal ist teilweise unterkellert, und bietet im UG Platz für Technik- und Abstellräume sowie eine Krypta. Auf Höhe der Tonregie im OG des Gebäudes befindet sich ein umlaufender, unbestuhlter Balkon mit erhöhter mittiger Orgelempore unterbrochen durch die Regie.

Die Kapazität umfasst in den beiden aktuell verwendeten Bestuhlungsvarianten „längs“ (Standard) bzw. „quer“, siehe Abb. 1, jeweils 199 Sitzplätze. Die Grundfläche beträgt ca. 470m^2, die Höhen von ca. 7m an den Wänden bis ca. 14m im Scheitel. Daraus wurde ein architektonisches Volumen von ca. 5.000 m^3 ermittelt.

Der Boden besteht aus Beton mit Parkett, die Wände bis zu einer gewissen Höhe aus Ziegelmauerwerk, darüber aus einer Holzverkleidung mit Hohlraum vor der Massivwand, die sich im gesamten Deckenbereich fortsetzt, s. Abb. 1. Der Hohlraum zwischen Verkleidung und Außenwand beträgt an einer einsehbaren Stelle ca. 13 cm, darin befindet sich ca. 3 cm dicke Mineralwolle. Diese Verkleidung findet sich auch auf der gesamten Rückwand des Saals mit Ausnahme des mit Spanplatten versehenen Tonregiedachs. Auf dem betonierten Balkonboden, der auf der Unterseite eine abgehängte Gipskartondecke aufweist, befindet sich ein dünner Nadelfilz-Teppich, die Brüstung besteht in Richtung Außenwand aus aufgedoppelten Holzplatten und besitzt in Richtung Saalmitte mit ca. 3 cm Mineralwolle gefüllte Lochbleche. Die Bestuhlung besteht aus Stapelstühlen mit ca. 2 cm dünnem Sitz- sowie Rückenlehnenpolstern. In der Regel befinden sich außerdem stets ca. 45 Notenpulte im Saal, sowie drei Flügel, ein Cembalo, sowie eine kompakte Hausorgel im vorderen Saalbereich, welche die auf dem Balkon befindliche Orgel ergänzt.

3.3. Nutzungsspektrum

Der Partika-Saal ist der größte Veranstaltungsraum der RSH, in der es außerdem einen kleineren Kammermusiksaal gibt. Da die Hochschule keinen weiteren Veranstaltungsraum besitzt, erfüllt der Raum eine Vielzahl an Funktionen. Er dient als Proben- und Veranstaltungssaal des Hochschulorchesters (mehrere Probenphasen pro Semester) sowie als Prüfungsraum und wird u.a. auch für Hochschulveranstaltungen, Vorträge oder Installationen benutzt sowie als Konzertsaal (1 Opernproduktion p.a.). Darüber hinaus dient er als Aufnahmeraum. Die meisten Veranstaltungen des als Basis betrachteten Jahres 2016 hatten kammermusikalischen Charakter, es folgten in absteigender relativer Häufigkeit: Orchesterproben, Opernproduktionen und Aufnahmen bzw. Unterricht des IMM; Vorträge o.ä. Veranstaltungen kommen ebenfalls vor, jedoch seltener.
Aufgrund des vorliegenden Nutzungsspektrums kann keine eindeutige Hauptnutzung definiert werden, da der Saal viele verschiedene Funktionen erfüllt, die unterschiedliche Anforderungen an die Raumakustik stellen.
Die Verschiedenartigkeit der hier vorliegenden Nutzungen legt somit eine variable Akustik über variable akustische Maßnahmen nahe, zumal der Saal auch Ausbildungsraum für Toningenieure ist. Solche sind für diesen Raum jedoch im Bestand nicht standardmäßig vorhanden noch verfügbar.

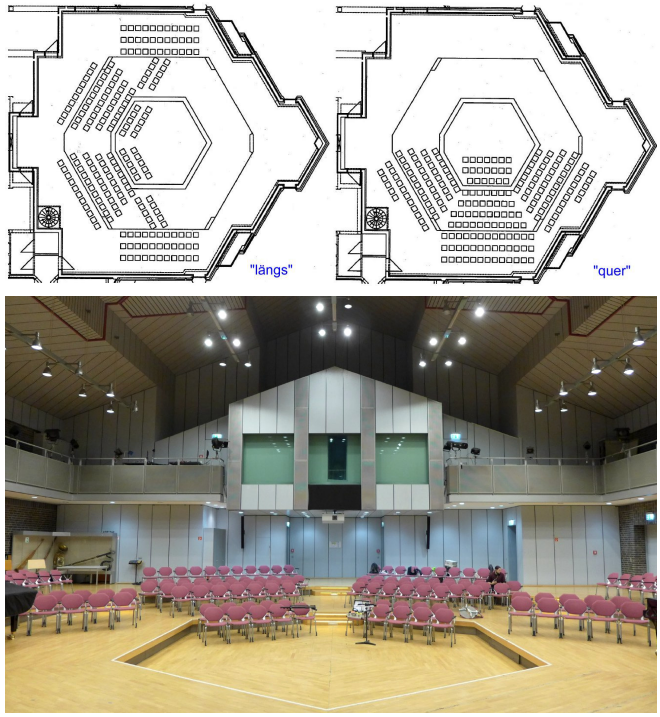


Abb. 1: oben: Grundrisse mit den beiden Bestuhlungsvarianten „längs“ und „quer“; unten: Innenansicht des Partika-Saals („längs“)

4. Bewertung durch die Nutzer

4.1. Umfrage zur Problemstellung

Um die Bewertung der derzeit im Saal vorhandenen raumakustischen Eigenschaften objektivieren zu können, wurde online eine hochschulinterne Umfrage zur subjektiven nutzungsabhängigen Bewertung der Raumakustik mit 113 Probanden durchgeführt. Den Antwortmöglichkeiten lagen fünfstufige verbale Rating-Skalen zugrunde. 75 Teilnehmer beantworteten die Umfrage vollständig. Nur diese Antworten wurden ausgewertet.
Alle Teilnehmer der Befragung waren Nutzer des Saals.

4.2. Zusammenfassung der Ergebnisse

Die Auswertung der Umfrage zeigte, dass die Raumakustik des Partika-Saals von den Nutzern für alle untersuchten Veranstaltungsarten als mittelmäßig beurteilt wird. Die Klangfarbe wird überwiegend als mittelmäßig und nur für Kammermusikkonzerte als gut empfunden. Die Nachhallzeit wird (mit Ausnahme der Bewertung für Orchesterkonzerte) als zu lang beurteilt. Der Nachhallpegel wird überwiegend als zu hoch beziehungsweise deutlich zu hoch beurteilt. Insgesamt lässt sich feststellen, dass die subjektive Bewertung verschiedener Aspekte der Raumakustik des Saals für die untersuchten Nutzungen durch die Nutzer überwiegend suboptimal ausfällt.
Die Hörsamkeit des Saals für die derzeitigen Nutzungen ist nach den Ergebnissen der subjektiven Bewertung durch die Nutzer im Bestand suboptimal. Es ist daher zu überlegen, raumakustische Maßnahmen zu erarbeiten, um die Hörsamkeit zu verbessern. Dabei ist empfehlenswert, eine variable Raumakustik, zu realisieren, um für die Nutzungen eine jeweils angemessene Hörsamkeit einstellen zu können.

5. Nutzungsabhängige Anforderungen

5.1. Probensaal des Hochschulorchesters

Für die Nachhallzeit T eines Orchesterprobensaals ist in der einschlägigen Literatur als Empfehlung zu finden, dass diese etwas geringer ausfallen sollte als in Konzertsälen. Bei bestehenden Probesälen liegen mittlere Nachhallzeiten zwischen 1,0 und 1,6 s vor [2].
Für das Stärkemaß G in Orchesterprobensälen werden Werte von unter 10 dB als optimal angesehen. Da in Probensälen in der Regel ein geringeres Volumen als in Konzertsälen vorliegt, sollte das Stärkemaß jedoch nicht zu hoch ausfallen. Aus der erweiterten Diffusfeldtheorie nach Kuttruff und Barron folgt für ein Volumen von 5.000 m^3 und einem angestrebten Stärkemaß < 10 dB, dass die Nachhallzeit in einem diffusen Schallfeld maximal 1,6 s betragen sollte. Allerdings sollte der Raumeindruck in der Probensituation auch nicht zu stark von dem im Konzert abweichen. Da die Konzerte des Hochschulorchesters in der Regel in der Tonhalle Düsseldorf stattfinden, können deren raumakustischen Parameterwerte als Anhaltspunkt dienen. Die Tonhalle Düsseldorf besitzt nach ihrer Sanierung im Jahr 2005 eine mittlere Nachhallzeit von 2,1 s bei einem Volumen von ca. 15.000 m^3 [2,5]. Da das Volumen des Partika-Saals nur ca.1/3 des Volumens der Tonhalle beträgt, sind auch etwas geringere Nachhallzeiten angemessen, insbesondere wenn das Stärkemaß nicht höher ausfallen soll.

5.2. Proben-/Konzertsaal für Kammermusik

Eine wesentliche Nutzung des Partika-Saals ist als Proben- und Konzertsaal für Kammermusik. In der einschlägigen Literatur finden sich empfohlene Nachhallzeitwerte von 1,2 s bis 1,8 s. Beim Vergleich verschiedener bestehender Kammermusiksäle lassen sich Nachhallzeiten von 1,2 s bis 1,7 s finden [2]. Auch hier sollte das Stärkemaß begrenzt werden, auf vozugsweise < 10 dB, woraus widerum eine Nachhallzeit von bis zu 1,6 s resultiert.
Bestehende Kammermusiksäle weisen gemäß [2] Werte des Stärkemaßes zwischen +8,8 bis +10 dB auf.

5.3. Verstärkte Darbietungen

Um eine gute Verständlichkeit der Sprache zu ermöglichen, sollte der Artikulationsverlust von Konsonanten AL_{cons}< 9 % betragen. Hieraus folgt, dass die Nachhallzeit volumenunabhängig auf maximal 1 s begrenzt werden sollte. Das Deutlichkeitsmaß C_{50} sollte für Räume mit Sprachnutzungen Werte über 0 dB annehmen. Für die elektroakustische Verstärkung von Musik ist ebenso eine mittlere Nachhallzeit von unter 1,2 s anzustreben.

Es wird deutlich, dass aufgrund des verschiedenen Charaktere der im Raum stattfindenden Nutzungen und den jeweils dazu empfehlenswerten Anforderungswerten eine variable Veränderbarkeit der raumakustischen Bedingungen im Raum in Abhängigkeit von der jeweiligen Darbietung zielführend erscheint, gerade im Hinblick auf die Nutzung des Raums für Aufnahmen unterschiedlichster Art.

Für Proben und Aufnahmen ist es günstig, die akustischen Bedingungen pro Nutzung variabel einstellen zu können.

6. Raumakustische Bestandsmessungen

6.1. Durchführung der Messungen

Um die im Jahr 2016 vorliegenden raumakustischen Bedingungen erfassen und beurteilen zu können, wurden im Februar 2016 detaillierte raumakustische Messungen der Nachhallzeiten, Pegelverteilungen und Impulsantworten in verschiedenen Raumzuständen im Partika-Saal durchgeführt:

Kürzel	Beschreibung Raumzustand/Bestuhlungsvariante
UBL	unbesetzt, bestuhlt, "Längs", mit Orchestermobiliar
UBQ	unbesetzt, bestuhlt, "Quer", mit Orchestermobiliar
BBL	mit Besetzungssimulation, bestuhlt, "Längs", mit Orchestermobiliar
UU	unbesetzt, unbestuhlt, mit Orchestermobiliar
UBLPA	unbesetzt, bestuhlt, "Längs", mit Orchestermobiliar, mit aktiver Beschallungsanlage
UBQPA	unbesetzt, bestuhlt, "Quer", mit Orchestermobiliar, mit aktiver Beschallungsanlage

Tab. 1: bei den Messungen erfasste Raumzustände

Die messtechnisch erfassten Bestuhlungsvarianten der Raumzustände UBL, UBQ, BBL, UBLPA und UBQPA wurden entsprechend den von der Hochschule zur Verfügung gestellten Bestuhlungsplänen aufgebaut, als Orchestermobiliar wurden exemplarisch 17 Notenpulte und 16 Polsterstühle für eine Kammerbesetzung verwendet. Die während der Messungen im Saal gemessene Temperatur betrug 21° C, die Luftfeuchtigkeit lag bei 36%.

Mit den Bestuhlungsplänen war zuvor ein Messplan mit Messpositionen erstellt worden, siehe Abb. 2.

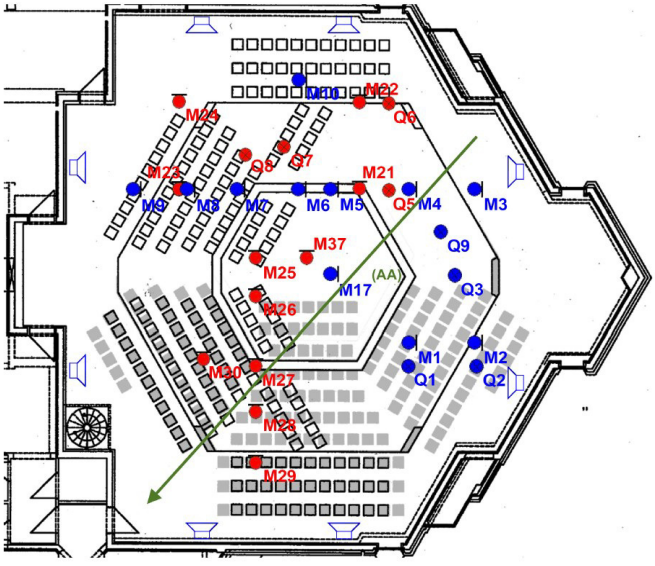


Abb. 2: Grundriss mit darin eingetragenen Messpositionen

Die Messpositionen wurden gemäß der Hauptnutzung Orchesterproben definiert, und zwar für beide verwendeten Orientierungen, so finden sich Quellpositionen entsprechend im jeweiligen „Bühnenbereich“, und Mikrofonpositionen im jeweiligen „Publikumsbereich“ (siehe Abb. 2).

Tabelle 2 zeigt die pro Zustand durchgeführten Messungen.

Raumzustände	Messung
UBL, UBQ, BBL, UU	Nachhallzeiten
UBL, UBQ, UBLPA, UBQPA	Raumimpulsantworten
UBL, UBQ	Indikatorwerte natürliche Sprachverständlichkeit
UBLPA, UBQPA	Indikatorwerte verstärkte Sprachverständlichkeit
UBL, UBQ, BBL, UU	Abnahme des Pegels mit zunehmendem Abstand von der Schallquelle
UBL	Hintergrundschallpegel

Tab. 2: pro Raumzustand durchgeführte Messungen

Um einen besetzten Zustand zu simulieren, wurden bei den Messungen mit Besetzungssimulation insgesamt 100 m^2 Publikums-Simulationsstoffe nach [2] über die 199 Stühle gelegt, siehe Abbildung 3.

Die Messungen wurden durchgeführt und ausgewertet nach den Angaben in DIN EN ISO 3382-1 und DIN EN ISO 14257. Die Bestimmung der Nachhallzeiten erfolgte dabei über das Verfahren des abgeschalteten Rauschens (Mittelung von je drei Abklingvorgängen je Messpfad), die der Raumimpulsantworten von den Quellpositionen 1 und 2 zu den Mikrofonpositionen 1 bis 10 und 17 und von den Quellpositionen 5 und 6 zu den Mikrofonpositionen 21 bis 30 und 37. Die Impulsantworten wurden mit dem Maximalfolgenmesssystem MLSSA 10WI Rev.9 erfasst mit einer Maximalfolgenlänge von 65535 Samples bei einer Abtastrate von 36 kHz und einer Amplitudenauflösung von 12 Bit. Zur Wiedergabe der Maximalfolgen wurde eine Punktschallquelle in 1,50 m Höhe verwendet, welche eine ziemlich omnidirektionale Richtcharakteristik aufweist (Q≈1), siehe [3].

Für die Messung der Schallpegelverteilung im Raum wurde die Pegelabnahme mit zunehmendem Abstand zu einer omnidirektionalen Schallquelle gemessen. Als Messpfad diente für beide Bestuhlungsvarianten die in Abb. 2 grün markierte Diagonale durch den Raum. Es wurden 24 logarithmisch verteilte Messpositionen in einem Abstand von 0,1m bis 25m von der Quelle verwendet. Als Schallquelle diente ein Horntreiber, dessen Austrittsöffnung durch ein Metallrohr mit einem Durchmesser von 1” und einer Länge von 30 cm vom Treiber räumlich getrennt ist, um eine punktförmige Schallquelle im Raum für das interessierende Frequenzband zu bilden.

6.2. Auswertung der Messungen

Die gemessenen Schalldruckpegel des Oktavbandes mit 1 kHz Mittenfrequenz wurden für die gemessenen Zustände und zunehmenden Abstände zur Schallquelle ausgewertet. Die gemessenen Schallpegel wurden anschließend auf den Schallleistungspegel normiert und sind in Abbildung 4 mit dem theoretischen Verlauf des Direktschalls, des Nachhalls und der daraus resultierenden theoretischen Abnahme des Pegels mit dem Abstand für eine omnidirektionale Schallquelle im Raum dargestellt. Die Berechnung der theoretischen Abnahme des Pegels mit zunehmendem Abstand erfolgt nach Untersuchungen von Peutz und van der Werff auf Basis der Sabine’schen Diffusfeldtheorie.

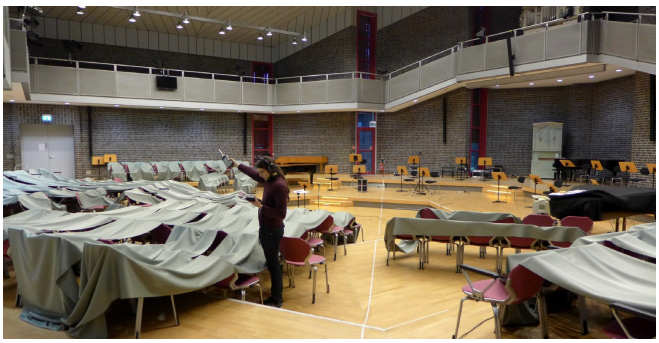


Abb. 3: Ansicht der Messsituation BBL mit Publikumssimulation

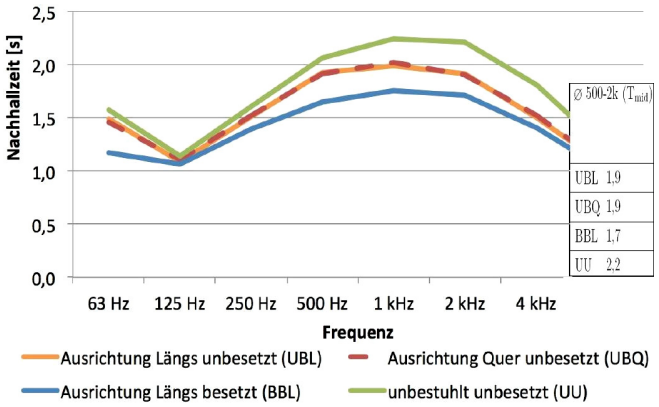


Abb. 3: Ergebnisse Nachhallzeitmessungen in den div. Zuständen

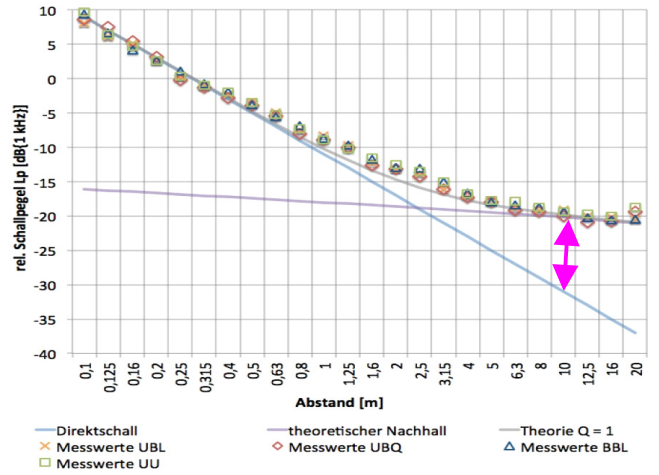


Abb. 4: Gemessene Abnahmen des Schallpegels mit zunehmendem Abstand von der Quelle für die Zustände UNL, UNQ, BBL und UU, Oktavband mit Mittenfrequenz 1 kHz. Pink: Stärkemaß G_{10mess}

Die Auswertung der Impulsantworten wurde konform den Angaben in DIN°EN°ISO°3382-1 in Oktavbändern durchgeführt. Die Oktavbänder mit Mittenfrequenzen 250, 500, 1.000 und 2.000 Hz wurden in Energie-Zeit-Kurven (ETC, Energy Time Curve) umgerechnet und in Grafiken dargestellt, in denen der Verlauf der Schallenergie in Abhängigkeit von der Zeit dargestellt ist. Zusätzlich wurde die ETC-Kurve zur Simulation der Ohrträgheit mit einem Tiefpassfilter der Integrationszeit 20 ms geglättet, siehe Abb. 5 (schwarze Einhüllende).

6.3. Messergebnisse

6.3.1. Nachhallzeiten

In Abbildung 3 sind die Ergebnisse der Messungen der Nachhallzeiten für die in Abschnitt 6.1 beschriebenen Raumzustände dargestellt. Im Bestand lagen mittlere Nachhallzeiten T_{MID} von 1,7 s für die Zustände mit Besetzungssimulation, 1,9 s für den unbesetzten, sowie 2,2 s für den unbestuhlten Zustand vor. Es lässt sich feststellen, dass die Ausrichtung der Bestuhlung im Saal bei den untersuchten Varianten keinen Einfluss auf die Nachhallzeit hat. Der Unterschied zwischen den Zuständen unbesetzt und mit Besetzungssimulation fällt mit 0,2 s relativ gering aus.

6.3.2. Schallpegelverteilung im Raum, Stärkemaße

Resultat der Messungen zur Schallpegelverteilung sind Werte des Stärkemaßes $G_{>2rH}$ von je nach Variante zwischen +11,1 (BB) und +11,9 dB (UU). Der Hallradius liegt bei ca. 2,9 m für die Zustände UBL/UBQ, bei 3,0 m für den Zustand BBL und 2,7 m für den Zustand UU. Es ergab sich auch, dass das akustisch wirksame Volumen des Raums mit ca. 3.200 m^3 geringer ist als das geometrische mit 5.000 m^3.

$$G_{diff,theo} = 45dB - 10 \cdot \log_{10}\left(\frac{V}{T}\right) \quad \text{[dB]} \quad (1)$$

Aus der Nachhallzeit und dem Raumvolumen folgt nach der erweiterten Diffusfeldtheorie nach Barron und Kuttruff gemäß Gleichung (1) ein theoretisches Stärkemaß $G_{diff,theo}$ von ca. +10,3 dB (BB) bis +11,4 dB (UU). Dies fällt geringer aus als die im Saal messtechnisch bestimmten Werte - Indiz dafür, dass nicht das gesamte architektonische Volumen gleichermaßen zur Nachhall-Generierung beiträgt. Dies bedeutet, dass gewisse Teile des architektonischen Volumens raumakustisch nicht wirksam werden, zum Beispiel der Raum unterhalb / hinter den Brüstungen des Balkons und die Nischen in Decke und Stirnwand.

6.3.3. Raumimpulsantworten

Die messtechnisch bestimmten Raumimpulsantworten wurden auf Artefakte und Echos untersucht. Zur Untersuchung der Reflexionsstrukturen wurden sie u. a. visuell auf Sprünge in der geglätteten ETC untersucht, welche auf Echos hinweisen können. In Abbildung 5 sind exemplarisch typische ETCs für beide Bestuhlungs-Varianten dargestellt. Diese zeigen eine gewisse wellige Grundstruktur, was darauf hinweist, dass der Raum im messtechnisch untersuchten Bestand in Bezug auf Grundform und Sekundärgeometrie breitbandig nicht optimal schalldurchmischend ist bzw. ausgestattet ist.

6.3.4. Klarheitsmaß C_{80}

Aus den gemessenen Raumimpulsantworten wurde das Klarheitsmaß C_{80} bestimmt: im Bühnenbereich der Bestuhlungsvariante „Längs“ ergeben sich Werte zwischen - 0,2 dB und +3,3 dB, für die Bestuhlungsvariante „Quer“ zwischen - 0,4 dB und +3,1 dB. Der Raummittelwert beträgt +1,5 für „Längs“ bzw. +0,9 dB „Quer“.

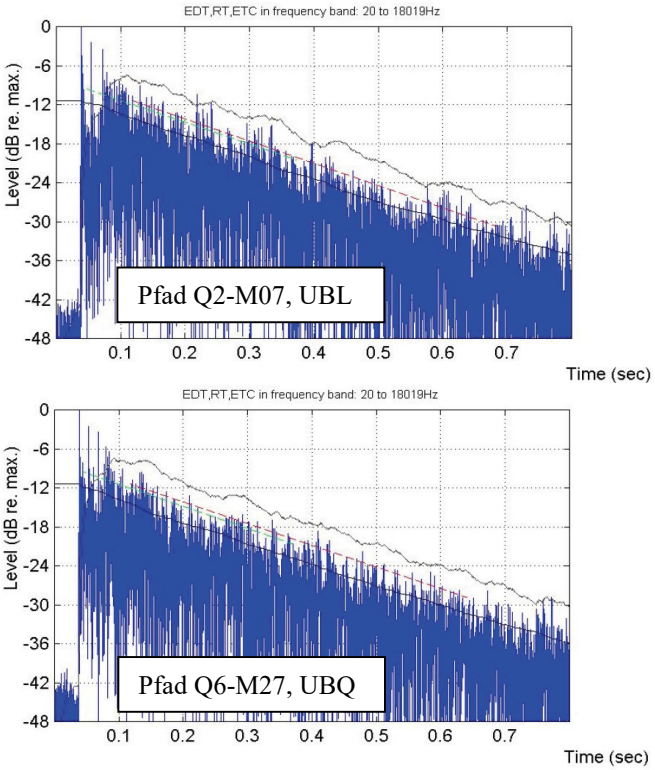


Abb. 5: ETCs im Bestand gemessener Raumimpulsantworten

Für den Publikumsbereich ergeben sich für die Bestuhlungsvariante „Längs“ Werte zwischen – 2,5 dB und +1,2 dB, der Raummittelwert beträgt – 0,6 dB, für die Bestuhlungsvariante „Quer“ zwischen - 1,5 dB und +1,5 dB mit einem Raummittelwert von – 0,3 dB.
Der Vergleich der beiden Bestuhlungsvarianten zeigt, dass für die Ausrichtung „Quer“ mit im Raummittelwert von – 0,3 dB im Mittel um ca. 0,3 dB höhere Werte festzustellen sind als in Variante „Längs“ mit – 0,6 dB.

6.3.5. Hintergrundschallpegel

Die ohne Lüftungsanlage gemessenen Hintergrund-Perzentilpegel L_{95} betrugen in der Messsituation ca. 19 bis 19,5 dB(A). Mit Lüftungsanlage wurden Werte von ca. 32,5 dB(A) in der Raummitte und ca. 32 bis 36 dB(A) in der Nähe der Luftaustrittsöffnungen im Publikumsbereich gemessen. Die spektrale Verteilung ist Abb. 6 zu entnehmen.

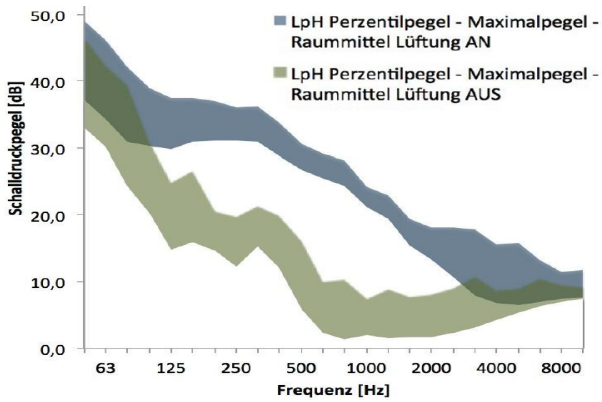


Abb. 6: Spektrale Verteilung der gemessenen Hintergrundpegel L_{F95} und L_{FMax} mit und ohne Lüftungsanlage im Zustand UBL

7. Bewertung der Messergebnisse

Die einzelnen Bewertungen aller verschiedenen Nutzungen zeigen, dass die Bestuhlungsvariante „Quer“ für alle betrachteten Parameter messbar günstigere Werte der raumakustischen Parameter aufweist als „Längs“.

Die Werte der raumakustischen Parameter der Bestuhlungsvariante „Quer“ erfüllen die in Abschnitt 5 definierten Anforderungen zumindest teilweise - mit Ausnahme der Nachhallzeit und ihres Frequenzverlaufs sowie der Regelmäßigkeit der Reflexionsstrukturen, wohingegen die Werte der raumakustischen Parameter der Bestuhlungsvariante „Längs“ zum großen Teil aus den Zielwertebereichen herausfallen.

Die Bewertung der Werte des Stärkemaßes zeigt, dass diese für verschiedene Nutzungen im Bestand etwas zu hoch ausfallen. Dies kann auf ein zu geringes Raumvolumen für diese Nutzungen hinweisen. Das vorhandene Volumen von 5.000 m^3 liegt für einen guten Orchesterprobenraum für große Besetzungen eher an der unteren Grenze.

Für einen Kammermusiksaal mit einer möglichen Bestuhlung für 200 Personen ist ein Volumen von ca. 5.000 m^3 tendenziell als vergleichsweise groß zu bewerten, wenngleich angemerkt werden muss, dass das akustisch wirksame Volumen etwas kleiner ausfällt.

Der Vergleich der Messergebnisse mit den subjektiven Bewertungen der Nutzer, die mit Hilfe der durchgeführten Umfrage erfasst wurden, zeigt, dass die überwiegend als zu lang empfundenen Nachhallzeiten in Bezug auf die den Nutzungen entsprechenden Anforderungen auch messtechnisch höher ausfallen als die Zielwerte.

Ebenso kann die in der Mehrzahl der Antworten als mittelmäßig eingestufte Klangfarbe mit der Nichtlinearität des Frequenzverlauf und den unregelmäßigen Reflexionsstrukturen erklärt werden. Der überwiegend als zu hoch beurteilte Nachhallpegel korreliert mit den die Wertebereiche der Anforderungen übersteigenden Werten des Stärkemaßes. Da der Nachhallpegel überwiegend als zu hoch, die Lautstärke jedoch als angemessen beurteilt wird, lässt sich darauf schließen, dass frühe Schallanteile im Verhältnis zum Diffusschallpegel tendenziell als zu gering beurteilt werden.

Die mittelmäßige Bewertung des Kontakts zu anderen Musikern lässt sich durch die Unregelmäßigkeiten in den Reflexionsstrukturen sowie dem relativ hohen Stärkemaß erklären.

Die Lüftungsanlage des Saals ist, wenn aktiv, für akustische Darbietungen mit Anspruch an die Hörsamkeit im Bestand als deutlich zu laut zu beurteilen.

Insgesamt sind die raumakustischen Bedingungen im Bestand des Partika-Saals sowohl, was die Messergebnisse der raumakustischen Bestandsmessungen betrifft, als auch, was die Bewertung durch die Nutzer angeht, insgesamt als suboptimal, aber optimierungsfähig einzuordnen.

8. Erstellung 3D-Computer-Modell

Zur Ermittlung des architektonischen Volumens und zur Bestimmung der Flächengrößen der unterschiedlichen Materialkategorien wurde im CAD-Programm SketchUp Vers. 16 ein 3D-Computermodell der im Bestand vorliegenden Raumgeometrie erstellt, s. Abb. 7 (Var. V1).

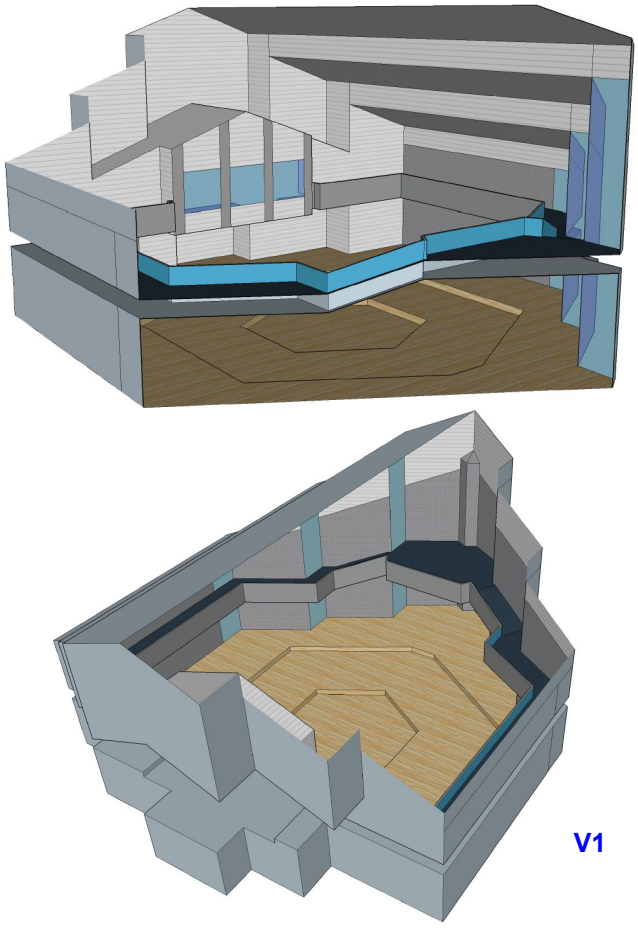


Abb. 7: Ansichten des 3D-Computermodells des Bestands (V1)

Die Modellierung erfolgte auf Grundlage der von der RSH zur Verfügung gestellten Grundrisse, Seitenansichten, während der Messungen aufgenommenen Fotos und Messdaten. Zur Erstellung des Modells wurden die Planunterlagen maßstabsgetreu in SketchUp eingelesen und nachgezeichnet. Daraus konnte durch extrudieren der gezeichneten Flächen und anschließendem Verschneiden ein vereinfachtes geschlossenes Modell der Raumgeometrie erstellt werden. Nach der Erstellung der Grobgeometrie wurden Details mit Abmessungen größer als ca. 0,25 m wie beispielsweise die Tonregiekabine oder der Balkon ergänzt.

Zur Ermittlung des geometrischen Volumens des Raums wurde die in SketchUp integrierte Berechnung des Rauminhaltes für geschlossene Körper verwendet. Diese ermittelt das umschlossene Volumen einer Gruppe oder von Komponenten, solange alle darin enthaltenen Flächen einen allseits geschlossenen Körper bilden.
Die Berechnung ergab ein auf 1 m^3 abgerundetes Volumen des Saals im Bestand von 5.008 m^3.

9. Erarbeitung Verbesserungsprinzipien

9.1. Ziele akustischer Verbesserungen

Das Ziel möglicher raumakustischer Verbesserungen ist es, eine möglichst gute Hörsamkeit für die im Partika-Saal vorliegenden Nutzungen zu erreichen. Da die hier zu berücksichtigenden Nutzungen teilweise gegensätzliche raumakustische Parameterwerte zur Erreichung dieses Ziels voraussetzen (z. B. Orchester- vs. Sprachdarbietungen), werden hier beispielhafte Prinzipien dargestellt, durch welche auf die Anforderungen der verschiedenen Nutzungen des Raums durch entsprechende *variable* raumakustische Maßnahmen eingegangen wird. Des Weiteren gibt es gemeinsame Anforderungen, die für alle Nutzungen gelten, beispielsweise eine bessere Linearität des Frequenzverlaufs der Nachhallzeit und das Ziel möglichst regulärer Reflexionsstrukturen sowie die Vermeidung von Echos, insbesondere Flatterechos und ähnlichen Artefakten.

9.2. Randbedingungen für Maßnahmen

Zur Umsetzung der Ziele und zur Verbesserung der raumakustischen Situation über das hinaus, was bei einer rein organisatorischen Bevorzugung der Bestuhlungsvariante „Quer“ erreichbar ist, sind umfangreichere, nicht nur punktuelle Maßnahmen erforderlich. Diese sollten zur Erhaltung des Gesamteindrucks des Saals optisch möglichst unauffällig sein beziehungsweise sich in das optische Konzept integrieren lassen. Zudem müssen variable raumakustische Maßnahmen einfach und verständlich bedienbar sein, um eine Verschlechterung der Hörsamkeit durch Fehlbedienung komplexer Maßnahmenkonzepte zu vermeiden. Wegen der Urheberrechte des Architekten wurden nur Maßnahmen temporärer Natur untersucht.

9.3. Prinzipien für Maßnahmen

Im Folgenden werden Prinzipien für beispielhafte Maßnahmen zur Verbesserung der Hörsamkeit erarbeitet.

9.3.1. Permanente Maßnahmen

Wie erläutert, ist für keine derzeitige Nutzung eine Nachhallzeit >1,7 s zu empfehlen. Daher könnte eine Reduzierung der Nachhallzeit auf 1,7 s durch permanente Absorptionsmaßnahmen erfolgen. Grundsätzlich ist eine gleichmäßige Verteilung der Absorptionsflächen über die Wände und Decke des Raums zu empfehlen, um eine gleichmäßige Schallverteilung im Raum zu begünstigen; Wandpositionen um Darbietende und Hörende herum sind dabei i.d.R. am wichtigsten (bei nicht-horizontalen Decken). Um die Optik jedoch möglichst wenig zu beeinträchtigen, sollte ein für das Publikum nicht unmittelbar auffälliger Bereich im Saal für die Positionierung der permanenten absorbierenden Flächen verwendet werden. So könnte die Dachfläche der Tonregie für diesen Zweck verwendet werden. Diese Fläche mit ca. 65 m^2 reicht jedoch auch bei Verwendung hochabsorbierender Materialien allein nicht aus, um die Nachhallzeit im mittleren Frequenzbereich auf 1,7 s zu senken. Durch eine geschickte Auswahl der hier zu verwendenden Absorptionsmaterialien ist es jedoch möglich, die Frequenzabhängigkeit der Nachhallzeiten zu verringern.

9.3.2. Verbesserung der Reflexionsstruktur

Wie die Ergebnisse der Auswertung der Impulsantworten zeigen, ist es empfehlenswert, die Reflexionsstrukturen des Raums zu verbessern. Im Bestand konnten in einigen Messpfaden echoartige Reflexionen festgestellt werden, siehe u.a. Abbildung 5. Dies ist auf die Geometrie des Raumes zurückzuführen. Grundsätzlich sollten parallele reflektierende Wandpaare vermieden werden, um Flatterechos zu vermeiden. Die Oberflächen bestehen überwiegend aus relativ glatten Flächen und weniger aus schallstreuenden Materialien. Zur Erzeugung eines regelmäßig abklingenden Nachhalls und zur Verringerung von Energielücken und Häufungen in der ETC eignen sich prinzipiell Streukörper und Reflektoren, die den Schall in verschiedene Richtungen reflektieren und streuen. Dabei muss die Dimension des Reflektors im Bereich der zu reflektierenden Wellenlänge liegen. Aufgrund dieser frequenzselektiven Eigenschaften von Reflektoren müssen entsprechend große Dimensionen gewählt werden, wenn eine tiefe Grenzfrequenz erzielt werden soll. Hier wäre zu berücksichtigen, dass die Reflektoren zusätzlich ein relativ hohes Flächengewicht aufweisen, damit sie als Reflektor und nicht als Absorber (im tieffrequenten Bereich, wegen des lokalen Minimums der Nachhallzeit) wirken. Zur Schallstreuung können verschiedene geometrische Strukturen verwendet werden: Eine einfache Möglichkeit zur Realisierung möglichst breitbandig streuender Körper sind konvexe Geometrien, wie beispielsweise Pyramiden, Kalotten oder Zylinderabschnitte. Streukörper können prinzipiell in Wandflächen integriert sowie darauf oder davor angebracht werden. Wegen der optischen Konsequenzen kommen im vorliegenden Fall fest installierte Streukörper kaum in Frage, jedoch könnten diese auch mobil realisiert werden. Eine Möglichkeit besteht darin, einzelne rollbare Elemente mit beispielsweise jeweils drei vertikalen konvexen Zylinderabschnitten mit einer Höhe von z. B. minimal 2,5 m und einer Gesamtbreite von ca. 3 m auf dem Boden (Eingangsniveau) unterhalb des Balkons vor den zur Verfügung stehenden Wandflächen zu positionieren. Hierdurch ließen sich auf einfache Weise z. B. dreizehn identische und somit austauschbare und ggf. frei positionierbare Elemente auf Rollen mit einer Gesamtfläche von ca. 95 m^2 realisieren. Grundsätzlich sollten zum Erreichen der bestmöglichen Reflexionsstrukturen so viele Streukörper wie möglich verwendet werden, insbesondere für ein diffuses Horizontalfeld im Wandbereich zwischen ca. 0,5 und 2,5 m über FOK. Die Rückseite dieser Streukörper könnte darüber hinaus, wenn variabel positionierbar, mit Breitbandabsorbern versehen werden, z. B. dicken porösen Absorbern (Mineralwolle). Es ist außerdem denkbar, darüber hinaus zusätzliche Elemente zur Verfügung zu stellen, die zu einer Eingrenzung eines Bühnenbereichs im Falle kleiner Besetzungen verwendet werden können. So könnte bei Bedarf das Klarheitsmaß um die Bühne noch erhöht werden.

9.3.3. Variable Absorptionsmaßnahmen

Zur weiteren Verringerung der Nachhallzeiten für die jeweiligen Nutzungen sind zusätzlich bedarfsabhängig

variable absorbierende Flächen erforderlich. Diese sollten ein möglichst breitbandiges Absorptionsspektrum aufweisen und einfach verwendet werden können. Hierfür eignen sich am besten schwere Vorhänge, die mit einer Faltenzugabe von 50 bis 100 % in ausreichend großem Abstand vor den Wänden aufgehängt werden. Solche Vorhänge können als poröse Absorber vor allem für mittlere und hohe Frequenzen sehr hohe Absorptionsgrade erreichen. Um diesen zu tieferen Frequenzen hin zu erweitern, sollten die Vorhänge mit einem Abstand zur Wand angebracht werden, der im Schallschnellemaximum der angestrebten unteren Grenzfrequenz liegt. Da das Schnellemaximum in einem Abstand von $\lambda/4$ vor der Wand liegt, ergibt sich bei einer Schallgeschwindigkeit von 340 m/s und einem Frequenzbereich von ca. 350 Hz bis 400 Hz ein erforderlicher mittlerer Wandabstand von ca. 20 bis 25 cm. Zur einfachen Verwendung der Vorhänge sollten sich diese an leicht zugänglichen Wandabschnitten befinden. Hierfür stehen unterhalb des Balkons bis zu 212 m^2 Wandflächen zur Verfügung, oberhalb theoretisch zusätzlich weitere Flächen.

10. Raumakustische Simulationen

Die oben erarbeiteten Prinzipien für die Hörsamkeit verbessernde Maßnahmen waren nun im Rahmen einer Computersimulation bezüglich ihrer Wirksamkeit darauf hin zu evaluieren, ob damit die in Abschnitt 6.3 festgestellten Abstriche an der Hörsamkeit des Saals im Bestand unter den dargestellten Randbedingungen merklich verbessert werden können. Grundlage dafür ist das 3D-Computermodell aus der diesem Beitrag zugrundeliegenden Abschlussarbeit [1], welches nun als Modell der Simulationsrechnungen verwendet und modifiziert wurde.

10.1. Arbeitsweise

Marktübliche akustische Computersimulationsmodelle arbeiten nach dem Prinzip der geometrischen Raumakustik mit Hilfe von Schallstrahlenmodellen. Hierbei werden von einer Schallquelle Schallstrahlen (Normalen der Wellenfronten) ausgesandt, an den Begrenzungsflächen gemäß den geometrischen Gesetzmäßigkeiten der Reflexion reflektiert (gültig für Frequenzen, deren Wellenlängen klein gegenüber der Begrenzungsflächen und Gegenständen sind). Sobald eine bestimmte Schallquelle als Sender gewählt wird, variiert die Anzahl der in den verschiedenen Richtungen ausgesandten Strahlen in Abhängigkeit von der Richtcharakteristik der Quelle. Die ausgesendeten Strahlen nehmen bei jeder Reflexion in Abhängigkeit des gewählten Absorptionsgrades der getroffenen Oberflächen an Energie ab. Zählvolumina an den Empfängerpositionen registrieren die dort eintreffenden Schallstrahlen in Abhängigkeit von Laufzeit und Einfallsrichtung. Im Simulationsprogramm können darüber hinaus für die Geometrien und Flächen des Raummodells verschiedene Oberflächentypen eingestellt werden (frequenzabhängige Absorptionseigenschaften), sowie welcher Anteil der Schallstrahlen an den Oberflächen in Spiegelrichtung und welcher diffus (modelliert über einen Zufallsprozeß) reflektiert wird.

10.2. Beschränkungen

Solche Simulationsprogramme weisen derzeit minimal folgende Beschränkungen auf:

- Die Absorption der modellierten Oberflächen ist nicht winkelabhängig;
- Beugung an Kanten kann im Strahlenmodell nicht adäquat berücksichtigt werden;
- Es findet keine Modellierung der Schallstreuung an Oberflächenrändern statt;
- Computermodelle ersetzen gekrümmte Geometrien durch mehrere Segmente aus polygonen Flächen;
- Schallkonzentrationen an konkaven Flächen werden hierdurch nicht korrekt simuliert.

Aufgrund der Beschränkungen von akustischen Computersimulationen und von der Realität geringfügig abweichenden Formgebungen können die simulierten Werte von der Realität frequenzabhängig abweichen [5].

Für die Ziele der vorliegenden Validierung der Maßnahmen werden solche Computersimulationen dennoch als ausreichend geeignete Annäherung an die Wirklichkeit bewertet - eine geeignete Modellierung vorausgesetzt - insbesondere, wenn konkave Krümmungen keine Rolle spielen, wie es hier der Fall ist.

10.3. Durchführung der Simulationen

Mit Hilfe des erstellten 3D-Computermodells wurden im Simulationsprogramm "CATT-Acoustics" verschiedene Varianten simuliert, wobei die im unbesetzten Raum gemessenen Impulsantworten und Nachhallzeitwerte zur Kalibrierung für ein erstes Modell des derzeitigen (unbesetzten) Zustandes dienten (V1), siehe Abschnitt 10.4. Die Simulationen wurden durchgeführt für die Oktavbänder mit Mittenfrequenzen von 125 Hz bis 8.000 Hz mit 1.500.000 Strahlen pro Oktavband.

In das 3D-Computermodell sind die geometrischen Randbedingungen des Saals und seiner Ausstattung eingegangen, soweit dies für diese Betrachtungen in diesem Rahmen erforderlich und sinnvoll war.

10.4. Kalibrierung Simulationsmodell

Zur Kalibrierung des Simulationsmodells wurden die Eingangsdaten der Materialeigenschaften des Bestands (Variante V1) sukzessive iterativ modifiziert, dass die Ergebnisse der Simulation des Bestandes mit den Messergebnissen des Bestandes ausreichend gut übereinstimmten. Weitergehend wurden ETC-Graphen übereinstimmender Messpfade der Simulationsvariante V1 (unbesetzt, bestuhlt) und der Messungen des realen Zustandes (unbesetzt, bestuhlt) verglichen.

Der Vergleich der Impulsantworten zeigte, dass sich die globale Struktur der Reflexionen der Bestandsmessungen in der Computersimulation des Bestandes in einigermaßen ausreichender Genauigkeit wiederfinden lässt.

Es lag somit nach der Kalibrierung des Simulationsmodells eine für die vorliegende Untersuchung ausreichende Übereinstimmung von Simulationsmodell und Wirklichkeit vor, so dass davon ausgegangen werden kann, dass eine darauf aufbauende Untersuchung von Änderungen des Bestandes in der Simulation valide Ergebnisse zeigen kann.

10.5. Untersuchte Simulationsvarianten

Basierend darauf wurden die variablen Maßnahmen aus Abschnitt 9.3.3 und 9.3.2 als Varianten V2 und V3 sowie zusätzlich mit permanenten Maßnahmen als V4 im Simulationsmodell realisiert, siehe Abb. 8, und simuliert: Für die Variante V2 wurden die in Abschnitt 9.3.2 erläuterten 97 m^2 Diffusoren vor den Wänden modelliert und simuliert, siehe Abbildung 8, in Variante V3 zusätzlich die variablen 212 m^2 Absorptionsvorhänge unterhalb des Balkons aus Abschnitt 9.3.3, mit frequenzabhängigen Absorptionsgraden eines üblichen Beispielprodukts dieser Materialgattung. Für die Variante V4 wurden die Diffusoren um 180° gedreht, und deren als absorbierend modellierte Rückseiten wirksam sowie permanente Absorption auf dem Dach der Regiekabine modelliert. Ansichten der simulierten Modellvarianten sind in Anlage 8 abgebildet.

10.6. Untersuchte Simulationspfade

Die im Computermodell berücksichtigten Quell- und Empfangspositionen für die Impulsantwortbestimmung wurden analog den Messpositionen der Bestandsmessung wie in Abbildung 2 dokumentiert gewählt. Um Variationen durch Verschiebungen der Schallwandler zuvorzukommen, wird bei den Simulationsen mit festen Positionen gearbeitet.

10.7. Auswertung der Simulationen

Die Auswertung der Simulationsergebnisse erfolgt analog der Auswertung der Messungen in der Realität, siehe Abschnitt 6.1. Direkten Vergleiche der simulierten Varianten bei konstanten Positionen minimieren systematische Fehler.

10.8. Simulationsergebnisse

10.8.1. Nachhallzeiten

Abbildung 9 zeigt die bei den Computersimulationen resultierenden Nachhallzeiten der verschiedenen Varianten. Unter Berücksichtigung eines geeigneten Materials und der zur Verfügung stehenden Wandflächen unterhalb des Balkons ließen sich demnach Nachhallzeit um ca. 1,5 s erreichen (V3). Würden zusätzlich die Dachflächen oberhalb der Regiekabine mit porösen Absorbern versehen und die Rückseiten der Diffusoren absorbierend, wären auch Nachhallzeiten von bis zu ca. 1,3 s erreichbar (V4).

Für noch geringere Nachhallzeiten könnten zusätzliche absorbierende (Vorhang-)Flächen auf der Balkonebene sorgen. Damit wäre es möglich, die Nachhallzeit an die Erfordernisse der Nutzungen in nahezu beliebigen Zwischenstufen anzupassen.

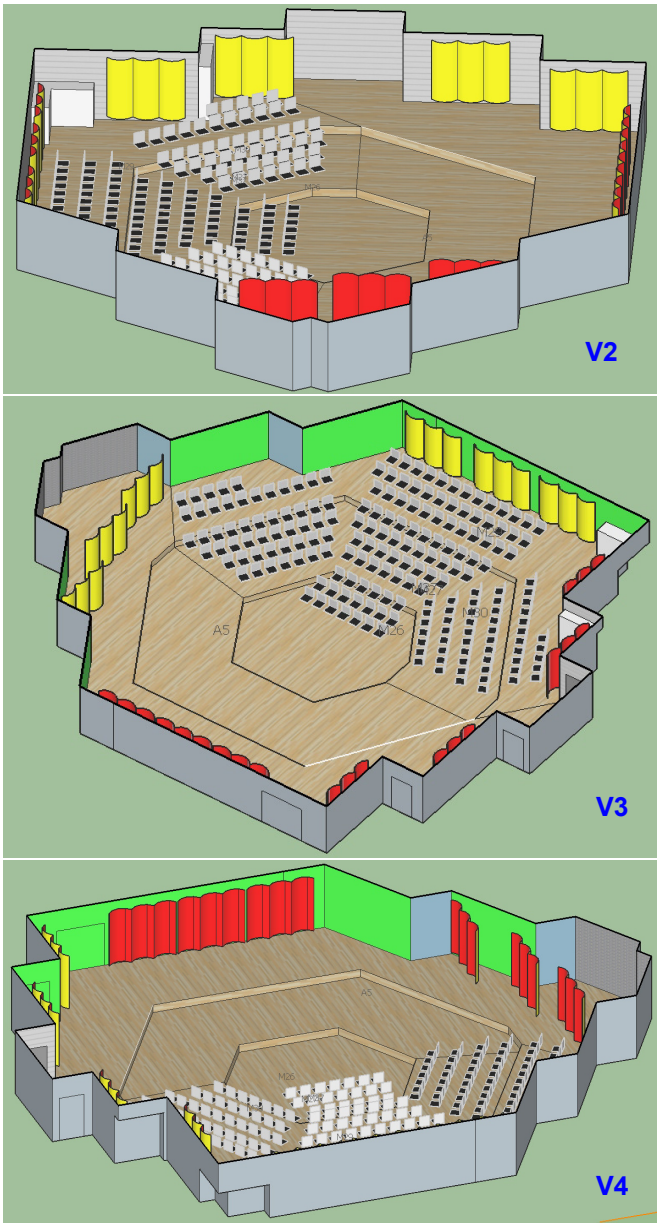


Abb. 8: Ansichten der Simulations-Varianten V2 (oben) V3 (mitte), V4 (unten) Orientierung "quer" mit Maßnahmen; gelb: Diffusor-Paneele; grün: absorbierende Vorhänge; rot: Absorption an Rückseite der Diffusoren

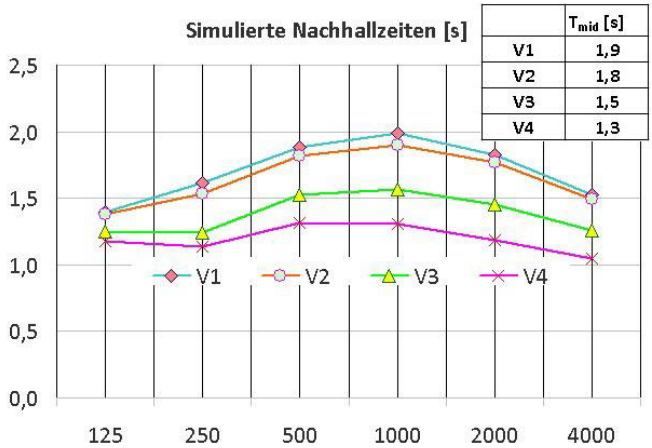


Abb. 9: Ergebnisse simulierte Nachhallzeiten der Varianten V1-V4

10.8.2. Impulsantwortstrukturen

Abbildung 10 zeigt typische Ergebnisse der simulierten Impulsantworten in ETC-Darstellung. Aus den Simulationserbenissen lässt sich im Vergleich zu Abb. 5 ablesen, dass die Reflexionsstruktur durch die im Simulationsmodell untersuchten zusätzlichen Maßnahmen effektiv verändert werden kann, und zwar selbst durch Maßnahmen unterhalb der Balkonebene in merklichem Ausmaß. Diese erste Übersichtsbetrachtung zeigt Optimierungspotenziale als Basis weiterer Planungen auf.

Artefakte mit Ursachen in der Raumgeometrie oberhalb der Balkonebene können mit Maßnahmen unterhalb dieser Ebene natürlich nur in gewissen Grenzen optimiert werden.

11. Conclusio und Ausblick

In der diesem Beitrag zugrundeliegenden Abschlussarbeit [1] eines Studierenden des Studiengangs Ton- und Bildtechnik konnte gezeigt werden, dass die Hörsamkeit im Partika-Saal des IMM Düsseldorf für die derzeitigen Nutzungen im Bestand suboptimal ist und Verbesserungen der Hörsamkeit darin erreicht werden können:

Über eine Verwendung der Bestuhlungsvariante „Quer“ als zu empfehlender neuer Standardausrichtung hinaus wurden als Möglichkeiten für Verbesserungsmaßnahmen Prinzipien zur Optimierung der raumakustischen Situation des Saals erarbeitet, welche die Basis für eine Planung von raumakustischen Verbesserungsmaßnahmen bilden können.

Die erarbeiteten Vorschläge für Maßnahmen wurden im Rahmen von raumakustischen Computersimulationen auf ihre Wirkung hin überprüft. So konnte gezeigt werden, dass mit geeigneten Maßnahmen unter Berücksichtigung der Randbedingungen im bestehenden Raum für die derzeitigen Nutzungen eine bessere Hörsamkeit erreicht werden kann.

Für die Dimensionierung der konkret zu realisierenden Maßnahmen im Rahmen einer raumakustischen Beratung und Planung sind damit die Grundlagen gelegt.

Darüber wäre noch eine - den Rahmen einer solchen Abschlussarbeit deutlich übersteigende - genauere Bestandsaufnahme und Analyse der vorhandenen und maßgeblichen Wand- und Deckenkonstruktionen des Saals erforderlich, sowie eine Feindimensionierung der zu realisierenden permanenten und variablen Maßnahmen.

Eine sorgfältige Planung, überwachte Umsetzung und messtechnische Validierung der Wirkung von solchen raumakustischen Maßnahmen ist dabei empfehlenswert.

12. Literatur

[1] Pfändner, B.: Untersuchungen zur Verbesserung der Raumakustik des Partika-Saals der Robert Schumann-Hochschule Düsseldorf unter Berücksichtigung der nutzungsspezifischen Anforderungen, unveröffentlichte Bachelorarbeit im Studiengang Ton und Bildtechnik der Hochschule Düsseldorf und der Robert Schumann Musikhochschule, Januar 2017

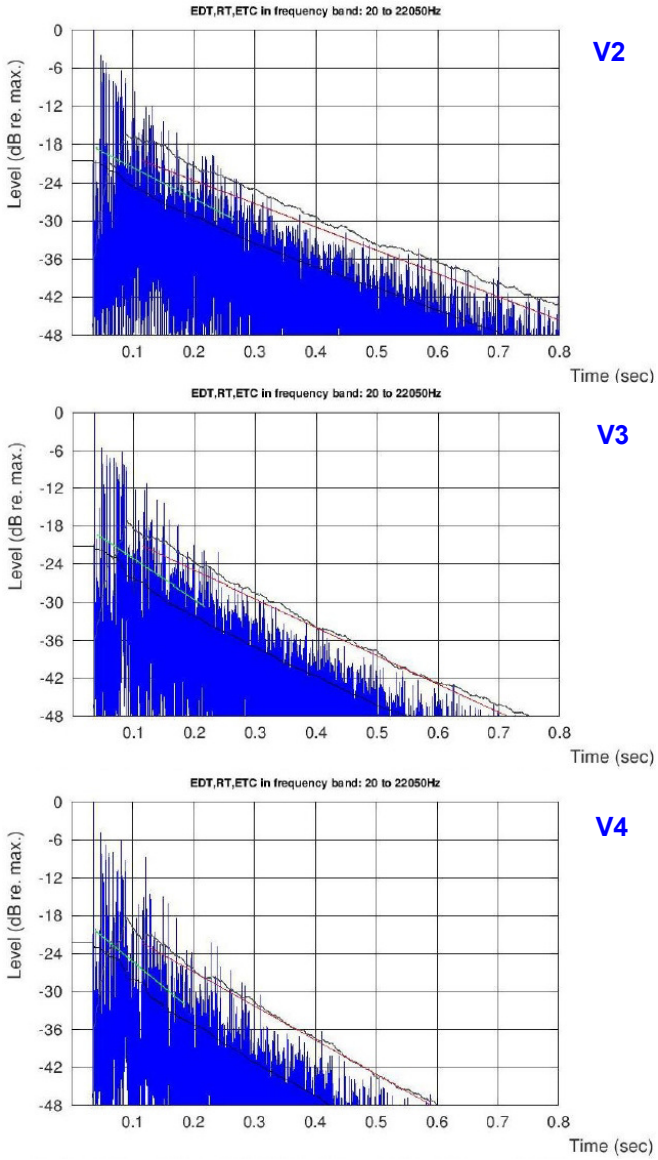


Abb. 10: ETCs der simulierten Raumimpulsantworten, Pfad Q6-M27, UBQ; oben: V2, mittig: V3, unten: V4

[2] Lorenz-Kierakiewitz, K.-H.; Pfändner, B. et al.: Mehrdimensionale visuelle Clusteranalyse raumakustischer Parameterwerte: Erweiterung auf andere Raumtypen und Vergleiche Vorher/Nachher, DAGA-Fortschritte der Akustik, München, 2018

[3] Lorenz, K.-H.; Breuer, F.: Simulation des Publikums durch spezielle Polyester-Textilien bei der raumakustischen Messung von Konzertsälen in der Praxis, DAGA-Fortschritte der Akustik, Aachen, 2003

[4] Lorenz-Kierakiewitz, K.-H.; Vercammen, M.: Entwicklung eines verbesserten raumakustischen Messlautsprechers für Raumimpulsantwortmessungen, DAGA-Fortschritte der Akustik, Berlin, 2010

[5] Heine. M., Lorenz-Kierakiewitz, K.-H. et al.; Vergleich von Maßstabsmodell- und Computersimulationsmethode bezüglich der akustischen Parameter C_{50} und D_{80} in einem Konzertsaal, DAGA-Fortschritte der Akustik, Stuttgart, 2007

Reproducing Low Frequency Spaciousness and Envelopment in Listening Rooms

D. Griesinger

David Griesinger Acoustics, Cambridge, MA 02138 USA Email: dhgriesinger@gmail.com

Abstract

Envelopment – the perception of being surrounded by a beautiful acoustic space – is one of the joys of great concert halls. Low frequencies in such spaces swirl around the head of a listener, giving the sound a life of its own. But in many playback rooms, particularly those with a single low frequency driver, low frequency reverberation is flat, frontal, and without motion. In this presentation we will show how with careful placement full range stereo loudspeakers or two independent subwoofers can bring envelopment back to a playback room, while at the same time making bass more uniform across the listening area. Contrary to current practice we choose loudspeaker positions that minimize room modes with high pressure at the listening position while maximizing modes with strong lateral components.

1. Introduction

This paper is a collection of ideas that might improve the sound in studios and playback rooms. But it is also a plea that we abandon the belief that low frequencies are not localizable in playback rooms, and need only one driver to reproduce.

While it is true that low frequency sine waves are difficult or impossible to localize, musical signals behave quite differently, particularly reverberation. Pluck a string bass in a large hall and the hall springs to life, with the sound moving and dancing around your head. It is not possible to capture that sound with a single microphone, or to reproduce it with a single loudspeaker.

The concept that low frequency envelopment gives a living quality to sound was suggested to me by the conductor Asher Fisch during a test of an acoustic enhancement system in the Vienna Volksoper. “Jetzt lebt der Ton” he said ecstatically after a beautiful pizzicato from the string basses. For example, in a recording of the Fauré requiem in Boston’s fabulous Trinity Church in Copley Square the low notes – including ones below 30°Hz - swim around your head, both in the church and in my playback room. The effect is marvellous. But the motion disappears with a single bass driver, no matter how powerful. This type of sound is one aspect of envelopment; the perception of that sound is all around a listener.

Newly developed and widely advertised systems for playback of music and films surround the listeners with multiple loudspeakers, ideally enabling the localization of sound events in three dimensional space. The goal is an “immersive” experience.

But a truly immersive reproduction needs to do more than reproduce localization. It needs to reproduce the spatial quality and the envelopment that we hear in nearly any space.

Most rooms have a characteristic sound, particularly at low frequencies. Sometimes it is just an omnipresent low frequency rumble. In this paper we are interested in the low frequencies in music that swirl around the head of a listener. Capturing the movement of air that creates envelopment is easy. But reproducing it in a small room with speakers takes care.

2. Localization and Envelopment

It is well known that localization in the horizontal plane at all frequencies depends on detecting the interaural time delay (ITD) and level difference (ILD) between the two ears. Localization in the vertical plane depends on timbre patterns created by individual pinnae at frequencies above 1000°Hz. (Loudspeakers in a 3D system above and below a listener need only reproduce the high frequency components of individual sounds. Such speakers can be very small if they are properly combined with speakers in the horizontal plane.)

Although we are mostly concerned with low frequencies in this paper, envelopment is also audible at mid and upper frequencies. Surprisingly perhaps the perception depends not only on the sound around and behind the listener, but also on the ability to distinctly localize individual sources.

Reverberation in large rooms arrives at the listener from all directions, but it is usually stronger in the direction of the sound sources, because that is where it is created. When the direct sound from the instruments cannot be distinctly heard and localized we hear only an inseparable mixture of direct sound and reverberation. The sound impression is blurry, frontal, and not enveloping.

However when individual voices and instruments can be distinctly localized into independent sound streams the ear and brain assigns the leftover sounds to a single background stream that we perceive as fully surrounding. This is the envelopment we hear in an excellent seat in a hall. [1]

In most spaces there is a distinct distance from the sources where sharp localization becomes impossible. We call this the Limit of Localization Distance, or LLD. Forward of the LLD reverberation is enveloping. Just a few feet farther back the direct sound and the reverberation blends together into a

frontal fuzzy ball of sound. The change in envelopment and sound quality is dramatic. Frequencies below about 120Hz behave differently. ITD and ILD can be detected for tones, and in a free field these sounds can be localized in the horizontal plane. The perception of spaciousness and envelopment arises when ITDs and ILDs are detectable but not stable. Room reflections arrive at the listener chaotically from all directions, causing the ILDs and ITDs at the listener's ears to fluctuate randomly at rates that depend on the frequency of the music and the size and reverberation time of the space. Fluctuations lower than about three Hertz are heard as motion – a kind of wander or swirl. Rates faster than about 20°Hz are heard as a kind of tone or buzz. But between 3°Hz and 20°Hz these fluctuations are perceived as a sense of enveloping space. Fluctuations above 20°Hz are simply perceived as a widening of the sound image.
These observations are not new. Blauert observed them as a graduate student, and ascribed the perception to what he called "localization lag".

To reproduce low frequency envelopment with either loudspeakers or headphones the reverberant fluctuations present in the original venue must be recorded and reproduced. Recording is not difficult with proper technique. However microphone techniques such as ORTF have too much correlation at low frequencies to be useful for reverberation, and sometimes well-meaning broadcasters will reduce the separation of a recording when they think it is excessive.

For example, years ago we made an impressively spacious organ recording. But when a test pressing came back the bass was flat and lifeless. A series of frantic phone calls revealed that the cutting engineer had rolled off the bass of the vertical component to cut a narrower groove. The reverberant fluctuations were mono, and there was no spaciousness or envelopment. The master had to be re-cut.

Figure one shows an L/R plot of the bass below 110°Hz from a randomly chosen two second segment of the Brahms requiem. The extreme variation in phase, typical for recordings of a string bass section playing arco, is very spacious on headphones. Note that the trace is not noise-like. The structure you see is important.

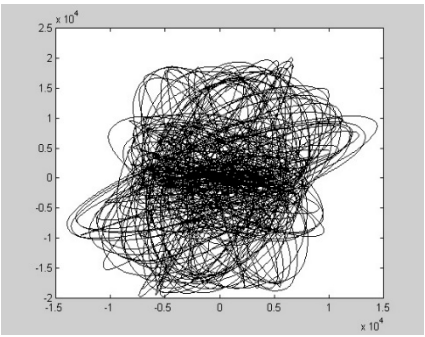


Fig. 1. An x/y plot of the amplitude and phase below 100°Hz in a two second segment from the Brahms Requiem, just before the chorus enters.

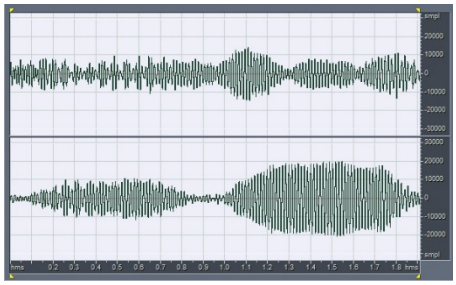


Fig. 2. The amplitude waveforms of the same segment. Note the amplitude fluctuations at ~20°Hz, followed by fluctuations at ~3°Hz.

Figure two shows the waveforms of this segment. The string basses are transitioning from a D at ~37°Hz to a C#. We can see that in the first half of the segment there is a rapid amplitude fluctuation at a rate of about 20°Hz. In the second half the fluctuation stabilizes to about a 3°Hz rate. With headphones or a successful playback room the sound is rich and enveloping. The spectrum shows strong peaks at ~37°Hz and 74°Hz. A playback system with a single subwoofer and a cross-over at 80°Hz would miss this effect entirely.

3. Low Frequency Envelopment in Small Rooms

From first principles we know that to reproduce a chaotic acoustic field in the horizontal plane you need at least three independent drivers. If you want a chaotic field in just the lateral direction you need at least two. We will consider a two-driver situation first, which assumes the listener is facing forward, and the main speakers are placed laterally.

To hear envelopment in this case we must create pressure differences at the two ears of a listener that depend on the phases and amplitudes of the low frequency reverberation on a recording.

To make pressure differences between the ears you need to find room modes or other methods that produce a lateral pressure gradient that varies with the phases of the recording. At low frequencies these pressure differences can be small, but they need to be heard. Unfortunately most modes in rooms have no lateral pressure gradients, and the gradients of those modes that have them are largest at the nulls of the sound pressure. To be able to hear the differences these gradients provide we need to limit the strength of competing modes.

All symmetric lateral modes, the ones that are excited by in-phase signals in the drivers, typically produce maximum pressure at the listening position. In addition all up-down and front-back modes have no lateral gradient, and also usually have maxima at the listening position. All mixed modes, of which there are a great many, have little or no lateral gradients. We need to somehow keep all of these under control.

But modes are not the only way bass travels from loudspeakers to the listener. Direct sound from powerful subwoofers can fill in frequencies between the room modes,

and reproduce frequencies well below the lowest room modes.

In our experience it is usually but not always possible to find loudspeaker positions that work adequately in a particular room. Successful installations in small rooms tend to follow a pattern.

When low frequency speakers and amplifiers were expensive and recordings were mono it made sense to put a low frequency loudspeaker in a corner where it would maximally excite all the room modes.

When stereo arrived we needed two full-range speakers, and corners were not a reasonable place to put them. Speakers moved away from the corners toward the center of a room, typically on stands. With this speaker arrangement the spatial quality of the bass was sometimes surprisingly good.

Modal resonances in small rooms are broader than they are in large rooms, and they are more likely to overlap more in frequency than those in large rooms.

In addition if we place speakers against a long wall at points 1/3rd and 2/3rds of the length in a rectangular room they are at nulls of the first symmetric lateral mode in the room – the one that would usually be excited when the two channels of the recording are in phase. But the speakers are at a node, so this mode will be attenuated.

The first asymmetric lateral mode, one octave lower, has a maximum pressure gradient at the center of the room. This mode is excited by the anti-phase component of the recording. If the room is reasonably well damped these modes will overlap. The in-phase signal creates some pressure, and the out-of-phase signal creates the pressure difference. The sound can be enveloping.

If the speakers are on stands at about half-way between the floor and the ceiling the lowest frequency will also be inhibited, reducing excess pressure from this mode. Envelopment is likely to improve further.

Several of my listening rooms in the past used this speaker arrangement with success. In one – a small control room at a music festival – had dimensions of about 12 long, 8°feet wide, and 8°feet high. The previous engineer had mounted the monitors in the corners of the narrow end. The sound was terrible. By experiment we found that putting the speakers at the 1/3 and 2/3s positions half-way up the long wall made an enormous difference. The sound was rich, spatially precise, and enveloping.

In preparing this paper I tried the same speaker arrangement in my bathroom, which is about the same size and well damped. The sound was good.

4. Low Frequency Envelopment in Larger Rooms

I decided to take this idea one step further. If we put the speakers on stands near the middle of the room at 1/3rd the distance between the front wall and the back wall the first front-back symmetric mode will also be inhibited. I tried this in my current studio, which is 20 feet wide, 17 feet deep, and 8 feet high. It is damped by large tube traps in the front corners.

The full-range main speakers were 10 feet apart and about 4 feet from the front wall. I had been quite happy with the sound, but I usually use a seven channel Logic 7 setup with side speakers that go down to 40°Hz. Spaciousness was very good.

When I turned off the L7 the sound was still somewhat spacious but it was not really great. So I decided to put the main speakers at 1/3 and 2/3 points of both the width and the length. This was not a good idea. The sound in the listening position was spacious – but I had to boost the bass by 12°dB to be able to hear it! So I put the speakers back where they were.

That the system sounded good with L7 switched on indicated that the side and rear speakers were contributing something important. The side speakers were nearly full-range, with good response to 40°Hz. When there is no rear steering Logic 7 runs the sides in parallel with the fronts, but with a delay of about 7ms. What would happen if I replaced the side speakers with two subwoofers?

To experiment with this idea I put two subwoofers at the sides of the listening position about twelve feet apart and ran them in parallel with the front speakers. The sound was marvelous. I have found before that putting subwoofers that augment frequencies between 30 and 150°Hz at the sides of the listening position can solve spatial problems in many rooms!

5. Envelopment in Two Dimensions

Earlier in this paper I mentioned that at least three low frequency drivers were needed to reproduce envelopment in the horizontal plane. In the previous sections I have assumed that the listener was always facing forward, and that only two drivers were necessary.

One of the goals of the Logic 7 system was to add enough drivers that envelopment could be heard regardless of how you turned your head. When you have five or more full range speakers available, you might as well use them! We mentioned above that the side speakers in a 7 channel system enhanced the envelopment for a listener facing forward. We did not mention that they can also, by adding a modal component between the front and rear of a room, create a sense of envelopment in the front-back direction. I did not test this in preparing this paper, but it was important to me in the past.

6. Square Rooms can be Difficult

There was a room at the former Harman Specialty Group building in Bedford Massachusetts that was nearly square. The marketing department had built it to demonstrate the most expensive speakers and audio equipment. But the bass sound was always very poor. Putting the speakers in the positions described above did not help at all. If we look at the modal patterns it is easy to see why.

Marketing insisted that the listening position was in the center of the room. The full range speakers being demonstrated were near the 1/3 and 2/3rds position in the width of the room. They were forward of the front wall, but not at the 1/3rd distance of the length. The lowest asymmetric mode in both the front/back and lateral directions was at 27°Hz, and the null was just at the listening position. Bass near this frequency was inaudible. The first modes with pressure in the listing position were at 54°Hz, and were both excited by the in-phase component of the recording. There was no way a pressure difference across the ears could be produced.

We could have put two powerful subwoofers at the sides of the listening area. They could be driven by the same signal as the front speakers, although ideally they should be driven from only the L-R component of the stereo signal. We did not try this. Marketing just turned up the bass. It is my impression that relatively square rooms are not uncommon for both music listening and for sound mixing. This is unfortunate. Sound mixers may not realize what they are missing.

7. Envelopment in Popular Music

Rooms intended for mixing and mastering are not necessarily designed to provide envelopment, although they generally include full-range main speakers. I have not worked in enough of them to know how they are set up and used.

Kick drums and bass guitars have powerful onsets, and a little reverberation can produce some interesting effects at low frequencies. Some of my favorite pop tracks have considerable envelopment, but most do not. Back when five channel mixing was popular recordings from artists such as Queen, Prince, and Michael Jackson enhanced the bass guitar and drums with lovely spatial effects.

In preparing this paper I spent a morning listening to a popular music radio broadcast while looking at the frequencies below 120°Hz on an x/y oscilloscope. About 25 percent of the recordings had at least some stereo bass. The bass in the rest was pure mono, alas.

8. Testing a System for Envelopment

It is not obvious how to determine whether a particular loudspeaker arrangement is better than another. A paper by Morimoto and Plosselt used de-correlated octave band noise as a test signal. For some time I had thought that such a test signal could be used in this way.

But in preparing this paper I found that it is not. De-correlated noise is quite difficult to localize in a room. Interference between the two channels results in level fluctuations at high rates. If you listen to monaural low frequency octave band noise with headphones you will hear that although you can tell it is balanced between the two ears, it is not sharply confined to the center. As frequency goes down the sound image gets wider and wider. Listening to such a sound in a room and trying to determine if it gets wider or enveloping when it there is a stereo bass signal is possible but not easy.

Music is not noise. It has structure, as can be seen in figures one and two. We need a test signal that is more musical. I made a set of tracks that have noise that that has been sharply filtered into five 4°Hz bands at the standard 1/3 octave frequencies of 40°Hz, 50°Hz, 63°Hz, 80°Hz, and 100°Hz. Figure 3 shows the waveform of the 63Hz track.

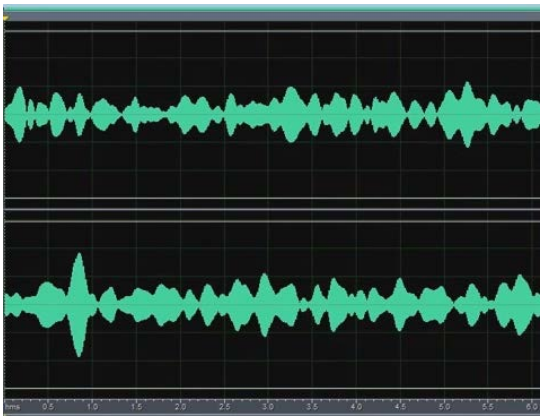


Fig. 3. An example of two uncorrelated low frequency tones formed by filtering noise with a 1/10th octave filter at 63°Hz. It is perceived as a low frequency tone that fluctuates in amplitude while fully surrounding the listener.

I played these five tracks in my studio using the two lateral subwoofers described previously. The sound was moving and spacious in all five bands. I then switched on the main speakers and their subwoofers. The lowest four bands were just as spacious as they were without the front speakers, but the highest band, at 100°Hz, was not spacious. It seems the 7ms delay in the side channels of L7 has an important function.

The five test bands are in my opinion a successful method of checking a system for its ability to reproduce envelopment. I have put these tracks and other interesting test material in a Dropbox folder. The address is here:

https://www.dropbox.com/sh/0xxrfcseeexrizn/AAB9nwIFNJgtYIwQ490g5txta?dl=0

9. Conclusions

This paper was inspired by an audio reviewer who had discovered an AES preprint I had written many years ago. I re-read it, and was mystified. [2] Since that time we have learned much more about how the brain processes sound, and I thought it would be brain processes sound, and I thought it would be a good idea to revisit the subject. The process has been both interesting and humbling. I had hoped that following some simple rules might make big improvements to the sound in most playback rooms, but the situation is more complex. Small rectangular rooms with speakers along the long wall can sound surprisingly good, but larger rooms have more numerous modes, and the modes have sharper peaks and valleys.

Low frequency drivers at the sides of the listening area are more important than I had realized, as is the usefulness of making the side speaker in a Logic 7 system capable of low frequency reproduction. Some rooms do not work at all with standard speaker arrangements. Fortunately for these spaces powered subwoofers that work to 30Hz or below near the sides of the listeners can work remarkably well.

The frequent lack of low frequency spatial ambience in popular recordings is unfortunate. There is no longer a need to worry about cutting a vinyl record, and high-quality monaural playback is becoming obsolete. These days popular music is mostly played on headphones, and bass attractiveness is both audible and appreciated. Modern tools can combine punch with an engaging sense of place. Engineers should use them more often.

Surround systems with at least four low frequency drivers can provide spatial bass throughout a space that is audible no matter how listeners turn their heads, but only if the recordings contain de-correlated low frequencies in at least four of the horizontal channels. This is often not the case. One of the design goals of Logic 7 systems was to provide de-correlated bass at the sides and the rear of a space, even if it was absent in the source material. It can work quite well.

9. References

[1] Griesinger, D.: The psychoacoustics of apparent source width, spaciousness & envelopment in performance spaces. Acta Acustica Vol. 83 (1997) 721-731. Available at the author's website. http://www.davidgriesinger.com/SPAC7A.DOC

[2] Griesinger, D.: Loudspeaker and listener positions for optimal low-frequency spatial reproduction in listening rooms. A paper on loudspeaker placement for reproducing envelopment, given at the Acoustical Society meeting in Vancouver, May 2005. Available at the author's website. http://www.davidgriesinger.com/asa05.pdf

Is single microphone position enough for immersive system equalization and level calibration in production monitoring?

A. Mäkivirta[1], T. Lund[1]

[1] *Genelec Oy, Iisalmi, Finland. Email: aki.makivirta@genelec.com*

Abstract

Immersive monitoring systems contain a large number of loudspeakers, each strongly and differently influenced acoustically by the room boundaries. This creates uncertainty regarding the actual in-room frequency response, results in large acoustical differences between individual loudspeakers, and can reduce the transparency and reliability of monitoring on such system. The aim of this work is to demonstrate this influence and then study the importance of using multiple measurement microphone positions instead of one measurement microphone position when preforming in-room frequency response calibration and system alignment. An NHK 22.2 immersive monitoring setup demonstrates the variability typical of immersive layouts, and the influence individualized in-room monitoring loudspeaker equalization can bring. After this, various spatial distributions in the vicinity of the main listening location in 12 rooms are compared to the typical situation of using single microphone location at the listening location. When using equalization to remove response differences between loudspeakers, no significant differences were found between single point measurements and spatial averages for small spatial averaging displacements in rooms having low reverberation times.

1. Introduction

Loudness-based production in broadcast relies on accurate monitoring for judging balance, formats, speech intelligibility, and audio quality. Reproduction systems for monitoring immersive audio presentations [1],[2] share certain assumptions about the monitoring system setup. These include flat frequency response at the listening location, the same level and time of flight from all monitors at the listening location, and normalized monitoring level [3],[4]. Room-related acoustic effects modify frequency responses of loudspeakers. For immersive monitoring layouts, these loadings are typically unique to loudspeaker locations. Because of this, loudspeakers for critical monitoring are equalised to reduce the acoustical room influence to improve accuracy in high quality professional listening [5].

The goal in professional monitoring is to hear the audio recording content without change. The monitoring system, the loudspeaker in the room, ideally approximates an allpass system within the audible frequency range.

$$\begin{cases} \left|H(e^{j\omega})\right| = 1 \\ -\frac{\mathrm{d}\arg\{H(e^{j\omega})\}}{\mathrm{d}\omega} = \text{const.} \end{cases} \quad (1)$$

Then, the monitoring system does not change the spectral content and merely causes a constant delay at all frequencies while delivering the audio for the engineer's ears. Recommendations consequently call for flat in-room response at the listening location, constant reverberation time in the room for audible frequencies, and sufficiently short room reverberation to not interfere excessively with the direct sound radiated by the loudspeaker. Minimum-phase implementations of loudspeakers lead to group-delay variation which, however, can be compensated to fulfil Equation 1 [6]. As loudspeakers are three-dimensional radiators, recommendations for directivity characteristics are also given, e.g. [5].

The obvious method for improving the sound colour at the listening location is electronic equalization. This is not a new idea. H. Tremaine's *Audio Cyclopedia* claims that RCA's John Volkman was the first to use electronic filters for room response equalization already in the 1920s [7]. Adjustable filters for room equalization have been available in the vacuum tube era (e.g. [8]). Parametric equalizer filters were introduced in the 1970s [9],[10] and digital signal processing has made both measurements and room equalization easy and precise.

While there is a consensus that a neutral loudspeaker has a flat anechoic frequency response, some researchers have questioned the suitability of the allpass target for the in-room frequency response, suggesting a down sloping character across the audio band. However, this largely seems to be motivated by listener preference [11],[12],[13],[14],[15],[16],[17],[18] while possibly too little attention has been paid to solving the problem with the 'circle of confusion'. This refers a self-referenced system where existing recordings are used to evaluate the room sound [18],[19].

1.1. One or more measurement locations

The present work concentrates on enabling reliable measurement of the in-room frequency response at the listening location, but not on the preferred in-situ response, which we regarded a separate issue.

The sound pressure measurable in a room at a single microphone location is local. At high frequencies the typical notching due to multi-way propagation and reverberation in rooms particularly at high frequencies varies when the microphone location changes. This effect is usually reduced by applying in-frequency smoothing with a sliding variable-width averaging window to the frequency response. Part of the late reverberation may also be rejected using a time domain window on the impulse response but this leads to loss

ISBN
978-3-9812830-9-9

The authors hold the copyrights of all of content in this manuscript. They have granted the VDT the right to publish it in print, online and as PDF download on https://tonmeistertagung.com

of low frequency information. Frequency smoothing reduces acoustic comb filtering and locality of measurement while extracting subjectively important low Q colorations in the midrange frequencies.

Spatial averaging of responses at multiple microphone positions can extract the common frequency response features visible in the measurement points. Spatial sampling of the room response in vicinity of the listening location has been suggested to improve the reliability of measurement and is assumed to produce a better representation of room acoustics over single point measurement, and thereby be more reliable starting point for system equalization.

Spatial sampling is done with several motivations. In the context of large seating area applications, such as cinema theatre equalization, even spatial sampling across the floor area aims to accomplish the best average equalization [5]. Similarly, spatial sampling can create information about the average coloration within an area, such as the large mixing console.

1.2. Immersive audio

Immersive reproduction systems are intended to create a spatially stable presentation of at least the frontal virtual sound images for a wide seating area [20] by having a large number of real loudspeaker sources correctly localizable at all seating position. Virtual sound images in immersive reproduction systems are created using two, three or more loudspeakers together, and localize the virtual images within the area span by these speakers, even if location error occurs for off-centre listening positions.

The combination of equalizing individual loudspeaker responses and careful alignment of playback levels and delays can reduce acoustic differences and can improve virtual sound image precision. Controlled in-room spectral response is the foundation of meaningful level calibration [4].

Immersive reproduction systems place loudspeakers on multiple height layers around the listener. The acoustical influence of the room differs more than for stereophonic or single-layer multichannel (surround) layouts, where it is easier to design room acoustics affecting the monitors to be similar. Increasing the number of height layers and loudspeaker channels complicates the acoustical playback system and increase acoustical differences between monitors.

1.3. Paper structure

The purpose of this work is to study the sufficiency of the single microphone position acquisition of the acoustical response at the listening position. Frequency domain smoothing (averaging in frequency) may in itself be able to increase spatial stability in a measured frequency response for features relevant to subjective experience of sound colour and for compensating the sound colour to reduce such colorations. First, a case study of the effects of electronic room calibration of an immersive NHK 22.2 system is presented. Then, research reported partially in [21] and [22] is revisited and additional analysis is provided. This study uses stereo loudspeaker setup in several monitoring rooms to study the benefit of spatial averaging for room calibration. The aim in both parts is to demonstrate the magnitude response at the nominal listening location. The study has been conducted in the context of professional monitoring spaces, typically having relatively low reverberation time and low level of the early reflections at the listening position.

Figure 1a. Method of suspending the upper layer monitors.

Figure 1b. Method of installing the lower layer monitors.

Figure 1c. Side view to the 22.2 immersive speaker layout.

2. Case study

A three-layer immersive monitoring system was set up according to the NHL 22.2 format. The installation room reverberation time in mid frequencies ($f > 250$ Hz) is 0.25 s on average and at low frequencies ($f < 200$ Hz) 0.5 – 0.8 s, increasing towards low frequencies. The room has an isolated prominent modal resonance at 44 Hz.

The loudspeaker system width is 3.22 m (from FL to FR, as well as from BL to BR). The channel name abbreviations are according to the SMPTE 2036-2 [2],[20]. The room height is 2.5 m, width 5.56 m and length 7.2 m. The listening position (measurement position) is 2.8 m from the left and right walls and 2.70 m from the front wall. The microphone height is 1.2 m. The upper layer monitors are suspended on rails attached

to the ceiling (Fig. 1a-1c). This places the upper layer loudspeaker front at 2.16 m height. The median distance to loudspeakers is 2.28 m, ranging from 0.96 m (TpC, room height limitation) to 3.01 m (TpFL). The time-of-flight (ToF) variation range was 4,97 ms (5,96 ms when TpC is included), see Table 1.

Table 1. Loudspeaker arrangement for the case example. Loudspeaker distance (DIST) and distance-related time-of-flight difference (DTOF) are given.

ROLE	TYPE	LAYER	DIST. (cm)	DTOF (ms)
BtFC	8330A	lower	211	2.62
BtFL	8331A	lower	257	1.28
BtFR	8331A	lower	261	1.16
BC	8250A	middle	130	4.97
BL	8240A	middle	208	2.70
BR	8340A	middle	209	2.67
FC	8330A	middle	213	2.56
FL	8330A	middle	272	0.84
FLc	8351A	middle	230	2.06
FR	8330A	middle	265	1.05
FRc	8351A	middle	230	2.06
SiL	8430A	middle	248	1.54
SiR	8430A	middle	226	2.18
TpBC	8320A	upper	139	4.71
TpBL	8330A	upper	223	2.27
TpBR	8330A	upper	214	2.53
TpFC	8330A	upper	247	1.57
TpFL	8330A	upper	301	0.00
TpFR	8330A	upper	292	0.26
TpSiL	8320A	upper	233	1.98
TpSiR	8320A	upper	214	2.53
TpC	8320A	upper	96	5.96
Sub L	7380A	sub	280	
Sub R	7370A	sub	280	

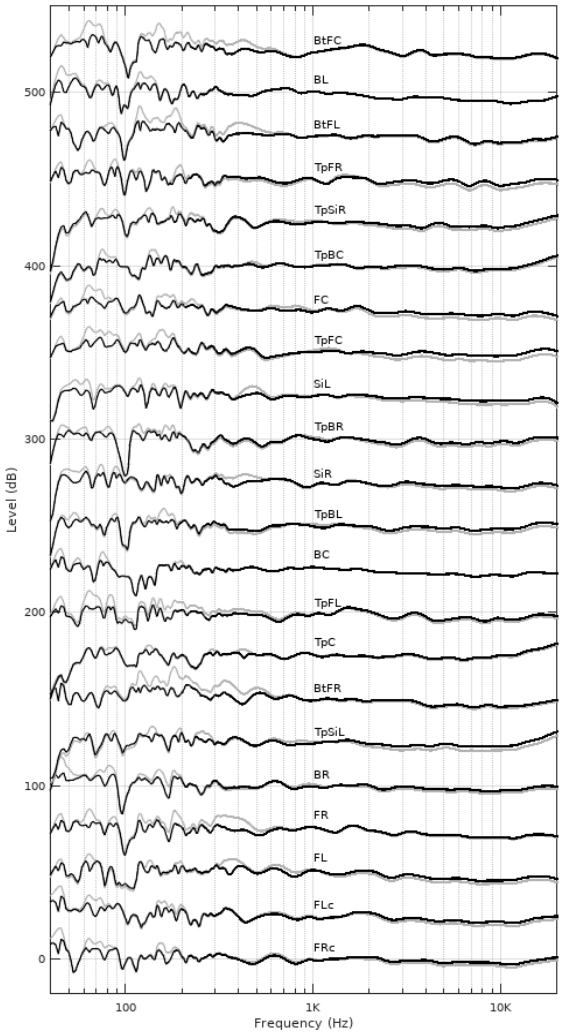


Figure 2. All response before (grey) and after equalization (black), aligned at 0.5 – 2 kHz and staggered by 25 dB to aid visibility.

Loudspeakers were measured at the listening position. In each loudspeaker, 6…10 second-order parametric notch filters and two low-frequency shelving filters were optimized for frequencies below 2 kHz, to reduce peaking above the reference level (mean level on log(f) scale in the range 0.5 – 2 kHz).

The outputs were adjusted to give the same mean output at all speakers. The times-of-flight were set the same using electronic delays in speakers.

Figure 2 presents shows the bi-smoothed loudspeaker responses (1/24 smoothing f < 200 Hz, then 1/3 octave). The 1/24 octave smoothing gives more detailed presentation of the SPL variation due to room modes at low frequencies. There, large level variation is typical due to sparse modal resonances compared to higher frequencies.

The maximum and mean levels recorded (Fig. 3, N=22) have been reduced. The audibility of mid and high Q peaks is higher than similar notches [23],[24],[25]. Equalizing the peaks reduces audible colorations and thereby renders loudspeaker responses more similar.

The equalization does not significantly change the amount of level shifting needed to bring the loudspeaker levels in alignment to the reference levels (Fig. 4). The magnitude equalization reduces values exceeding the reference level, seen in reduction of the maximum and median response level.

Looking at the pooled values exceeding the reference for all monitors (N=22), before equalization the level distribution shows a higher incidence of large deviations toward high pressure from the reference level (Fig. 5) than after the equalization (Fig. 6).

The case example demonstrates typical acoustical effects and level shifting seen in realistic, slightly acoustically suboptimal monitoring rooms. Equalization, level alignment and time-of-flight alignment can reduce acoustical differences and help create a predictable and reliable monitoring system.

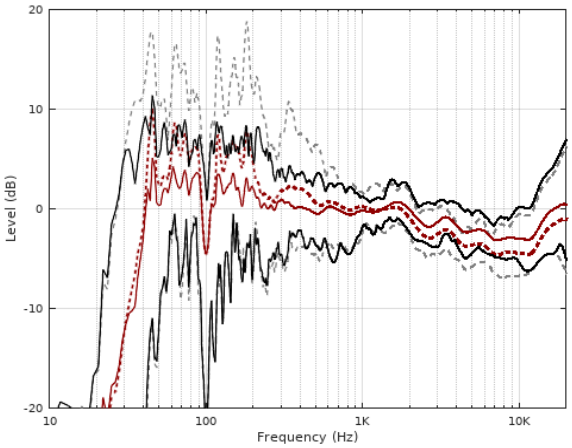


Figure 3. Minimum, maximum and mean responses after aligning the mean value calculated in log(f) in the range 0.5 – 2 kHz (N=22). The level scale is relative. The responses have been smoothed with a 1/3 octave window above 200 Hz and with a 1/24 octave window below this frequency.

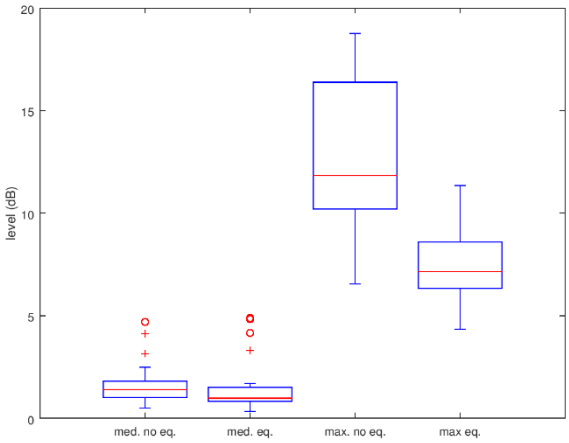


Figure 4. The median and maximum of the values exceeding the reference level, before and after calibration.

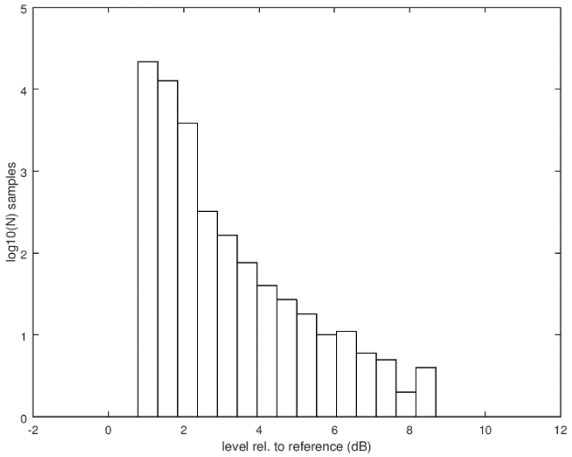


Figure 5. Histogram of the values exceeding the reference level before equalization (N=22).

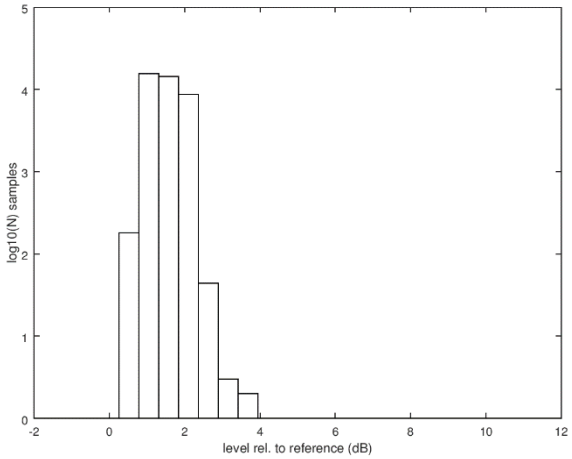


Figure 6. Histogram of the values exceeding the reference level after equalization (N=22).

3. Sufficiency of single measurement

3.1. Method

This section studies the ability of single position magnitude response measurement to render essentially the same information as spatial averaging. Methodology, choice of rooms and measurement data set used in this study has been reported in [21].

The measurement microphone has an electret capsule with omnidirectional characteristics. The main microphone position and 17 offset positions located at increasing distances are measured (Fig. 7 and 8). The offset distances are 0.1 meters to the front, 0.1, 0.25, and 0.5 meters to the side and back, and 1.5 meters to the back, see Figure 1. Two-way loudspeakers were used (5 in woofer, 2/3 in tweeter, type Genelec 8330 or similar).

The impulse response is measured using a log sine sweep with length 256 kilo-samples (2.97 s, fs = 44.1 kHz). Two sets of 18 measurements are taken for each room. In total, 432 measurements in 24 series are taken in the study.

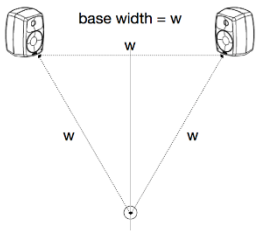


Figure 7. The principle of the loudspeaker layout follows the standard stereo pair placement. The main microphone position is marked.

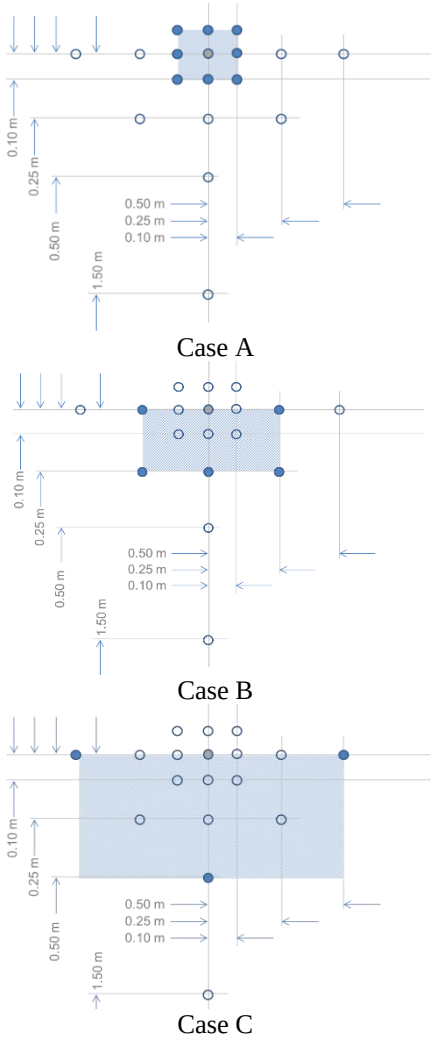


Figure 8. Definitions of the spatial average cases A to C showing the spatial averaging areas (shaded) and the numbers of measurement positions used (dark colour).

The magnitude response measurements are smoothed to standard 1/3 octave and 1/12 octave. Differences in distances are removed by level-normalizing responses. Measurements are therefore normalized in level using the mean response level across the 500 Hz to 5 kHz. The impulse responses are processed to generate (1) the main microphone position alone, (2) main position response averaged with the microphone responses at 0.1 m distance offset, (3) main response averaged with the responses at 0.25 m offset, and (4) main response averaged with the responses at 0.5 m offset.

Each spatial average response is calculated using the 1/3 octave smoothed data. All measurements in the spatial average receive the same weight. The weight of an individual measurement depends on the number of measurements included in the spatial average. After smoothing, the differences of the spatial averages to the single microphone position set at the listening location are calculated. This results in three sets of difference pressure graphs for each room, one for each spatial averaging case.

The mean pressure within an octave is calculated, resulting in 11 octaves from 16 Hz to 16 kHz. This results in 24 data points for each octave.

The spatial averages are compared to the single microphone position measurement. Measurements in all rooms and for all speakers are pooled. Box-and-whiskers plots are generated for each octave and for each spatial averaging distance. These describe the variability of deviation in the pressure from the single microphone position, in the octaves across rooms and loudspeaker position.

room	vol. (m3)	base width (m)	RT60 (s)	TER (ms)		LER (dB)	
				L	R	L	R
A	29.6	1.30	0.49	8.9	2.7	-11.9	-11.8
B	115.1	1.30	0.66	1.5	1.7	-10.5	-10.3
C	39.5	1.17	0.43	8.6	8.7	-10.8	-14.0
D	39.5	1.32	0.46	2.4	2.5	-13.5	-13.2
E	39.5	1.20	0.44	3.3	3.3	-15.9	-16.7
F	16.6	1.15	0.31	0.85	0.85	-10.5	-11.8
G	97.6	2.27	0.19	5.6	5.7	-10.8	-11.7
H	22.6	0.86	0.50	4.0	3.7	-16.7	-17.7
I	18.2	1.01	0.56	5.9	7.5	-17.2	-12.2
J	69.4	1.30	0.55	4.6	11.0	-14.8	-20.9
K	80	1.31	0.48	4.1	4.2	-17.2	-16.7
L	90.5	1.31	0.48	4.6	6.2	-20.9	-20.0

Table 2. Listening rooms used in the study. Reverberation time (RT60), highest early reflection level (LER) and the associated early reflection delay (TER) at the listening location for left (L) and right (R) loudspeakers.

3.2. Octave decay time in rooms

The twelve rooms (Table 2) share low background noise, acceptable reverberation time, and reasonably low level and acceptable direction of early reflections at the listening position.

The 12 rooms included in the study were evaluated for octave decay time. Three measurements were taken, at the reference location, 0.5 meters to the right and 1.5 meters to the back of the reference location.

The octave decay is compared to the recommendations [5], Fig. 9 and 10. The rooms fit recommendations for high

resolution listening location in terms of the median octave decay time, but the quartile limits exceed slightly the recommended flatness for octave decay time.

Notably, the lowest octave evaluated (63 Hz) shows large variability and two rooms show large reduction in decay time at high frequencies. However, the rooms are not untypical of the acoustic in real monitoring environments [26], Fig. 11, making the data collected in these environments useful for practical evaluations of in-room acoustics and the effect of spatial sampling of responses close to the listening location.

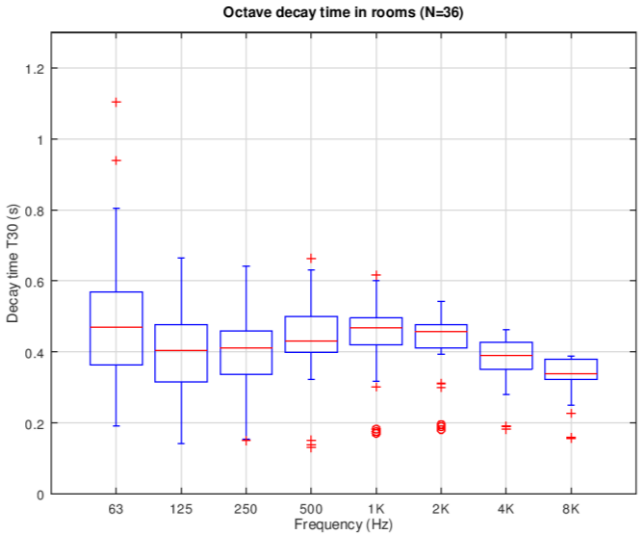


Figure 9. Distribution of the octave decay time in rooms. Three measurements were taken in each room.

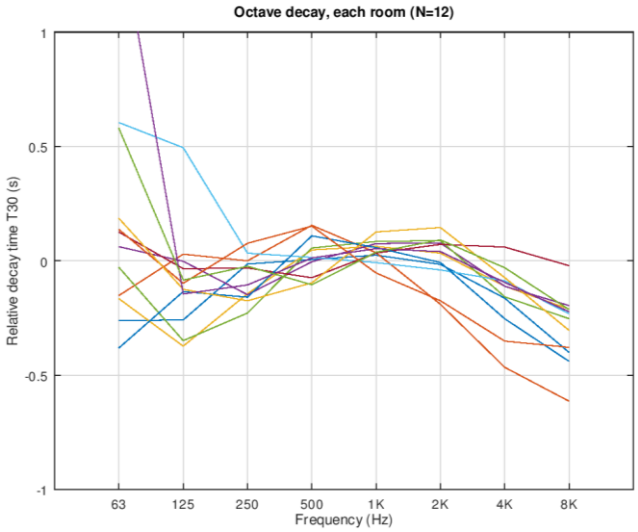


Figure 10. Deviations in octave decays in each room. Room average decay time in range [200Hz…4kHz] has been scaled to zero for each room, to enable comparison.

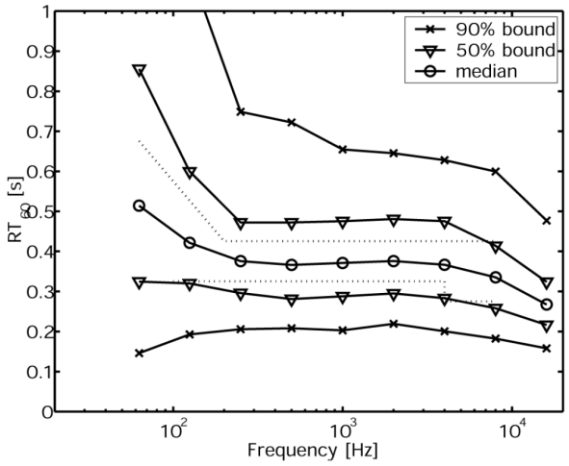


Figure 11. Distribution of decay time in professional monitoring rooms according to [26].

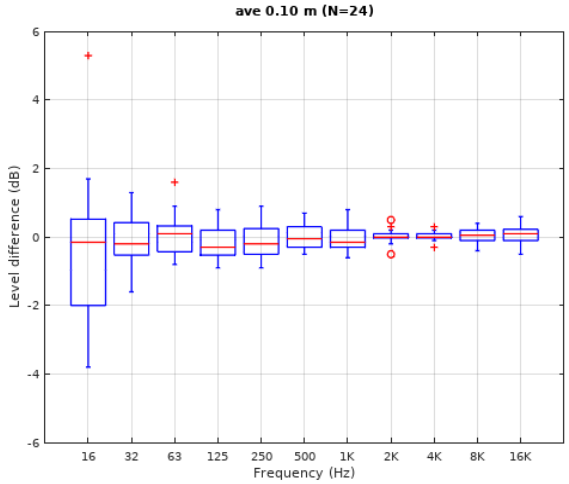


Figure 12. 0.1 m average vs. single measurement

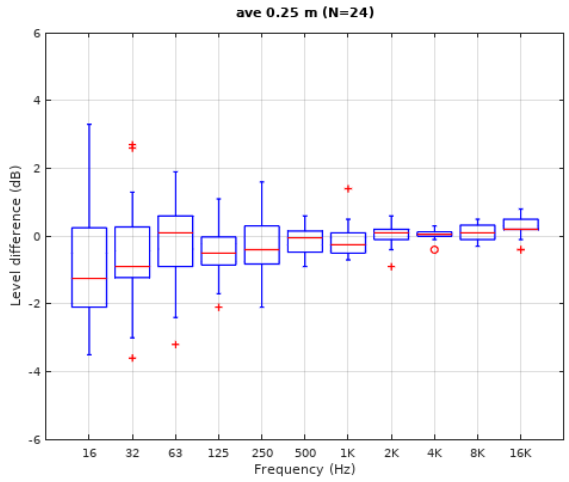


Figure 13. 0.25 m average vs. single point measurement

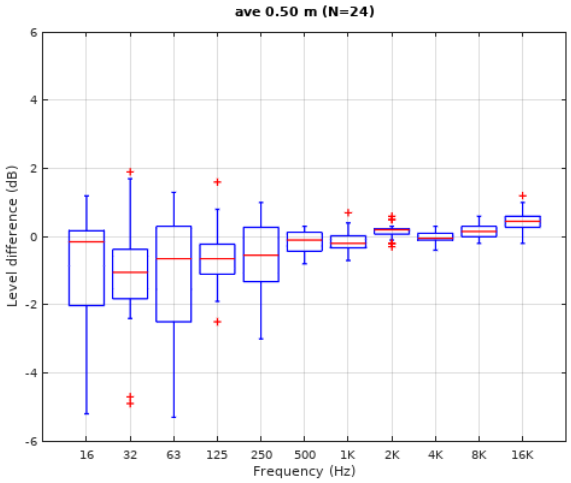


Figure 14. 0.5 m average vs. single point measurement

3.3. Frequencies f > 500 Hz

Differences to the single microphone position increase with increasing distance of the offset points.

The 0.1 m offset case (Fig. 12) shows small differences to the single measurement. The median difference remains less than 0.5 dB for all offset cases in frequencies f > 500 Hz (Fig. 13 and 14). Quartiles, describing 50% of the deviation, are within ± 0.5 dB. Extremes (minimum, maximum in octave) remain less than ± 1 dB for f > 500 Hz. These are close to the just noticeable limit (JND) [24].

As a conclusion, using single measurement appears to result in similar equalization as using a spatial average within the spatial offset range not exceeding 0.5 m.

3.4. Frequencies f < 500 Hz

For low frequencies f < 500 Hz, acoustic interaction of the room with the loudspeaker radiation leads to larger room-dependent variability.

Spatial averages taken using smaller offset distance show smaller variability, as expected. The octave median levels across the rooms and speakers show good agreement with the single measurement position for small displacement spatial averaging (0.1 m).

For frequencies f > 63 Hz the medians in all spatial average cases remain in agreement with the single microphone measurement to ± 1 dB. The lowest frequency medians (32 and 16 Hz) show larger deviation, but even then, the deviation to single microphone measurement does not exceed 2 dB, close to the just noticeable limit (JND) [23].

Quartiles range covering 50% of data in each octave remains within ± 2 dB while extremes (minimum, maximum) are seen deviating by as much as 6 dB from the single point measurement at frequencies f < 125 Hz and this happens mainly toward smaller pressure, indicating that this may be related to notches in the frequency response at low frequencies. Notches at low frequencies are known to have poor audibility [23].

	spatial average offset		
	0.1 m	0.25 m	0.5 m
mean	0,074	0,190	0,327
std	0,67	0,89	1,02
min	-3,50	-2,65	-3,00
max	4,30	3,30	4,80

Table 3. Pooled octave band statistics across all rooms for the full audio bandwidth [21].

	spatial average offset		
	0.1 m	0.25 m	0.5 m
mean	0,0	0,016	0,011
std	0,16	0,27	0,24
min	-0,45	-0,80	-0,35
max	0,40	0,60	0,70

Table 4. Pooled octave band statistics across all rooms for the octave bands 500 Hz and higher [21].

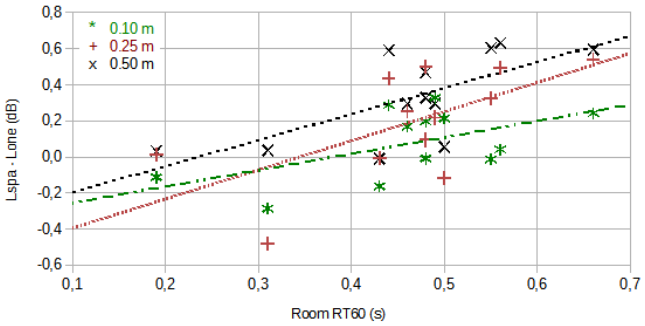


Figure 15. Mean level difference in octave bands, averaged across the audio band for the spatial offset cases 0.1, 0.25 and 0.5 m. The decay time is the maximum within the frequencies 0.4-4 kHz [21].

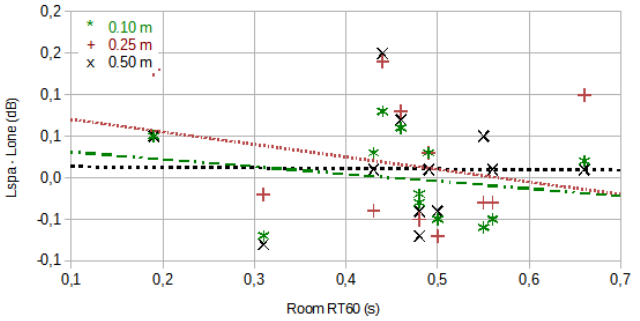


Figure 16. Mean level difference in octave bands 500 Hz and above, for the spatial offset cases 0.1, 0.25 and 0.5 m. The decay time is the maximum within the frequencies 0.4-4 kHz [21].

3.5. Significance of spatial averaging

Spatially averaged responses do not deviate significantly from the single point measurement at the listening position for small spatial averaging displacements (± 0.5 m) except for very low frequencies. The largest differences between the single point octave mean levels and the spatial average octave mean levels occur at low frequencies, in octave bands with

frequency lower than 500 Hz. The differences are small at 500 Hz and higher frequency octave bands (Tables 3 and 4).

When full bandwidth data is considered, a slight trend exists where increasing distance from the reference position and increasing reverberation measured inside the octave bands 0.4-4 kHz is linked to larger difference between the single point measurement and spatial average (Fig. 15) but this trend is mainly created by low frequency room effects and if octave bands at and above 500 Hz are considered, the trend disappears (Fig. 16). The reverberation time appears to have a stronger effect on this than the spatial average distance.

4. Discussion

For immersive reproduction systems, controlled in-room spectral response is the foundation for meaningful loudness calibration and accurate monitoring [2],[3],[4]. For example, the ATSC recommended practice A/85 [1] devotes an entire chapter to monitoring system calibration to dismantling what they call the 'circle of confusion' for audio monitoring. Reproduction system calibration is the key to achieving this.

We have studied the ability of single position magnitude response measurement to render essentially the same information as spatial averaging. The design of the present research is suitable for modelling typical audio engineering workflow, where decisions are taken primarily in seated position. The spatial average collection microphone position layout in our work consequently reflects the single engineer situation. Fixed microphone height at the listener's ear height was used, therefore the present work cannot reveal sound colour stability in vertical orientation although similar findings can be expected also for vertical direction.

In-frequency smoothing is the typical tool used for in-room frequency response measurements and it can attenuate very local features at mid and high frequencies, to extract the subjectively significant wider bandwidth features in the system response. Excessive smoothing easily leads to loss of precision at very low frequencies.

Spatial averaging can extract the common spectral features visible in all the spatial measurement points, rejecting local differences. Spatial averaging is typically used for very wide seating area applications, such as cinema auditoria, to remove the shared spectral coloration. Nevertheless, spatial averaging has also been offered as a more reliable starting point for system equalization over single-point measurement even for single-person working environments.

Magnitude responses measured have been smoothed to *de facto* standard 1/3 octave smoothing. This may limit the accuracy low frequency for narrow band resonances, audible particularly when associated with long decay time. This was demonstrated by the case example reported in this paper, where significantly less smoothing was applied to low frequencies, offering better visibility to modal resonance effects in the room.

Octave-band averaging of pressure can also affect extremes but is suitable for modelling the wideband sound colour variation as a function of frequency.

When in-frequency smoothing is applied mostly non-significant differences were found in our study between the single-point measurement and spatial averages taken with distances less than 0.5 meters for listening rooms similar to professional rooms in decay characteristics. Large differences between a single point measurement and a spatial average are connected to strong modal resonances in the room, typically happening at low frequencies where the natural room absorption is not sufficient to reduce the decay time. However, at low frequencies the audibility of such medium-to-high Q resonances is much reduced comparing to mid frequencies. Studio monitoring rooms typically have relatively short reverberation times, reducing local variation in pressure and further decreasing significance of spatial averaging.

5. Conclusion

The use of single point equalization appears to be a safe choice for measurement of studio monitoring rooms typically having relatively low reverberation times and well controlled room modes. The use of spatial average is not likely to significantly change or improve the outcome of equalization in the case of equalizing an audio reproduction system for a limited listening area essentially intended for one person.

6. References

[1] International Telecommunication Union: "Advanced sound systems for programme production", ITU-R BS.2051-1 (2017).

[2] International Telecommunication Union: *Multi-channel sound technology in home and broadcasting applications*, Report ITU-R BS.2159-6 (2014).

[3] Advanced Television System Committee (ATSC): *ATSC Recommended Practice: Techniques for Establishing and Maintaining Audio Loudness for Digital Television (A/85:2013)*, ATSC A/85

[4] European Broadcasting Union (EBU): *R128 Loudness Normalisation And Permitted Maximum Level of Audio Signals*, (2014).

[5] International Telecommunication Union (ITU): *Recommendation ITU-R BS.1116-3*, (2015).

[6] Mäkivirta A., Liski J., Välimäki V.: *Effect of Delay Equalization on Loudspeaker Responses.* AES 144 Convention (2018).

[7] Tremaine, H.: *Audio Cyclopedia*, 2nd. ed., H.W. Sams, Indianapolis, (1973).

[8] Villard, J.R.: *Adjustable frequency selective apparatus*, US Patent 2672529 (1954).

[9] Flickinger, D.: *Amplifier system utilizing regenerative and degenerative feedback to shape the frequency response*, US Patent 3752928 (1973).

[10] Massenburg, G.: *Parametric Equalization*, AES 42th Convention (1972).

[11] Pedersen, J.: *Loudspeaker-Room Adaptation for a specific Listening Position using Information about the Complete Sound Field*, AES 121 Convention (2006).

[12] Olive, S. et al.: *The Subjective and Objective Evaluation of Room Correction Products*, AES 127 Convention (2009).

[13] Pedersen, J. et al.: *Natural Timbre in Room Correction Systems*, 122nd AES Convention, (2007).

[14] Pedersen, J., El-Azm, F.: *Natural Timbre in Room Correction Systems (Part II)*, 32nd International AES Conference: DSP For Loudspeakers (September 2007).

[15] Toole, F.: *Loudspeaker Measurements and Their Relationship to Listener Preferences: Part 1*, J. Audio Eng. Soc., vol. 34, 4 pp. 227-235, (1986).

[16] Toole, F.: *Loudspeaker Measurements and Their Relationship to Listener Preferences: Part 2*, J. Audio Eng. Soc., vol. 34, 5 pp. 323-348, (1988).

[17] Toole, F.: *Sound Reproduction*, Focal Press (2008).

[18] Toole, F.: *The Measurement and Calibration of Sound Reproduction Systems*, J. Audio Eng. Soc., Vol. 63, No. 7/8, pp. 512-541 (2015).

[19] Toole F.: The Acoustics and Psychoacoustics of Loudspeakers and Rooms – The Stereo Past and The Multichannel Future, AES 109 Convention (2000).

[20] Society of Motion Picture and Television Engineering, *Ultra High Definition Television, audio characteristics and audio channel mapping for programme production*, SMPTE 2036-2-2008 (2008).

[21] Mäkivirta A, Lund T.: *Spatial stability of the Frequency Response Estimate and the Benefit of Spatial Averaging,* 141 AES Conv. (2016).

[22] Mäkivirta A., Lund T.: *Efficacy of Using Spatial Averaging in Improving the Listening Area of Immersive Audio Monitoring in Programme Production*, AES Int. Conf. Spatial Reproduction (2018)

[23] Olive S., Schuck P., Ryan J., Sally S., Bonneville M.: *The Detection Threshold of Resonances at Low Frequencies,* J. Audio Eng. Soc. 45 (3) pp. 116 – 128 (1997).

[24] Moore B., Oldfield S., Dooley G.: *Detection and discrimination of spectral peaks and notches at 1 and 8 kHz*, J. Acoust. Soc. America, 85 (2) 820 - 836 (1989).

[25] Bahritkar S., Robinson C., Poulain A.: *Equalization of Spectral Dips Using Detection Thresholds,* 140 AES Convention (2016).

[26] Mäkivirta A., Anet C.: *A Survey Study of In-Situ Stereo and Multichannel Monitoring Conditions*, AES 111 Convention (2001).

Eine frei zugängliche Stimulus-Datensammlung für Anwendungen in der Virtuellen Akustik und zum Testen von 3D-Audio-Technologien

Dieter Leckschat[1], Franz Zotter[2], Christian Epe[1]

[1]*Hochschule Düsseldorf - University of Applied Sciences, Germany, Email: tost.medien@hs-duesseldorf.de*

[2] *Kunstuniversität Graz, Institut für Elektronische Musik und Akustik (IEM)*

Kurzfassung

Nachdem die Notwendigkeit spezieller Testsignale für die Evaluierung von Algorithmen der virtuellen Akustik erkannt worden ist, wurde seit 2016 das im Titel genannte Projekt der Deutschen Gesellschaft für Akustik (DEGA) durchgeführt. Nach einer Konzeptphase wurden die nicht trivialen urheberrechtlichen Fragen geklärt, um das Ziel einer Gemeinfrei-Stellung von Audiomaterial und Metadaten zu erreichen. Hierzu wurde ein Mustervertrag erstellt, mit dem Künstler ihre Rechte an die DEGA abtreten. In der Folge wurden Tonaufnahmen mit sehr weitgehenden Dokumentations-Erfordernissen erstellt und ggf. tonmeisterlich nachbearbeitet. Dabei ist z.B. eine hohe Kanaltrennung zwischen Instrumenten erforderlich, da einzelne Quellen in virtuellen Umgebungen prinzipiell beliebig positioniert werden können. Das Material wird in eine Datenbank eingepflegt und durch die DEGA für die allgemeine, freie Nutzung zur Verfügung gestellt. Bereits nutzbare Produktionen umfassen etwa die Bereiche Klassik, Jazz/Funk oder Orgel. Daneben sind einfache Sprecheraufnahmen und synthetische Testsignale verfügbar.

1. Motivation und Einführung

Die Entwicklung neuer, räumlicher Audioformate, die eine gegenüber den Konventionellen deutlich verbesserte räumliche Wiedergabe und damit ein intensiveres Hörerlebnis ermöglichen, ist in den vergangenen Jahren forciert und erfolgreich weiterentwickelt worden. Für eine umfassende Evaluierung bedürfen die zusätzlichen Merkmale einer verbesserten räumlichen Audiowiedergabe, aber auch der Generierung neuer Strategien und Daten, die insbesondere gemeinschaftlich nutzbar sein sollen.

Im Fachausschuss Virtuelle Akustik der DEGA ist die Notwendigkeit eines Pools gut beschriebener Testsignale, welche in wissenschaftlichen Experimenten Verwendung finden sollen, als dringlich erkannt worden. Es handelt sich dabei vorrangig um Audiosignale, die in objektorientierter Weise verarbeitet werden können. So sind beispielsweise bei musikalischen Ensembles (Chor, Band, Orchester) die Einzelstimmen ohne Übersprechanteile anderer Musiker vonnöten. Ein weiteres Entwicklungsgebiet ist das Feld „Game Audio“ mit ebenfalls spezifischen Sound Design – Anforderungen.

Auf einen Antrag hin hat die Deutsche Gesellschaft für Akustik im März 2016 das Vorhaben, solche speziellen Testsignale zu erzeugen, zu dokumentieren und eine Stimulusdatenbank aufzubauen, als “DEGA-Projekt” genehmigt und unterstützt dieses, auch finanziell. Im Zuge des Projektes wurde 2018 das Memorandum VA 1201 verabschiedet [11].

2. Urheberrecht

Das Ziel im Projekt besteht auch darin, dass die produzierten Stimuli der Allgemeinheit zur Verfügung gestellt werden können, ohne dass Urheberrechte verletzt werden. Dies ist nicht so leicht wie es zunächst erscheint, denn “*in Deutschland, Frankreich oder Österreich … ist es generell nicht möglich, auf das Urheberrecht zu verzichten und das geistige Eigentum am Werk aufzugeben. Das Urheberrecht wird als eine Art Menschenrecht betrachtet, das weder verzichtbar ist, noch für sich genommen übertragen werden kann*” [1]. Aus diesem Grunde wurde mit Hilfe eines Medien-Fachanwalts ein Lizenzvertrag erstellt, welcher der DEGA weitestgehende Rechte an den Tonaufnahmen einräumt. Dazu ist es insbesondere notwendig, dass den Künstlern, die z.B. Musik einspielen, ein angemessenes Honorar gezahlt wird. Weiterhin müssen sie über den Zweck und die weitere Verwendung der Aufnahmen klar informiert werden. Mit diesem ersten Schritt, dem Schließen des DEGA-Lizenzvertrags, erwirbt die DEGA das Recht zur gemeinfreien Veröffentlichung.

Die Veröffentlichung durch die DEGA kann anschließend unter dem Creative-Commons-Modell “CC0” geschehen [2][3]. Dieses “*CC0 ist … eine standardisierte Erklärung eines … Verzichts und kann von jedem verwendet werden, der sein Werk für gemeinfrei erklären möchte*” [1].

3. Randbedingungen der Aufnahmen

Für den avisierten Einsatzzweck ist es nicht ausreichend, „normale“ Tonaufnahmen zu machen, denn es soll zum Einen gewährleistet sein, dass Schallsignale pegelrichtig, also mit dem ursprünglichen Schalldruckpegel wiedergegeben werden können. Dies wäre beispielsweise bei der Schätzung eines Entfernungseindrucks relevant. Zum Zweiten muss es möglich sein, die einzelnen Anteile eines Ensembles als „Objekte“ weiterzuverarbeiten, beispielsweise an beliebigen Stellen im Raum zu platzieren, oder auch bei einer Wiedergabe schlicht wegzulassen. Dies alles bedingt, dass Instrumente eines Ensembles separat aufgenommen werden müssen, damit akustisches Übersprechen komplett vermieden wird. Speziell für Gruppen, die es gewohnt sind, interagierend

miteinander zu spielen (Streichquartett, Jazzband) ist diese Forderung eine Schwelle, die es zu überwinden gilt, bspw. durch „guide tracks“. Die künstlerische Qualität von Einspielungen darf im Idealfall nicht unter diesen für die Musiker ungewöhnlichen Bedingungen leiden.

Im Vorfeld der Aufnahmen und als weitreichend gültige Empfehlung wurden so genannte Dokumentationserfordernisse zunächst in tabellarisch gesammelter Form, später als Prosatext zusammengestellt. In Kürze umfassen diese Erfordernisse folgende Punkte:

1. Äußere Aufnahmebedingungen

Hier soll die Raumgeometrie festgehalten werden, als vermaßte Skizze und ggf. mit Fotos. Sodann ist die Aufstellung während der Aufnahme zu dokumentieren, wo sich also Mikrofone und Instrumente bei der Aufnahme befunden haben.

Manche Aufnahmen finden sinnvoll in reflexionsarmer Umgebung statt. Nicht immer ist dies allerdings optimal oder überhaupt möglich, etwa bei einer Kirchenorgel. Dann ist die Akustik des Aufnahmeraums von Belang; hier sollte mindestens eine frequenzabhängig gemessene Nachhallzeit in Oktavbändern vorliegen, und optional ein berechneter Hallradius.

Insgesamt sollen die Aufnahmen mit sehr hochwertigem Equipment durchgeführt werden. Gerade bei älteren Produktionen zu Testzwecken wird teilweise ein zu hohes Rauschen bemängelt. Um dies einschätzen zu können, soll der Ruhegeräuschpegel im Aufnahmeraum bestimmt werden und ein kurzes Stück „Hintergrundgeräusch“ mit aufgezeichnet werden.

2. Aufnahmeequipment

Die verwendeten Studio-Mikrofone und Referenz/Mess-Mikrofone sollen genau dokumentiert werden mit ihren Einstellungen wie Richtcharakteristik, Trittschallfilter etc. Nützlich sind an dieser Stelle auch Angaben zum Mikrofon-Eigenrauschen und dem Übertragungsfaktor (in V/Pa).

Die verwendete Studioelektronik, insbesondere Vorverstärker und A/D-Wandler sind mit den vorgenommenen Einstellungen aufzuführen. Hierzu zählen auch Parameter wie Abtastrate und Bittiefe der Aufnahme. Sofern machbar ist es immer geraten, zusätzlich zu den Studiomikrofonen die Aufzeichnung eines kalibrierten Messmikrofons „mitzunehmen“.

3. Dieser Punkt ist optional: Es können Impulsantwortmessungen und Vergleichsmessungen mit Referenzmessmikrofon und Lautsprechern durchgeführt werden. Es wird also nach der Aufnahme z.B. eines Cellos das Instrument durch einen Messlautsprecher ersetzt und der Weg von dieser Quelle zum Mikrofon (Messmikrofon und/oder Studiomikrofon) samt aller Raumreflexionen erfasst. So kann später trocken aufgenommenen oder artifiziellen Signalen diese Raumimpulsantwort per Faltung aufgeprägt werden.

4. Musikalischer/sprachlicher Inhalt

Hier sind z.B. Noten, Texte oder Regieanweisungen festzuhalten. Bei manchen Sprecheraufnahmen wurden später sowohl professionelle Sprecher als auch Laiensprecher aufgenommen. Es ist hier immer elementar, dass das Material frei von Komponistenrechten o.ä. ist.

5. Instrumente und Quellen

Benennung (z.B. „Bassklarinette“), Hersteller, Typ, Baujahr, ggf. Seriennummer sollen angegeben werden, um eine Identifikation eventuell später vornehmen zu können, oder um Aufnahmen wiederholen zu können.

6. Besonderheiten bei der Aufnahme

An dieser Stelle können besondere Umstände der Tonaufnahme festgehalten werden. Beispielsweise wurden im Rahmen dieses Projekts Orgelaufnahmen gemacht, welche sich eine ganze Nacht hindurch erstreckten. Dieser Umstand kann Auswirkungen auf den Störschallpegel gehabt haben und verdient deswegen erwähnt zu werden.

4. Beispielhafte Tonproduktionen

4.1.1. Kirchenorgel

In Abbildung 1 ist das oben bereits erwähnte Anwendungsbeispiel illustriert: Aufnahmen der Orgel in der Kirche St. Jakobus in Hilden wurden gemäß den Anforderungen durchgeführt und dokumentiert. Es wurde für die Kirche sogar eine 3D-Animation hergestellt, so dass ein Anwender der Tonaufnahmen sich durch das Kirchenschiff bewegen kann und dabei die Mikrofonpositionen “ablaufen” kann. In der Abbildung zu sehen ist eine Klein-A/B-Anordnung von zwei (AKG C-414) Kondensatormikrofonen, nebst unmittelbar daneben angebrachten Referenzmikrofonen zur Erfassung des absoluten Schallpegels.

Abbildung 1: 3D-Modell eines Kirchen-Innenraums; Mikrofonierung der Orgel siehe Text

4.1.2. Akustikgitarre

An der Kunstuniversität Graz entstanden Aufnahmen einer Akustikgitarre [6]. Dabei wurde “klassisch” mikrofoniert, mit einem Studiomikrofen, das auf den Hals der Gitarre ausgerichtet war, und einem weiteren Studiomikrofon, welches den Klanganteil des Korpus´ schwerpunktmäßig erfasste. Wie in der Abbildung 2 zu erkennen ist, wurde auch hier jedem Aufnahmemikrofon ein Messmikrofon zur Seite gestellt, um absolute Schalldruckpegel zu erfassen

Abbildung 2: Gitarre mit Mikrofonierung

In der Abbildung 3 ist die nachfolgende Messung einer Raumimpulsantwort dargestellt; die Richtcharakteristik des Messlautsprechers kann zwar nicht identisch zu jener der Akustikgitarre sein, jedoch ist die Messung eine gute Orientierung bezüglich der Raumakustik des Sudios.

Abbildung 1: Messung von Raumimpulsantwort/Nachhallzeit

4.1.3. Streichquartett

Ein weiteres Beispiel zeigt die Abbildung 4, wo Studierende der Musikhochschule Düsseldorf ein Stück für Streichquartett einspielen. Auch hier ist wieder zu sehen, dass z.B. das Cello (rechts) sowohl mit einem rauscharmen und breitbandigen Kondensatormikrofon (Sennheiser MKH800) aufgenommen wird, als auch mit einem – in diesem Fall preiswerten – kalibrierten Messmikrofon. Gezeigt ist hier eine Probe, bei der noch kein Monitoring zum Einsatz kommt. Für die spätere Aufnahme wurde zunächst das Stück gemeinsam eingespielt und dann jeweils von einem/einer einzelnen Instrumentalisten, wobei der *Guide Track* mit geschlossenen Kopfhörern abgehört wurde.

Abbildung 4: Aufnahme des Streichquartetts im Tonstudio der Hochschule Düsseldorf

4.1.4. Nachhallfreie Orchesteraufnahme

An der TU Berlin entstand im Reflexionsarmen Raum eine Aufnahme unter nachhallfreien Bedingungen [7]. Eingespielt wurde die 8. Symphonie von Ludwig van Beethoven, von Orchestermusikern der Wiener Akademie. Da bei den Aufnahmen lediglich 21 Musiker anwesend waren, wurden vielfach Stimmen von einzelnen Musikern mehrfach eingespielt, so dass die komplette Aufnahme auf insgesamt 61 Stimmen/Tonspuren kommt. Die für die Musiker sehr ungewöhnliche Situation im Reflexionsarmen Raum wurde dadurch entschärft, dass ein Kopfhörer-Monitoring zum Einsatz kam, und es jeweils eine Videoeinspielung des Dirigenten und des Konzertmeisters gab. Durch ein gruppenweises, halbsequentielles Einspielen sind die einzelnen Instrumentengruppen nicht vollständig frei von Übersprechanteilen. Durch Präparieren des schalltoten Raums mit Stellwänden und optimierte Aufstellung der Instrumentalisten beträgt die Übersprechdämpfung aber mindestens 35 dB, was für fast alle Anwendungen mehr als ausreichend ist.

4.1.5. Funk/Pop-Produktion

Die zehnköpfige Düsseldorfer Funkpop-Formation “Rhythmussportgruppe” (Abbildung 5) steuerte zum Projektvorhaben zwei Songs bei, von denen einer eigens für diese Produktion komponiert wurde. Aufgenommen wurde wiederum im großen Tonstudio der HS Düsseldorf.
Besonderheiten ergeben sich hier beim Bereitstellen bzw. Exportieren der Signale aus der Protools-DAW. Für den hier angestrebten Zweck ist es sinnvoll, sowohl Rohspuren zur Verfügung zu stellen, die einfach nur die aufgenommenen Mikrofonsignale enthalten, und andererseits bearbeitete Spuren, die so oder ähnlich in eine Stereosumme oder anderweitige Mischung einfließen würden. Es muss nämlich angenommen werden, dass nicht jeder Nutzer der Datenbank die tontechnische Bearbeitung nachvollziehen kann oder will.

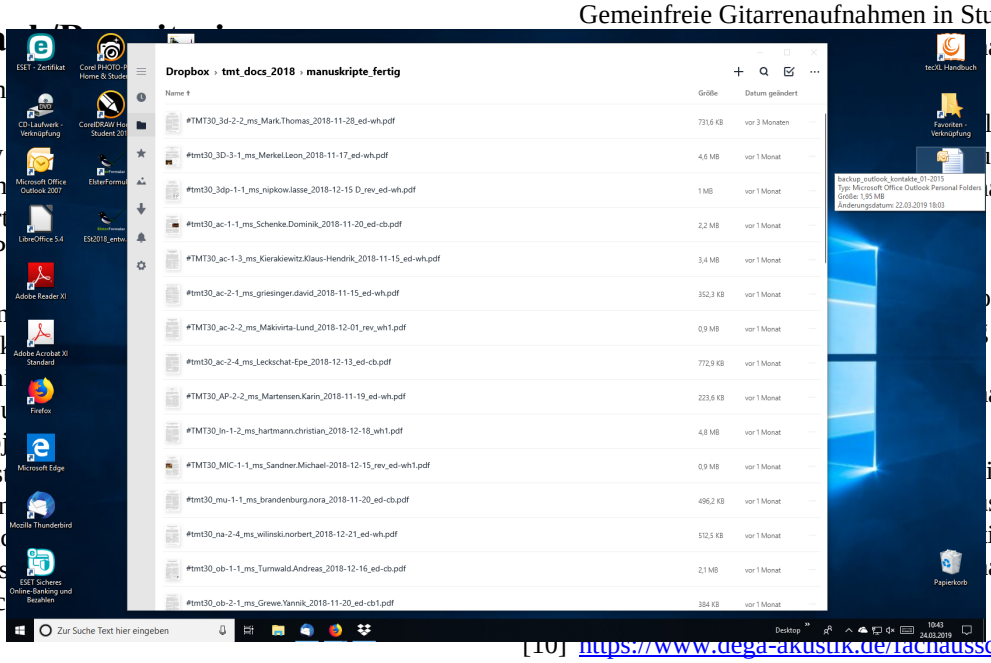
Abb. 5: Live-Auftritt der Funkpop-Band “Rhythmussportgruppe”

Eine Herausforderung liegt darin, dass eine fertige Tonmischung nach heutigem Standard nicht aus einem simplen Zusammenführen bearbeiteter Einzelspuren bestehen wird (Beispiel: *ducking* während Gesangspassagen). Hier wurde angestrebt, bearbeitete Tonspuren anzubieten, die einfach aufaddiert den ebenfalls zur Verfügung gestellten Referenzmischungen einigermaßen nahe kommen.

Eine weitere Hilfe für die Nutzer ist die Bereitstellung von *stems* (ein *stem* ist eine Vormischung von einer Gruppe zusammengehörender Tonspuren, z.B. aller Schlagzeugspuren. Ein *stem* klingt durch zahlreiche Maßnahmen wesentlich „reifer“ als die einfache Zusammenmischung der Einzelsignale). Die *stems* wurden in Stereo erstellt und so kann ein Klavierflügel oder das Schlagzeug mit einer gewissen Ausdehnung (*ASW, auditory source width*) in virtuellen Räumen positioniert werden.

5. Datenba

In Berlin wird ein
aus der Akustik
Fach-Community
Repositorium kön
fristig gespeichert
elektronischen P
Identifier (DOI)
Verfügbarmachun
ergebnissen, Dok
rithmen und ein
System für die Nu
vorliegenden Proj
zum OPERA-Syst
über eine Erhebur
das Web-Interfac
wurden. Vorerst s
Webseite des Fac
[10] abrufbar.



6. Danksagung

Die Deutsche Gesellschaft für Akustik fördert das vorliegende Projekt ideell und finanziell, so dass beispielsweise den beitragenden Sprechern, Komponisten und Musikern adäquate Honorare gezahlt werden konnten. Für die Erarbeitung des daraus resultierenden DEGA-Memorandums [11] wird den Fachausschussmitgliedern der virtuellen Akustik, musikalische Akustik und Hörakustik für ihre aktive Unterstützung gedankt, insbesondere Stefan Weinzierl, Malte Kob, Bernhard Seeber und Sascha Spors.

7. Literatur

[1] T. Kreutzer: Open Content – ein Praxisleitfaden zur Nutzung von Creative-Commons-Lizenzen, ISBN 978-3-940785-78-7

[2] https://creativecommons.org/publicdomain/zero/1.0/deed.de

[3] https://creativecommons.org/publicdomain/zero/1.0/legalcode

[4] S. Weinzierl
An Open Repository for Research Data in Acoustics (OPERA), Jahrestagung der Deutschen Gesellschaft für Akustik (DAGA), 2017, Kiel

[5] C. Epe, D. Leckschat, O. Kreuer, R. Reinhard
Erstellung und Dokumentation von Kirchenorgelaufnahmen zur Verwendung als Stimuli in virtuellen Umgebungen
Jahrestagung der Deutschen Gesellschaft für Akustik (DAGA), 2018, München

[6] D. Rudrich, F. Zotter
Gemeinfreie Gitarrenaufnahmen in Studioqualität
aft für Akustik

l

fnahmen
aft für Akustik

n Musikstücken
als Stimuli in

aft für Akustik

inzierl, F. Zotter
s-Datenbank für
ik”
aft für Akustik

[10] https://www.dega-akustik.de/fachausschuesse/va/

[11] DEGA Memorandum VA1201 „zur Durchführung und Dokumentation von Audio-Produktionen für wissenschaftliche Anwendungen in der Akustik“, 2018

Was soll klingen?
Der Einfluss des Tonmeisters auf das klangliche Ergebnis

Dr. Karin Martensen[1]

[1] *Musikwissenschaftliches Seminar Detmold/Paderborn, Email: karin.martensen@uni-paderborn.de*

Abstract

Wer steht eigentlich im Mittelpunkt einer Tonaufnahme? Das Werk? Der Raum? Der Künstler? Welche Rolle spielt der Digitalschnitt? Und wer entscheidet über alle diese Fragen? Mein Vortrag will zu diesen Überlegungen einen musikwissenschaftlichen Beitrag leisten. Ein Aspekt hierbei ist, dass der Musik die Fähigkeit fehlt, „Ich" sagen zu können. Sie ist daher immer auf ‚Umsetzende' angewiesen. Hierzu werde ich Aussagen verschiedener Tonmeister miteinander vergleichen und Beispiele für verschiedene Aufnahmephilosophien aufzeigen. Wie können sich alle Beteiligten darüber, was klingen und wie dies aufgenommen werden soll, angemessen miteinander verständigen? – Sicherlich sind die gestellten Fragen nicht letztgültig zu beantworten. Mein neues, an der TU Berlin angesiedeltes DFG-Forschungsprojekt möchte hierzu Grundlagenarbeit leisten.

1. Einführung

Wer oder was steht eigentlich im Mittelpunkt einer Tonaufnahme? Das Werk? Der Raum? Der Künstler? Ist es der Hörer? Oder ist es der Tonmeister? In meinen nachfolgenden Ausführungen soll es nicht um die technische Realisierung einer Aufnahme gehen. Vielmehr möchte ich versuchen, zu diesen Fragen einen musikwissenschaftlichen Beitrag zu leisten.

2. Der Werkbegriff und die Tonaufnahme

Um sich dem Thema anzunähern, soll hier der Begriff des „Werks" herausgegriffen werden, der bereits oben kurz angesprochen wurde. Seit langer Zeit gilt es als selbstverständlich, dass der Wille des Komponisten oder der Komponistin die Aufführung eines Werkes bestimmt. Jede Musikerin, jeder Musiker ordnet sich prinzipiell diesem Anspruch unter und bemüht sich darum, die Intentionen des Komponisten zu erkennen und das Stück ihnen gemäß zu interpretieren. Hier steckt der Teufel aber im Detail: Was sind die Intentionen des Komponisten, wo sind diese zu finden und wie interpretiert man sie angemessen?

Mit dem bloßen Hinweis auf die Partitur des jeweiligen Werkes greift man in der sogenannten Klassischen Musik sicherlich zu kurz, da sie die klingende Realisierung des Werkes und die Interpreten selbst nicht berücksichtigt. Zugleich ist hiermit eine Besonderheit der erklingenden Musik selbst angesprochen: es fehlt ihr die Möglichkeit, auf sich selbst zu verweisen, also „Ich" sagen zu können [1]. Wem die Definitionshoheit über ein bestimmtes Werk und seine Ausführung zusteht, kann daher aus der Partitur nicht hervorgehen.

Jeder Musikausübende, ob nun als Dirigent, Instrumentalistin oder als Sänger bzw. Sängerin, muss sich folglich mit der Tatsache auseinandersetzen, dass hier mindestens zwei Stimmen ‚sprechen': nämlich sie selbst als Interpreten mit ihren Wünschen, Ideen und ihrem jeweiligen Erfahrungshorizont. Und eben der Komponist, der durch seine Partitur ‚spricht'. Mit der aufnahmetechnischen Realisierung schließlich kommt eine dritte Dimension zum Werk hinzu, die folgende Frage lösen muss: Soll die Aufnahme das konkrete Ereignis einer Aufführung abbilden? Um dies mit einem bekannten Schlagwort zu umschreiben: „Are we recording the score or the hall?" [2].

Wie man an diesen kurzen Ausführungen schon erkennen konnte, ist der Werk- und der Autorschaftsbegriff in der klassischen Musik nicht unproblematisch und auch nicht leicht fassbar – und zwar schon deshalb nicht, weil diese Begrifflichkeiten tief in ästhetische Traditionen zur Klanglichkeit zurückreichen. Zwar zeichnet sich in diesem Genre seit einigen Jahren die Erkenntnis ab, dass man mit den etablierten Begriffen von ‚Werk' und ‚Autor' nicht sehr weit kommt bzw. dass sie nicht trennscharf genug sind, um die klingende Realisierung eines Musikstücks hinreichend genau zu beschreiben [3]. Ferner lässt die Forschung über die Wichtigkeit des Interpreten bzw. der Interpretation schon seit längerem keinen Zweifel mehr, und zwar sowohl in der New Musicology [4] als auch in der deutschsprachigen Musikwissenschaft [5]. Doch, wie Bettina Schlüter feststellte, „widmen sich bislang nur einzelne Studien der Notwendigkeit, neue Beschreibungskategorien und Terminologien zu entwickeln, die die Dimensionen von Schall, Geräusch und Klang mit in ihre Systematik einbeziehen und der Begrenztheit eines genuin musikanalytisch geprägten Vokabulars Alternativen entgegensetzen." Zur Erläuterung verweist Schlüter darauf, „dass über die Konzeption von

Musik als ‚Tonkunst' sich Kompositions- und Technikgeschichte über viele Jahrzehnte voneinander entkoppeln." Die daraus ihrer Meinung nach folgende „lange historische Wirksamkeit eines autonomieästhetisch begründeten Verständnisses von Musik als ‚Tonkunst', die auch die Entstehung des Fachs Musikwissenschaft im 19. Jahrhundert entscheidend prägt, [...] basiert im Wesentlichen auf einer diskursiven Konstellation, die von Beginn an versucht, ihre eigene ästhetische Kontingenz und damit mögliche alternative oder konkurrierende Musikbegriffe auszublenden [...]" [6]. Zwar wird diese Denkungsart, wie erwähnt, bereits seit den 1990er-Jahren von der sogenannten ‚New Musicology' kritisiert und darauf hingewiesen, dass das Zusammenspiel zwischen inner- und außermusikalischen Faktoren selbst historisch und kulturell variiert und sich auf je unterschiedliche Weise gestaltet, doch bietet auch dieser Ansatz keine Begrifflichkeiten oder Methoden an, wie man sich etwa dem Stimmgebrauch beschreibend und analytisch nähern könnte. Aus der Musikethnologie kam die Erkenntnis, dass zahlreiche, insbesondere außereuropäische Musikstrukturen nicht durch Notation fixiert werden können und daher dem hergebrachten europäischen Werkverständnis nicht zugänglich sind.

Auch wenn die deutschsprachige Musikwissenschaft auf all diese Erkenntnisse durch Öffnung ihrer Gegenstandsfelder reagiert und zumindest zu Teilen auch die wissenschaftlichen und wissenschaftstheoretischen Implikationen mitzudenken begonnen hat, sind Methoden, sich diesen auch tatsächlich nähern zu können, noch immer Mangelware. Es erscheint daher geradezu geboten, sich auch in der Historischen Musikwissenschaft mit den neuen Möglichkeiten zu beschäftigen, die sich – so formuliert es Schlüter – „der experimentellen Aneignung der seit dem 19. Jahrhundert kontinuierlich fortentwickelten Forschungen zur Klanganalyse, Klangsynthese, Klangspeicherung und Klangverarbeitung verdanken und die sich damit direkt dem ‚Unaufschreibbare[n] an der Musik und unmittelbar ihre[r] Technik' widmen" [7].

Eine allgemein akzeptierte Methode, um diese „Soundtechnologien" sowie das Tun der Beteiligten und die Ergebnisse auch in der klassischen Musikproduktion angemessen zu beschreiben, gibt es aber noch immer nicht. Auch wenn die Begrifflichkeiten zur Beschreibung eines Klangereignisses seit längerer Zeit etabliert sind und von Technikern bei ihrer Arbeit auch verwendet werden [8], ist unklar, ob dieses Vokabular auch von allen Beteiligten benutzt und auf dieselbe Weise verstanden wird. Zum anderen ist bis jetzt unerforscht, wie sich die Beteiligten unter- und miteinander über ihre Ziele, Wünsche und ästhetischen/künstlerischen Absichten sowie nicht zuletzt über den jeweiligen räumlichen Klangeindruck verständigen (dies beschreibt Weinzierl [9] als ein Desiderat), damit am Ende einer Aufnahmesitzung ein für alle befriedigendes klangliches Ergebnis steht.

Die Tonaufnahme bildet insofern einen Spezialfall, als das Tun des Tonmeisters und des Toningenieurs für das auf Platte aufgenommene klangliche Ergebnis ebenfalls essenziell ist. Auch bei diesen Beteiligten finden sich natürlich Annahmen, Erwartungen und Vorerfahrungen, die jeweils spezifische Auswirkungen in Bezug auf das aufzunehmende Werk haben. Eine der sich hier stellenden Fragen ist die schon erwähnte: Are we recording the score or the hall? Soll die Tonaufnahme das konkrete Ereignis einer Aufführung abbilden oder die Intention des Komponisten?

3. Are we recording the score or the hall?

Um diese schwierigen Fragen in einer Tonaufnahme anzugehen, wurden bekanntlich verschiedene Aufnahmephilosophien entwickelt. In einem im Januar 2017 im VDT-Magazin veröffentlichten Artikel werden einige dieser Philosophien, also Annahmen und Erwartungen von Tonmeistern, gut auf den Punkt gebracht. In dem Artikel wird gleich zu Beginn die folgende Frage an verschiedene Tonmeister gestellt: „Does a music recording represent an incident of a musical performance that takes place in a certain space at a certain time – or an interpretation of the musical score where the space and performance are manipulated to bring a piece of music to life?"

Zur Beantwortung dieser Fragen kamen in dem Artikel verschiedene Tonmeister zu Wort, deren Aufnahmephilosophien nachstehend kurz referiert werden sollen. Damit soll natürlich nicht gesagt sein, dass hiermit letztgültige Antworten auf diese Fragen gefunden wären. Und noch weniger soll es hier um ein Entweder/Oder gehen. Vielmehr soll hierdurch nur kurz angedeutet werden, dass es überhaupt verschiedene Herangehensweisen gibt. Und vor allem soll darauf hingewiesen werden, dass sich die Beteiligten an einer Tonaufnahme auf eine Herangehensweise bzw. eine Aufnahmephilosophie verständigen müssen, wenn am Ende eine für alle zufriedenstellende klangliche Realisierung stehen soll.

Die Unschärfe der Begrifflichkeiten und zugleich die Unmöglichkeit, zwischen der einen oder der anderen Alternative zu entscheiden – also zwischen Aufführung und Komponistenintention – zeigt meines Erachtens bereits die Einlassung des ersten der beteiligten Diskussionsteilnehmer. Er wird in dem genannten VDT-Artikel mit folgender Aussage zitiert: „He wants to present the music and uses the hall and score as tools to do so." Ein weiterer Gesprächsteilnehmer umging eine Antwort auf diese schwierige Entscheidung und stellte stattdessen den Hörer in den Mittelpunkt seines tonmeisterlichen Tuns. Über seine Aufnahmephilosophie heißt es in dem Artikel: „For him the listener and his enjoyment of the music are in the center of the approach to a recording." Für diese Zugangsweise führte der Tonmeister folgende interessante Begründung an: „Since the listener of a recording is lacking the visual sensation, it is the task of the recording team to enhance and interpret the score to create the desired effect just by listening." Folglich geht es um die Frage, wie Musik – losgelöst von ihrer körperlichen Aufführung – in der Tonaufnahme dargestellt werden könnte.

4. Theorie der Musikproduktion

Wir haben es hier also mit einer regelrechten Gemengelage zu tun: Das Werk, der Raum, der Interpret, der Tonmeister. Es gilt daher, sich über eine Philosophie oder mindestens eine Vorgehensweise zu verständigen. Wie aber können sich alle Beteiligten darüber, was klingen und wie dies aufgenommen werden soll, angemessen austauschen? Und ferner stellt sich die Frage, wie dies konkret geschieht? Musikwissenschaftlich gesehen ist an dieser Frage auch interessant, dass es hierzu noch so gut wie keine empirische Forschung gibt.

Auf den ersten Blick scheint es eine Binsenweisheit zu sein, dass auch in der sogenannten Klassischen Musik keine befriedigende Tonproduktion möglich ist ohne den durchdachten Einsatz von Aufnahmetechnik. Während jedoch bei Tonaufnahmen im Bereich der Pop- und Rockmusik nicht nur seit langem bekannt und akzeptiert ist, dass dort Aufnahmetechnik zur Erzeugung und Wiedergabe von Sound eingesetzt wird, sondern auch umfangreiche theoretische Überlegungen hierzu angestellt wurden, [10] ist dies im Bereich der klassischen Musik noch nicht geschehen.

Zahlreiche Studien zum Einfluss der Tonaufnahme auf das klangliche Ergebnis wurden sowohl von technischer als auch von musikwissenschaftlicher Seite bereits vorgelegt.

Stefan Weinzierl und Christoph Franke etwa erläutern ganz generell, dass „der Charakter der Musikaufzeichnung durch kaum ein anderes technisches Gestaltungsmittel so grundlegend verändert [worden sei] wie durch den Schnitt. Dies gilt für die Herangehensweise und den Ablauf einer Produktion ebenso wie für die Ästhetik des produzierten Tonträgers im Verhältnis zu der auf ihm abgebildeten musikalischen Aufführung. Durch die Möglichkeit, das aufgenommene Tonmaterial zeitlich neu anzuordnen und zu montieren entsteht die Illusion einer musikalischen Aufführung, die in dieser Form nie stattgefunden hat und vielleicht nie stattfinden könnte“ [11].

Hans-Joachim Maempel hält zudem fest, dass „Musik bei ihrer Vermittlung in mehrfacher Hinsicht transformiert [wird]: akustisch, interpretatorisch, klanglich, räumlich, zeitlich und kontextuell. Diese Transformation verändert natürlich das Erleben von Musik, wie schon Hans-Peter Reinecke in einem Hörversuch zeigen konnte. Und sie beeinflusst längerfristig Umgang, Einstellungen und Bewertungsmaßstäbe aller an musikalischer Kommunikation Beteiligten. Insofern lassen sich auch musikwirtschaftliche, technologische, musikhistorische und rezeptionspsychologische Zusammenhänge kaum unter Vernachlässigung dieser Transformation erschließen“. [12] Maempel weist ferner darauf hin, dass dies in besonderem Maße für den Klang gelte.

Dennoch kommt Maempel zu dem Schluss, dass „der Prozess der medialen Umsetzung von Musik bislang kaum wissenschaftlich untersucht worden“ sei. Andrew Hallifax, Tonmeister von Beruf, formuliert sogar: „Yet while the influences that affect musicians are increasingly subject to musicological investigation those of the engineer are not“, zumindest „outside the German Tonmeister tradition“. [13]

Auch von musikwissenschaftlicher Seite wurden bereits wichtige Aspekte dieses Prozesses beleuchtet [14], ohne jedoch ein theoretisches Fundament bereit zu stellen [15]. Forschung hierzu gibt es auch im Bereich der elektronischen bzw. elektroakustischen Musik [16].

Doch der entscheidende Unterschied ist, dass sich diese Autoren mit Komposition und Performance beschäftigen, aber nicht mit deren Produktion für ein medientechnisches Abbild. Genannt seien ferner die Forschungen der Sound Studies [17], die sich aber ebenfalls nicht mit der klassischen Musikproduktion beschäftigen.

Umgekehrt wurde von den methodischen Anregungen, die von dieser Forschung ausgehen, bislang für die klassischen Musik bzw. der Historischen Musikwissenschaft noch zu wenig Gebrauch gemacht. Im Gegenteil wird der Einfluss der Medialität auf dieses Genre bestritten [18].

Man könnte zuspitzend sagen, dass die verschiedenen Forschungsgebiete bislang nicht oder nicht genügend voneinander Kenntnis genommen haben.

Vor diesem Hintergrund verwundert es nicht, dass der Ruf von Martin Elste aus dem Jahr 1987, die Dominanz des „Partiturwissens“ in der Historischen Musikwissenschaft durch eine „Klangwissenschaft“ zu ersetzen [5], immer noch folgenlos geblieben zu sein scheint. Will man aber mit der „Klangwissenschaft“ Ernst machen, bedarf es nicht nur neuer Impulse hinsichtlich einer methodischen Fundierung. Es bedarf, worauf ebenfalls schon Elste hinwies, einer ganz neuen Art philologischer Quellenforschung: aus dem Klanggeschehen hinter dem Mikrophon muss quasi das Klanggeschehen vor dem Mikrophon herausgefiltert werden [19].

Es wurde bislang zudem kein Versuch unternommen, die Mechanismen des Aufnahmeprozesses in einen größeren Zusammenhang zu stellen. Oder anders formuliert: Es gibt im Bereich der klassischen Musik noch immer keine Theorie der Musikproduktion, in der – ähnlich wie in der Filmwissenschaft – Gestaltungsprinzipien herausgearbeitet sind sowie eine wissenschaftliche Basis zu deren technischen Umsetzung und diskursiven Verhandlung zwischen den beteiligten Akteuren gelegt wurde.

Zudem ist zwar bekannt, dass es sich bei der Tonaufnahme um eine Gruppenleistung handelt. Jedoch wurde bisher nicht untersucht, wie diese denn genau aussieht bzw. zustande kommt, und es wurde nicht gefragt, welche Folgerungen dies etwa für den Autorschafts- und den Werkbegriff haben könnte, der oben problematisiert wurde. Erst recht gibt es, wie schon gesagt, keine umfassende empirische Untersuchung, die alle Beteiligten einbezieht, also die Künstler auf der Bühne und die Künstler hinter der Bühne.

Mein neues Forschungsprojekt, dessen Finanzierung von der Deutschen Forschungsgemeinschaft erfreulicherweise zwischenzeitlich bewilligt wurde [20], und das an der TU Berlin angesiedelt ist (Fachgebiet Audiokommunikation), will hier Grundlagenarbeit leisten.

5. Forschungsprojekt „Das Tonstudio als diskursiver Raum“

Zwar wurde die eben skizzierte Thematik schon von verschiedenen Personen behandelt, die im Prozess der klassischen Musikproduktion künstlerisch und technisch beteiligt sind. Und selbstverständlich bilden sie die Basis der tonmeisterlichen Ausbildung. Zum Teil haben sich Überlegungen hierzu, wie dargelegt, in der Literatur niedergeschlagen. In der (musikwissenschaftlichen) Öffentlichkeit jedoch und zum Beispiel auch im Diskurs um die Frage der Natürlichkeit / Echtheit / Authentizität / Liveness bzw. der Realitätsnähe einer Tonaufnahme [14] sowie um die Frage, ob es sich bei einer Tonaufnahme um ein künstlerisches Produkt (ähnlich einer professionellen Fotografie oder einem Film) handelt, gehen die Meinungen nach wie vor auseinander [21]. Oder eigentlich müsste man sagen: es ist weithin überhaupt nicht bekannt, dass die genannten Diskurse auch im Medium Tonaufnahme ausgetragen werden [22]. Ferner darf man wohl sagen, dass die Bedingungen und Vorgänge einer Musikübertragung in der gegenwärtigen Forschungsdiskussion noch zu wenig berücksichtigt werden, insbesondere was die Ästhetik der Klangbildgestaltung betrifft.

Vielmehr kann man den Eindruck gewinnen, dass die Geschichte der Tonaufnahme dominiert wurde (und wird) von der Suche nach einer Wiedergabetreue, die die Zuhörer glauben lässt, das musikalische Geschehen, wie es wirklich war‘ vor sich zu haben. Das Aufnahmemedium als solches hingegen verschwindet dahinter. Die Gründe hierfür mögen im Geniediskurs sowie dem Autorschafts- und Werkdiskurs zu suchen sein, aber vermutlich auch in Marketingstrategien, Firmengeheimnissen [12], dem Natürlichkeitsdiskurs [23] und sicher auch in Abgrenzungsbestrebungen zwischen dem Klassik- und dem Popmusiksektor [24]. Aber untersucht ist dies nicht. Vielmehr ist festzustellen, dass es in der wissenschaftlichen Literatur und auch in der empirischen Forschung äußerst wenige Überlegungen zu einer Klangästhetik von Tonträgern gibt bzw., dass diese „unterbelichtet“ ist [11].

Mein Forschungsprojekt will sich diesen Fragestellungen dadurch nähern, dass das Tonstudio als diskursiver Raum beschrieben wird. Es soll also untersucht werden, wie über Annahmen und Wünsche bei einer Tonaufnahme gesprochen und gedacht wird. Auf diese Weise soll Klarheit über die getroffenen aufnahmetechnischen bzw. künstlerischen Entscheidungen gewonnen werden: Welche Annahmen liegen in Bezug auf die Tonaufnahme vor? Schlagen sich etwa vorhandene Konzepte von ‚Natürlichkeit‘ ggf. auch auf die konkrete Konzeption der Aufnahmesitzung nieder? Wie haben diese und andere Annahmen auf technischer Seite auf die Aufstellung der Mikrophone Einfluss genommen? Fühlen sich die Musiker gut abgebildet? Ist ihre künstlerische Idee angemessen wiedergegeben worden?

5.1. Theoretische Vorüberlegungen

Wie könnte dies nun in einen größeren Rahmen eingeordnet und dort beschrieben werden? Im Rahmen dieses Aufsatzes können die theoretischen Vorüberlegungen freilich nur skizziert werden (und sie müssen sicherlich auch im Verlaufe des Forschungsprozesses weiter profiliert und ausgearbeitet werden). Sie reichen von Michel Foucault (der formulierte, dass Wissenselemente mit einem System spezifischer Regeln und Zwänge konform gehen [25]), bis zu Siegfried Jäger, von dem folgende Erkenntnis stammt: „Wissen ‚haust‘ auch im Handeln von Menschen und in den Gegenständen, die sie auf der Grundlage von Wissen produzieren“ [26].

Mein Projekt basiert also auch auf der Annahme, dass man mit einer Beschränkung auf Sprechakt(theorien) [27] den diskursiven Raum ‚Tonaufnahme‘ nicht zureichend beschreiben und seiner habhaft werden kann. Vielmehr bin ich der Meinung, dass man die Tonaufnahme als diskursiven Raum, ja geradezu als ein Netzwerk der Diskurse auffassen kann, und dass diese Diskurse auch in den verwendeten Gegenständen, also etwa den Mikrophonen, wie in Knotenpunkten eines Netzes zusammenlaufen.

Dieses Netzwerk ist aber nicht nur eines der Diskurse, sondern zugleich auch eines der soziokulturellen Kontexte, in der die subjektiven Wahrnehmungsweisen der Bedeutungszuweisungen der handelnden Akteure ihre Aussagekraft erhalten. Hierzu beziehe ich mich auf Bruno Latour [28], Christopher Small [29] sowie Pierre Bourdieu [30] und bin der Meinung, dass alles, was im Zusammenhang mit dem Tonaufnahmeprozess gewusst und getan wird, ihn und alle daran Beteiligten formt und beeinflusst.

Mit Hilfe der so skizzierten Überlegungen ergibt sich die Möglichkeit, den gesamten (kulturellen) Prozess einer Tonaufnahme, an dem unterschiedliche wissenschaftliche, künstlerische und institutionelle Akteure beteiligt sind, im Auge zu behalten und dementsprechend zu beschreiben.

5.2 Fragestellungen

In dem Forschungsprojekt soll daher den einer Tonaufnahme im Studio zugrunde liegenden Prozess, die Handlungsweisen der Beteiligten, ihre Annahmen, ihre Kommunikation und die Umsetzung der musikalischen und technischen Ideen genauer beleuchtet werden. All dies kann im Einzelnen und im Detail betrachtet werden, wenn man sich auf die Tonproduktion konzentriert, also die Aufnahmesitzungen im Tonstudio.

Ich werde mich also nicht mit einer Livedarbietung eines Klangereignisses beschäftigen, also einem Konzert, sondern eben mit der Tonaufnahme als solcher. Dabei werde ich mich aber wiederum nicht mit ihrem medialen Endprodukt befassen, also einer CD, sondern mit dem Weg dorthin. Zwar bietet die CD im Gegensatz zum Konzert auf den ersten Blick die Möglichkeit, des klanglichen Ereignisses überhaupt habhaft zu werden, bevor dieses aufgrund seiner Transitorik wieder im Off verschwindet. Doch muss zum einen die CD stets in ihrer Medialität berücksichtigt werden; sie gibt als solche nicht die – wie auch immer geartete – klangliche ‚Wirklichkeit‘ wieder. Zum anderen berücksichtigt eine (nur scheinbare) Beschränkung auf den Tonaufnahmeprozess, dass dieser (wie schon oben an-

gedeutet) eine Gruppenleistung ist, an dessen erfolgreichem Zustandekommen mehrere Personen beteiligt sind. Ihr Tun ist aber auf der fertigen CD so miteinander verschmolzen, dass unauflösbar ist, wer was wie eingebracht hat, welchen Effekt und welche Relevanz dies jeweils hatte. Eine Detailbetrachtung ist jedoch möglich, wenn man Aufnahmesitzungen im Tonstudio begleitet.

Dabei sollen nicht nur die Abläufe, sondern auch die Kommunikation der Beteiligten über ihre Ziele, Wünsche und ästhetischen/künstlerischen Absichten sowie über den jeweiligen räumlichen Klangeindruck einer Prüfung unterzogen werden. Die Fragestellung lautet: Was tragen die an der Tonaufnahme beteiligten Menschen auf und hinter der Bühne zur Entstehung einer CD bei?

Hierbei geht es zum einen um die Kommunikation der beteiligten Gruppen (Ingenieure und Musiker) untereinander, zum anderen um die Kommunikation dieser Gruppen miteinander. Erforscht werden soll, wie die diese Gruppen etwa über Klangfarbe oder musikalischen Ausdruck miteinander sprechen und wie die Wünsche und Anregungen von der einen zur anderen Gruppe transportiert werden. Dadurch soll Einblick in die Diskussionskulturen und die Verständigungsprozesse aller Beteiligter gewonnen werden. Die Fragestellung lautet dabei: Wie wird Musik in Technik umgesetzt und umgekehrt?

Wie bereits angedeutet, kann der diskursive Raum Tonaufnahme nicht hinreichend genau beschrieben werden, wenn man sich nur auf die Kommunikation der Beteiligten beschränkte. Daher sollen auch reale Handlungen im Aufnahmeprozess berücksichtigt werden: Welche Mikrophone werden wie, warum und mit welcher ästhetischen Idee aufgestellt? Eine Überprüfung dieser Fragestellung geschieht in der Überzeugung, dass die Dinge, also das verwendete Equipment, sicherlich technisch-physikalische Eigenschaften besitzt. Wie diese Eigenschaften jedoch interpretiert werden, ja dass sie überhaupt interpretiert werden können, gilt es zu erhellen. Anders formuliert: die Geräte (also hier vor allem die Mikrophone) transportieren mehr als nur ihre technischen Eigenschaften: sie erzählen auch Geschichten über ihre Nutzung und über die Personen, die sie genutzt haben [31].

Die Forschungsfrage lautet daher: Inwiefern wird Kulturgeschichte in den Mikrophonen gespeichert? Was erzählen sie zum Beispiel über die berufliche Ausbildung und die berufliche Erfahrung ihre Nutzer? Von welchen Annahmen etwa über ‚Natürlichkeit' wird die Aufstellung der Mikrophone geleitet?

5.3 Vorgehensweise

Um allen diesen Fragen genauer nachzugehen, möchte ich Aufnahmesitzungen im Tonstudio begleiten und qualifizierte Interviews mit den Beteiligten führen.

Tatsächlich besteht eine der Annahmen qualitativer Forschung darin, durch die Einzeläußerungen hindurch das zugrunde liegende Muster oder Konzept identifizieren zu können, denn – wie Cornelia Helfferich schreibt – „einerseits sind die Einzeläußerungen Ausdruck dieses zu Grunde liegenden Musters, andererseits wird das Muster durch die Vielzahl seiner Äußerungen erfasst, es ist demnach keine dauerhaft fixierte Struktur." [32] In Vorstudien konnte bereits gezeigt werden, dass in Experteninterviews nicht nur die hinter den Aussagen liegenden Konzepte hinsichtlich einiger Teilmomente der Tonaufnahme (Klangeinstellung, High Fidelity) herausgelesen werden können [33].

Es konnte darüber hinaus gezeigt werden, dass Informationen und Deutungswissen der befragten Personen gar nicht voneinander zu trennen sind, sondern im Gegenteil ineinanderfließen bzw. dass die Information stark durch die Deutung geprägt wird. Es konnte sogar verifiziert werden, dass bei den Beteiligten unterschiedliche Erwartungen und Einstellungen etwa in Bezug auf die ‚Natürlichkeit' einer Tonaufnahme vorliegen, die sich dann auch auf die (gewünschte) konkrete Konzeption der Aufnahmesitzung niederschlug. Auf diese Weise konnte nicht nur das reflexive, sondern auch das handlungsleitende Wissen der Akteure und damit ihre Handlungspraxis herausgefiltert werden.

Ferner sollen Ausbildungsbedingungen von Tonmeistern und Toningenieuren mit der Frage untersucht werden, inwiefern ihre Ausbildung, ja ihr „Geschmacksurteil" [30] in Bezug auf die Ästhetik des Klangs ihr Denken und Handeln bestimmt.

Die durch empirische Forschung gewonnenen Daten sollen anschließend zusammengeführt werden. Dies dient nicht nur Erkenntnissen über die Abläufe von Produktionen und ihren Strukturen, sondern soll auch Fakten liefern etwa für Überlegungen zum Werk- und zum Autorschaftsbegriff. Auf dieses Weise soll auch versucht werden, mit einer „Klangwissenschaft" Ernst zu machen.

6. Zusammenfassung

Wer steht also im Mittelpunkt einer Tonaufnahme? Das Werk? Der Raum? Der Künstler auf der Bühne oder der Künstler hinter der Bühne? Was kann getan werden – oder vielmehr: Was wird getan, damit am Ende einer Aufnahmesitzung ein für alle befriedigendes klangliches Ergebnis steht? Sicherlich sind diese Fragen nicht letztgültig zu beantworten. Mein Forschungsprojekt möchte sich jedoch auf den Weg machen, hier Lösungen vorzuschlagen. Es ist angesiedelt an der Nahtstelle von Produkt und Prozess und will musikwissenschaftliches sowie praktisch-technisches Wissen zusammenführen.

Und vielleicht ist die skizzierte Fragestellung auch für die an der Tonaufnahme Beteiligten interessant. Es könnte sich hier eine Möglichkeit auftun, über das eigene implizite Erfahrungswissen nachzudenken, dass das Tun und Denken bei der Aufnahme ganz wesentlich bestimmt, aber bislang nur unzureichend erforscht ist. Möglicherweise könnte meine Arbeit auch eine Grundlage für eine Klärung der Frage sein, ob dem Tonmeister auch im Bereich Klassik ein eigenes Urheberrecht zusteht.

7. Literatur

[1] Grotjahn, R.: Die Teufelin und ihr Obrister. Primadonnen, Komponisten und die Autorschaft in der Musik. In: Gerards, M., Grotjahn, R.: Musik und Emanzipation. Oldenburg 2010, 131-140.

[2] Schwarz Anderson, U.: Are we recording the score or the hall? vdt 1/2017, 20-22.

[3] Kittler, F.: Der Gott der Ohren. In: ders.: Draculas Vermächtnis. Technische Schriften, Leipzig 1993, 130-148; Janz, T.: Qualia, Sound, Ereignis. Musiktheoretische Herausforderungen in phänomenologischer Perspektive. ZGMTH Sonderausgabe, Hildesheim 2010, 217-239.

[4] Cook, N.: Between Process and Product. Music and/as Performance. Music Theory Online 2001, Volume 7, No. 2.

[5] Elste, M.: Von der Partiturwissenschaft zu einer Klangwissenschaft. Überlegungen zur Schallplattenforschung. Jahrbuch des Staatlichen Instituts für Musikforschung Preußischer Kulturbesitz 1983/84, Kassel 1987, 115-144.

[6] Schlüter, B.: Musikwissenschaft als Sound Studies. Fachhistorische Perspektiven und wissenschaftstheoretische Implikationen. In: Volmar, A., Schröter, J.: Auditive Medienkulturen. Techniken des Hörens und Praktiken der Klanggestaltung, Bielefeld 2013, 207-226.

[7] Ebd. 222 mit Verweis auf Kittler: Der Gott der Ohren [FN 3].

[8] Kemper, E.: Realisierbarkeit und Beurteilung ästhetischer Klangkonzepte bei klassischen Musikaufnahmen, Diplomarbeit, Erich-Thienhaus-Institut, Detmold 2006; Binas-Preisendörfer, S.: Rau, süßlich, transparent oder dumpf. Sound als eine ästhetische Kategorie populärer Musikformen. Annäherung an einen populären Begriff. PopScriptum 10. Das Sonische. Sounds zwischen Akustik und Ästhetik, Berlin 2008; Weinzierl, S., Maempel, H.J.: Zur Erklärbarkeit der Qualitäten musikalischer Interpretationen durch akustische Signalmaße. In: Loesch, H.v., Weinzierl, S.: Gemessene Interpretation, Mainz 2011, 213-236; Martensen, K.: Der Soundbegriff – eine Methode zur Beschreibung des Phänomens ‚Operngesangsstimme'?, Archiv für Musikwissenschaft 2017/2, 140-159.

[9] Weinzierl, S.: Wirklich dabei sein. forschung 3/2012, 4-9.

[10] Smyrek, V.: Die Geschichte des Tonmischpults. Die technische Entwicklung der Mischpulte und der Wandel der medialen Produktionsverfahren im Tonstudio von den 1920er-Jahren bis heute, Berlin 2013; Porcello, T. G.: Sonic Artistry: Music, Discourse, and Technology in the Sound Recording Studio. Diss. Austin 1996.

[11] Weinzierl, S., Franke, C.: Lotte, ein Schwindel. Geschichte und Praxis des Musikschnitts am Beispiel von Beethovens 9. Symphonie. Vortrag auf der 22. Tonmeistertagung, Hannover 2003.

[12] Maempel, H.J.: Technologie und Transformation. Aspekte des Umgangs mit Musikproduktions- und Übertragungstechnik. In: de la Motte-Haber, H., Neuhoff, H.: Musiksoziologie, Laaber 2007, 160-180.

[13] Hallifax, A.: The engineer as stylist, 2008. URL: http://charm.cchcdn.net/redist/pdf/s6hallifax.pdf.

[14] Saxer, M.: Grammophon-Konzerte. Historische Medienkombinationen mit Schallplatte und der Wandel der Live-Ästhetik. In: dies.: Spiel (mit) der Maschine, Bielefeld 2016, 121-135.

[15] Schmidt Horning, S.: Engineering the Performance: Recording Engineers. Tacit Knowledge and the Art of Controlling Sound. Social Studies of Science 34/5 [2004], 703-731.

[16] Ungeheuer, E.: Elektroakustische Musik, Laaber 2002; Harenberg, M., Weissberg, D.: Klang (ohne) Körper. Spuren und Potenziale des Körpers in der elektronischen Musik, Bielefeld 2010.

[17] Pinch, T, Bijsterveld, K.: The Oxford Handbook of Sound Studies, New York 2012; Fetzner, D., Dornberg, M.: Intercorporeal Splits. Künstlerische Forschung zur Medialität von Stimme Haut Rhythmus, Leipzig 2015.

[18] Pfeiffer, K.L.: Operngesang und Medientheorie. In: Kolesch, S., Krämer, S.: Stimme. Annäherung an ein Phänomen, Frankfurt 2006, 65-84.

[19] Elste, M.: Tonträger und Tondokumente: In: Die Musik in Geschichte und Gegenwart, Sachteil Bd. 9, Kassel 1998, Sp. 646-662.

[20] DFG-Forschungsprojekt „Das Tonstudio als diskursiver Raum: Theorie, ästhetisches Konzept und praktische Umsetzung" (Eigene Stelle). Projektnummer: 419911712.

[21] Cook, N.: Methods for analysing recordings. In: Cook et al.: The Cambridge Companion to Recorded Music, Cambridge 2009, 221-245.

[22] Reinecke, H.P.: Musik im Original und als Reproduktion. Kommunikationstechnik 77 [1978], 140-151; Stolla, J.: Abbild und Autonomie. Zur Klangbildgestaltung bei Aufnahmen klassischer Musik 1950–1994, Marburg 2004.

[23] Harenberg, M.: Leiblichkeit und das Körperspiel der Maschinen. In: Saxer, M.: Spiel (mit) der Maschine, Bielefeld 2013, 227-244.

[24] Pfleiderer, M.: Musikanalyse in der Popmusikforschung. Ziele, Ansätze, Methoden. In: Bielefeld, C., Dahmen, U., Grossmann, R.: PopMusicology. Perspektiven der Popmusikwissenschaft, Bielefeld 2008, 153-171.

[25] Foucault, M.: Was ist Kritik?, Berlin 1992.

[26] Jäger, S.: Kritische Diskursanalyse. Eine Einführung, Münster 2012, 73.

[27] Austin, J.: Zur Theorie der Sprechakte, Stuttgart 1998.

[28] Latour, B.: Die Hoffnung der Pandora, Frankfurt 2000.

[29] Small, C.: Musicking: The Meanings of Performing and Listening, Hanover 1998.

[30] Bourdieu, P.: Die feinen Unterschiede. Kritik der gesellschaftlichen Urteilskraft, Frankfurt 1984.

[31] Hahn, H.P.: Materielle Kultur. Eine Einführung, Berlin 2014.

[32] Helfferich, C.: Die Qualität qualitativer Daten. Manual für die Durchführung qualitativer Interviews, Wiesbaden 2011, 22.

[33] Martensen, K.: Überlegungen und Interviews zum Einsatz von Aufnahmetechnik und zur Gruppenleistung Tonaufnahme im Bereich der Klassischen Musik. Acta Musicologica 2/2017, 145-170.

Loudness in radio broadcasting - a successful implementation illustrated by the example of Bayerischer Rundfunk

C. Hartmann[1]
[1] *Bayerischer Rundfunk / IVZ, Germany, Email: christian.hartmann@extern.br.de*

Abstract

"Loudness levelling is not beneficial to radio and only makes sense for television", or "loudness is a major competitive disadvantage in the context of FM transmission". These and many other preconceptions revolve around the subject of loudness normalization in the field of radio broadcasting. A successful project at Bayerischer Rundfunk shows, that a considerable part of these simply belongs in the world of myths and legends. In a joint effort of editorial departments, technical planning, system technology and production, it was possible to fruitfully introduce loudness normalization in all of BR's radio stations after several years of planning and realization. The following paper details the objectives and motives pursued by the project, explains the implementation concept and discusses the experiences gained and results achieved.

1. Introduction

Loudness levelling in television broadcasting according to EBU R 128 and ITU-R BS.1770 is used in more than 30 countries [2] and can safely be described as a worldwide success story. It owes its rapid spread and popularity not least to the fact that its advantages are clearly apparent to end customers: Unwanted loudness jumps while switching between channels, but also annoying deviations from the average loudness within a running program, e.g. during commercial breaks, have been reduced considerably. In countries where loudness levelling has already established, television has thus become more enjoyable for many viewers - the need for corrective actions with the remote control has become much less common. Positive viewer reactions and a significant decrease in complaints clearly confirm this tendency.

In radio broadcasting, however, loudness levelling is only very slowly gaining momentum. This is probably due to the fact that the main advantages of loudness levelling in radio differ considerably from those in television and are generally less obvious to the end customers. In addition, the boundaries between editorial and production staff in radio are often less distinct and becoming increasingly vague. For the introduction of loudness levelling in a radio workflow, it is therefore necessary to involve a large number of areas and stakeholders, which makes such a project a communicative challenge. A further complication is that a complex, psychoacoustically motivated subject has to be explained to many decision-makers with little or no audio knowledge. It is therefore not particularly surprising that the discussion about loudness levelling in radio braodcasting is often characterized by preconceptions and misunderstandings. A successful implementation project at Bayerischer Rundfunk shows that a considerable part of these belongs in the world of myths and legends. By a joint effort of technical planning, editorial departments, system technology and production, BR was one of the first major radio providers in the world to switch all of its corporate stations to loudness levelling according to EBU R 128. The following paper details the objectives and motives pursued by the project, explains the implementation concept and discusses the experience gained and the results achieved.

2. Underlying conditions in the ARD

ARD's internal debate on loudness levelling in radio broadcasting started at the beginning of 2012 with a first ad-hoc working group on this topic. This was followed by a second working group under the participation of the editorial departments in 2013. Assignment of the respective groups was the investigation of chances and risks of loudness levelling as well as its feasibility in a radio broadcasting environment. The ratings of both groups were positive, but the targeted benefits were mainly considered in other areas compared to television. As a result of the positive recommendations, the Radio Commission and the Production and Technology Commission of ARD passed resolutions at the beginning of 2014 on the introduction of loudness levelling in ARD radio broadcasting. A mandatory switchover date was not agreed due to the different encumbrances of the state broadcasting corporations.

3. Objectives

The BR project "Introduction of loudness measurement according to EBU R 128 in radio broadcasting" is based on the objectives and opportunities identified by the ARD working groups. The primary goal of loudness control in radio broadcasting is to support self-operating moderators and editorial producers. The idea behind it: If already loudness normalized source material is played out to the radio studio or is available in the basic audio editors, the journalists can focus on their core task, the content-related design of the program, and are mostly relieved of the more technical issue of correcting levelling errors. This facilitates the work for moderators and producers, especially in demanding live situations.

Moreover, unwanted loudness jumps between individual pro-

ISBN
978-3-9812830-9-9
The authors hold the copyrights of all of content in this manuscript. They have granted the VDT the right to publish it in print, online and as PDF download on https://tonmeistertagung.com

grams and program elements (intra-program loudness) should be reduced by loudness levelling. These can occur in radio broadcasting for various reasons. Examples for this are on-location sounds recorded by correspondents or contributions produced by editorial staff, which often have to be created under great time pressure and with a strong focus on content. However, even when elaborate productions, e.g. from the fields of cultural spoken word and serious music, are presented in sequence, there can be strong fluctuations in the average loudness. Although such productions are usually correctly levelled by experts, in the past there was no suitable electroacoustic measure to define a common operating point for the subjective loudness perception. The resulting average loudness was therefore different from production to production and strongly dependent on factors such as the selected listening level, the listening conditions, signal dynamics and not least the listening preferences of the producer. Unwanted loudness jumps in radio are often addressed by means of strong dynamic processing of the broadcast signal sum. This approach is merely a symptom treatment and is not technically suitable for eliminating the underlying leveling deficiencies. It is therefore usually accompanied by qualitative compromises in sound processing and is not applicable to programs with high signal dynamics.

Loudness control is also an important prerequisite for intensifying bi- and tri-media cooperation. A further goal is therefore the creation of a common in-house standard for audio levelling, which facilitates the exchange of material between radio, television and the online domain and minimizes the effort and costs for manual adaptation work. In the medium term, the BR project also aims to simplify the cross-ARD radio program exchange as well as the EBU-wide exchange of radio programs via Euroradio by means of a harmonized international levelling standard.

Explicitly no objective of the project was the distribution-side adaptation of the loudness of particular stations (inter-program loudness), which was the main motive for the introduction of loudness levelling in television. It would have the effect that less disturbing loudness jumps would occur when switching between stations. In order to achieve this, however, the average loudness of some stations had to be lowered. In difference to television, which is exclusively digitally distributed in Germany today, a reduction of the baseband power in analog FM transmission due to the decreasing signal-to-noise ratio can, however, lead to considerable losses of coverage. In light of the impending loss of listeners and advertising customers in the highly competitive radio market, a reduction of the average loudness is hardly feasible. Furthermore, as a result of the regulatory constraints in the FM domain, there are only comparatively small loudness differences between individual radio stations. An IRT study from the year 2012 [1] shows a deviation of the average loudness of the ARD-wide loudest FM channel compared to the softest of only about 5 LU.

4. Project overview

In order to accomplish the aforementioned objectives, an investment project for the implementation of loudness levelling in radio broadcasting was launched at BR at the beginning of 2014. After several delays caused by necessary IT adjustments and organizational changes, the project was successfully completed in December 2017 after nearly four years of operation. As early as in the conception phase, many specialist departments had to be involved in order to be able to examine the entire signal chain of the radio production and distribution with regard to levelling aspects. For this purpose, a core project team consisting of seven interdisciplinary experts was formed, which, if necessary, was supported by sub-working groups on the following subjects:

- spoken word production,
- music production,
- radio studio,
- distribution,
- tools and
- training courses.

As shown in **Fig. 1** up to 35 participants were occasionally involved in the extended project team and the steering group.

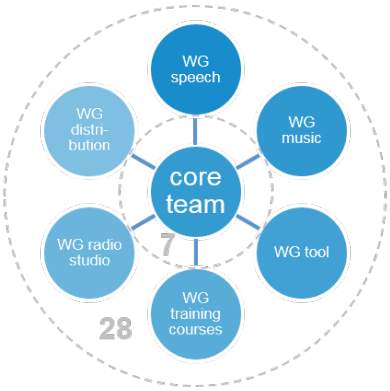


Fig. 1: Overview of the project team organization

In the course of the project, a number of key assignments were identified and processed. The project started with an assessment of the technical workflows from the production stage through to distribution. At the same time, initial market research and the evaluation of suitable technical tools for measuring and adjusting loudness were carried out. After a thorough evaluation of various technical approaches, new workflows were designed and, together with DAVID Systems, a specification for the extension of the Production and Broadcast System was created. After estimating the future maintenance and operating efforts, a schedule for the migration was developed and the technical implementation initiated. Simultaneously about 200 staff members (self-operating moderators, producers, audio-visual media designers and sound engineers) were trained in cooperation with the ARD.ZDF Medienakademie. The project was accompanied by various marketing activities, such as press releases on the BR intranet, to increase awareness and acceptance among the future users.

The required investment for the migration of all stations amounted to approx. 380k EUR, divided into three main cost items:

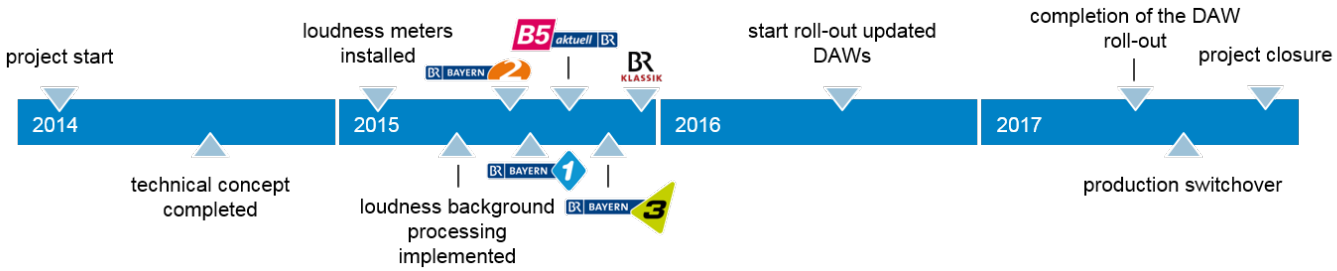


Fig. 2: Project timeline

- Licenses and functional extensions of the Production and Broadcast System,
- loudness meters and
- sound processors with loudness support.

5. Migration Scenario

For the change from peak to loudness levelling, a three-step migration concept was developed. In a first step, the backend was converted, thus the radio Content Management System and the Broadcast System were extended by a background process for loudness analysis and an automatic loudness adjustment by the playout server. The functional extensions were first tested in a reference system and then successfully validated in a pilot test under simulation of the radio show Bayern 2 "radioWelt am Morgen".

After being transferred to the Production Systems, a second migration step was carried out in which the radio studios of the individual stations were successively switched over, meaning that the loudness adjustment during playout was activated. The process was started with Bayern 2, a cultural station with a high percentage of spoken word content. The large differences in the average loudness of individual program segments of this station made the effects of loudness normalization particularly clear. This was followed by the switchover of all other stations over a period of several months in accordance with the sequence shown in **Fig. 2**. Finally, the central news studio was migrated in December 2015. During this phase, the radio production continued to level after quasi-peak.

In the last step, peak levelling was replaced by loudness levelling in the entire radio production at a specific point in time, both in the production areas supporting the radio studio and the elaborate production.

The advantage of the chosen migration strategy was that the late switchover of the production still allowed an unhindered exchange of all self-produced contents between the stations. It was also possible for the central news studio to continue supplying all stations. Loudness adjustment for the migrated channels was initially achieved through automatic processing and selective manual intervention. The qualitative benefits of loudness control could therefore not yet be fully exploited, and some additional effort was required. On the other hand, it was possible to successively switch over the delicate live operation, so that the transition dates could be well prepared, supported by experts and the associated risks were significantly reduced. Moreover, the automatic loudness normalization in the backend achieved good results in a short period of time, which eased the pressure on the technically and operationally more demanding switchover of the production.

6. Technical implementation

BR's DAVID DigAS Content Management System (CMS) and BCS Broadcast System have been modified to trigger a DPE-controlled background process when importing audio files, which analyzes them and returns all essential loudness metadata to the system databases. This principle is shown in **Fig. 3**. In addition to a numerical representation in the Database Manager (DBM) and the Broadcast Scheduling System DigAIRange, colored ear symbols indicate the analysis results as shown in **Fig. 4**. A green ear means that the Integrated Loudness already meets the target value, while an orange ear indicates that the loudness of a file has not yet been adjusted, but can be normalized by a simple gain in- or decrease. A red ear, however, signals that there is not enough headroom available to adapt the loudness without limiting the signal peaks.

Fig. 4: Colored ear symbols for indication of the loudness analysis results in DigAS

At BR it was determined that a manual adaptation has to be made in this case. This means that the actual audio essences are not altered, but the content is simply enriched with additional metadata. The DAVID Turboplayer interprets the metadata during playback and determines on the basis of the Integrated Loudness measure whether a gain increase or decrease is required in order to meet the preconfigured target value of -23 LUFS. Automatic normalization can be deactivated for specific program elements via a checkbox. These are then marked with a crossed-out ear symbol (see

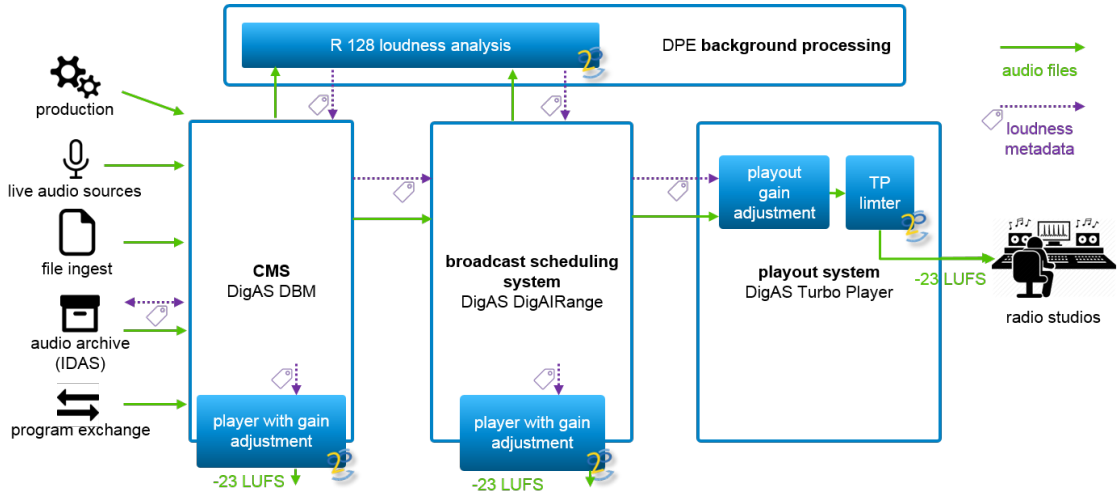


Fig. 3: Schematic representation of the radio backend (CMS and Broadcast System)

Fig. 4) and played without any gain changes. This is useful, for example, for short layout elements such as jingles or music beds, since the EBU target value does not have to be applied to each individual audio element, but merely describes an average loudness value for an entire program. A proper use of soft and loud program elements as a stylistic feature is still highly desired and the basis of a natural sounding program dynamic. The Turboplayer was additionally equipped with a True Peak-capable limiter. The pre-listening players of the CMS and Broadcast System have been extended by a metadata-based loudness adaptation, too, so that it is possible to pre-listen at all workstations in a loudness normalized way. Even though loudness adjustment without limiting ensures a quality-neutral and reversible signal modification, a metadata-based loudness adaptation can be advantageous. For storage and performance reasons, MPEG1 Layer II encoded audio material is still widely used in Production and Broadcast Systems. In order to adapt the loudness in this case, typically a complete decoding and recoding cycle is needed, which may affect the audio quality. Audio files in the archive as well as in the production and broadcast storages therefore remain unchanged and retain their original loudness. The loudness metadata of already analyzed files, however, are transferred to BR's audio archive "IDAS" and ARD's radio database (HFDB) and can be retrieved into the Production or Broadcast Systems when the associated audio file is restored. Multiple analyses of the same files are therefore not necessary thanks to consistent metadata management.

In the radio and production studios, various RTW loudness meters were installed, depending on the intended application. For self-operating moderators and editorial producers, mainly the TM3 was provided, while TM7 and TM9 were installed in pre-production and production studios. **Fig. 5** shows a TM3 meter in the self-operating radio studio of Bayern 3. Standardized presets were developed for all meters to enable consistent workflows at all workplaces. Furthermore, the voice processing had to be adapted so that a target loudness of -23 LUFS was achieved. A real-time capable loudness processor is usually not required at this point, so that existing devices can be continued to be used. The gain-staging of external feeds and from local sources was also considered.

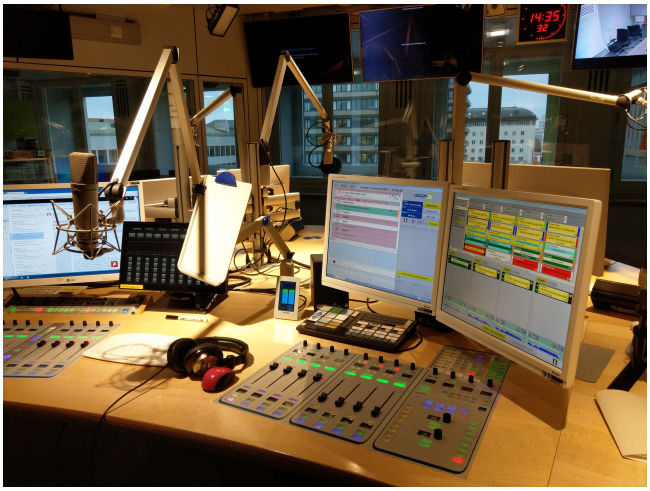


Fig. 5: TM3 loudness meter in a self-operating radio studio of Bayern 3

For example, a static gain reduction had to be applied to the feed coming from the central news studio as long as this area was still levelled after the quasi-peak standard. For sources with largely unknown loudness, such as ENG feeds, telephone hybrids, CD players and the microphones of studio guests, the integration of a real-time loudness processor into the broadcast console can be a useful enhancement. This has not yet been implemented with BR due to technical limitations. A schematic representation of the signal flow in the radio studio can be seen in **Fig. 6**.

Since the average loudness of the broadcast sum signal according to the new in-house levelling target of -23 LUFS is usually some LU below the previously achieved average loudness, the sum signal had to be statically increased in gain by 3-5dB prior to sound processing, for the signal to be transmitted with the same loudness as before. The setting was made individually for each station. Likewise, the sound processing of all stations was optimized in such way that a loudness normalized signal is not affected again by this processing step, e.g. by an AGC that does not comply with loudness. Most manufacturers of sound processors offer models that are compatible with loudness control. For this purpose, some of BR's older Jünger devices have been replaced by new

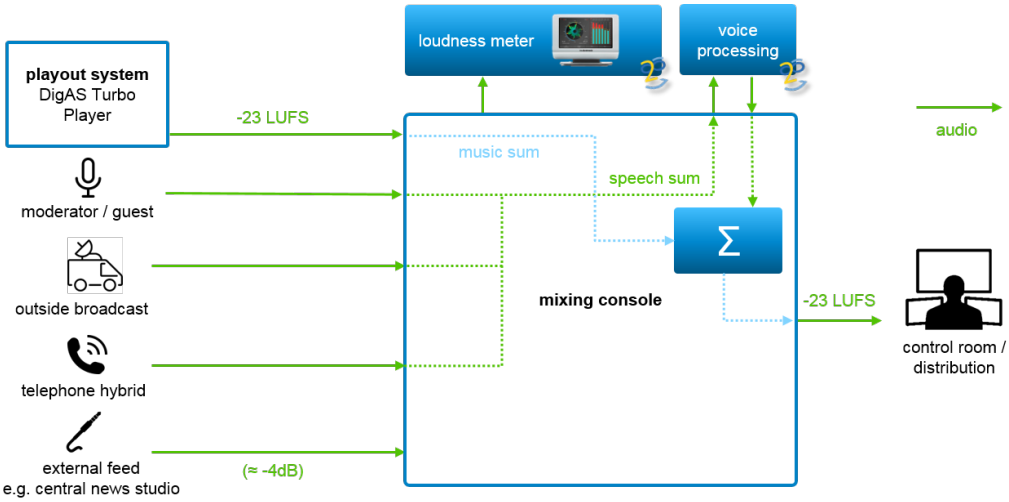


Fig. 6: Schematic representation of the signal flow in the radio studio

loudness compliant DAP*4 units. A schematic representation of the signal flow in distribution is shown in **Fig. 8**.

For the technical implementation of loudness levelling in production, software meters and tools for file-based loudness adaptation had to be integrated into the audio editors. Within the scope of the project various loudness analysis functions and file-based loudness normalization for the Takelist of MultiTrackEditor 6 were specified and tested together with DAVID. Prior to the switchover date, all DigAS editors had to be updated to the current version. Music production at Bayerischer Rundfunk mainly relies on the MAGIX Sequoia and AVID ProTools DAWs. While Sequoia already comes with an extensive feature set for metering and loudness adjustment by default, ProTools was complemented by the plug-in ”LUFS Meter” from Klangfreund. Although the BR production is generally aiming at a target level of -23 LUFS, an exception has been defined for the studio recording of serious music, since the recorded signal dynamics in this field are often so high that dynamic reduction has to be applied in order to achieve this target value. While this is usually an unproblematic approach for live broadcasting, it is usually undesirable for a secondary use on audio carriers, hi-res streaming and for long-term archiving. This is why the levelling for serious music is still made on the basis of peak levels; the loudness adjustment is performed automatically in post-processing by the Turboplayer. If dynamic reduction is required before loudness normalization, this is handled manually in the production areas supporting the radio studio. The exact levelling distinction for serious music can be derived from **Fig. 7**.

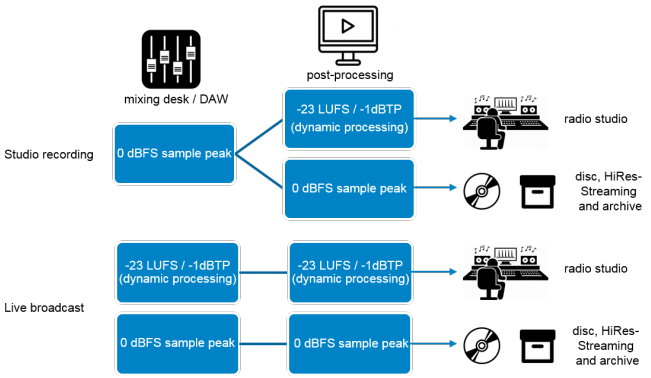


Fig. 7: Levelling procedure for serious music

7. Critical aspects

In the course of the project implementation, various issues were found to be particularly critical to the project success and should be taken into account when planning comparable schemes. The early involvement of all concerned areas and stakeholders, especially the editorial departments, has proven to be highly beneficial. This entailed high communication efforts, which could only be achieved by a professional project management, but was justified by the interest and enthusiasm of the later users. Furthermore, committed supporters on management level showed themselves to be indispensable for the overall success of the project, as a large number of decision-makers had to be engaged and convinced. Furthermore, the decision to include all of BR’s radio stations in the concept right from the beginning and to switch them over to loudness control soon after one another was advantageous. Despite the higher initial complexity and resistance, significant interfacing problems and costs could be avoided this way, which would have compromised many of the benefits and objectives of the project. It should also be taken into account that for stations with regional windows, the technical infrastructure at remote locations (state broadcasting houses, regional studios) has to be migrated in parallel, which involves particular challenges. Last but not least, unexpected technical problems and evolved workflows that interfere with the introduction of loudness levelling should be considered in time planning. In the BR project, for example, it was noticed late that production computers in various correspondent offices and other remote locations had not yet been updated to Windows 7. However, this was an important prerequisite for the introduction of MultiTrackEditor 6. The need to replace these IT components significantly delayed the switchover of production.

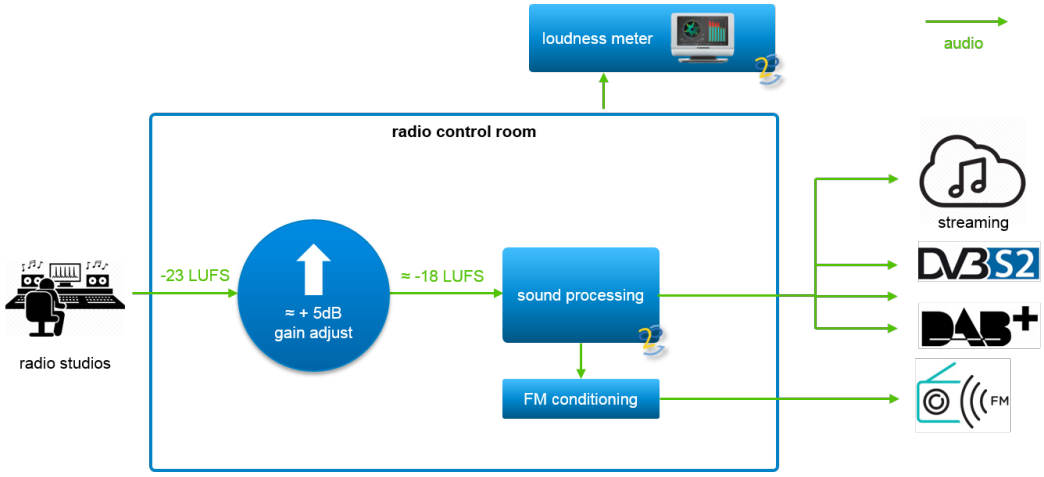


Fig. 8: Schematic representation of the signal flow in control room and distribution

8. Lessons Learned

Various factors, such as the involvement of many decision-makers, the cultural change associated with the project, the need for customer-specific software development, the re-design of evolved workflows, the introduction of new technology and processes across multiple locations, and internal reorganization had a considerable impact on the time needed to complete the project. For similar ventures, these aspects should be considered in advance. Moreover, many broadcast technicians initially opposed the project because they felt their area of responsibility was in danger. Such concerns need to be constructively addressed at an early stage and common solutions have to be developed. Another finding of the project was that, as expected, the concept of file-based loudness normalization reaches its limits for pre-produced contents. Loudness inconsistencies in pre-produced programs cannot be corrected this way. For stations with a particular orientation, e.g. news stations, an automated analysis and adaptation of the loudness range measure should be considered to address this issue. Moreover, it was found that real-time loudness processing of external feeds and specific sources in the radio studio can further relieve the self-operating moderators and improve the overall levelling quality. It should be emphasized in particular that the introduction of loudness levelling in radio broadcasting must also involve the online distribution. Again, the distribution of content with production loudness level might be too soft as it often cannot be adequately reproduced on mobile devices. Consequently, a technical possibility should be provided for increasing the volume of podcasts and radio live streams in analogy to the FM distribution path.

9. Outcomes

Positive feedback from self-operating moderators confirmed the expected relief effects of loudness levelling. In the field of editorially produced contents, the new tools support both the quality of production as well as the work efficiency. But also in the professional production and pre-production, the acceptance of loudness levelling has meanwhile reached a high level and the loudness-associated functionalities of the audio editors have enabled new, more time-efficient workflows. In the course of the introduction process, the entire radio signal chain from production through to distribution had to be analyzed. This allowed to identify and consolidate a considerable number of provisional technical facilities and obsolete workflows. Moreover, BR's technical subscriber service confirms a trend towards less loudness-related complaints in radio broadcasting.

On the other hand, there is a particular need for improvement in the area of online distribution, as podcasts are still published with too low and sometimes inconsistent loudness. Long-term optimization opportunities are the increasing use of real-time loudness processing for external feeds in the radio studio as well as the more consistent application of the loudness range measure to identify and optimize poorly produced contents.

10. Conclusion

As part of a successful project at Bayerischer Rundfunk, all corporate radio stations were switched over to loudness levelling according to EBU R 128. In a project period of almost 4 years, a three-step migration scenario was applied to first implement a metadata-based, automated loudness normalization in the radio Content Management and Broadcast Systems. Subsequently, the radio studios of all stations were successively migrated to loudness levelling. Finally, on a specific date, loudness levelling replaced the traditional peak levelling in the entire radio production. The primary objective of supporting self-operating moderators and editorial producers was achieved, and so was the reduction of unwanted loudness jumps between individual shows and program elements. The introduction of loudness levelling in radio broadcasting also provided an important basis for the intensification of bi- and tri-media cooperation. However, the optimization of loudness levelling for the online distribution paths requires further consideration.

11. Acknowledgement

A special thank-you goes to Monika Hartberger, Maria Scherr, Frank Lott and the project team of Bayerischer Rundfunk, who enabled the realization of the project in the first place and provided advice to the author during the preparation of

this manuscript.

12. References

[1] IRT: Ergänzungsbericht zur Lautheit gemäß EBU R 128 von Hörfunkwellen über den UKW-Verbreitungsweg. München, 2012

[2] BEAM international specifications - regional variations from the generic specification URL: https://www.beam.tv/content/user-guides/beam_international_specifications.pdf

Transparenz bei Orchesterproduktionen

Neue Erkenntnisse zur Laufzeitkompensation von Stützmikrofonen

Michael Sandner[1]

[1] *Erich-Thienhaus-Institut der Hochschule für Musik Detmold / SWR Stuttgart*

Abstract

Das Thema Laufzeitkompensation der Stützmikrofone zeigt sich heute, fast 35 Jahre nach den ersten Versuchen von G. Theile, in einem anderen Licht. Wir können durch verbesserte technische Möglichkeiten viel genauere Verzögerungswerte einstellen und diese auch im laufenden Programm verändern. So werden geringere Unterschiede hörbar und die Auswirkungen können gezielter untersucht werden. Dies führte zu einer differenzierteren Betrachtungsweise, in der den einzelnen Situationen unterschiedliche Schwerpunkte zugeordnet werden konnten.

Meine Versuche haben ergeben, dass sich die Verzögerung der Stützmikrofone je nach Art der Signale anders auswirken und auch unterschiedlich konfiguriert werden kann. Eine primäre, in allen Situationen gleiche Wirkung der Laufzeitkorrektur gibt es nicht (mehr), vielmehr müssen wir sie als vielseitiges Werkzeug begreifen, das mal besser, mal weniger geeignet ist, die Transparenz der Mischung zu erhöhen.

Abb. 1: Typische Mikrofonierung bei Kammermusik: Hauptmikrofon als Decca-Tree, Stützen u. Raummikrofone (nicht sichtbar)

Einleitung

Genau dreißig Jahre ist es her, dass ein Team um Günther Theile (IRT) und Martin Wöhr (BR) ihr Konzept zur „Raumbezogene Stütztechnik - eine Möglichkeit zur Optimierung der Aufnahmequalität“ [1] auf der Tonmeistertagung vorgestellt haben. Ausgehend von den akustischen Laufzeiten zwischen den Mikrofonen stellten sie das Konzept der Verzögerung der Stützsignale dar, das seitdem bei unzähligen Aufnahmen angewandt wird. Ich gehe davon aus, dass dieses hinreichend bekannt ist, habe aber einen Aspekt davon in meine Untersuchung einbezogen.

Bis heute gibt es dennoch Diskussionen um das Thema, und nicht alle Kollegen sind von dem Konzept der verzögerten Stützen überzeugt. Hat sich die Aufnahmetechnik so verändert, oder sind gar die klangästhetischen Maßstäbe verschoben? Vielleicht beides, bewegen wir uns doch so gut wie immer im digitalen Bereich mit hochauflösenden Signalen und verlustfreier Speicherung und streben nach mehr Klarheit und Durchhörbarkeit. Die Werkzeuge sind zudem viel präziser und verfügbarer geworden - Verzögerung, aber auch künstliche Raumeffekte stehen in großer Anzahl auf Knopfdruck zur Verfügung.

Für mich war das ein Grund, das Thema neu aufzugreifen und differenzierter zu betrachten. Ich wollte den hörbaren Auswirkungen einer Laufzeitkompensation auf den Grund gehen und herausfinden, ob so die Transparenz von Orchestermischungen mit vielen Stützmikrofonen erhöht werden kann. Außerdem wollte ich wissen, bei welchen Signalen ein Gewinn zu erwarten ist, und wie sich dieser äußert. Deshalb habe ich im Sommersemester 2017 an der Hochschule für Musik Detmold ein Forschungssemester gemacht.

Anstoss und Voraussetzung

Das Ringen nach möglichst transparenten Orchesteraufnahmen beschäftigt mich schon lange, vor allem bei Stereoaufnahmen ist ein klares und durchhörbares Klangbild ab einer gewissen Klangfülle sehr schwer zu bewerkstelligen. Dies liegt zum Einen daran, dass fast alle Signale über Phantomschallquellen dargestellt werden müssen, die über nur 2 Lautsprecher gebildet werden, zum Anderen addieren sich meist Signale, die nicht frei von Übersprechen sind, sodass es zu gegenseitigen Beeinflussungen kommt.

Die Aufnahmesituation bei Orchestern ist zum Glück meist an entsprechend geeignete Räume gebunden. Der Grund dafür ist in erster Linie die Verknüpfung mit dem Konzertbetrieb, aber auch die Hörsamkeit innerhalb des Ensembles und die Anforderungen an eine gute Aufnahme-Akustik. Dies erlaubt uns i.d. Regel die Aufstellung eines Hauptmi-krofons nahe oder am Hallradius, damit dort eine angenehme Raumbalance erfolgt. Allerdings ist mit nur einem Haupt-

ISBN
978-3-9812830-9-9

Der Autor hält die Urheberrechte für den gesamten Inhalt seines Manuskripts. Er überlässt dem VDT das Recht auf Veröffentlichung in gedruckter Form, online oder als PDF-Download von https://tonmeistertagung.com

mikrofon meist keine perfekt partiturgerechte Instrumentalbalance möglich, sodass die jeweiligen Instrumentalgruppen „gestützt“ werden müssen.

Die akustischen Gegebenheiten beim Einsatz von Stützmikrofonen sind aber etwas komplexer. Wir nutzen aus klanglichen Gründen einen gewissen Abstand zu den Instrumenten, der uns aber ein deutliches Übersprechen beschert, leider verstärkt durch die Tatsache, dass die lauteren Instrumente hinten im Orchester sitzen. Vor allem Blech und Schlagwerk schlagen dann gerne bei den leiseren Instrumenten durch. Bei einer akribisch angewandten Verzögerung muss man dieses berücksichtigen, und da kam die Anregung meines Kollegen Prof. Oliver Curdt (HdM Stuttgart) wie gerufen: Er schlug vor, die Signale der Pauke auch als Grundlage für die Laufzeitkompensation der davor sitzenden Holzbläser zu verwenden.

Gängige Aufnahmepraxis

Bei Orchesterproduktionen werden Abstände vom Stützmikrofon zum Instrument bevorzugt, die etwa bei ein bis zwei Metern liegen, wie gesagt, hauptsächlich aus klanglichen Gründen. Größere Abstände sind v.a. bei chorischen Stimmen (wie den Streichern) gewünscht, um einen flächigen Klang zu erreichen und keine Einzelstimmen zu fokussieren. Die Kanaltrennung erfolgt durch Position, Abstand und über die Wahl der Mikrofonrichtcharakteristik, neben Nieren sind hier auch gelegentlich Supernieren und Acht-Charakteristiken üblich, bei chorischen Stimmen aber auch breite Nieren.

Bei einem klassisch-romantischen Orchester kommen oft über 40 Stützen zusammen, dazu noch zwei oder vier Raummikrofone. Die Ausdehnung kann durchaus 25 Meter in der Breite und 15 in der Tiefe betragen (Bühne der Liederhalle Stuttgart), je nach Besetzung und Bühnenraum auch mehr. Das führt zu akustische Laufzeiten von bis zu 35 ms, die durchaus schon Echowirkung hervorrufen können. Bei akustisch aktiveren Bühnen kommt ein deutlicher Raumanteil auch über die Stützen zustande, v.a. wenn die Instrumente oder Sänger den Raum stark anregen, und die Stützmikrofone einen etwas größeren Abstand haben.

Heutige Mischtechniken

Der Hauptbestandteil des Klangs kommt eigentlich immer von einem Hauptmikrofon (z.B. Decca-Tree), dieses ist auch in der Zeitstruktur der Signale die Referenz. Dazu werden dann Stützen für alle Instrumentengruppen hinzugemischt (typischerweise mit einem Pegel von ca. -6 dB oder mehr), um die Instrumentalbalance beeinflussen zu können. In der Position werden diese meist mit dem Panoramapotentiometer und künstlichem Raum in der Horizontalen und Tiefe ins Klangbild eingepasst. Bei guten Räumen runden Raummikrofone (wegen paralleler Surround-Mischung oft 2 x 2 Mikrofone) die Integrität ab.

Die unverzögerte Addition der Signale kann zu unerwünschten Nebenwirkungen führen. Zum Einen ist das der hinlänglich bekannte Präzedenzeffekt, auch Gesetz der ersten Wellenfront, der die Ortung der Signale nach dem früheren Signal bestimmt. Dann erzeugen ähnliche Signale mit unterschiedlichen Laufzeiten Kammfiltereffekte, die v.a. bei Vokalstimmen als Verfärbung zu hören sind. Als weiteres Phänomen sind verschleierte Einschwingvorgänge zu beobachten. Dies äußert sich natürlich bevorzugt bei Impulshaltigen Instrumenten wie der Pauke oder Schlagzeug. Durch die unterschiedlichen Laufzeiten und Übersprechen wird hier ein Cluster von vielen Impulsen erzeugt, das ein unpräzises diffuses Hörerlebnis zur Folge hat.

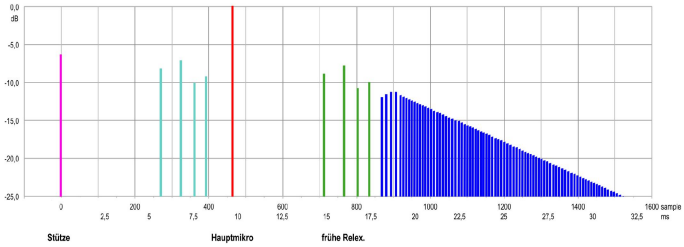


Abb. 2: Reflexionsstruktur bei unverzögerter Stütze

Als Werkzeuge stehen uns neben Pegel- und Panoramaregelung auch individuell einstellbare Verzögerungen, Filter zur Frequenzgangkorrektur; künstliche Halleffekte und die Dynamikbeeinflussung durch Regelverstärker (v.a. Limiter) zur Verfügung. Bei den üblichen Digitalmischpulten sind dazu noch nicht mal externe Geräte nötig, da alle Bearbeitungsschritte latenzkompensiert in dem digitalen Signalprozessor gerechnet werden. Die Verzögerung dient dabei der Integrität des Klangbilds, man möchte zu vordergründige Stützsignale vermeiden. Aus Zeitgründen werden im Rundfunkbetrieb allerdings meist nur Näherungswerte bzw. die Werte einer vergangenen Messung eingesetzt, die Auswirkungen solcher Abweichungen sind aber nicht zu vernachlässigen, wie meine Versuche zeigten.

Verzögerte Stützen

Voraussetzungen

Die unverzögerten Stützen eilen dem Signal aus dem Hauptmikrofon voraus. Um sie in die Raumstruktur einzubetten, sollte die exakte Laufzeit des Signals bekannt sein, diese ist nicht gleichzusetzen mit der Entfernung der Mikrofone voneinander, sondern deren Differenz.

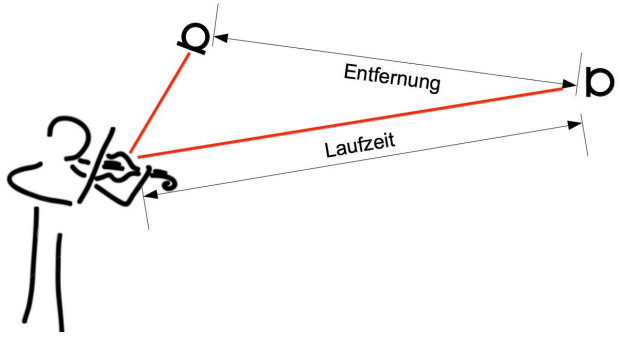


Abb. 3: Laufzeitenskizze

Diese Laufzeiten sollten möglichst genau ausgemessen werden, dazu verwendet man am Besten einen Impuls (Händeklatschen, Knackfrosch u.ä.), den man in der DAW anhand der Wellenform gut erkennen kann. Natürlich ist ein samplegenaues Erfassen optimal, aber eine Genauigkeit von etwa

1ms ist meist ausreichend, zumal in den meisten Mischpulten ms-Werte einfach und eindeutig einzustellen sind.

Bei der Laufzeitkompensation müssen natürlich auch mögliche Fehlerquellen benannt werden:

- Der Ort der Impulsgenerierung ist nicht identisch mit dem der Schallerzeugung.
- Die Zuordnung zu den Hauptmikrofonen ist nicht eindeutig (siehe unten).
- Das Instrument (der Musiker) bewegt sich.
- Unterschiedliche Tonhöhen werden unterschiedlich abgestrahlt (Richtwirkung Instrument), und treffen so mit frequenzspezifischen Veränderungen im Hauptmikrofon ein.

Diese Phänomene können nicht ausreichend kompensiert werden, sodass eine gewisse Toleranz und Unschärfe hingenommen werden muss. Die Frage stellt sich nur, in welche Richtung eine Toleranz angelegt werden sollte: ist ein Mittelwert anzustreben, oder sollte sicherheitshalber eine kleine Zugabe geschehen? In der schon öfter erwähnten Arbeit von Wöhr, Theile u.a.1 wird diese Frage dadurch umgangen, dass generell eine Zugabe von ca. 20ms geschieht, sodass das Stützsignal als künstliche Raumreflexion angesehen wird.

Die Impulsgenerierung sollte möglichst am Ort der Schallabstrahlung geschehen, bei stationären Instrumenten wie Flügel oder Pauke kann man auch einen Ton anspielen (wenn man darf), nur so wird die reale Zeitdifferenz erfasst,. Eine aus Zeitgründen gelegentlich anzutreffende Methode, direkt beim Hauptmikrofon zu klatschen, um so alle Stützmikrofone gleichzeitig zu erfassen, ist nicht ausreichend, denn hier wird nur die Entfernung der Stützen gemessen, nicht aber die reale Zeitdifferenz.

Im realen Spiel werden sich die Musiker je nach Instrument und Temperament bewegen - und mit ihnen das Instrument und der Ort der Schallabstrahlung. Um die Größenordnung der Bewegung abzuschätzen, möchte ich als Beispiel folgende Rechnung für eine sich bewegende Geige machen:

angenommene Bewegung:	± 10cm
max. Laufzeitdifferenz:	0,6 ms
max. Sample bei 48 KHz:	28
mögliche Kammfilterfrequenzen:	860, 2580, 4300, 6020 Hz etc.

T**abelle 1:** Auswirkung der Bewegung der Musiker

AutoAlign Plugins

Der Gedanke, dass diese Aufgabe auch durch Plugins erledigt werden könnte, ist verlockend. Ich habe mir deshalb vier vielversprechende Kandidaten angeschaut, zwei davon können allerdings nur eine manuelle Korrektur vornehmen.

AutoAlign von SoundRadix

Das Plugin von SoundRadix hat sich auf die Fahnen geschrieben, den zeitaufwändigen manuellen Prozess des "Time alignment" zu automatisieren. Dazu wird es einmal in die Referenzspur wie das Hauptmikrofon eingeschleift und einem send-bus zugeordnet (es stehen max. 9 sends zur Verfügung) und ein zweites Mal in den Zielkanal eingesetzt. Dort kann dann mittels die Schaltfläche "detect" das Signal analysiert werden, unerwünschtes Übersprechen wird mit den "noise-floor" Reglern unterdrückt.

AutoAlign hat zwei Anzeige-Modi (siehe Anhang), der erste zeigt die Phasenbeziehung zwischen den beiden Signalen an, wobei die Mitte das Ideal wäre. Währen des detektierens oder nach Betätigung des "display"-Buttons wird der zweite Modus eingeschaltet, bei dem die Übereinstimmung der Signal über die Zeit dargestellt wird. AutoAlign sucht sich die wahrscheinlichste Spitze aus, reagiert aber oft unberechenbar auf den Gehalt des analysierten Ausschnitts. Über die Pfeiltasten können aber auch benachbarte Werte eingestellt werden, so dass in geringem Maße eine Kor- rektur möglich ist. Die Zeitanzeige kann von Sample auf ms, cm oder inch umgestellt werden.

MautoAlign von MeldaProduction

Das zweite Plugin mit einer Analysemöglichkeit für die automatische Anpassung kommt von MeldaProduction. MautoAlign hat zwar nicht so eine spektakuläre Optik, wie AutoAlign, dafür aber mehr Einstellmöglichkeiten und eine etwas technischere Herangehensweise. Das Plugin wird in alle zu bearbeitenden Kanäle eingeschleift, wobei für die Analyse Gruppen gebildet werden können (es stehen 36 Gruppen zur Verfügung). Bei der Analyse kann sowohl die "capture"-Dauer, als auch ein Tief- und Hochpass und der "noise-floor" eingestellt werden, bei eingeschalteter spektraler Phasenanalyse zusätzlich die Auflösung und ein weiterer Schwellwert.

Die Analyse setzt eine etwa 2 bis 4 Sekunden Passage voraus, in der möglichst nur das eine Instrument/Stimme zu hören ist. Dennoch ergeben sich oft irritierend andere Werte, v.a. weil MautoAlign mitunter auch die Referenzspur verändert. Es ist durchaus sinnvoll hier (wie auch bei AutoAlign) schon vorab den Zielwert im groben abzuschätzen, um bei Ausreißern gleich eine zweite Analyse anzuschließen. Nützlich ist bei MautoAlign auch die mögliche Vorgabe einer maximalen Verzögerung (diese kann bis zu 1000 ms betragen).

InPhase von Waves

InPhase ist ein manuell einzustellendes Plugin, das die Phasenbeziehung zwischen zwei Kanälen oder eines Stereosignals korrigieren kann. Waves benutzt ein Allpass-Filter, um die Phasenverschiebung über den Frequenzgang anzupassen. Zusätzlich hat es ein time-delay: bei einer Latenz von 20 ms (Live-Komponenten haben keine Latenz) kann ±20 ms Verschiebung realisiert werden. Als Hilfsmittel dient eine hochauflösende Wellenformanzeige, in die 2 Sekunden Audio zur Analyse ge"captured" werden können, und ein Korrelationsgradmesser, der die Maximal- und Minimalwerte über farbige Reiter festhält.

Die Bedienung ist etwas gewöhnungsbedürftig, da in beiden Zweigen (Kanälen) die gleichen Einstellungen vorgenommen werden können, diese aber gegensätzliche Auswirkung haben. Dennoch erzielt InPhase bei sich ähnelnden Signalen

recht genaue und dank der genauen Wellenformdarstellung nachvollziehbare Ergebnisse.

Voxengo PHA-979

Das Plugin ermöglicht die manuelle Justage von Delay und Phase, jeweils für den linken und rechten Kanal getrennt. Als Unterstützung gibt es dazu einen Multiband-Korrelationsgradmesser, der im Idealfall über alle Frequenzen im grünen (positiven) Bereich sein sollte. Eine schöne Idee, kann man doch so die Phasenbeziehungen leichter bewerten und sich auf den primären Frequenzbereich beschränken. Im Ausgangsmodul ist auch eine M/S Matrix, Monoschaltung und Seitentausch möglich, ein A-B Vergleich zwischen zwei Einstellungen ist ebenfalls hilfreich.

Der Voxengo PHA-979 ist ein professionelles Plugin, um einem Audiosignal Phasenverschiebungen zuzufügen … Außerdem kann das Plugin für positive oder negative Zeitverschiebung eingesetzt werden. Dadurch ist es nicht mehr notwendig Audiomaterial im Sequenzer manuell zeitlich anzugleichen. Der PHA-979 verfügt ebenfalls über einen Regler zur Einstellung der Mitten-/Seiten-Balance sowie für das Panning …[1)]

Zuverlässigkeit der automatischen Zeitmessung

Bei der Zeiterfassung der Laufzeiten meiner Beispiele benutzte ich auch drei der Plugins. Im Großen und Ganzen stimmen die gemessenen Werte mit den automatisch generierten Werten von AutoAlign und MautoAlign überein, Abweichungen gab es aber bei Verzögerungswerten über 21 ms, offensichtlich rechnen die Hersteller nicht mit hohen Laufzeiten (im Pop Bereich sind diese auch unüblich). Die offensichtlichen Abweichungen der beiden "Auto"-Plugins wurden mit mindestens drei Messungen bestätigt und dann noch manuell überprüft, dabei half das PlugIn „InPhase“ von Waves. Hier zeigte sich, dass die manuelle Methode bei Weitem am zuverlässigsten war. Eine visuelle Erkennung der Hüllkurve ist eindeutig im Vorteil, und es wird nur eine Frage der Zeit sein, bis die automatischen Algorithmen dieses auch berücksichtigen.

Als Fazit des Vergleichs dieser Plugins kann man leider nur eine Warnung abgeben: Es gibt keine zuverlässige Laufzeiterkennung auf Knopfdruck! Ohne Plausibilitätscheck und manuelle Überprüfung ist deren Einsatz nicht sinnvoll. Auch sollte man sich der Obergrenze von ca. 20 ms bewusst sein.

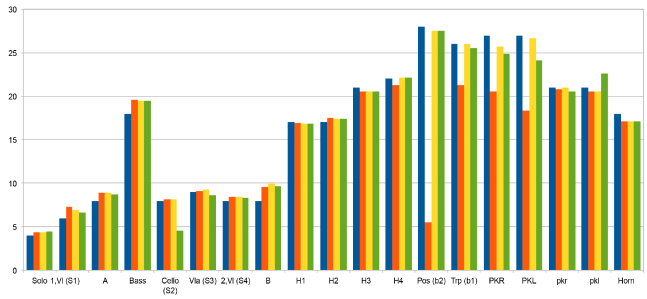


Abb. 4: Pluginvergleich bei den Laufzeit-Messungen einer Orchesterproduktion (Manuell, AutoAlign, MautoAlign, InPhase)

1) Bedienungsanleitung Voxengo PHA-979

Bezugspunkt und Übersprechen

Bei Orchesterproduktionen wird als Hauptmikrofon gerne ein Decca-Tree oder eine AB-Kombination verwendet (dies ist die Standardkonfiguration im SWR). Dann stellt sich die Frage, zu welchem der Mikrofone die Laufzeit ermittel werden sollte, denn je nach Basisbreite und Abstand der Mikrofone gibt es erhebliche Differenzen. Ich habe in meinen Untersuchungen für die Aufnahme des Staatsorchesters Stuttgart alle drei Werte einzeln erfasst und exemplarisch verschiedene Einstellungen ausgewertet (siehe Kapitel 10.6). Die Unterschiede sind bei den Gesangssolisten deutlich wahrzunehmen, haben jedoch dort andere Auswirkungen als bei den entfernten Instrumenten (wie Pauke), bei denen die Laufzeitdifferenzen auch nicht so groß sind.

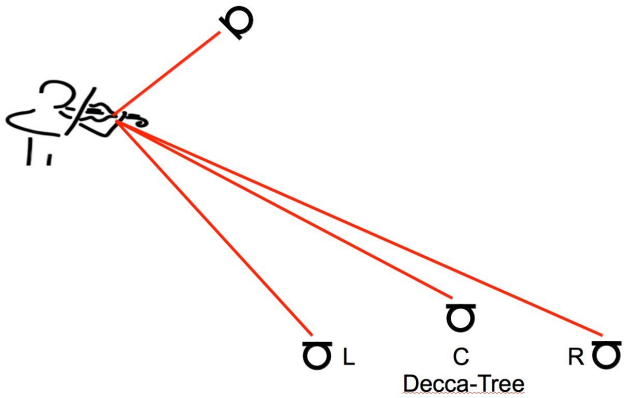


Abb. 5: Unterschiedlicher Bezugspunkt zum Hauptmikrofon

Deckung der Einschwingvorgänge

Pauke und Holzbläser

Die Pauken befinden sich im Orchester meist direkt hinter den Holzbläsern und schlägt auf deren Stützmikrofone kräftig durch. So wird bei korrekt ausgemessenen Laufzeiten ihr Impuls zusätzlich und mit abweichenden Werten in die Mischung eingehen, es entstehen ungewollte Verdoppelungen, wie in vorigen Absatz dargestellt. Schaut man sich die realen Differenzen an, so sind das im typischen Fall der Liederhalle Stuttgart, Pauken direkt hinter den Holzbläsern, nur wenige Sample (2 bis 10) bei den einzelnen Holzstützen. Nur beim Holz-Hauptmikrofon ergeben sich größere Abweichungen (bis zu 100 Sample), da dieses ja deutlich höher hängt und sich somit näher an der Pauke aber weiter weg von den Holzbläsern befindet.

Zu dieser Problematik habe ich einige Beispiele gemacht. Die Unterschiede sind deutlich wahrnehmbar und eine für die Pauke optimierte Lösung wäre, dessen Signal auch auf den Holzbläserstützen richtig zu verzögern. Allerdings verlagern sich dann die Abweichungen in die Holzbläser, ich habe jedoch den Eindruck gewonnen, dass eine Fehlanpassung bei den Holzbläsern aufgrund der andersgearteten Signale weit weniger stört, als bei der Pauke.

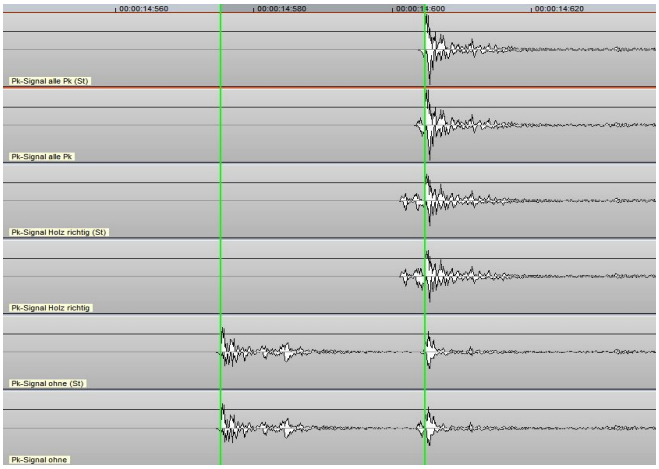


Abb. 6: Impulse des Paukensignals komplett, Pauke und Holz richtig und ohne Laufzeitkompensation (je 2 von oben nach unten)

Pegelerhöhung

Die korrekte Verzögerung der Stützsignale führt zu deckungsgleichen Signalen. Dies hat auch Auswirkungen auf den Pegel, und zwar sowohl auf den Spitzenpegel, als auch auf die Loudness-Werte. Ich habe deshalb bei allen relevanten Beispielen der CD während des Bounce eine Loudness- und True-Peak Analyse gemacht. Die stärkste Erhöhung ließ sich naturgemäß bei den Pauken beobachten, z.B. TP -3.1dB (ohne) zu -0.7dB (mit) bei der Pauke allein und TP -4.7dB (ohne) zu -2.7dB (mit) im Kontext der Mischung der Siegessinfonie. Die Loudness-Werte stiegen dort auch an, und zwar um 0.9 und 2.1 LU (Short-Term), was angesichts der alleinigen Veränderung der Verzögerung der Pauke schon signifikant ist.

Pauke Signal (ohne delay):	**Pauke Signal (delay auf Pk.):**
Loudness: -9.4 LU (-32.4 LUFS)	Loudness: -8.9 LU (-31.9 LUFS)
Max. M.: -5.4 LU (-28.4 LUFS)	Max. M.: -4.5 LU (-27.5 LUFS)
Max. ST: -14.0 LU (-37.0 LUFS)	Max. ST: -13.1 LU (-36.1 LUFS)
LRA: 2.5 LU	LRA: 3.6 LU
Max. True Peak: -3.1 dBTP	Max. True Peak: -0.7 dBTP
L: -6.3 dBTP	L: -2.1 dBTP
R: -3.1 dBTP	R: -0.7 dBTP
Siegessinfonie (ohne delay):	**Siegessinfonie (delay auf Pk.):**
Loudness: 5.0 LU (-18.0 LUFS)	Loudness: 5.8 LU (-17.2 LUFS)
Max. M.: 10.0 LU (-13.0 LUFS)	Max. M.: 12.6 LU (-10.4 LUFS)
Max. ST: 7.6 LU (-15.4 LUFS)	Max. ST 9.7 LU (-13.3 LUFS)
LRA: 9.6 LU	LRA: 11.8 LU
Max. True Peak: -4.7 dBTP	Max. True Peak: -2.7 dBTP
L: -4.7 dBTP	L: -3.7 dBTP
R: -4.8 dBTP	R: -2.7 dBTP

Tabelle 2: Pegelanalyse (Auszug)

Chorische Stimmen

Eine andere Rolle spielen chorische Stimmen, wie die Streicher. In einer großen Orchesterbesetzung spielen da z.B. 14 bis 16 erste Violinen, die wir in der Regel mit zwei Stützmikrofonen aufnehmen. Und auch wenn ich meist versuche, jedes Stützmikrofon in einer Vierergruppe mit gleichem Abstand zu platzieren, sind damit nur die Hälfte der Instrumente annähernd gleich in der Laufzeit abgebildet. Aber das heißt noch lange nicht, das diese auch exakt synchron spielen, und das ist auch gut so, denn so kommt ja erst der chorische Effekt dieser Stimmen zustande. Nur wird es in der Musik daher auch keine deckungsgleichen Einschwingvorgänge geben, die optimal zum Hauptmikrofon ausgeglichen werden können.

Dazu kommt natürlich noch das teils erhebliche Übersprechen der Streicher-Stützen untereinander, z.B. bei der ersten und zweiten Violine in einer amerikanischen Orchesteraufstellung. Dennoch ist die Laufzeitkompensation nicht ganz unnütz, v.a. bei großen Entfernungen vom Hauptmikrofon, wie sie bei einem Chor auftreten können. Hier spielt der Effekt der räumlichen Integrität vielleicht auch noch eine Rolle.

Räumliche Integrität

Bei dem Egmont-Projekt bot sich noch eine etwas andere Untersuchung an. Der beteiligte Sprecher war, anders als sonst bei Konzerten üblich, in großer Entfernung auf der Chorbühne platziert und über das Headset mit einer naturgemäß sehr trockenen Stütze eingebunden. Im Hauptmikrofon war, auch wegen der notwendigen Beschallung, eigentlich kein Direktsignal zu hören, zudem bewegte sich Herr Michalek über die ganze Chorbühne hin und her und vor und zurück. Man könnte also meinen, eine Laufzeitkompensation wäre nicht gewinnbringend, schwankten doch die ermittelten Delay-Werte zwischen 45 und fast 70 ms.

Dennoch stellte ich eine Beispiel-Sequenz her, in der die Verzögerung zwischen 32 und 100 ms in 5 Stufen erhöht wurde. Vergleicht man nun die Eckwerte ohne und mit 100ms Verzögerung, so hört man den Sprecher beide Male mit einer unnatürlichen Räumlichkeit - im ersten Fall mit viel zu später Raumantwort (quasi wie ein Echo), im zweiten Fall mit einer vorauseilenden Hallfahne. Mit einer Kompensation auf einen Mittelwert von 54 ms hört man dann einen angenehmen, wenig aufdringlichen, plausiblen Raum. Diesen Effekt möchte ich die räumliche Integrität nennen, die durch die Laufzeitkompensation ggf. hergestellt werden muss.

Auswirkungen auf das Hören

1 Es ergibt sich Transparenzgewinn durch die Bewahrung des ITDG, weil keine Stützensignale dem Hauptmikrofonsignal vorauseilen. Die räumliche Integrität wird gewahrt.

2 Die Stützen springen einen nicht an (z.B. beim Schlagzeug), dies möchte ich den Präzedenzeffekt der Tiefe nennen.

3 Die Stabilität der Ortung wird nicht zerstört (z.B. für Solostützen nahe am Decca-Tree Hauptmikrofon)

4 Kammfiltereffekte werden minimiert, diese treten vorrangig bei ähnlichen Signalen und Pegeln auf, und sind v.a. bei kleinen Verzögerungszeiten & Vokalstimmen hörbar.

5 Die Basswiedergabe gewinnt durch phasengleiche Addition der Stützsignale.

6 Impulse werden verstärkt, dies zeigt sich v.a. bei der Pauke - dadurch ergibt sich auch ein Lautheitsgewinn.

Die vorigen beiden Punkte führen zu einer Steigerung der Effektivität der Stütze: Der Pegel kann etwas geringer gewählt werden, um den gleichen Präsenzgewinn zu erreichen.

7 Für chorische Gruppen (Streicher, Chor) ist kein signifikanter Gewinn festzustellen, es sei denn, die räumliche Integrität muss justiert werden.

Feedback

Studenten

Das erste Feedback holte ich mir am 20.11.2017 an der HdM (Hochschule der Medien) Stuttgart im Rahmen eines Masterseminars ein. Ich stellte in einem kurzen Vortrag die Arbeit und einige Beispiele aus den Voruntersuchungen vor. Dann wurden die Studenten aufgefordert, Mischungen des Liedes "Freudvoll und Leidvoll" und der Siegessinfonie mit und ohne Delay herzustellen, die Trackliste und die Werte der Laufzeiten habe ich Ihnen als Arbeitsgrundlage mitgegeben. Die Mischung erfolgte dann in zwei Gruppen in unterschiedlichen Studios. Es wurden ausdrücklich auch Experimente mit veränderten Verzögerungswerten angeregt, jedoch entschieden sich beide Gruppen für eine exakte Kompensation.

Die Auswirkung der Laufzeitkompensation wurde durchweg als positiv bewertet, und die Mischungen als klarer und durchhörbarer charakterisiert. In einem anschließenden Blind-Hörtest mit beiden Gruppen konnten die Unterschiede eindeutig herausgehört werden, also auch, wenn es nicht die eigene Mischung war. Die Laufzeitkompensation wurde immer gleich erkannt, abgehört wurde dabei auf den gewohnten Regielautsprechern des Studios.

Tonmeister

Im SWR Stuttgart machte ich Anfang des Jahres eine Präsentation meiner Arbeit vor 6 Tonmeisterkollegen mit anschließender Vorführung ausgewählter Beispiele. Alle sind auch mit der Produktion des SWR Sinfonieorchesters befasst und haben große Erfahrung in ästhetischer Klangregie. Gehört wurde über die gewohnten Studio-Lautsprecher Genelec 1038b, den im SWR üblichen Stax-Kopfhörer zog nur ein Kollege zurate.

Das Ergebnis überraschte mich, denn der überwiegende Tenor war: Die exakte Kompensation klingt tot, langweilig, ohne Verzögerung, und bei den Beispielen im Solo-Gesang mit +20 ms dagegen „geht der Raum auf“ (ein Kommentar eines älteren Kollegen, der deutliche Rückwürfe gewohnt ist). Fast immer wurde die nicht-kompensierte Version bevorzugt, bei den unterschiedlichen Laufzeiten der Sänger die maximale oder unverzögerte Fassung.

Vielleicht ist dies auf Gewohnheit zurückzuführen, denn fast alle Kollegen haben ihre Karriere noch ohne selbstverständliche Möglichkeiten zur Delay-Einstellung angefangen, oft zu Analogzeiten. Ich denke aber, dass, wie in jeder Kritik, auch hier ein Fünkchen Wahrheit drin steckt. Bei einer exakten Kompensation gibt es signifikant weniger Impulsverdoppelungen (virtuelle Rückwürfe sozusagen) in der Mischung, die wahrgenommene Räumlichkeit reduziert sich merkbar. Hier wäre es folglich Aufgabe der Mischung, anderweitig für genügend Reflexionen und Rauminformation zu sorgen.

Toningenieure

Zwei Wochen später machte ich eine weitere Vorführung im SWR im gleichen Studio, diesmal mit neun Toningenieuren, die alle über viel Erfahrung mit Orchestermischung verfügen. Wieder stellte ich die Arbeit kurz vor und spielte dann die vorbereiteten Beispiele über unsere Studiolautsprecher ab. Einige Kollegen setzten auch den Kopfhörer (Stax SR-L700 Pro) auf, dabei lief allerdings die Lautsprecherwiedergabe ungemindert weiter, was ein gewisses Übersprechen bedingt.

Interessant war, dass von dieser Berufsgruppe ganz andere Prioritäten gesetzt wurden, und die Diskussion in eine andere Richtung lenkte. Der Nutzen einer Laufzeitkompensation wurde hier nicht bestritten, einige Kollegen ließen sogar durchblicken, dass sie dies auch gegen die Meinung ihrer Tonmeisterkollegen stillschweigend durchsetzten. Scheinbar lässt sich mit den Verzögerungen die Mischung einfacher handhaben, die Pegelverhältnisse und Räumlichkeit eindeutiger einstellen.

Allerdings gab es eine gewisse Skepsis, was die Einordnung der Solisten im Panorama anging. Mit Laufzeitkompensation und darüber hinausgehenden Verzögerungswerten geht die Kontrolle der Position auf das Hauptmikrofon über, die Stütze kann nicht mehr korrigierend die Richtung verschieben, so dass dann doch oft ohne Delay gearbeitet wird.

Schlussfolgerungen

Aus den Untersuchungen zeigt sich, dass die Auswirkungen der Laufzeiten unterschiedlich sein können.

1 Eine möglichst exakte Kompensation ist immer noch der beste Ausgangswert.

2 Stützen als Reflexionen zu betrachten, ist Schnee von gestern, oft verschlechtert sich das Klangbild sogar.

3 Die Optimierung hängt von der primären Eigenschaft des Signals, v.a. von dem Verhältnis Impuls (Einschwingen) zu Ton (stationärem Teil), und von der Ähnlichkeit der Signale ab (z.B. abhängig vom Abstand zu dem Hauptmikrofon).

4 Das Übersprechen der Signale ist nicht zu vernachlässigen, allerdings ist hier keine Gesamtlösung möglich. Bewährt hat sich die Optimierung der Pauke bei den davorstehenden (Holzbläser-)Stützen, relevant sind Mikros mit einem Übersprechpegel größer als -18dB.

5 Chorische Stimmen (Streicher, Chor) profitieren wenig bis gar nicht von einer Laufzeitkompensation.

Literatur

[1] „Raumbezogene Stütztechnik - eine Möglichkeit zur Optimierung der Aufnahmequalität“, Martin Wöhr, Günther Theile, Hans-Jürgen Goeres, Alexander Persterer; IRT, Bericht zur TMT 1988(20) S. 302, ISBN 3-598-20356-X

[2] „On the Audibility of Comb Filter Distortions“, Stefan Brunner, Hans-Joachim Maempel, Stefan Weinzierl, TU Berlin AES Convention:122 (May 2007) Paper Number:7047

[3] „Applied Delay on Support Microphones in Multi Microphone Recording Technique“, Michael Kaaber Harrit, Danmarks Radio, Copenhagen, Denmark, AES Convention 1996 May 11-14 Copenhagen, Paper Number: 4159 (D-2)

[4] „Digital Time Coherent Recording Technique“, T. Anazawa and Y. Takahashi, AES Paper 4909, Preprint2, 493(H-2) (October 1987), 1-8

[5] Forschungsarbeit „Transparenz bei Orchesterproduktionen“, Untersuchungen zur Erhöhung der Transparenz mit den Mitteln der Laufzeitkompensation, Prof. Michael Sandner, Erich-Thienhaus-Institut der Hochschule für Musik Detmold

Die richtige Intonation – eine Geschmacksfrage?

Nora Brandenburg
Saarländischer Rundfunk, Email: kontakt@norabrandenburg.de

Abstract

Künstlerische Aufnahmeleitung im Bereich klassischer Musik bedeutet permanente Urteilsbildung über die Richtigkeit von Notentext, Balance, Dynamik und – natürlich – Intonation. Doch kann die Frage nach „zu hoch“ oder „zu tief“ wirklich immer zweifelsfrei beantwortet werden? Ein Aufnahmevergleich zeigt zwei offenbar entgegengesetzte Geschmacksrichtungen des Intonierens, die sich bei genauerem Hinsehen theoretisch begründen und einem ästhetischen Konzept zuordnen lassen. Als Produzenten müssen wir Intonation also unbedingt unter Einordnen der gewählten Ästhetik und vor dem Hintergrund der Gesamtinterpretation beurteilen und gegebenenfalls korrigieren.

1. Ist Intonation eine Entscheidung?

Wenn uns im Zusammenspiel mehrerer Instrumente etwas unsauber erscheint, hat das natürlich sehr oft technische Ursachen. Abgesehen von spieltechnischen Unfällen gibt es eine Reihe instrumentenspezifischer Probleme, die das Intonieren erschweren: So wird z.B. bei Oboen und Fagotten die Tonhöhe im Decrescendo tiefer, bei Klarinetten und Saxophonen dagegen höher; Blasinstrumente verändern ihre Tonhöhe bei Temperaturschwankungen in genau entgegengesetzter Richtung zu Saiteninstrumenten; und bei Sängern und bestimmten Blasinstrumenten weicht die Tonhöhen-Selbstwahrnehmung stark von der Außenwahrnehmung ab [1].

Gute Musiker sind sich dieser Schwierigkeiten bewusst und haben gelernt, sie zu kompensieren. Doch wie lässt es sich erklären, dass Weltklasse-Musiker, von denen wir annehmen, dass sie bei einer Einspielung ihre Vorstellung von Intonation zu hundert Prozent realisieren konnten, dennoch zu sehr unterschiedlichen Ergebnissen gelangen – unter Umständen so unterschiedlich, dass es mehreren Musikern nicht gelingt, „sauber“ zusammen zu spielen?

2. Ein Interpret – zwei Meinungen

Der Vergleich zweier verschiedener Bach-Einspielungen von YoYo Ma aus den Jahren 1983 und 2018 (die *Allemande* aus der V. Suite für Violoncello Solo, BWV 1011) zeigt nicht nur große Interpretationsunterschiede hinsichtlich Tempo, Klang, Phrasierung und Artikulation, sondern auch eklatante Unterschiede in der Intonation.

Grundsätzlich scheinen in der frühen Einspielung die dritte und sechste Tonleiterstufe von c-Moll tiefer und der Leitton (*h*) höher intoniert als in der späten Einspielung. Das bedeutet in harmonischer Hinsicht, dass die Mollterzen in der frühen Einspielung kleiner gespielt sind als in der späten, die Durterzen dagegen größer; in melodischer Hinsicht bedeutet es, dass die Halbtonschritte in der frühen Einspielung erheblich enger genommen sind als in der späten Einspielung.

Die beiden Aufnahmen sollen hier paradigmatisch für zwei grundsätzliche Richtungen des Intonierens stehen, deren Vergleich noch wesentlich extremer ausfiele, wenn man Pablo Casals und einen Vertreter der historischen Aufführungspraxis gegenüberstellen würde. Doch wie lassen sich diese unterschiedlichen Entscheidungen, die selbst ein einzelner Interpret im Laufe seines Lebens trifft, begründen?

3. Reine und pythagoreische Intonation

3.1. Schwebungsfreiheit

Ein Hörvergleich dreier verschieden großer Durterzen (386, 400 und 408 Cent) ergibt, dass sich die Mehrheit der Befragten für die reine Terz (386 Cent) entscheidet, da sie schwebungsfrei klingt. Das Phänomen der Schwebung – ein regelmäßiges An- und Abschwellen der Lautstärke, hervorgerufen durch die mal gleichphasige, mal gegenphasige Überlagerung von Schwingungen mit benachbarter Frequenz – tritt deswegen auch bei Terzen und nicht nur bei Primen auf, weil sich außer den Grundtönen auch die gemeinsamen Teiltöne überlagern (**Abb. 1**). Stimmt das Teilungsverhältnis der Grundtöne, addieren sich die gemeinsamen Teiltöne, und der Klang wird lauter. Man erhält den Höreindruck, dass der Klang „steht“. Weicht das Intervall vom reinen Teilungsverhältnis ab, schweben die gemeinsamen Teiltöne gegeneinander – je größer die Abweichung, desto schneller die Schwebung. Da bei einer Oktave alle Teiltöne des oberen Intervalltons auch Teiltöne des unteren Tons sind, erzeugt schon eine sehr kleine Abweichung vom reinen Teilungsverhältnis 1:2 eine unangenehme Schwebung, weshalb es kaum Meinungsverschiedenheiten darüber gibt, ob eine Oktave sauber ist oder nicht. Bei Terzen, die weniger gemeinsame Obertöne haben, ist der Diskussionsspielraum wesentlich größer.

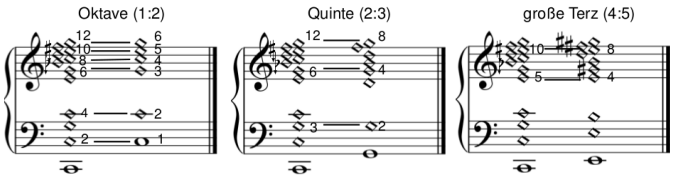


Abb. 1: Überlagerung gemeinsamer Teiltöne bei reinen Intervallen.

ISBN
978-3-9812830-9-9

Der Autor hält die Urheberrechte für den gesamten Inhalt seines Manuskripts. Er überlässt dem VDT das Recht auf Veröffentlichung in gedruckter Form, online oder als PDF-Download von https://tonmeistertagung.com

Außerdem stört die Schwebung umso stärker, je obertonärmer das Spektrum ist. Bei Streichinstrumenten, besonders in einer chorischen Gruppe, ist die Toleranzschwelle relativ hoch, während im obertonarmen Spektrum von Klarinetten und Flöten jede Schwebung empfindlich hörbar ist. Holzbläser versuchen daher immer, Dreiklänge rein auszustimmen.

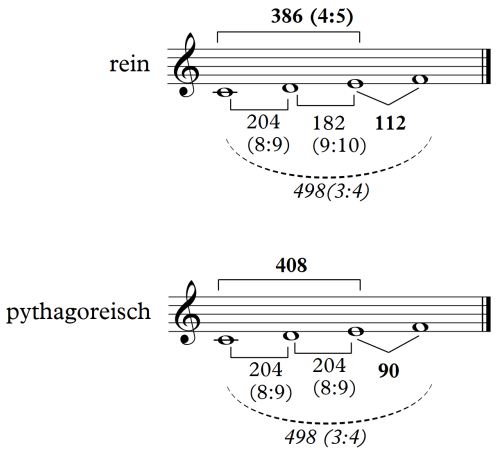


Abb. 2: Dur-Tetrachord in reiner und pythagoreischer Stimmung (Intervallgrößen in Cent).

3.2. Syntonisches Komma und Temperatur

Die reine Terz wirkt im melodischen Zusammenhang ganz anders als im harmonischen. Schreitet man in einem diatonischen Tetrachord von der Durterz zur IV. Stufe weiter (**Abb. 2**), bevorzugen die meisten Hörer die pythagoreische Terz (408 Cent), weil ein kleinerer Halbtonschritt entsteht und beide Ganztonschritte gleich groß sind. Ganzton und Halbton lassen sich besser unterscheiden, und es entsteht ein stärkerer Eindruck von melodischer Richtung.

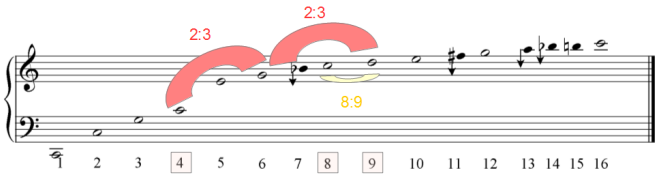


Abb. 3: Naturtonreihe und Teilungsverhältnisse von Intervallen

Eine pythagoreische Terz lässt sich ebenso durch das Addieren von zwei sog. „großen“ Ganztönen (8:9) wie durch das Addieren von vier reinen Quinten (2:3, 702 Cent) herstellen (**Abb. 3**). Nichts anderes passiert beim Einstimmen der leeren Saiten eines Streichquartetts (**Abb. 4**). Streicher wissen, dass die Dezime c-e" zwischen der tiefsten und der höchsten Saite nur dann als Zusammenklang nutzbar ist, wenn anstelle der schwebungsfreien Quinten verengte Quinten eingestimmt werden. Andernfalls ist die Terz wegen der starken Schwebung nicht brauchbar. Die Differenz zwischen der reinen und der durch Quintenschichtung erzeugten Terz heißt *syntonisches Komma* und beträgt etwa 21,5 Cent.

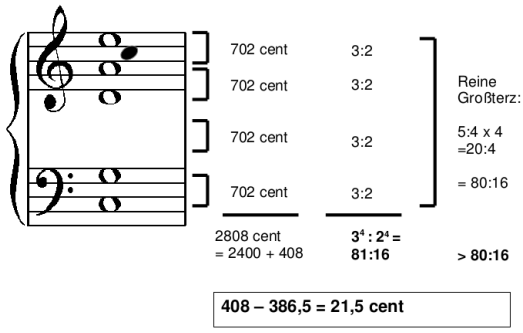


Abb. 4: Das syntonische Komma ("Terzen-Komma").

Jede Art von Temperatur basiert darauf, dieses Komma irgendwo im System zu verstecken oder zu verteilen (**Tab. 1**), also einen Kompromiss zu finden zwischen reinen Terzen und reinen Quinten. Je nachdem, für welche Musik und welches Instrumentarium sie geeignet sein soll, kann sich eine Temperatur also für ein Tonsystem entscheiden, das zugunsten einiger wichtiger schwebungsfreier Intervalle einen unbrauchbaren Rest in Kauf nimmt, oder ein Tonsystem, das unter grundsätzlichem Verzicht auf Schwebungsfreiheit das Komma so gleichmäßig verteilt, dass es keine ganz unbrauchbaren Intervalle gibt.

	rein	**mitteltönig**	**gleich-schwebend**	**pythagoreisch**
Quinte	702 (2:3)	697	700	702 (2:3)
Großterz	386 (4:5)	386 (4:5)	400	408
Ganzton	204 (8:9)/ 182 (9:10)	193	200	204 (8:9)
Eigen-schaften	Wolfsquinte zwischen II. + VI.	acht Terzen sind rein	alle Intervalle schweben	alle Terzen schweben

Tab. 1: Temperatur als Kompromiss: Intervallgrößen in verschiedenen Stimmungssystemen (in Cent).

3.3. Pablo Casals' *expressive intonation*

Die pythagoreische Durskala (**Abb. 5**) wirkt im Vergleich zur reinen geradezu pädagogisch, weil sie den Unterschied zwischen Halb- und Ganztönen so stark deutlich macht. Die Differenz zwischen dem großen reinen und dem kleinen pythagoreischen Halbtonschritt entspricht dem syntonischen Komma. In der Praxis der melodischen Intonation von Streichern ist dieser Unterschied im Allgemeinen noch weitaus größer. Pablo Casals, gehandelt als größter Cellist des 20. Jahrhunderts und sicher einflussreichster Lehrer über mehrere Generationen, hat sein Ideal des Intonierens als *expressive intonation* bezeichnet. Mit Bezug auf Bach fordert Casals, im schnellen Tempo die diatonischen Halbtonschritte auf bis zu einen Achtelton (!) zu verengen [2] und die chromatischen Halbtonschritte entsprechend weit zu spielen. Große und übermäßige Intervalle sollen erweitert, kleine und verminderte Intervalle verengt werden.

Casals' ästhetische Begründung für diese Methode ist das funktionsharmonische Spannungsfeld, in dem zwischen Tonika und Dominante, zwischen Grundtönen und Leittönen magnetische Anziehungskräfte wirken.

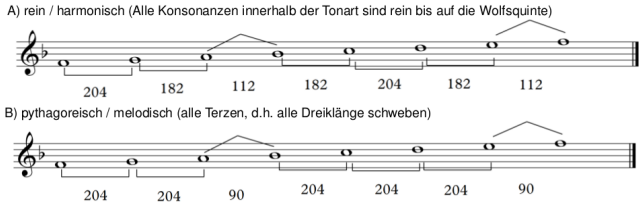


Abb. 5: Die Durtonleiter in verschiedenen Stimmungssystemen (Skalenschritte in Cent).

melodische / pythagoreische Intonation	**harmonische / reine Intonation**
„Streicher-Intonation“	„Bläser-Intonation“
Pablo Casals: *expressive intonation*	historische Aufführungspraxis
enge diatonische Halbtonschritte („leittönige“ Intonation)	weite diatonische Halbtonschritte (spannungsarme Melodik)
weite Durterz (408 Cent), enge Mollterz (294 Cent)	enge Durterz (386 Cent), weite Mollterz (316 Cent)
großer Unterschied zwischen Dur und Moll	geringer Unterschied zwischen Dur und Moll
gis ist höher als *as*	*gis* ist tiefer als *as*
schwebende Dreiklänge	schwebungsfreie Dreiklänge
erweiterte Sprünge, dadurch erhöhte hohe Lage	Oktavidentität, neutrale Intonation

Tab. 2: Gegenüberstellung von pythagoreischer und reiner Intonation.

3.4. Intonation als Kompromiss

Tabelle 2 zeigt eine bewusst polemisierende Gegenüberstellung von pythagoreischer und reiner Intonation. Natürlich intoniert kein Streichquartett ausschließlich pythagoreisch, sondern strebt im harmonischen Kontext schwebungsfreie Dreiklänge an. Ebenso unmöglich ist es aber auch, ausschließlich rein zu intonieren. Ein Experiment, bei dem in einer Hassler-Motette konsequent alle horizontalen und vertikalen Intervalle rein gestimmt wurden, zeigt, dass die Gesamtstimmung zwangsläufig schon nach 16 Takten um einen Halbton absinkt. Grund dafür ist die Wolfsquinte (das Komma), die sich typischerweise zwischen der II. und VI. Tonleiterstufe als Problem einschleicht. Intoniert man die VI. Tonleiterstufe in C-Dur (*a*) als reine Terz der Subdominante (also 16 Cent tiefer als in Klavier-Stimmung) und schreitet von F-Dur aus zu einem d-Moll-Dreiklang fort, kann man den zum tiefen *a* als reine Quinte gestimmten Ton *d* (18 Cent tiefer) nicht mehr als reine Quinte der Dominate G-Dur benutzen, ohne die übrigen Dreiklangstöne von G-Dur um 22 Cent nach unten zu korrigieren.

3.5. Schwierigkeiten in der Praxis

Ein Beispiel aus der Orchesterliteratur soll abschließend noch einmal die Schwierigkeiten des Intonierens aufzeigen. Der gewählte Ausschnitt, von dem ich noch keine sauber intonierte Einspielung gefunden habe, stammt aus dem Dritten Akt aus Wagners *Siegfried* (T. 1067-1072): Brünnhildes Erwachen nach dem Schlaf auf dem Felsen und ihre Begrüßung von Erde und Himmel. Die Geste wird markiert durch einen e-Moll-Dreiklang in den Holzbläsern, der sich dann zu einem zwanzig Sekunden andauernden, himmelwärts strebenden C-Dur Dreiklang wendet. Der Akkord beginnt in den Bläsern und wird über aufsteigende Harfen-Arpeggien an die Geigen weitergegeben, die ihn in sehr hoher Lage beschließen. In der gewählten Referenzeinspielung aus Bayreuth (Teldec 1991) ist zu hören, wie die Holzbläser den zunächst unsauber angesetzten e-Moll-Dreiklang innerhalb weniger Augenblicke rein ausstimmen. Der C-Dur-Akkord beginnt zunächst sauber, wird aber durch den Einsatz der Harfen gestört, deren gleichschwebende hohe Durterzen nicht in den Akkkord passen und die zu diesem Zeitpunkt sowieso schon seit Längerem nicht mehr nachstimmen konnten. Die Geigen schließlich, im Bestreben, einen möglichst strahlenden Klang zu erzeugen, enden fast einen Viertelton zu hoch.

4. Fazit

Es ist deutlich geworden, dass es keine universal gültigen Kategorien von „zu hoch“ und „zu tief“ geben kann. Als Produzenten müssen wir Intonation immer vor dem Hintergrund der gewählten Ästhetik und Gesamtinterpretation beurteilen, was zur Voraussetzung hat, dass wir diese Ästhetik erkennen und einordnen können.

5. Literatur

[1] Geller, D.: Praktische Intonationslehre. Bärenreiter, Kassel, 1997.

[2] Blum, D.: Casals and the Art of Interpretation. Univ. of California Press, Berkeley, 1977, S. 107.

Konzeption einer SMPTE 2110 konformen Broadcast-Infrastruktur
am Beispiel eines transmedialen Verkaufsfernsehsenders

NorbertWilinski
HSE24 Ismaning, Germany: n.wilinski@hse24.de

Abstract

Eine zukunftssichere Broadcast-Infrastruktur ist für den Erfolg eines Fernsehsenders mit vorwiegendem Live Programm entscheidend. Parallel zu den verbreiteten TV-Formaten kann man eine überproportional steigende Bedeutung von E-Commerce und Social Media beobachten, das gilt auch für Teleshopping. Dabei muss die Produktionsumgebung flexibel gestaltet werden, um diese schnell an veränderte Marktbedingungen anzupassen. Die simultane Produktion für TV und Social Media muss ebenso möglich sein wie das sukzessive Arbeiten für verschiedene Medien, möglichst ohne Zeitverlust durch technische Umbauten.An einer IP-basierten Lösung führt kein Weg vorbei, jedoch müssen aus Gründen der Sendesicherheit auch weiterhin konventionelle Ansätze berücksichtigt werden. SMPTE 2110 bietet erstmals eine Lösung, die speziellfür den Live Sendebetrieb geeignet ist. Vorgestellt wird eine Konzeption, die eine klassisch konzipierte Broadcast-Infrastruktur IP-basiert weiterentwickelt.

1. Einführung

Der Broadcastbereich wurde durch die IP-Technologie in den letzten 12 Jahren revolutioniert.Audio- und Videoapplikationen verschmelzen immer mehr.In der Planung für neue Systeme lassen sich diese Bereiche nicht mehr trennen.

Die Praxis zeigt, dass konventionell aufgebaute Broadcast Systeme sukzessive durch IP-basierte ersetzt bzw. erweitert werden.

Einen möglichen Übergang zur IP-basierten Infrastruktur im Hinblick auf SMPTE 2110 soll hier am Beispiel eines marktführenden Shoppingsenders dargestellt werden.

2. Milestones Broadcast IP

Ein kurzer Rückblick auf die Entwicklung der letzten Jahre im Bereich Broadcast IP zeigt, dass sich die Digitalisierung der Broadcast Systeme seit den 1990er Jahren mit der Einführung von AES 10 [1] noch verhältnismäßig langsam vollzog.Nach der Digitalisierung der Audiosignale wurde zunächst die Steuerung für Router und andere Endgeräte im Broadcastbereich IP-basiert. Serielle Schnittstellen verschwanden allmählich. Nach der Einführung von IP-Audio 2006 mit *Dante* und *Ravenna* nahm die Entwicklung Tempo auf. Damit wurde Processing in dezentralen IP-Netzwerken möglich und setzte sich schnell durch.

Der Durchbruch im Videobereich begann im Jahre 2012 mit der Einführung von SMPTE 2022-6 [2]. Ein Jahr später wurde im Audiobereich mit der Einführung von AES 67 [3] ein weiterer wichtiger Schritt vollzogen.

Erst mit SMPTE 2110 im Jahre 2017 schließlich wurde ein Standard erreicht, der für den Live Betrieb voll einsetzbar ist [4]. Die Streams für Video, Audio und Metadaten werden hier separat geführt, im Gegensatz zu SMPTE 2022, das im Wesentlichen für den Beitragsaustausch konzipiert wurde.

Da das Audiostream-Format in SMPTE 2110 auf AES 67 basiert, wird damit flexibles, bandbreitenoptimiertes Routing möglich. Für Sender mit hohem Live-Anteil und komplexem Routing sind das die geeigneten Voraussetzungen, um diese Technik einzuführen.

3. Produktionsumgebung

Die Produktionsumgebung am Beispiel von HSE24 besteht aus drei Studios und einem Aussenset. Daneben gibt es drei Bildregien, drei Tonregien sowie ein Playout Center. Bild- und Tonregien sind räumlich voneinander getrennt. Ausgespielt wird synchron über drei TV Kanäle und Social Media, jedoch mit jeweils unterschiedlichem Content.

Zugeschaltet werden können noch eine Dolmetscher Kabine für Simultanübersetzungen sowie Skype, LiveU (ein Übertragungssystem für Video via Mobilfunk und Internet) und andere externe Quellen.

Die Regien und Studios sind den Programmen variabel zuordenbar. Das garantiert eine hohe Flexibilität. Es kann zum Beispiel Programm°1 (PGM1) aus Studio°1 in Verbindung mit Bildregie°3 und Tonregie°3 mit zugeschalteter Dolmetscherkabine bespielt werden, während PGM2 aus Studio°2 mit Splitt-Regie°1/2 und Skype bespielt wird (Abb.1).

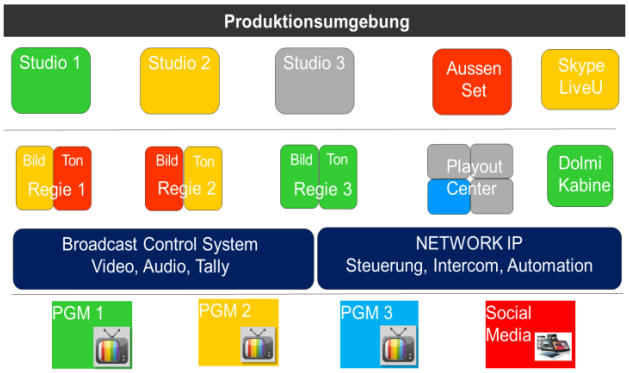


Abb. 1: Produktionsumgebung und Delegation: Studios, Aussenset, Dolmetscherkabine, LiveU, Skype und ein Playout Center sind den drei Programmen und Social Media über die drei Bild- und drei Tonregien frei zuordenbar.

Die Delegation von Audio- und Videosignalen, sowie Tally-Signalisierung (Rotlicht) funktioniert dabei klassisch über einen Broadcast Controller und wird von der Regiecrew selbst über endsprechend eingerichtete Bedienoberflächen und Panels durchgeführt. Das gleiche gilt für die Delegation von Kamerasteuerung, Interkom und Automation über IP-Network.Auf diese Weise wird ein transmediales Programm Management ermöglicht.

Im Folgenden wird zunächst die Broadcast-Infrastruktur dieser Produktionsumgebung analysiert, danach wird auf die IP-basierte Erweiterung eingegangen.

4. Broadcast-Infrastruktur

Der Aufbau der Broadcast-Infrastruktur ist hier mit einem Baseband Video Router und einem dezentral aufgebautem AudioRouter klassisch ausgeführt (Abb. 2).

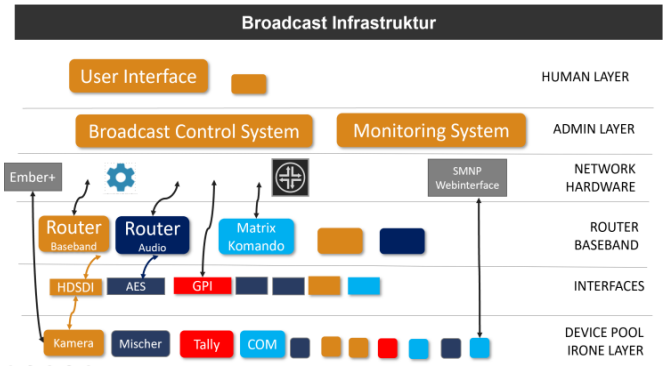


Abb. 2: Aufbau einer klassischen Broadcast-Infrastruktur

Im Device Pool oder *Irone Layer* der untersten Ebene befinden sich die Endgeräte der Studios, Regien und Technikräume. Diese sind über Interfaces mit den Routern verbunden, undin der Regel dezentral aufgebaut,z.B. bei Medion- oder Nexus-Installationen.

Darüber hat sich in klassischen Installationen häufig eine Netzwerkebene mit einem Coreswitch etabliert, der als Backbone für das Broadcast Control System dient. Damit werden u. a. die Router gesteuert und Endgeräte gemonitort und konfiguriert.

Auf der Admin-Ebene befindet sich der Broadcast Controllerzur Steuerung der Kreuzschienen und der Signalisierung. Außerdem findet man hieroft auch ein Monitoring System mit Zugriff auf die Device Ebene via SNMP, *Simple Network Management Protocol*, einem Überwachungsprotokoll zur Übertragung von Fehlermeldungen, sowie via Webinterfacezur Konfiguration [5].

On Top befindet sich der sogenannte Human-Layer mit den User Interfaces: Bedienpanels, Oberflächen und Kontrollanzeigen.Bedienoberflächen für Delegation und Havarie-Schaltungen werden hier unter Verwendung des Steuerprotokolls Ember+ realisiert. Dabei handelt es sich um ein Open Source Steuer- und Monitoring Protokoll für Broadcast-Anwendungen, das von vielen Herstellern unterstützt wird [6]. Man kann so Bedienoberflächen gestalten, die u.a. Routing, Kommando und DSP-Funktionen verschiedener Geräte vereinen und damit den Workflow erheblich vereinfachen.

Im Folgenden wird die Optimierung des Network Hardware Layer in Bezug auf SMPTE 2110 betrachtet.

5. Spine-Leaf Netzwerk

5.1. Aufbau

Es wird von einem Network Hardware Layer mit einem SingleCore switch und angeschlossenen Nodes ausgegangen (Abb.3).

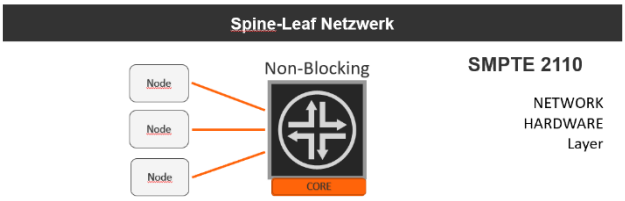


Abb. 3: Non-Blocking Monolith Netzwerk Core

Der Single Coreswitch ist zu einem sogenannten *Non-Blocking Monolith Network Core*mit IGMP-Management konfiguriert und erfordert deshalb keine weitere Kontrolle hinsichtlich der Bandbreiten mehr.

Non-Blocking bedeutet, dass die Bandbreite im Netzwerk in jeden Fall ausreichend ist, um *Clean Switching*zu gewährleisten. Es kann also ohne Latenz auch unter Volllast geschaltet werden, was insbesondere für den Livebetrieb erforderlich ist [7].

IGMP,*Internet Group Management Protocol*‘ ist ein Netzwerkprotokoll der Internetprotokollfamilie und dient zur Organisation von Multicast-Gruppen, also zur Verteilung von Streams [8]. Dies ist die einfachste Konfiguration eines SMPTE 2110 konformen Netzwerks.

Bei geeigneter Wahl der richtigen Plattform kann dieses Single-Core-Netzwerk einfach auf ein leistungsstarkes *Spine-Leaf Netzwerk* erweitert werden [9].

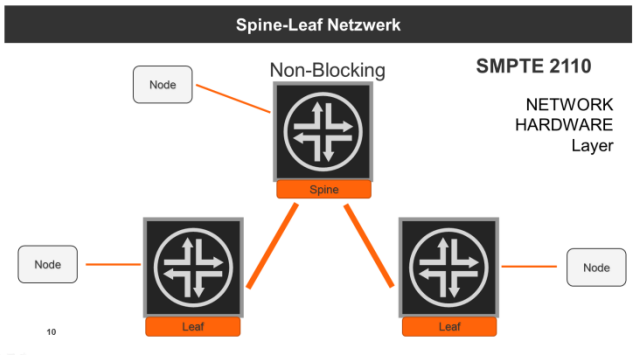


Abb. 4: Spine-Leaf Netzwerk

Wird der Single Coreswitch zu einem Spine Switch, kann das Netzwerk mit Leaf Switches ausgebaut werden. Alle oder einige Nodes können dann auch umgezogen werden. Auf diese Weise ist das System stetig erweiterbar (Abb. 4).

Eine Spine-Leaf Topologie besteht aus zwei Ebenen: Spine und Leaf. Diese Konstellation bietet den Vorteil, dass die Streams von einem Sender zu einem Empfänger immer den gleichen Weg nehmen und damit auch stets die gleiche Bandbreite benötigen.

Dadurch ergibt sich eine einfache Kontrolle der Streams Engpässe können so ausgeschlossen und Latenz kann minimiert werden.

5.2. Bandbreite im Netzwerk

In einem Spine-Leaf Netzwerk für Live Betrieb muss stets genügend Bandbreite zur Verfügung stehen, um *Clean Switching* zu gewährleisten. Eine wichtige Voraussetzung dafür ist, dass genügend Trunk-Ports vom Spine zum Leaf Switch vorhanden sind.

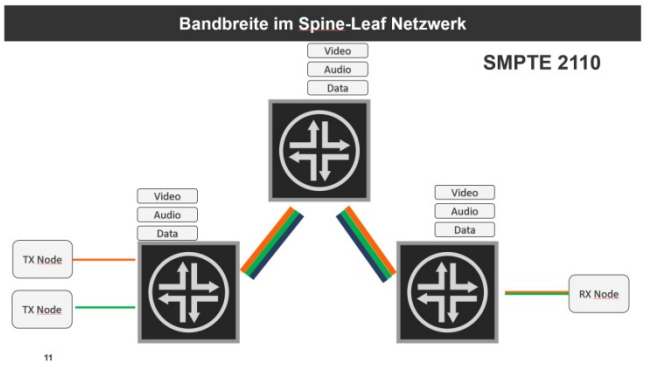


Abb. 5: Bandbreite im Spine-Leaf Netzwerk

Dazu ein Beispiel: Um eine Frame-genaue Bildumschaltung zwischen 2 Kamerasignalen auf einen Bildmischer zu erreichen, ist zunächst die doppelte Bandbreite erforderlich:Beide Signale liegen am Bildmischer an (Abb.°5).

Bei schnellen Umschaltvorgängen kann die *Leave Message* eines Streams (Abkündigung eines Streams) zusätzliche Zeit erfordern, sodass die Anforderung eines neuen Streams,noch nicht ausgeführt werden kann. Es besteht dann die Gefahr, dass die verfügbare Bandbreite überzeichnet wird. Eine dritte Trunkverbindung wird für *Clean Switching* notwendig [10].

Um die Non-Blocking Funktion im Netzwerk zu gewährleisten, muss hier die dreifache Bandbreite vorgehalten werden. Die Trunk-Verbindungen haben entweder genügend Bandbreite oder müssen mit Hilfe eines sogenannten *Orchestrators* verwaltet werden (s.u.).

Die notwendige Bandbreite ist nicht nur abhängig vom Signal, sondern auch von der Stream Anzahl, die gleichzeitig durch das Netzwerk geroutet werden muss.

An einem Beispiel soll veranschaulicht werden, mit wieviel Streams in einem SMPTE 2110 konformen Netzwerk zu rechnen ist, wenn die Video-, Audio- und Metadaten in separaten Streams geführt werden.

5.3. Streams imNetzwerk

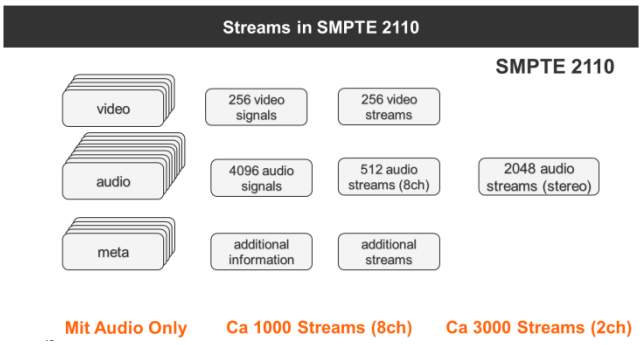


Abb. 6: Berechnung der Stream-Anzahl in einer SMPTE 2110 Installation

Ausgehend von 16 Audiokanälen pro Videosignal ergeben sich bei einer mittelgroßen Installation mit 256 Videoquellen und Senkeninsgesamt 4096 Audioquellen und Senken nur für den Videopart. Dazu kommen die Metadaten (Abb. 6).

Die AudioStream Anzahl kann auf Empfehlung von SMPTE 2110 zwischen 2 und 64 Kanälen gewählt werden. Gehen wir von 8 Audiokanälen pro Stream aus, erhält man 512 Audiostreams für die 256 Videostreams. Bei einer Stereoinstallation erhält man dann 2048 Audiostreams alleine fürs Video.

Mit Metadaten und Audio Only kann man also mit ca. 1000 Streams bei 8 Audiokanälen oder 3000 Streams bei einer Stereokonfiguration rechnen.

Eine kleine Veränderung im StreamDesign kann also eine große Veränderung der Stream Anzahlzur Folge haben.

6. Broadcast-Infrastruktur IP

6.1. Aufbau

Wie ändert sich die klassische Broadcast Infrastruktur wenn diese SMPTE 2110 konform ausgebaut wird?

Im Device Pool befinden sich neue kompatible Endgeräte. Diese können an einen Leaf-Switch angeschlossen werden. Hier ist es bereits redundant ausgelegt und verfügt über ein cloudbasiertes Processing. Die klassischen Baseband Video- und Audiorouter werden falls möglich mit SMPTE 2110

bzw. AES 67 kompatiblen Schnittstellen nachgerüstet und können so auch in die neuen Infrastruktur voll integriert werden.

Human und Admin Layer bleiben unverändert: Broadcast Control System, User Interfaces und das Monitoring, welches jetzt auch das Netzwerk überwacht (Abb. 7).

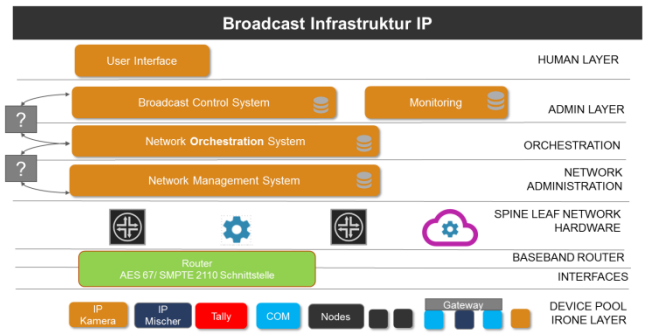


Abb. 7: Aufbau der IP erweiterten Broadcast-Infrastruktur

Um auf Netzwerkebene schalten zu können ist in der IP Infrastruktur das *Network Orchestration System* von zentraler Bedeutung. Man kann den *Orchestrator* mit einem Navi vergleichen, welches ein Regelwerk enthält, wie IP-Ströme durchs Netzwerk geroutet werden, damit keine Engpässe bezüglich der Bandbreite bzw. der Anzahl der Streams entstehen.

Nur wenn das Broadcast-Netzwerk in einem Verbund mit anderen Netzwerken betrieben wird, benötigt man an dieser Stelle ein *Network Management System*. Hier würde man beispielsweise auch die Schnittstelle zu einem *Medien Asset Management* (MAM) administrieren [11].

Jedes dieser Systeme enthälteine Datenbank mit Informationen über das Netzwerk und den angeschlossenen Geräten, mit jeweils unterschiedlichem Fokus.Netzwerk Management und Orchestration sind IP fokussiert. Die Datenbanken enthalten vor allem Routing Informationen und sind mit der Broadcast-Welt nicht unbedingt kompatibel.

Die Datenbank des Broadcast Controllers hingegen enthält die Namen der Mikrofone, Kameras, Server, Monitore usw.In der Praxis besteht die Schwierigkeit die Datenbanken der beteiligten Systeme zu verwalten und zu synchronisieren.Hier beobachtet man auf dem Markt interessante Entwicklungen. Man versucht, die Datenbanken und die Funktionalität der Komponenten zusammenzuführen um die Administration wesentlich zu vereinfachen.

Der *Orchestrator* hat also als zentraler Bestandteil des Broadcast IP Netzwerkes zentrale Funktionalität, im Folgenden ein kurzer Überblick.

6.2. Network Orchestration System

Der *Orchestrator* führt die Schaltvorgänge im Netzwerk aus, indem Streams vom Sender zum Empfänger adressiert werden. Er simuliert den Router für das Broadcast Control System.

Hier können den Streams dynamisch bestimmte Trunk-Ports zugewiesen werden, damit die dreifache Bandbreite beim Schalten zur Verfügung steht: Stichwort *Clean Switching*.

Der *Orchestrator* dient auch zur Überwachung und Administration des Netzwerks, beispielsweise von PTP, *Precision Time Protocol* (Synchronisation-Protokoll) [12].

Über ihn wird ebenfalls das Havarie- und Konfliktmanagement durchgeführt.

Er gibt Rückmeldung und Infos an den Broadcast Controller, die dieser dann an die angeschlossenen User Interfaces signalisieren kann.

Im besten Falle kann er auch mit den EndNodes kommunizieren um u.a. die Fehlersuche zu vereinfachen [13].

7. Zusammenfassung

Am Beispiel eines Verkaufssenders wurde gezeigt, wie weit die Integration von IP Systemen im Broadcastbereich fortgeschritten ist.An einer aktuellen Produktionsumgebung wurde gezeigt wie flexibel die Produktions-Ressourcen eingesetzt werden können.

Die vorhandene Broadcast-Infrastruktur wurde IP-basiert erweitert und dazu der Netzwerk-Layer neu designed. Dabei wurde die Spine-Leaf Topologie vorgestellt und die Bandbreite im Netzwerk in Verbindung mit Trunk-Lines untersucht.

Des Weiteren wurden anhand eines Beispiels die Anzahl der Streams in einem SMPTE 2110 Netzwerk berechnet, sowie die Funktionalität neuer Komponenten erläutert, wobei dem Network Orchestration System eine zentrale Bedeutung zukommt.

8. Fazit

Ein IP-basierter Ausbau einer Broadcast-Infrastruktur bringt erhebliche Vorteile. Wichtiges Kriterium derartiger Systeme ist die einfache Bedienbarkeit. Bei der Auswahl adäquater Systeme für den Einsatz in einer Live Broadcast Umgebung ist das entscheidend. Hier muss immer die Sendesicherheit im Fokus stehen.

Die entscheidenden Punkte, um das Potential der IP Technologie im Live-Broadcastbereich voll auszuschöpfen, sind:

- intelligente Userinterfaces
- übersichtliche Bedienoberflächen
- Tools zur Fehlersuche und Korrektur
- redundantes Netzwerk Design

In diesem Zusammenhang ist auf Seiten der Hersteller noch viel Entwicklungsarbeit zu leisten und als Anwender sollte man sich genau informieren.

Dennoch eröffnet IP interessante Alternativen zur klassischen Infrastruktur. Mit SMPTE 2110 wird die IP-basierte Broadcast-Infrastruktur zweifelsohne in naher Zukunft weiter an Bedeutung gewinnen.

9. Links

[1] AES 10:Audio Engineering Society, Inc.;URL:<http://www.urbancom.us/aes10-1991.pdf>; Abrufdatum: 20.12.2018

[2] SMPTE 2022:SMPTE Professional Development Academy, Thomas Edwards; URL: <www.smpte.org/sites/default/files/2017-08-17-ST-2022-Edwards-V4-Handout.pdf> Abrufdatum: 20.12.2018

[3] SMPTE 2110:SMPTE LTD, URL:<https://URL:<www.smpte.org/st-2110> Abrufdatum: 20.12.2018

[4] AES67: Audio Engineering Society, Inc.URL:<http://www.aes.org/standards/blog/2013/9/aes67-2013-audio-over-ip-130911> Abrufdatum: 20.12.2018

[5] SMNP: “Simple Network Management Protocol“ IETF, Network Working Group; J. Schoenwaelder, Jacobs University Bremen, 2008; URL: <https://tools.ietf.org/html/std78> Abrufdatum: 20.12.2018

[6] Ember+:Lawo Group, L-S-B Broadcast Technologies GmbH; URL: <https://github.com/Lawo/ember-plus/wiki> Abrufdatum: 20.12.2018

[7] Non Blocking Network: N. Desai; P. Balaji, P. Sadayappan, M. Islam; Ohio State University, Columbus; URL: <https://www.mcs.anl.gov/~balaji/pubs/2008/cluster/cluster08.netspots.pdf > Abrufdatum: 20.12.2018

[8] IGMP: „Internet Group Management Protocol“, Cisco Systems, Inc.; URL: <https://www.cisco.com/c/en/us/td/docs/ios-xml/ios/ipmulti_igmp/configuration/xe-3se/5700/imc-igmp-xe-3se-5700-book/imc_constr_sw_eth_nw.pdf> Abrufdatum: 20.12.2018

[9] Spine-Leaf:Cisco Systems, Inc; URL: <https://www.cisco.com/c/en/us/products/collateral/switches/nexus-7000-series-switches/white-paper-c11-737022.pdf> Abrufdatum: 20.12.2018

[10] Leave Massage: Cisco Systems, Inc; URL: <https://www.cisco.com/c/en/us/td/docs/ios-xml/ios/ipmulti_igmp/configuration/xe-3se/5700/imc-igmp-xe-3se-5700-book/imc_igmp.pdf>; Abrufdatum: 20.12.2018

[11]MAM:„Medien Asset Management“, wikepedia; URL:<https://de.wikipedia.org/wiki/Digital-Asset-Management>; Abrufdatum: 20.12.2018

[12] PTP:“Precision Time Protocol” IEEE-SA. URL: <https://ieeexplore.ieee.org/document/1048550>; Abrufdatum: 20.12.2018

[13] Network Orchestration System: PersönlicheBeratung; Humphrey Hoch, Senior Innovation Director, Broadcast Solution GmbH, URL:<https://broadcast-solutions.de/en/>; Abrufdatum: 20.12.2018

Eurovision Song Contest 2018 – immersive and interactive

Andreas Turnwald, Christian Simon, Ulli Scuda
Fraunhofer IIS, Erlangen, Germany, Email: andreas.turnwald@iis.fraunhofer.de

Abstract

With about 186 million viewers worldwide the Eurovision Song Contest (ESC) is one of the largest music TV events. In 2018 it took place in the Altice Arena in Lisbon, Portugal, in front of more than 10,000 spectators. With its combination of music and crowd reactions, the ESC is very suitable for a three-dimensional audio production. Along with the legacy 2.0 and 5.1 broadcast sound, the European Broadcast Union (EBU) in collaboration with Fraunhofer IIS produced an immersive audio version, which reflects the impressive atmosphere of this live event.

By using MPEG-H Audio, several commentaries in different languages can be broadcast in the same audio stream. Moreover, user interactivity can be provided, offering additional audio representations to the viewer. This paper reports about technical aspects of the recording, the aesthetic decisions during the mix, and about authoring and encoding to MPEG-H Audio.

1. Introduction

With the specification of MPEG-H 3D Audio in 2015 a new audio standard has been released, offering both immersive and interactive experiences for content producers and consumers alike [1]. As it enables consumption on various playback devices – be it binaural over headphones, on a smartphone, or in an eclectic 3D home cinema – this universal delivery approach is attractive for broadcasters searching for a way to reduce format overhead and to distribute both via conventional broadcast platforms as well as IP-based streaming.

The EBU and the Fraunhofer Institute for Integrated Circuits (IIS) have worked together during the Eurovision Song Contest in May 2018 for a live production of interactive and immersive sound using MPEG-H Audio.

Alongside a 5.0+4H international bed, the object-based capabilities of MPEG-H Audio were used to offer 3-5 commentaries in multiple languages in a single bit stream. Additional presets offered enhanced speech intelligibility or a 'live' version without commentary.

2. Immersive bed mix

The sound design for the ESC 2018 has been supervised by Daniel Bekerman (Auditiv), while the 5.1 live broadcast sound was mixed by Robert Edwards (X-Factor, Olympic Games, FIFA world cups).

Following the ESC tradition, all voice elements of the songs have to be performed live, only the background music is being played back. Vocals effects have been created using a Reaper DAW with Universal Audio plugins, automated via the timecode of the music playback.

The Fraunhofer sound engineers received the music and vocals & effects as separate 2.0-stereo mix stems from the OB van. For the immersive mix, the music was kept stereo as it was, while the vocals & effects have been upmixed to 3.0+2H using an Illusonic IAP16 processor.

The main difference of Fraunhofer's immersive approach compared to the aired 2.0 and 5.1 broadcast sound is the handling of the ambience microphones: For crowd reactions 22 Sennheiser MKH416 microphones have been positioned around the arena: 14 mounted close to the audience on floor stands and 8 more next to the delayed line arrays of the PA speakers.

For immersive purposes, the microphones mounted on the PA system had too much bassy music spill on their signal, making it unpleasant to use for crowd ambience while the song was performed. Thus, 10 of the floor microphones have been selected and panned to the speaker positions of the 5.0 medium layer and to the upper back speakers of the height layer.

By the use of the floor microphones close to the audience, crowd reactions (clapping along, singing, cheering) could be mixed to the songs without blurring the music's sound image too much.

The height effect was generated by adding an additional 4-channel Hamasaki Square microphone array to the existing crowd microphones used for the EBU broadcast [2].

Fig. 1: Hamasaki Square array for the height layer. Detail view: Figure-of-eight capsules pointing downwards by 30°

The array with a side length of two meters consisted of four Schoeps MK8 capsules mounted on an Ambient A-Ray. An Ambient Floater was used to prevent the vibrations of the lighting and PA rig from spilling into the precious ambience signal.

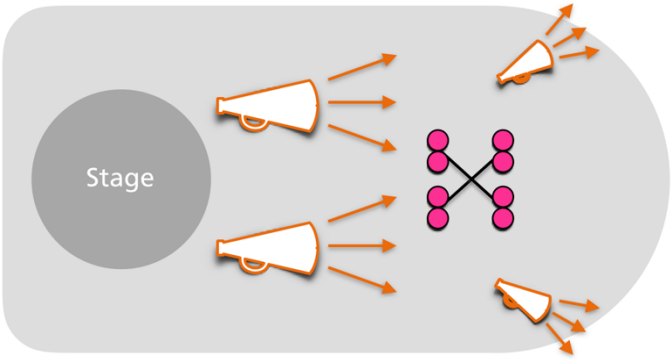


Fig. 2: Position of the Hamasaki square relative to the PA's line array speakers (much more delay lines had been in use, the drawing is reduced for clarity).

The Hamasaki Square was mounted in the middle of the arena in 26-meter height, above the coverage of the PA system. The figure of eight capsules, pointing outwards to the sides of the arena, showed their insensitive 90° angle towards the PA speakers in the front of the array (see Figure 2). This delivered a de-correlated ambience with as little music spill as possible and unproblematic behavior in possible 2.0 downmix scenarios.

Unlike Hamasaki's original concept, the figure of eight capsules were pointed slightly downwards by 30° to increase channel separation between left and right and to prevent the ambience from becoming too diffuse in this height above the audience (see detail, Figure 1).

The signals of the Hamasaki capsules were panned to the four corners of the height layer of the immersive mix. Their relatively diffuse sound adds beautifully to the dryer signals of the crowd microphones on the floor and complete the 'dome' effect of the immersive sound sphere.

Some effort has been spent on adjusting the correct delay compensation for each ambience signal. In a large space like the Altice Arena, all stage signals had to be delayed by 190ms to compensate for the arrival time to the last microphone in the back of the arena. All other ambience signals were placed accordingly inside this timing range. Caring about the delay values revealed itself to be sound design for the immersive acoustical space. Adjusting the delays cleared the sound image, removed a big portion of the flutter echoes present in the arena and also kept the audience clapping in rhythm with the music.

To monitor and mix the signals on location, an additional 5.0+4H control room was set up, receiving the OB van's sub-mixes (music, vocals & effects, hosts, video FX) and ambience microphones splits. All signal transmissions were realised by optical MADI connections.

Panning and mixing was done live on location inside a Steinberg Nuendo DAW, using its native VST Multipanner to create the immersive bed in a 5.0+4H channel layout.

Video playback (for example the 'video postcards' introducing the talents) and the preproduced sound effects (in combination with light effects at the beginning of each song) have been upmixed using the Nugen Halo Upmix plugin. Apart from the vocals- and video FX upmix, no further immersive effects, e.g. reverbs, have been used.

The resulting immersive mix can be monitored using existing MPEG-H Authoring and Monitoring Units (AMAU) or a DAW with MPEG-H enabled plug-ins to verify downmixes or a binauralized rendering.

3. Commentaries

The 37 commentators from various countries present at the ESC shows have been serviced and distributed by Riedel's MediorNet, independently from the EBU's international feed broadcast. While the national broadcasters received their respective commentary signals via SIP or ISDN, Riedel provided the whole bundle of 37 commentaries to the Fraunhofer control room via MADI.

All commentaries were filtered, dynamically compressed and levelled to have the same loudness impression, before up to five of them were selected to be part of the MPEG-H Audio scene.

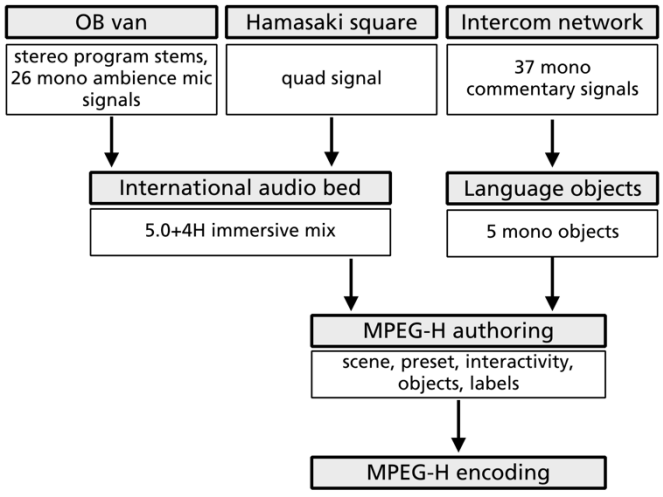


Fig. 3: Schematic audio production and authoring workflow of the ESC MPEG-H field test.

4. MPEG-H Audio

Often referred to as a "Next Generation Audio (NGA)" broadcast system, MPEG-H Audio has been adopted by ATSC [3], DVB [4] and SCTE [5, 6] specifications for usage of audio and video codecs within the respective system. Recommended bitrate values for broadcast can be found in [7].

For broadcast and streaming applications, MPEG-H Audio enables support for:

- 48kHz sample rate.
- Up to 32 audio elements (channels, objects, ambisonics) in one audio stream.
- Simultaneous rendering of up to 16 audio elements.

5. Authoring

To offer interactivity and personalization features to the consumer, metadata describing the MPEG-H Audio scene has to be authored: Different presets and possibilities for gain and position interactivity are created. The MPEG-H metadata is then embedded in the audio stream together with the audio content.

The audio scene for the ESC consists of the immersive 5.0+4H channel-based mix, which acts as an international bed, and up to 5 commentary objects in different languages (see Figure 3).

MPEG-H Audio is capable of handling static and dynamic objects as well as multichannel objects. For ESC each commentary signal was authored as a static mono object and placed at the centre speaker position.

As it is not intended to listen to several commentaries in different languages at the same time, the commentary signals are gathered in a so-called '*switch group*'. The receiving device will automatically select the language corresponding to the user's preference settings. Of course, the user can also select a different language in a specific programme.

Broadcasting several language objects alongside a common channel-bed significantly reduces the necessary bandwidth and enables broadcasters to efficiently offer multi-lingual services to their viewers.

Three presets were authored, featuring different sound representations of the event: *Default*, *Dialog+* and *Venue*. Presets are accessible in the MPEG-H on-screen menu of the receiver, offering a simple personalization option to the user (see Figure 4). For each preset the content creator can decide what features to enable. In regard to the ESC the *Default* preset contained the bed and the commentary switch group with no interactivity – just like in conventional TV broadcast. The *Venue* preset presented a version without commentary, offering the viewer a 'live' feeling of being present at the event.

Finally, the *Dialog+* preset offers enhanced speech intelligibility by lowering the background bed's volume by 9dB, while keeping the same overall loudness impression.

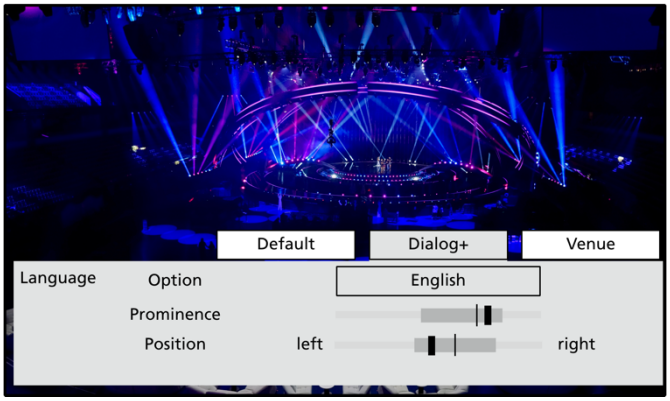


Fig. 4: Schematic of an MPEG-H receiver's on-screen menu offering user interactivity

In the advanced menu the viewer can access sliders to adjust the speech to background ratio to his own taste, or even pan the commentary (or an audio description object) away from the centre speaker position (left/right and up/down). All of these interactivity features are under full control of the content creator during authoring and can be enabled/disabled during live broadcast.

The MPEG-H Audio scene was authored on location using a standalone software tool developed by Fraunhofer IIS. Alternatively, different available DAW plugins with MPEG-H support could be used for this purpose.

For live broadcast purposes, hardware units like the Jünger MMA or Linear Acoustic AMS allow to author and monitor MPEG-H Audio scenes and generate the necessary metadata.

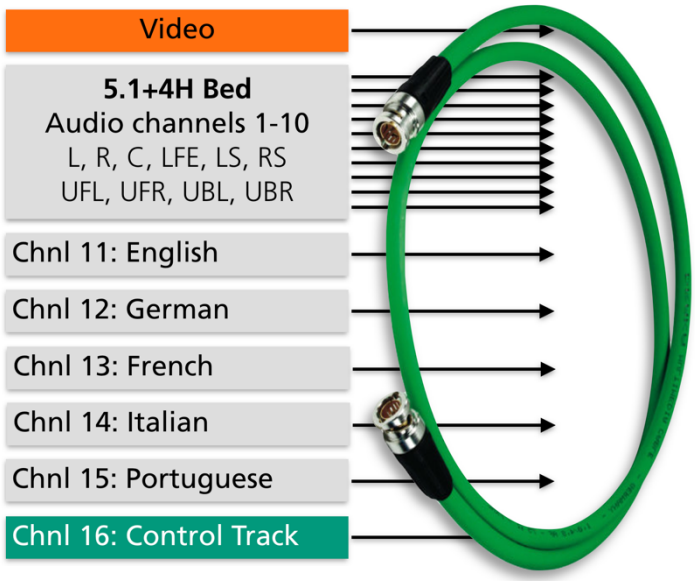


Fig. 5: Possible layout of a PCM multichannel audio signal, ready for encoding. MPEG-H metadata is carried in the Control Track on channel 16. In-house transmission is possible via SDI- or network infrastructure or by file-based workflows.

6. Encoding

The output of such MPEG-H authoring tools is plain PCM audio, with bed channels and objects on separate tracks. The MPEG-H metadata containing the description of the audio scene is being carried on the last audio channel, the *'control track'* (see Figure 5).

The control track – being a PCM audio signal – is robust enough to be fed through standard broadcast infrastructure and even survives level changes and sample rate converters.

The control track can be edited along with the audio and video on the borders of each video frame, making it possible to combine pre-authored audio scenes (e.g. in different video clips) into a broader programme context. The transition from one MPEG-H Audio scene to another is called *'configuration change'*.

In the case of the ESC, Fraunhofer IIS made use of such seamless configuration changes at the beginning of some songs: The commentary related presets are withdrawn and different music and ambience mix versions become available as presets for the consumer. All of this can be done 'on the fly' in live broadcast.

7. Conclusion

At ESC 2018 in Lisbon, Fraunhofer IIS in cooperation with the EBU conducted a field test for immersive and interactive audio production using the MPEG-H Audio codec.

Adding only four additional microphones, it was shown that capturing and producing immersive audio for live broadcast situations does not require completely new recording techniques.

The resulting immersive mix extends the emotional experience of the listener/viewer and increases the value of the content. Many visitors of Fraunhofer IIS' demos have stated that immersive audio offers a much deeper audio experience.

MPEG-H Audio can be easily enabled in existing production facilities. Content creators can offer advanced personalization features to the user, while maintaining full control over all interactivity options. Demands for speech intelligibility and audio description can be served by dedicated presets. Finally, multi-language transmission offers many possibilities to broadcasters and consumers alike.

MPEG-H Audio playback is available to the consumer market by receivers, soundbars and binaural headphone playback, making MPEG-H Audio an open and attractive format to produce for.

7. References

[1] International Organization for Standardization (ISO): "Information technology - High efficiency coding and media delivery in heterogeneous environments - Part 3: 3D audio", Geneva, Standard ISO/IEC 23008-3:2015, 2015.

[2] K. Hamasaki et al.: "Reproducing spatial impression with multichannel audio", in Proc. AES 24th Conference, Banff, 2003.

[3] Advanced Television Systems Committee: "ATSC Standard: A/342 Part 3, MPEG-H System", March 2017.

[4] European Telecommunications Standards Institute (ETSI): "Digital Video Broadcasting (DVB) – Specification for the use of Video and Audio Coding in Broadcasting Applications based on the MPEG-2 Transport Stream", TS 101 154 v2.3.1, 2017. Available: http://www.etsi.org/deliver/etsi_ts/101100_101199/101154/02.03.01_60/ts_101154v020301p.pdf

[5] Society of Cable Telecommunication Engineers (SCTE): "Next Generation Audio Coding Constraints for Cable Systems: Part 3 –MPEG-H Audio Coding Constraints". SCTE 243-2:2017

[6] A. Murtaza et al.: "MPEG-H TV Audio System for Cable Applications," in SCTE Journal of Digital Video Subcommittee, Volume 2, Number 1, July 2017.

[7] R. Bleidt et al.: "Development of the MPEG-H TV Audio System for ATSC 3.0," in IEEE Transactions on Broadcasting, vol. 63, no. 1, March 2017.

The Impact of Object-Based Audio on the Production of a Binaural Audio Play

Y. Grewe[1], J. Stuchlik[1], M. Ehrhard[1], M. Schill[1], M. Wagner[1]

[1] *Stuttgart Media University, Germany, Email: yg009@hdm-stuttgart.de*

Abstract

Object-based audio offers advantages over channel-based approaches such as format agnostic reproduction to improve the listening experience. An audio scene comprises of a number of objects each consisting of audio content and additional information which is affixed in the shape of metadata. This metadata is interpreted by a renderer, which creates the audio signals, according to the target reproduction system.
In an artistic context, recent work focused more on the use of audio objects played back through loudspeaker systems, while the influence of an object-based approach for binaural reproduction has been understudied.
This paper shows aspects of an object-based production, while using the example of an immersive, binaural audio play. In an object-based context, reconsiderations of the production process for binaural reproduction is required. Usually, audio plays consist of different elements such as dialog, music and effects, which were typically produced in multichannel-based formats. The results show, that correlated multichannel signals are not feasible for binaural reproduction due to coloration, localization and externalization issues. Furthermore, technical aspects as well as creative challenges are discussed and the developed production workflow is presented.

1. Introduction

Object-based audio (OBA) and its production and delivery has become a relevant topic for audio content producers in recent years. OBA differs from legacy channel-based approaches in several ways, such as the use of additional audio-scene describing metadata. OBA poses a number of challenges for the production workflow, such as object panning and monitoring through different classes of playback devices.

In contrast to OBA productions, channel-based binaural audio plays are popular in the audio industry. They can be produced by using several kinds of approaches such as a dummy head recording on location, re-recording of loudspeaker signals using a dummy head or the usage of rendering algorithms which are commonly based on signal convolution with pre-recorded Binaural Room Impulse Responses (BRIR) [1]. This results in challenges to an auditory composition for free-field like acoustical environment.

Usually, audio plays consist of different elements such as dialog, music and effects which are typically produced in channel-based formats. However, object-based approaches may have significant advantages over channel-based reproductions. Based on such considerations, a practical approach is combining channel-based signals with object-based sound sources, even for binaural reproductions. Combining channels and audio objects requires new tools during the production process, which are capable of panning or rendering. This paper investigates the workflow of an object-based audio production by using the example of a binaural audio play. During the production, technical approaches as well as creative workflows such as the feasibility of using traditional channel-based microphone setups such as OCT Surround [2] or Double MS [3] for playback over virtual loudspeakers, the combination of audio objects together with channel-based ambience recordings or binaural rendering in a free-field like acoustical environment were investigated.

Initially, this paper provides an overview of the production of an audio play using a combination of channel-based and object-based audio approaches. The remainder of this paper is structured as follows: chapter 2 provides technical basics on object-based audio in context of binaural rendering and differences to a channel-based approach. Chapter 3 provides the storyline of the audio play and shows the resulting challenges to the acoustical composition. Chapter 4 presents a preliminary informal listening test investigating the reproduction of combining channel-based ambience recordings with audio objects. Furthermore, the production workflow including recordings, music scoring and mixing is presented. Chapter 5 discusses the results of the production, based on the previous described findings. Finally, chapter 6 concludes this paper.

2. Audio Objects and Binaural Rendering

An object-based audio scene comprises a number of objects each consisting of semantic audio content and additional information which is appended in the shape of metadata. This metadata is interpreted by a renderer, which creates the audio signals according to the target reproduction system, such as loudspeaker configurations or binaural rendering using headphones.

The following chapters will detail some basic information about object rendering in context of binaural reproduction. Further specifications can be found in [4] or [5].

2.1. Using Audio Objects for Binaural Reproduction

Legacy channel-based binaural approaches which use convolution of pre-recorded BRIRs, lead to a fixed position

of loudspeaker information. These Impulse Responses (IRs) can be used for binaural reproduction [1]. Psychoacoustic principles such as phantom sources are generated to create a spatial impression.

Channel-based signals can be reproduced using this approach. In the context of this paper, this is called 'virtual loudspeakers.' Figure 1 shows a schematic illustration of sound source panning using virtual loudspeakers.

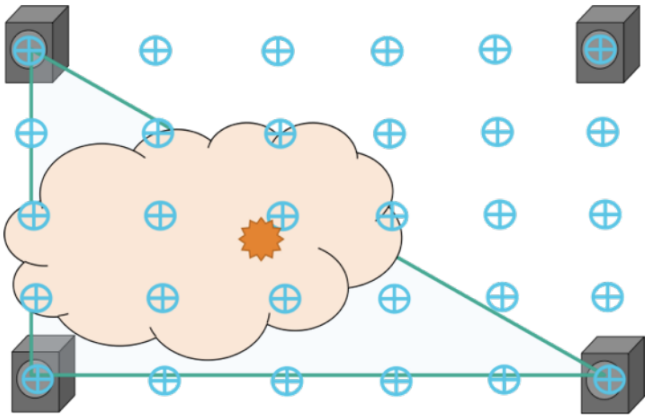


Fig. 1: Schematic illustration of sound source placement using virtual loudspeakers. It leads to less spatial resolution and blurred sound due to phantom sources [6].

In contrast, object-based audio metadata includes information such as loudness and most prominent, the position of each object in time. Thus, objects can be statically placed or dynamically panned to any position within a three-dimensional space, including positions below and above listeners ear level [7]. This allows an accurate positioning of sound sources which leads to higher localization accuracy compared to phantom source creation. Figure 2 shows a schematic illustration of object placement within a three-dimensional sound field.

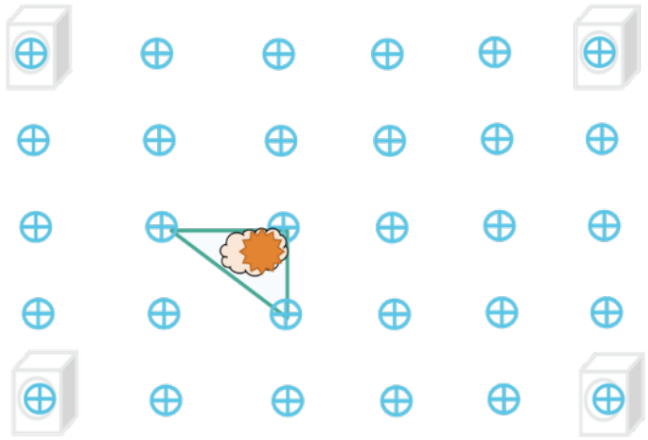


Fig. 2: Schematic illustration of sound source placement using object rendering. Higher spatial resolution than using virtual loudspeakers. Accurate localization due to metadata position and rendering [6].

In the context of this investigation, the audio objects metadata were described and transmitted using the MPEG-H 3D Audio standard [8] and later rendered using a binaural rendering algorithm which is able to interpret the metadata stream [8].

2.2. Rendering and Software Algorithm

Using OBA in combination with channels may lead to a powerful tool by combining the advantages of both techniques: already produced channel-based content for ambience and distant sound, which may lead to an enveloping base. Using additional audio objects improves the auditory spatial impression and localization accuracy by virtual loudspeaker independent placement.

For this investigation, a plugin solution was chosen which is able to render both approaches independently; channel-based signals using virtual loudspeakers as well as audio objects. In total, 15 channels or objects were able to be rendered simultaneously to regular stereo headphones. Filters were applied to equalize the headphones frequency response which were used during production. The use of additional head tracking [9] for the auditory scene was resigned due to artistic decisions.

3. Scope of the Project

3.1. Content of the Audio Play

The binaural audio play 'Meermenschen' is inspired by a true event. The story takes place after a shipwreck on the Mediterranean Sea. The surviving refugees Amila and Sahel drift on the open sea with dwindling hope. Sahel's baby alone makes them hope for a good outcome. The prehistory is represented by two flashbacks which tell the reason for the escape, as well as the actual ship accident.

3.2. Acoustical Environment

The audio play consists of three different acoustical environments. The main part of the audio play takes place in a free-field like situation on the open, calm sea. The first flashback scene takes place in an average living room while war noises can be heard. The second flashback scene takes place during a thunderstorm on a rough sea, whereby small parts even play under water.

Nevertheless, the free-field like scenes show major differences. Due to loud ambient noises caused by waves and wind during the shipwreck scene, it becomes difficult to precisely locate sound waves, due to masking effects.

4. Production Workflow and Preliminary Informal Listening Lest

The following chapter shows aspects of an OBA production that differ from ordinary stereo audio play productions. For speech recordings, a legacy spot microphone per actor has been chosen.

The goal was to be able to assign the speech signals as individual objects to make use of the investigated advantages of an object-based approach. Thus, the avoidance of crosstalk had to be observed even more precisely during the recording.

4.1. Preliminary Informal Listening Test

For sound design of ambience and effects, an initial concept was elaborated, using an ordinary production approach: mix a combination of multichannel recordings as well as foley sounds and effects to virtual loudspeakers according to their physical positions. In addition, specific auditory cues were intended to be assigned and panned as objects.

Furthermore, a preliminary informal listening test was conducted to prove the overall audio quality [10] of the concept. Fourteen attendees, all students and employees of the audiovisual media bachelor's and master's program of Stuttgart Media University, participated in the test and were asked to judge the overall audio quality without reference or anchor. In total, six parts of the audio play were selected, distinguished by their acoustical environment, such as a thunderstorm on a rough sea, free-field like ambience on calm sea or speech in an indoor environment. The verbal statements were summarized and led to the results, that this ordinary production approach is not efficient. The creation of a binaural ambience from multichannel recordings is challenging due to inter-channel correlation. This lead to worse spatial envelopment due to the creation of phantom sources and comb filtering [11]. However, a placement of audio objects in space achieved significantly better results. These results correspond to previous findings of [12] and [13].

4.2. Ambience & Effects

The auditory realization of the free-field scenes was one of the major challenges during production. Possible approaches were a reproduction of a realistic acoustical environment in contrast to an artistic design. The scenes on the open sea in 'Meermenschen' followed the latter approach and were more orientated towards the expectations of listeners in regard to modern media production. In order to design the ocean scenes according to their dynamics, the automation of volume and position was an important factor in the sound design process.

To be able to create the remaining acoustical scenarios (apartment and shipwreck scene) with audio objects as the main elements, the soundscapes had to be analyzed and built according to their specific acoustical requirements. The use of individual audio objects offers the advantage to place objects in a three-dimensional space, without altering the sound by using virtual loudspeakers. As stated in [14], the number of sound objects is decisive for immersion.

The flashback scenes (apartment and shipwreck) build a strong contrast to the ocean scenes in terms of immersion. The aim of the flashbacks was to fully utilize the three-dimensional panorama by using audio objects and their specific advantages in position and movement in time.

4.3. Music

For object-based binaural reproduction, the musical space must be completely rethought. The approach was therefore a scenic integration of music. The metaphorically chosen instruments were therefore evenly integrated into the scene in order to support the dramaturgical storyline. In addition, non-diegetic reverb was used in order to improve the auditory impression that music is a further level of meaning.

Audio objects that were carrying the musical content were either static or dynamically panned. During dynamical panning, the piano cluster moves and finally disappears around the listener and finally disappears into the virtual acoustic space with Amila's dwindling hope. Thus, the listener is surrounded by the instruments, which enhances the emotional impact and identification with the main character.

At the end of the audio play Amila sings a lullaby, which is played without accompaniment. Motifs from this song are hinted at throughout the entire score. As a spatial peculiarity, a very close playback position to the listener was chosen to transfer aspects of intimacy and security to the listener.

4.4. Mix

Audio objects were already arranged and panned during music composition and sound design process. Nevertheless, fine tuning in order to avoid leaps within the scene as well as across scenes was conducted.

Since binaural signals contain head-related transfer function (HRTF) [15] dependent frequency and level differences, these must be treated with greater caution regarding common mixing practices like equalization or levelling, in order to avoid changes in the spatial information.

However, the circumstance that HRTF-signals have to be treated with greater caution was also used to achieve a targeted amplification of the effect of three-dimensional panning. For example, strong filtering in the low and high frequency ranges was applied to wind sounds in order to achieve a more credible sound representation of distance.

Due to a limitation of 15 channels or audio objects per render instance, several parallel plugin instances had to be used during the composition and sound design process. An important task in the mix was therefore to integrate the individual modules with each other. It was noticeable that an overlapping of many elements caused a weaker localization sharpness. As a result, objects were completely removed or repositioned in the mix to ensure an immersive impression.

5. Results

In comparison to ordinary stereo productions, considerable differences were noticeable. The use of conventional multichannel recordings such as 5.1 surround over virtual loudspeakers led to highly audible filter artifacts during binaural rendering. A placement of audio objects in space achieved significantly better results.

The final result show, that the number of simultaneously rendered objects has a strong influence on the auditory spatial impression. A high number of simultaneously rendered objects lead to spatial artifacts due to inter-channel phase cancellations. Soundscapes can be created with less than ten sources. However, it also turned out that the border between too many and too few sound sources, which may lead to spatial artifacts, strongly depend on the reproduced signals.

Since binaural rendering relies on BRIR, a true-to-life representation of free-field like acoustical environments is

challenging. An approach lays in an artistic realization by treating parameters such as sound movement (e.g. used for ocean waves) and distance.

On a musical level, it could be shown that in addition to instrumentation and composition, the spatial distribution can increase the emotional impact of a musical score. OBA can therefore help to discover new content-related concepts, such as a stronger link between music and space.

Despite many advantages, general problems with binaural playback were noticeable. In addition to the problem of non-individualized HRTFs, it was noticeable that sound sources from the front are particularly poorly representable. This can be attributed to the lack of structure-borne sound, head tracking, and the absence of a visual equivalent that people expect from front sources [15]. Since there are already numerous approaches to these topics, there is also a broad field for future research in the audio play context.

6. Conclusion

The impact of object-based audio (OBA) on the production of a binaural audio play was presented. Combining audio objects with channel-based approaches over virtual loudspeakers lead to a higher spatial impression due to improved localization accuracy. An informal listening test proved these findings. However, it was figured out that the number of simultaneously rendered objects has a strong influence on the auditory spatial impression due to possible inter-channel phase cancellations on HRTF signals. These findings strongly depend on the reproduced signals.

Furthermore, the production workflow has to be adapted to a more holistic approach, since the boundaries of the individual production steps, especially sound design, composition and mix can no longer be clearly separated from each other.

7. Acknowledgment

The authors would like to thank Prof. O. Curdt and J. Bauer for their support during production, H. R. Price for his valuable feedback and B. Böröcz, M. Hepp, S. Kablan, W. Schorrig for their participation in the project. Furthermore, the authors would like to thank all partners and supporters who made this project possible.

8. References

[1] Stade, P., Ahrend, J.M., Pörschmann, C.: Perceptual Evaluation of Synthetic Early Binaural Room Impulse Responses Based on a Parametric Model. Proceedings of 142nd AES Convention (2017), Berlin

[2] Theile, G.: Multichannel Natural Music Recording based on Psychoacoustic Principles. Proceedings of 108th AES Convention (2000), Paris

[3] Wittek, H., Haut, C., Keinath, D.: Doppel-MS: eine Surround-Aufnahmetechnik unter der Lupe. Proceedings of 24th Tonmeistertagung (2006), Leipzig

[4] Rumsey, F.: Immersive Audio: Objects, Mixing, and Rendering. Journal of AES Volume 64, Issue 7/8 pp. 584-588, 2016

[5] Karapetyan, A., Fleischmann, F., Plogsties, J.: Elevation Control in Binaural Rendering. Proceedings of 140th AES Convention (2016), Paris

[6] Nordmann, J., Grewe, Y.: Immersive VR Audio Production - Capture, Mixing and Delivery. 26th Korea International Broadcast Audio & Lightning Equipment Show (2016), Seoul

[7] Grewe, Y., Simon, C., Scuda, U.: Producing Next Generation Audio Using the MPEG-H TV Audio System. Proceedings of Broadcast Engineering and Information Technology Conference (2018), Las Vegas

[8] ISO/IEC 23008-3: High efficiency coding and media delivery in heterogeneous environments - Part 3: 3D Audio, MPEG-H 3D Audio Phase 2. Geneva, 2016

[9] Laitinen, M.-V., et al: Influence of Resolution of Head Tracking in Synthesis of Binaural Audio. Proceedings of 132nd AES Convention (2012), Budapest

[10] Silzle, A.: 3D Quality Evaluation: Theory and Practice. Proceedings of 2nd ICSA (2014), Erlangen

[11] Theile, G., Wittek, H.: Principles in Surround Recordings with Height. Proceedings of 130th AES Convention (2011), London

[12] Scuda, U.: Immersive Sound Recording for Virtual Reality Documentary. Proceedings of 29th Tonmeistertagung (2016), Cologne

[13] Pike, C., Taylor, R., Parnell, T., Melchior, F.: Object-Based 3D Audio Production for Virtual Reality Using the Audio Definition Model. Proceedings of AES Conference on Audio for Virtual and Augmented Reality (2016), Los Angeles

[14] Görne, T.: Sound Design - Klang, Wahrnehmung, Emotion. Hanser, München, 2017

[15] Rumsey, F.: Spatial Audio. Focal Press, Oxford, 2001

Immersive Object-Based Mastering

S. Hestermann[1], M. Seideneck[1], C. Sladeczek[1]
[1] *Fraunhofer Institute for Digital Media Technology IDMT, Email: christoph.sladeczek@idmt.fraunhofer.de*

Abstract

To date, object-based spatial audio has been subject to much research, and a few production tools have entered the market. The main difference to the channel-based audio paradigm is that object-based spatial audio enables engineers to create three-dimensional audio scenes regardless of any particular loudspeaker setup. To playback audio scenes, a separate rendering process calculates all loudspeaker signals in real time, based on the meta data of audio objects in a scene. While the authoring process of an audio scene can be easily connected to conventional audio recording and mixing processes, possibilities for mastering audio scenes are still very limited. The main limitation stems from the aspect that compared to channel-based mastering, no loudspeaker channels can be altered. However, instead of always altering audio objects individually, the meta data part of audio objects can be used to provide engineers with convenient tools for object-based spatial mastering as well. Beyond solely mastering the audio data of a scene, new tools can also alter the meta data of audio objects on a global level for meta data mastering. These new object-based spatial mastering techniques are presented.

1. Introduction

Object-based audio has gained much interest in recent years. In contrast to channel-based audio, no speaker signals, but audio objects are saved as the result of a spatial audio production. An audio object is a virtual sound source, based on an audio signal with additional metadata, for instance position and gain. An audio rendering unit reproduces these audio objects by generating loudspeaker or headphone signals. The audio rendering takes into account the position of loudspeakers or the position of the listener in the virtual scene.

The traditional process of audio content creation can be divided into three stages: recording, mixing and mastering. While all three steps are well defined for channel-based audio, object-based audio requires new workflows. The recording step does not have to be changed in general, even though future recording techniques could bring new possibilities [1]. The mixing process only differs slightly, as the sound engineer no longer pans signals to dedicated loudspeakers. Instead, all positions of audio objects are usually set in a spatial authoring tool, which allows to define the metadata component of every audio object. A complete mastering process for audio objects has not been developed so far, although such would have benefits similar to those in channel-based workflows. In [2], first thoughts on global sound alterations of audio scenes were presented. In this paper, a flexible alternative is introduced which enables partial adjustments on arbitrary parts of an audio scene.

2. Mastering

Channel-based audio mastering improves the overall sound of complete mixes without altering single elements of the production individually [3]. In this sense, the fixed output channels of the channel-based audio paradigm are of advantage for the mastering process, despite their limitations regarding output formats.

Mastering is based on the antecedent audio production steps. At the beginning of the mixing process, all audio tracks can be altered individually. For each output format, every audio track is panned to a set of particular output channels. For stereo mixes, for instance, this results in a left and a right channel, whereas for 5.1 surround mixes, the result are six channels for front and rear speakers, and one channel for the subwoofer [4]. In this sense, an arbitrary number of source audio tracks is always routed to a fixed output channel set. Mastering then only alters these output channels. This allows for efficient adjustments of the overall sound.

Suboptimal acoustic environments often cause mixes with an uneven frequency response that do not translate well to other playback systems. In mastering, these mix issues can be eliminated for a consistent sound quality across different listening setups [3]. The tonal consistency of a mix is usually improved by a combination of compression and equalization. Furthermore, an artistic use of such tools can noticeably shape the character of a mix [4].

Another goal of mastering is raising the overall perceived loudness of mixes. Today, however, many distribution channels compensate for differences in loudness between tracks with normalization algorithms [3]. In the context of different listening scenarios, mastering engineers further need to ensure a satisfying sound quality across different audio codecs. Hence, during mastering a variety of formats and bitrates is monitored. To date, however, this applies mostly to stereo mastering, as surround formats are uniformly standardized [4].

3. Object-Based Audio

In channel-based audio, loudspeaker signals are saved as the result of a spatial audio mix. Since the loudspeaker feed contains all the spatial information of an acoustic scene, loudspeakers always need to be placed at the exact same position

ISBN
978-3-9812830-9-9
The authors hold the copyrights of all of content in this manuscript. They have granted the VDT the right to publish it in print, online and as PDF download on https://tonmeistertagung.com

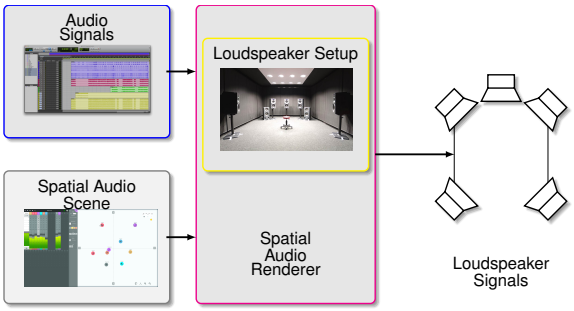


Fig. 1: Schematic of object-based audio reproduction. The rendering engine calculates all speaker signals in realtime by combining the input signals and metadata of the audio objects. The calculation is based on the properties of the given loudspeaker setup.

at which they were positioned during mixing. Only then the listener has the same intended spatial impression. Due to unsatisfactory automatic up- and downmix algorithms, in practice separate audio mixes are created for each loudspeaker configuration. This holds for the process of converting stereo to 5.1 surround sound, for instance, and for downmixing multichannel content to a reproduction system with a smaller set of channels. Since new multichannel formats introduce more loudspeaker channels, such as 22.2 with 24 loudspeakers, this conversion problem is worsening [5].

Object-based audio overcomes these channel limitations by following a different paradigm [6]. Audio objects are saved instead of loudspeaker channels. An audio object consists of an input signal (e.g. male vocals, a drum track) and associated metadata, like position, gain and further properties if applicable. In reproduction, a rendering engine calculates loudspeaker signals in real time using an object-based reproduction method, such as wave field synthesis [7]. This reproduction sequence is visualized in Figure 1. During production, a spatial authoring tool is used for generating the audio object metadata, as presented in [8]. The audio is produced in a digital audio workstation (DAW) or provided by a multichannel audio player. The finalized metadata and audio signals of a production can be packed in an object-based file format like MPEG-H [9]. As the reproduction of object-based audio content always requires rendering loudspeaker signals, the mastering process is completely undefined in this regard, since conventional mastering is based on modifying loudspeaker signals.

4. Object-Based Mastering

A potential object-based mastering process would have benefits similar to those in channel-based audio. Creators would no longer need to pay detailed attention to different distribution requirements in the mixing and authoring steps, even in the case that the spatial production is delivered in a channel-based format. A mastering process would also facilitate the essential aspect of conventional mastering, the tonal translation to other playback systems.

As of now, the entire spatial authoring and mixing process of object-based audio productions usually takes place in one location. This limits the quality of the spatial mix to the production environment. If the acoustic properties of the latter are not ideal, the spatial mix will not translate well to larger listening setups, such as cinemas, and may fall short of the listener's expectation.

Although it is already possible to send spatial projects to other production facilities for sound improvement, every audio object still needs to be treated individually, since no speaker channels can be altered. This is highly inefficient, as the arrangement of a large set of audio objects within a creative mixing process often results in complex audio scenes.

The presented concept for object-based spatial mastering combines the user requirement of overall adjustments with the technical aspect that every audio object needs to be altered individually. This limitation is a consequence of the rendering process, as described earlier. In order to make any kind of global adjustment possible, these have to be converted to individual changes on affected audio objects. [2] describes the basic sequence of how metadata can be used to derive object-specific parameters from global adjustments, which the proposed concept is based on.

One explicit challenge of Mastering Object-based audio scenes lies in the existence of both meta and audio data, as opposed to only audio material in channel-based mastering. Although tonal adjustments on the audio component of audio objects are likely to be of greater importance, tools for mastering metadata are necessary as well in order to optimize spatial scenes. The introduced concept is universally applicable to both metadata and audio signals.

Since different spatial audio renderers can deliver different results, a reference renderer should be defined for the object-based mastering process.

4.1. Mastering Objects

The suggested realization of overall adjustments are Mastering Objects. Mastering Objects enable the user to alter multiple audio objects, up to the entire scene, at once. Adjustments done via Mastering Objects can involve both the audio and metadata component of an audio object. Just as usual audio objects, Mastering Objects can be positioned anywhere in a scene and freely in real time.

On an abstract level, Mastering Objects can be regarded as containers that address an arbitrary number of audio objects and hold audio and/or metadata processing parameters for these. Whenever the user applies any type of processing to the Mastering Object, an algorithm derives individual processing parameters for the affected audio objects from the overall user settings. The user may switch between these algorithms.

4.1.1. Direct Connection

The direct connection approach is similar to grouping. A Mastering Object M is directly connected to audio objects S_i. By changing properties of the Mastering Object, all connected audio objects will be modified. This is visualized in Figure 2. Here, the Mastering Object M is connected to the sources S_1, S_3 and S_5. Modifying M will therefore have no effect on the other sources of the scene, in this case S_2, S_4. The position of the Mastering Object does not have any influence

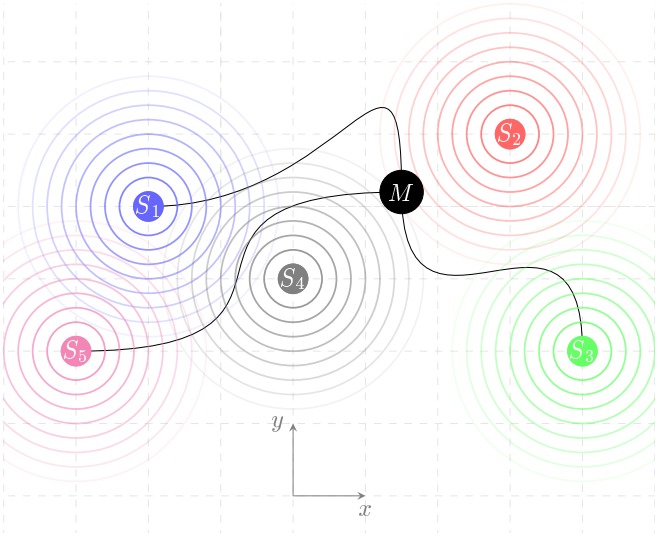


Fig. 2: Direct connection approach. A Mastering Object M is directly connected to multiple audio objects S_i. Modification of the properties of M will only influence connected audio objects.

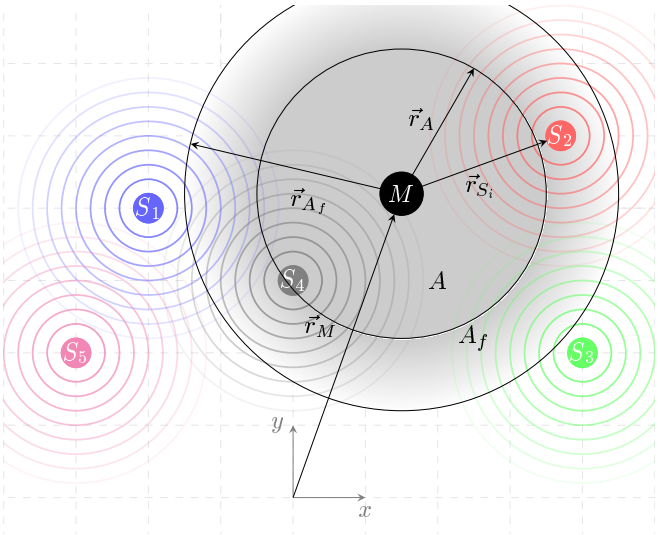


Fig. 3: Area of interest approach. The Mastering Object M has an ambit A and a fade area A_f. If an audio object is within A or A_f, it will be affected by parameter changes of M.

on parameter changes.

4.1.2. Area of Interest

Figure 3 depicts the area of interest approach. The user defines an area A and a fade area A_f around the Mastering Object M by choosing the radii $|\vec{r}_A|$ and $|\vec{r}_{A_f}|$. Depending on the position $\vec{r}_M$ of the Mastering Object M, a weighting factor $p_i(\vec{r}_M)$ for the audio objects S_i can be calculated as follows:

$$p_i(\vec{r}_M) = \begin{cases} 1, & S_i \in N_A \\ \frac{|r_{A_f}| - |r_{S_i}|}{|r_{A_f}| - |r_A|}, & S_i \in N_{A_f} \\ 0, & else. \end{cases} \quad (1)$$

where N_A and N_{A_f} is a set of audio objects within the areas A or A_f.

4.1.3. Object-Interspace

Another possible method in the context of object-based mastering is an interaction approach using the object interspace.

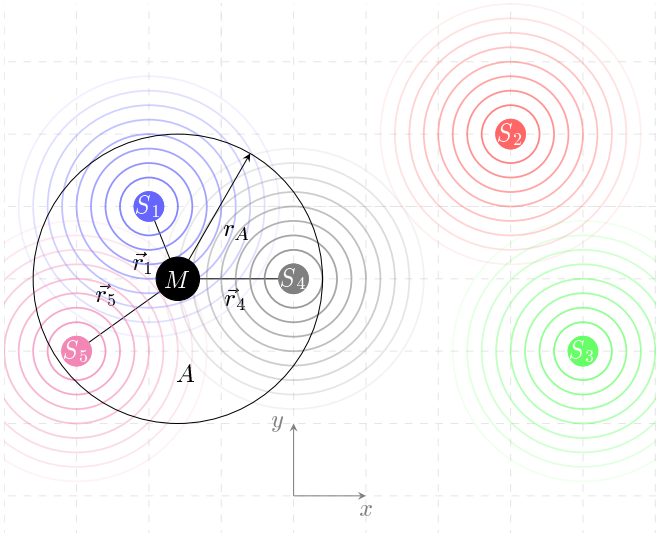


Fig. 4: Distance model. Depending on the radius $|\vec{r}_A|$ around the Mastering Object M audio objects will be selected. According to the distances $|\vec{r}_i|$ between Mastering Object and audio objects, the influence of parameter settings of M on the audio object will be determined.

This principle is visualized in Figure 4. The audio engineer places the Mastering Object M somewhere in the spatial audio scene. Then, the radius $|\vec{r}_A|$ around M can be modified so that all objects which should be influenced are inside area A. Finally, the user can modify the position of M within area A which leads to different distances $|\vec{r}_i|$ to the audio objects S_i. The degree to which the Mastering Object will influence the audio object can now be determined by the calculation of ratios according to the distances $|\vec{r}_i|$.

4.1.4. Angle of Interest

Humans can localize sound sources in a certain direction much more accurately than giving a precise statement about their depth [10]. From a user perspective, the angle between the Mastering Object and audio objects might sometimes represent their listening impression more accurately. In [2], a global control of any audio plugin parameter via the azimuth of audio objects is presented. This approach can be adopted by calculating the difference in angle θ_i between the Mastering Object and audio objects S_i in its ambit. This is depicted in Figure 5. Only audio objects inside area A will be influenced, which depends on the position of Mastering Object M and the spread defined by angle θ_M.

4.1.5. Angle Interspace

In conjunction with the object-interspace approach, the same method can be applied by using the angles between the Mastering Object and the audio objects. This is also visualized in Figure 5.

4.1.6. Scene Transformation

While the meta data part of audio objects can be used to calculate audio effect parameters, the meta data itself can also be altered to achieve global changes to the audio scene. So far, audio objects can only be moved individually and not in conjunction with adjacent objects. Instead of individual position changes, in mastering a global scene transformation

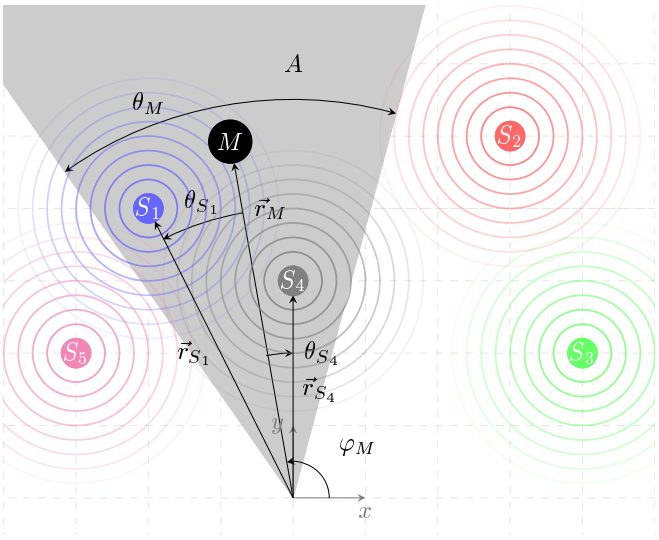


Fig. 5: Relative angle approach. Depending on the angle θ_{S_i} between an audio objects S_i and the Mastering Object M a weighting factor will be determined. Only audio objects within area A will be modified.

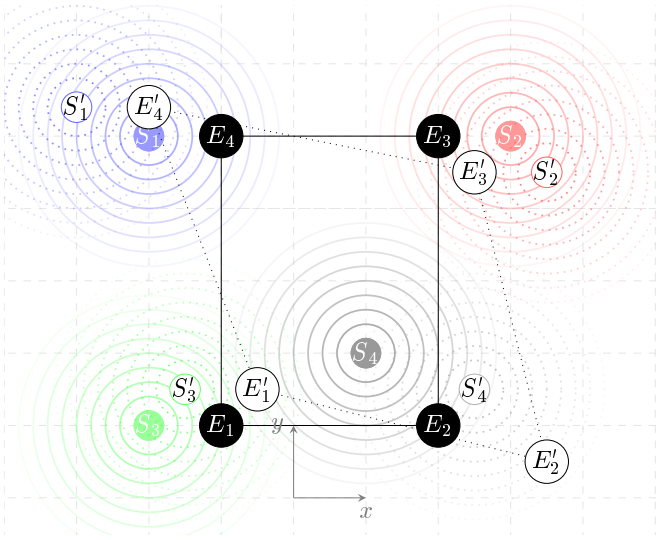


Fig. 6: Scene transformation scenario: Audio scene with a scene transformation canvas and audio objects S_1 to S_4. The user can drag each of the canvas corners C_1 to C_4 independently to new positions C'_1 to C'_1. This shifts all audio objects in the scene to new positions S'_1 to S'_4.

tool can be useful to conveniently transform the position of subsets of audio objects or the entire scene.

Figure 6 shows a scene with four audio objects S_1 to S_4. A distortion canvas with user-draggable corners is shown on top of the scene. By dragging the corners C_1 to C_4 of the canvas to new positions C'_1 to C'_4, a transformation matrix is calculated. The position vectors of the audio objects are then multiplied with the transformation matrix. This results in new audio object positions S'_1 to S'_4.

While the canvas overlay may control the position of all audio objects in a scene, this functionality can be attached to Mastering Objects as well. Then, the transformation canvas would appear on top of a Mastering Object and only audio objects in the scope of the Mastering Object would be moved. This variant is realized in the prototype which is introduced in 5.

From a mathematical perspective, a two-dimensional transformation of a scene in the horizontal plane can be realized using homography [11]. A transformation matrix H maps each audio object at position S_i to its new position S'_i:

$$H := \begin{pmatrix} h1 & h2 & h3 \\ h4 & h5 & h6 \\ h7 & h8 & h9 \end{pmatrix}, \; S'_i = HS_i. \tag{2}$$

The transformation matrix H is calculated using the coordinate changes of the canvas corners. The two-dimensional coordinates are the following:

$$C_i := \begin{bmatrix} x_i \\ y_i \end{bmatrix}, \; C'_i := \begin{bmatrix} x'_i \\ y'_i \end{bmatrix}. \tag{3}$$

This can be used to set up a linear system of equations for the coefficients of the transformation matrix H [11]:

$$\begin{pmatrix} x_1 & y_1 & 1 & 0 & 0 & 0 & -x'_1x_1 & -x'_1y_1 \\ 0 & 0 & 0 & x_1 & y_1 & 1 & -y'_1x_1 & -y'_1y_1 \\ x_2 & y_2 & 1 & 0 & 0 & 0 & -x'_2x_2 & -x'_2y_2 \\ 0 & 0 & 0 & x_2 & y_2 & 1 & -y'_2x_2 & -y'_2y_2 \\ x_3 & y_3 & 1 & 0 & 0 & 0 & -x'_3x_3 & -x'_3y_3 \\ 0 & 0 & 0 & x_3 & y_3 & 1 & -y'_3x_3 & -y'_3y_3 \\ x_4 & y_4 & 1 & 0 & 0 & 0 & -x'_4x_4 & -x'_4y_4 \\ 0 & 0 & 0 & x_4 & y_4 & 1 & -y'_4x_4 & -y'_4y_4 \end{pmatrix} * \begin{bmatrix} h1 \\ h2 \\ h3 \\ h4 \\ h5 \\ h6 \\ h7 \\ h8 \end{bmatrix} = \begin{bmatrix} x'_1 \\ y'_1 \\ x'_2 \\ y'_2 \\ x'_3 \\ y'_3 \\ x'_4 \\ y'_4 \end{bmatrix} \tag{4}$$

Once the coefficients of the transformation matrix H are calculated, this matrix can be applied to all audio objects of the scene or those within the scope of a Mastering Object. A new transformation matrix only needs to be calculated whenever the user distorts the transformation canvas again.

Depending on the context of an audio scene, it may be useful to prevent certain objects from being moved other than manually. Therefor a user-accessible option can be implemented which marks an audio object as locked. Then, the algorithm that applies the transformation to audio objects can skip locked audio objects which prevents them from being re-positioned.

5. Prototype Implementation

A prototype was implemented to evaluate the concept of object-based mastering. An existing system infrastructure was used as the development basis.

5.1. SpatialSound Wave

The prototypical implementation uses the *SpatialSound Wave (SSW)* technology developed by Fraunhofer IDMT [8]. SpatialSound Wave consists of two main components: a real time audio renderer and a web server. The audio renderer generates speaker signals based on speaker positions, audio signals and the metadata of all audio objects in the audio scene. The server provides web applications for the audio renderer configuration. This includes the production of audio

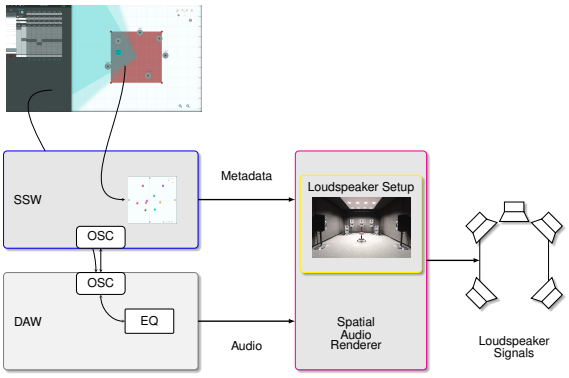


Fig. 7: Schematic of the system architecture used for the realization of the prototype implementation.

scenes by interacting with the contained audio objects, as well as the transfer of information to the rendering unit. The web server features a modular design for fast extensions and provides interfaces for communication with external devices, such as RS232 and OSC.

5.2. Mastering Unit

SpatialSound Wave's extensibility mentioned in 5.1 is taken advantage of in the prototypical implementation. A mastering unit is added to the system. The mastering unit provides and manages Mastering Objects and their influence on audio objects. As introduced in 4, this includes, but is not limited to, object selection, parameter calculation, and execution. The parameters can control both the metadata and the audio signal of audio objects. The internal function library of the SpatialSound Wave Server already provides control over the metadata, such as position and volume of an audio object. Changes to the audio signals, for example via equalization or compression, on the other hand, are part of the player; in this case a Digital Audio Workstation (DAW). Hence, a dedicated interface for communication is needed to access these functionalities inside the DAW.

Therefore, REAPER was used in the prototypical implementation, since it supports the OSC interface for data exchange [12]. The DSP parameters generated by the mastering unit are transmitted to REAPER via OSC and are then applied to the audio channels. The system architecture is shown in Figure 7.

5.3. Graphical User Interface

In addition to the backend, a web-based application was developed as the frontend of the prototypical implementation. Figure 8 shows the graphical user interface of the prototype. The application consists of the Mastering Object configuration parameters and the visualized audio scene. In this example, four Mastering Objects influencing different areas of the audio scene have been created. The turquoise objects with the wave symbol represent the freely movable Mastering Objects in the audio scene. The grey Mastering Objects have been disabled by the user, i.e. their audio and metadata changes are bypassed. On the left-hand side of the screen, the equalizer parameters of the currently selected Mastering Object are displayed. The selection is indicated by a thin, bright turquoise line around the object.

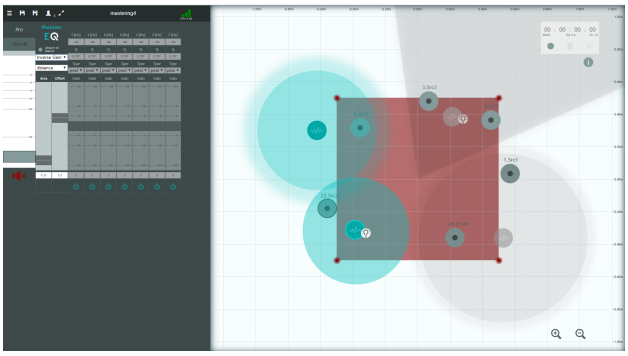


Fig. 8: Graphical user interface of the prototypical implementation.

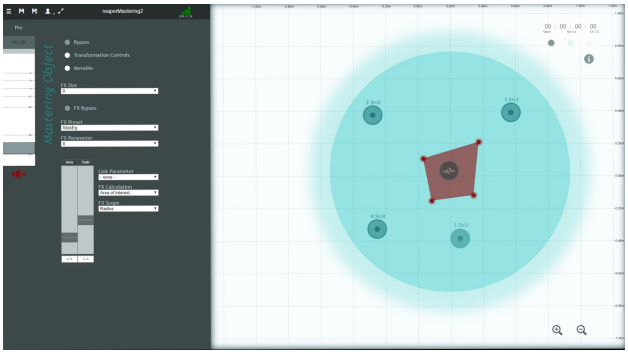


Fig. 9: Second prototype with general plugin controls and a scene transformation canvas on top of a Mastering Object.

The turquoise, transparent homogeneous areas around the left two Mastering Objects use the Area of Interest and Object Interspace approach respectively. Around the homogeneous circular area, the transparent shading indicates the area in which all parameters except for the gain parameters are taken over unchanged from the sources. The gain parameters of the equalizer, on the other hand, are weighted according to the distance between the sources and the Mastering Object. The "Area" slider controls the size of the radius of the circle around the Mastering Object. The "Feather" slider controls the size of the radius of the surrounding transition area. The upper right Mastering Object uses the Angle Interspace approach.

Figure 9 shows an updated graphical user interface with general plugin parameter controls. The control section on the left-hand side of the screen now enables the user to control any plugin inside REAPER for mastering. The user can specify which parameters are weighted by the interaction approach and which remain constant. It is further possible to link multiple plugins to the Mastering Object. Via the Transformation Controls checkbox, the user can toggle the scene transformation canvas on top of a Mastering Object, as introduced in 4.1.6. In this particular case, only one Mastering Object has been placed in the audio scene. The positions of the four audio objects surrounding the Mastering Object were distorted using the scene transformation canvas.

6. Conclusion

This paper presented ideas for object-based spatial audio mastering. In channel-based workflows, the mastering process is

the last production step where loudspeaker signals are altered. In the context of object-based audio production, mastering was completely undefined. Nonetheless, it is likely that sound engineers would also benefit from tools to optimize object-based mixes for optimal sound quality. This paper showed different interaction approaches for mastering using *Mastering Objects*. These were realized in a prototype to evaluate the concept as part of an interactive system.

Under https://youtu.be/cqC1rWSxT4c a video is available which summarizes the object-based mastering approach (last visit: 11/30/2018).

7. References

[1] Christoph Sladeczek, Annika Neidhardt, Martina Böhme, Markus Seeber, and Alejandro Gasull Ruis, "An Approach for Fast and Intuitive Monitoring of Microphone Signals Using a Virtual Listener," in *International Conference on Spatial Audio ICSA*, Erlangen, Germany, Februar 2014.

[2] Frank Melchior, Uwe Michaelis, and Robert Steffens, "Spatial mastering - a new concept for spatial sound design in object-based audio scenes," *Proceedings of the International Computer Music Conference 2011, University of Huddersfield, UK*, August 2011.

[3] Bob Katz and Robert A. Katz, *Mastering Audio: The Art and the Science*, Butterworth-Heinemann, Newton, MA, USA, 2003.

[4] B. Owsinski, *The Mastering Engineer's Handbook – The Audio Mastering Handbook*, Thomson Course Technology, Boston, MA, USA, 2008.

[5] Kimio Hamasaki, Koichiro Hiyama, and Reiko Okumura, "The 22.2 multichannel sound system and its application," in *Audio Engineering Society Convention 118*, May 2005.

[6] Alejandro Gasull Ruiz, Christoph Sladeczek, and Thomas Sporer, "A description of an object-based audio workflow for media productions," in *Audio Engineering Society Conference: 57th International Conference: The Future of Audio Entertainment Technology, Cinema, Television and the Internet*, Mar 2015.

[7] Edwin Nico Gerard Verheijen, *Sound Reproduction by Wave Field Synthesis*, Ph.D. thesis, Technische Universiteit Delft, January 1998.

[8] Gabriel Gatzsche and Christoph Sladeczek, "A flexible system architecture for collaborative sound engineering in object-based audio environments," in *Audio Engineering Society Convention 136*, Apr 2014.

[9] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "Mpeg-h 3d audio – the new standard for coding of immersive spatial audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 770–779, Aug 2015.

[10] J. Blauert, *Spatial hearing – The psychophysics of human sound localization*, The MIT Press, 1999.

[11] Elan Dubrofsky, "Homography estimation," M.S. thesis, University of British Columbia, March 2009.

[12] Andrew Schmeder, Adrian Freed, and David Wessel, "Best practices for open sound control," in *Linux Audio Conference 2010*, May 2010.

Ethernet AVB as a means to distribute and control sound for Immersive Audio applications

R. Foss, S. Devonport

Computer Science Department, Rhodes University, Grahamstown, South Africa, Email:r.foss@ru.ac.za

Abstract

This paper describes how Ethernet AVB is used to distribute sound from Digital Audio Workstation software to Ethernet AVB-capable endpoints. The AVDECC control protocol, an open standard, is used to control a signal processing capability on the endpoints. A particular use for this capability is in the field of immersive sound. The endpoints can be speakers connected to POE+ switches, thereby enabling the transmission of audio streams, control and power. This eases the configuration of immersive sound systems. The clocking and stream reservation capabilities of Ethernet AVB ensure that all devices work from the same clock, and that stream transmission is deterministic.

1. Introduction

There are a number of workstation-based approaches to generating multi-channel spatial sound. These vary from spatial audio plugins that are incorporated into Digital Audio Workstations (DAWs), to modular toolkits that have spatial audio capabilities added to them. In all cases these approaches depend on hardware audio interfaces and their associated workstation drivers to provide access to multiple audio output transmission channels. Application Programming Interfaces (APIs) enable applications to present audio output configurations to users.

With the advent of 'immersive sound', there has been a demand for increased speaker numbers, and correspondingly more audio output channels to feed these speakers. Audio interface manufacturers have responded in two ways:

- Providing more audio outputs on their interfaces.
- Providing network capabilities on their interfaces to enable multiple networked audio interfaces to enhance the channel count.

In this paper we present networking approaches to the provision of multiple audio output channels for sound spatialization. In particular, we focus on Ethernet AVB networking approaches. The paper will:

- Introduce Ethernet AVB technology
- Describe an approach to multichannel access via Ethernet AVB interfaces with multiple audio outputs
- Introduce an approach to multichannel access via speakers that incorporate Ethernet AVB functionality
- Provide a comparative assessment of the above two Ethernet AVB audio interface technologies.
- Describe an approach that offloads the mixing required for immersive audio to Ethernet AVB interfaces or speakers

2. Ethernet AVB overview

Ethernet AVB (Audio Video Bridging) is a collective name for a set of IEEE networking standards that determine a networked media transmission solution [1]. They specify an approach to real-time media transmission over an Ethernet network that ensures appropriate Quality of Service (QoS) delivery.

With regard to audio, an Ethernet AVB network will comprise one or more controllers, one or more endpoint devices, and one or more switches. The endpoint devices will be transmitters of audio (Talkers) and receivers of audio (Listeners). A typical Ethernet AVB network is shown in **Fig.1** below.

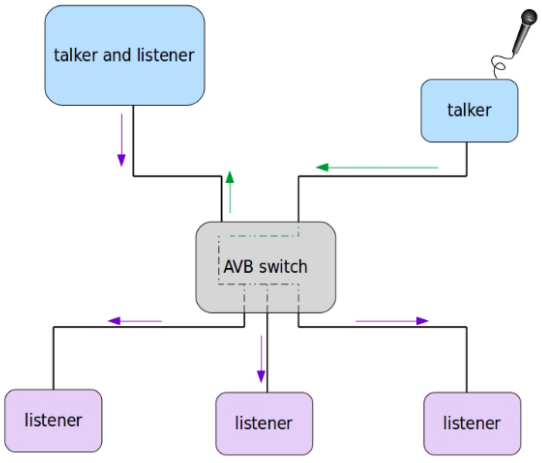


Fig. 1: A typical Ethernet AVB network

In **Fig. 1**, purple arrows correlate to audio streamed directly to Listeners attached to speaker outputs. The green routing path correlates to audio sent from one AVB device to another before being streamed to those same Listeners.

An endpoint device may have both Talker and Listener capability. Talkers will transmit their audio in the form of one

ISBN
978-3-9812830-9-9

The authors hold the copyrights of all of content in this manuscript. They have granted the VDT the right to publish it in print, online and as PDF download on *https://tonmeistertagung.com*

or more multichannel 'streams'. Listeners will have the capability of receiving one or more of these multichannel streams. Talker streams can be directed to Listeners via connection management commands. These commands are part of a comprehensive discovery, connection management and control protocol, known as AVDECC (Audio Video Discovery, Enumeration, Connection and Control) [2].

A Controller device, typically a workstation, will discover device endpoints, enumerate their streams, enable a user to make connections and control parameters. The connection process involves a process of bandwidth reservation for the Talker stream, from the Listener to the Talker. Along the path to the Talker, there will be one or more Ethernet AVB capable switches. These switches ensure that, in routing streams from input ports to output ports, there is not overloading and that low latency transmission is ensured. To do this, the switches cooperate in the bandwidth reservation process. These stream connections may be controlled via an AVDECC connection matrix as shown in **Fig.2**.

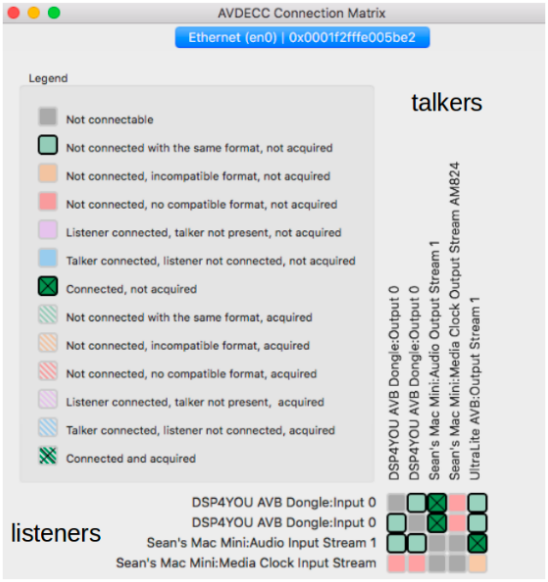


Fig. 2 An AVB connection matrix.

The connection matrix in Figure 2 is natively implemented within an Apple Mac. Columns denote device Talker outputs. Rows denote device Listener inputs. Active connections are shown as dark green with an 'x' and will generate appropriate AVDECC connection commands for devices being connected. The connection matrix shows 4 devices streaming audio to one another. Here we have an Apple Mac virtual AVB entity acting as both a Talker and Listener. The other devices are specifically Talkers or Listeners. Their connections are listed below.

1. TheAppleMac's virtual AVB Talker streams a single audio stream to 2 DSP4YOU network DACs.

2. The Ultralite AVB interface streams to the Apple Mac's virtual AVB Listener.

All device clocks on an Ethernet AVB network are synchronized. The Precision Time Protocol is one of the Ethernet AVB standard protocols. It incorporates a number of message types that ensure that a single 'grandmaster' clock is selected to be the source of the network wide clock, and that all device clocks are synchronized to this clock.

With regard to workstation controllers, Apple Macs with Thunderbolt transmission capability have Ethernet AVB protocols built in to the OSX operating system. The Apple Mac provides for the creation of a virtual entity with Talkers and Listeners and a controller that enables stream connections to be made by the user. There are two Ethernet AVB solutions for Windows machines, both requiring the insertion of an Ethernet AVB capable network card into a slot of a desktop machine. Echo Audio provide their 'Streamware' solution that depends on their own Echo Streamware card [3]. The Streamware software provides a connection matrix for Talker/Listener stream connections. AudioScience provide a Virtual Sound Card (VSC) that runs above an Intel I210-T1 Ethernet card [4]. The VSC provides a virtual entity with Talkers and Listeners. The Hono controller enables the connection of Talker streams to Listeners on the network. A third option, the RME Digiface AVB is currently on pre-order, and allows for connection to a USB 3 port, thereby enabling PC laptop control. In the following sections, we shall assume that the controller is an Apple Mac workstation. Our comparative study of channel accessibility will not be affected by workstation choice.

3. Multi-channel access via AVB audio interfaces

As an example, MOTU manufacture a number of Ethernet AVB capable audio interfaces [5]. These have varying numbers of analogue outputs ranging from 8 to 24. Correspondingly, they have a varying number of Ethernet AVB channels that they can receive as inputs. These channels are grouped into 8-channel streams.

More than one of these interfaces can be connected on an AVB Network via an AVB compliant switch (MOTU manufacture an AVB switch), thereby providing up to 64 channels of audio to an immersive sound application. In tests at Rhodes University, three of these MOTU interfaces have been used to create a 24 channel immersive sound configuration. These are shown connected to a MOTU switch in **Fig. 3** below.

Fig. 3. 3 x MOTU devices and a MOTU switch.

To accommodate the 8 channel stream capability of the MOTUs, the Apple Mac virtual entity needs to be configured to transmit and receive 8 channel streams. This is accomplished via the AVDECC Entity controller application that is one of the Apple OSX AVB utilities, and is shown in **Fig. 4** below:

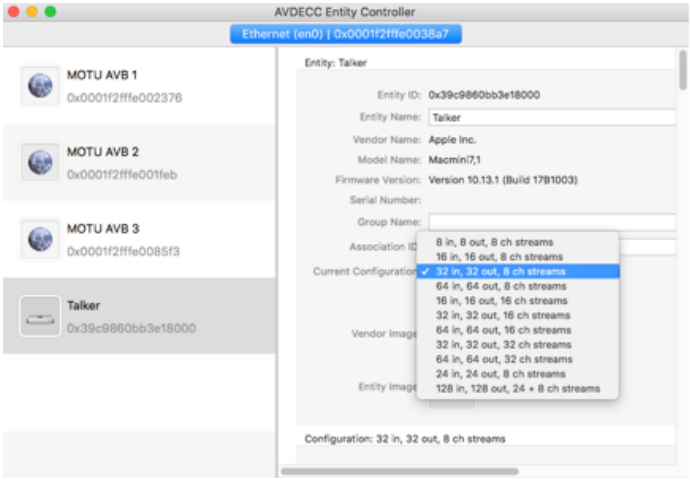


Fig. 4: The configuration of streams

Here a configuration has been selected that allows for 4 streams, with each stream containing 8 channels of audio. Configurations with 8 – 32 channels of audio can be chosen.

Stream and channel connections must be made to each of the MOTU interfaces. These can be made using a MOTU web application. A browser is used to retrieve the application by providing a URL that is the identifying IP address of the device. In the application a Talker stream must first be selected (this is an Apple virtual entity Talker stream). This selection is shown in **Fig. 5** below.

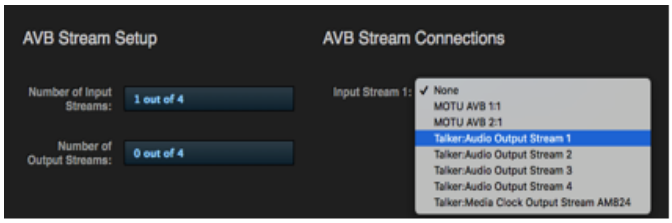


Figure 5. The selection of a Talker stream

Following this, a graphical routing matrix must be used to route the channels within the chosen AVB stream to the analogue audio outputs for connection to amplifiers or active speakers. A channel from the AVB stream could be routed to any audio output, however in our case we want to direct channel *n* to analogue audio output *n*. We are simply using the network and audio interfaces as a conduit for channels from our application to speakers. This routing is shown in **Fig. 6** below.

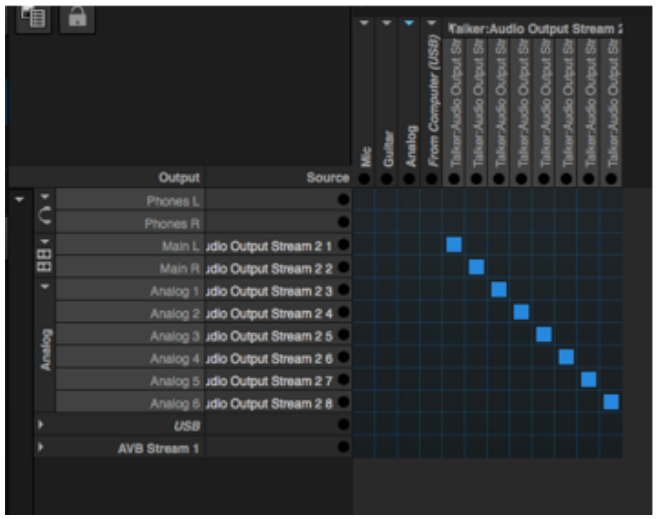


Fig. 6: Routing AVB stream channels

The above procedure of application access, stream allocation and channel routing must be performed for each of the MOTU devices. It is important that Talker stream 1 be selected for the MOTU device connected to the first 8 speakers, Talker stream 2 for the MOTU device connected to the second 8 speakers, and so on. This will lead to a configuration shown in an abstract way in **Fig. 7** below.

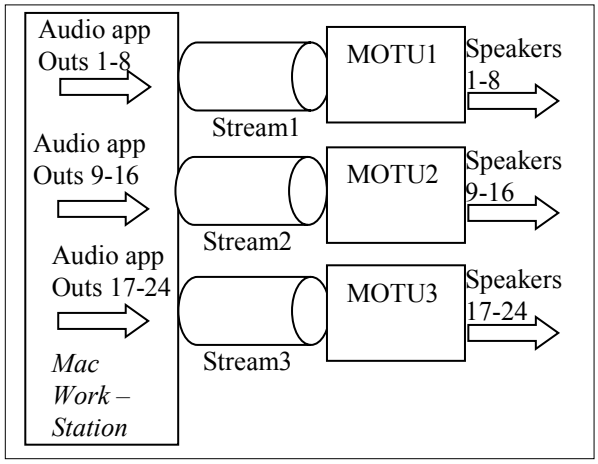


Fig. 7 Multichannel access via MOTU devices

MOTU devices with 16 and 24 analogue outputs will have 'banks' to which 8 channel streams can be allocated. An advantage here is that many more channels can be routed in a single routing matrix.

In order to connect the streams, AVDECC ACMP (connection management) commands are used. However channel selection is performed via HTTP messages.

4. Multi-channel access using AVB speakers

miniDSP manufactures an audio interface (NDAC-8) that receives an 8 channel Ethernet AVB stream and has 8 analogue outputs [6]. miniDSP also manufactures an active speaker (SPK-4P) with AVB capability and class D amplification that can be configured to receive a stream containing 8-32 channels. There is a further speaker output on the active speaker that connects to and provides amplification to a passive speaker [6]. The active speaker can be powered over Ethernet and provides about 30 watts of power in total. Both device types incorporate matrix mixers that enable the routing of AVB channels to their analogue outputs, with AVDECC-based level control over the matrix cross points.

AVBFlex is a program developed at Rhodes University, in partnership with miniDSP, which works with the above miniDSP devices. It enables the routing of audio output channels from a workstation audio application to multiple speakers. It uses AVDECC connection commands to connect the Apple AVB virtual entity Talker streams to the devices, and uses AVDECC mixer control commands to extract channels to the speakers. As with the MOTU devices, the Apple workstation's virtual entity needs to be configured to create Talker and Listener streams with the appropriate number of channels – see **Fig. 4**.

In order to make connections from the Apple virtual entity Talker streams to the NDAC-8 or SPK-4P Listener devices, AVBFlex needs to have access to the Ethernet addresses of the devices as well as know how many speakers are attached to each device. This is provided in the form of an XML speaker configuration file. Some example entries in such a file are shown in Listing 1 below:

```
<config>
  <amplifier macadr = "70:B3:D5:12:F3:B4">
    <speaker number = "1" x="0" y="0" z="0">
    </speaker>
    <speaker number = "2" x="0" y="0" z="0">
    </speaker>
  </amplifier>
  <amplifier macadr = "70:B3:D5:12:F3:AF">
    <speaker number = "3" x="0" y="0" z="0">
    </speaker>
    <speaker number = "4" x="0" y="0" z="0">
    </speaker>
  </amplifier>
.
.
</config
```

Listing 1. AVBFlex speaker configuration file

Each amplifier element represents a device, has an Ethernet (Mac) address attribute and contains one or more speaker elements. In Listing 1 there are just two speaker elements for each amplifier because the devices are SPK-4Ps. Each speaker element has a speaker number attribute and three-dimensional coordinate attributes, which are not used by AVBFlex, but are required for programs that perform rendering.

The user interface of AVBFlex allows for the generation of this file. Each device has a label indicating its Ethernet address. A user will read the Ethernet address of each device and enter it into a table, together with the number of speakers attached to the device. The ordering of devices and their speakers will determine their audio channel allocation within the workstation audio application. This speaker configuration table entry for 7 devices is shown in **Fig. 8** below.

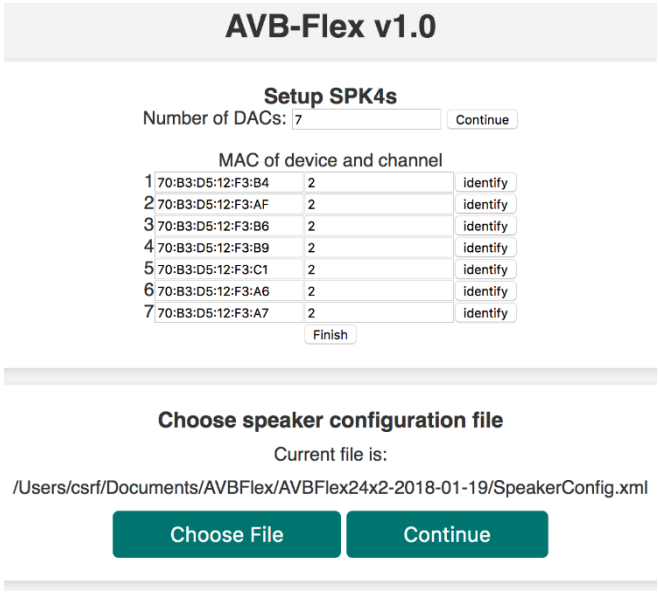


Fig. 8. AVBFlex device setup

A speaker configuration file will be generated as a result of this configuration process, and this file can be selected when this same configuration is used again. The setup procedure only has to be performed once for a particular speaker configuration.

AVBFlex uses the speaker configuration file to determine what Talker/Listener stream connections to make via AVDECC connection commands, and also what channels to extract from the streams via AVDECC mixer control commands.

After running AVBFlex, a user will:

- Confirm the listed configuration file or choose a different one.
- Check that the number of devices discovered by AVBFlex matches the number in the configuration file
- Press 'connect' and have AVBFlex go ahead and make the stream connections and channel selections.

The interface is shown in **Fig. 9** below for a 7-device speaker configuration.

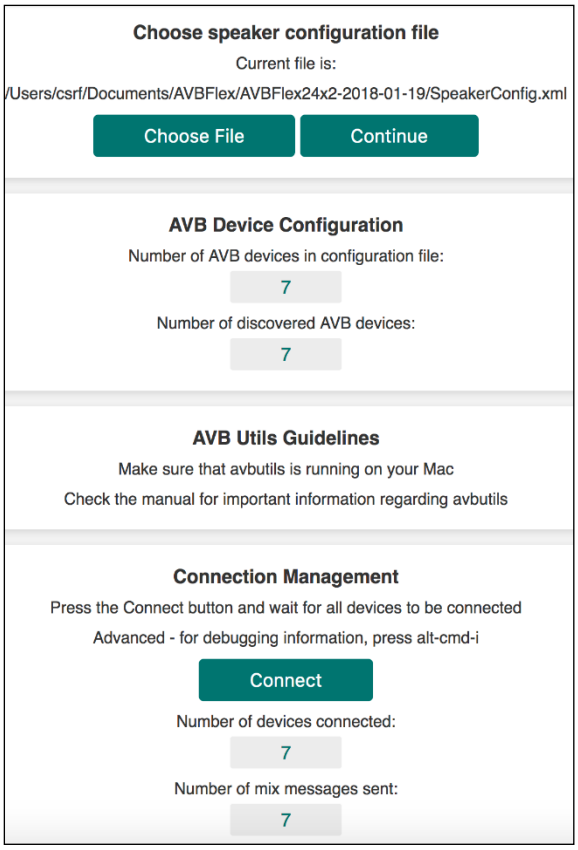


Fig. 9: AVBFlex device connection interface

There are two scenarios that AVBFlex needs to account for:

1. When the stream channel number exceeds the number of speakers attached to a device in the configuration. In this case the stream must be sent to a second device for it to extract the remaining channels.

2. When the total number of speakers in the configuration exceeds the number of channels in the Talker streams. In this case more than one stream must be used to carry the additional channels.

Fig. 10 below displays these two scenarios using SPK-4P devices. An assumption in Figure 10 is that the streams from the Apple Mac Talker each transmit 24 channels of audio.

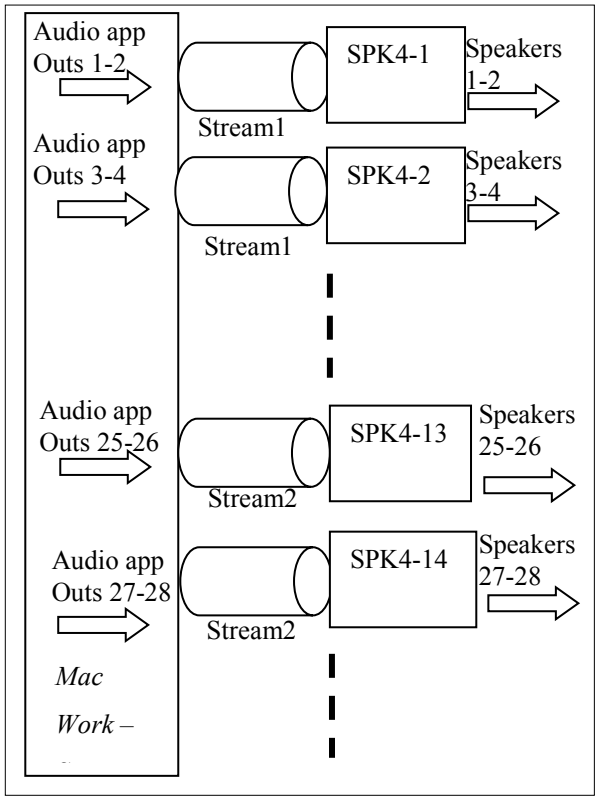


Fig. 10 Multi-channel access via miniDSP SPK4s

As can be seen in **Fig. 10**, the same stream can be sent to more than one speaker. The channel extraction process within the matrix mixer of the first SPK4 is shown in **Fig. 11** below.

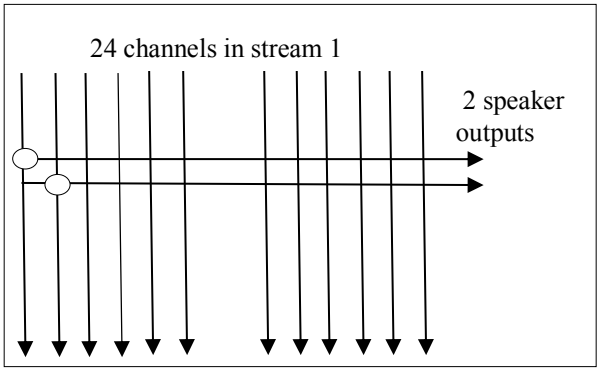


Fig. 11: Two channel extraction within SPK4-1

Fig. 12 below shows the channel extraction process for SPK4-14.

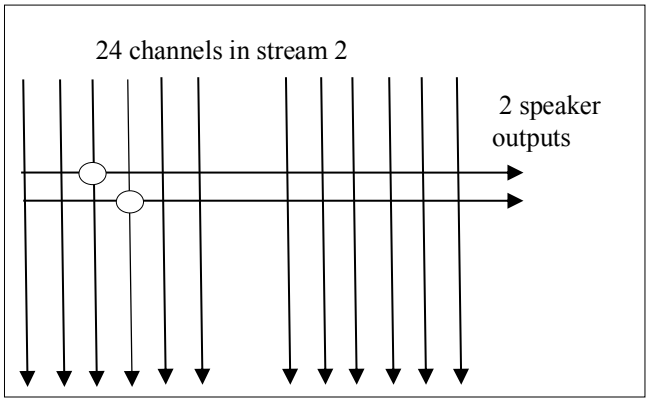


Figure 12. Two channel extraction within SPK4-14

AVBFlex uses AVDECC connection management commands to make the stream connections shown in Figure 10, and then uses AVDECC mixer control commands to extract the appropriate channels as shown in **Fig. 11** and **Fig. 12**. AVBFlex determines the appropriate stream connection and mixer control commands from the speaker configuration file, but only transmits commands to the devices that it determines to be present on the network. Device presence is determined via AVDECC discovery messages coming from the devices.

Fig. 13 below shows a configuration of 96 SPK-4Ps being set up for wavefield synthesis experimentation.

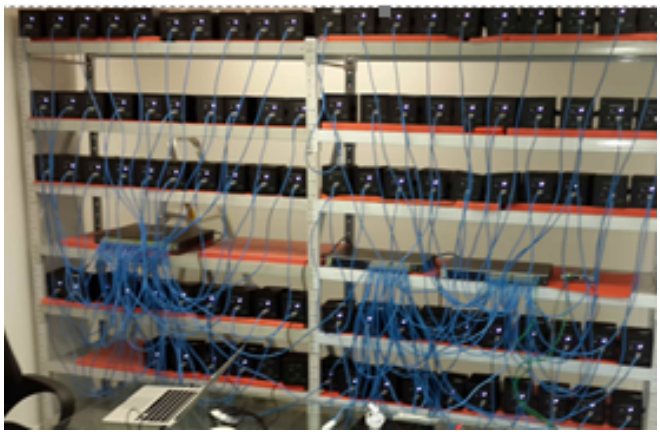


Fig. 13: AVBFlex control over 96 SPK-4Ps

In the diagram a MacBook Pro is running AVBFlex and three Ethernet AVB switches with Power over Ethernet capability have each of their ports connected to an SPK-4P.

If AVBFlex is used with the miniDSP NDAC-8 devices, then the process of stream connection and matrix control is much the same as for SPK-4Ps from a user perspective. Each NDAC-8 device can have 8 speaker outputs, and this will be reflected in the speaker configuration file. As with the MOTU devices, the NDAC-8s will only receive 8 channel streams. The matrix that enables channel extraction from the incoming stream to a Listener NDAC-8 device is an 8 x 8 matrix and will extract up to 8 channels from the stream.

5. Comparison of approaches to channel provision for immersive audio

With regard to the two approaches to using Ethernet AVB networking, they both provide the necessary output channels to an audio spatialization program. Both require the identification of devices, IP address in the case of MOTU devices and Ethernet address in the case of AVBFlex. Once the respective configurations are constructed, the necessary connection management and channel selection will happen automatically.

The two configurations are adapted to different contexts, the multi-output audio interface configuration to a more studio oriented context, and the speaker based configuration to an installation context. Once configured the output stream channels are accessible to a standard Digital Audio Workstation or to a more modular and experimental platform such as Max/MSP [7]. **Fig. 14** displays the routing matrix for the Reaper DAW [8], each track being assigned to its own output channel. Each channel would in turn be routed to the appropriate audio output of an AVB audio interface or Speaker via the AVB network. Spatial audio plugins can be used to process the channels appropriately, given that they are aware of the speaker positioning of the speakers allocated to the channels. An appropriate spatial audio renderer will be used to construct the channel output, knowing the speaker position and also the position of various virtual audio sources that need to be spatialized.

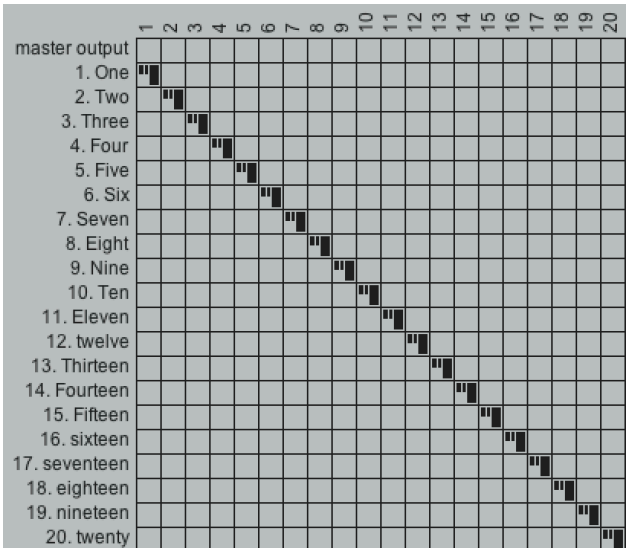


Fig. 14: Reaper track to AVB channel routings

In Max/MSP a DAC module would provide the interface to which processed channels can connect. **Fig. 15** below shows a configuration at Rhodes University, where multiple NDAC8s were used with Wave Field Synthesis (WFS) [9] modules, each module containing 8 tweeters and 2 woofers.

Fig. 16 then displays the WFS channels feeding the DAC module. In this project a Spat viewer [10] was used to position 6 audio sources.

Fig. 15 WFS modules, NDAC-8s and amplifiers

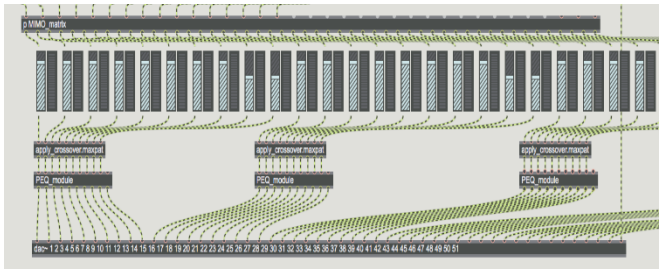


Fig. 16: WFS channels fed to DAC module

6. Beyond simple connection capability

All of the endpoint devices, described in the preceding sections, have processing capability. The sources of audio, whether they be DAWs or Max/MSP, often have quite substantial processing requirements, and the processing resources have to be shared with the immersive audio processing requirements. It makes sense then, to distribute some of the immersive audio processing to the endpoints. This demands a quite different routing of streams to endpoints. Each endpoint must be in a position to review all the audio channels transmitted by the audio source, and determine how it will process each of them. This determination can be made by the workstation and transmitted to the endpoint.

In this context, it is important for a workstation to be able to determine all the endpoints on the AVB network, what the processing capabilities of the endpoints are, what speakers could possibly be connected to the endpoints, and what their positions are in three-dimensional space. AVDECC provides a discovery protocol. Each endpoint will send out discovery messages at regular intervals, indicating its availability. Furthermore, each endpoint incorporates a number of descriptors that describe its connection capabilities and its processing capabilities. These capabilities can be enumerated by a workstation.

At Rhodes University, we have developed a client/server program called immerGo [11] that utilizes the miniDSP N-DAC8 and SPK-4P endpoints. This serves as an example of the utilization of:

1. AVDECC device discovery
2. AVDECC device control
3. Utilization of endpoint processing capability

A typical immerGo configuration is given in **Fig. 17** below.

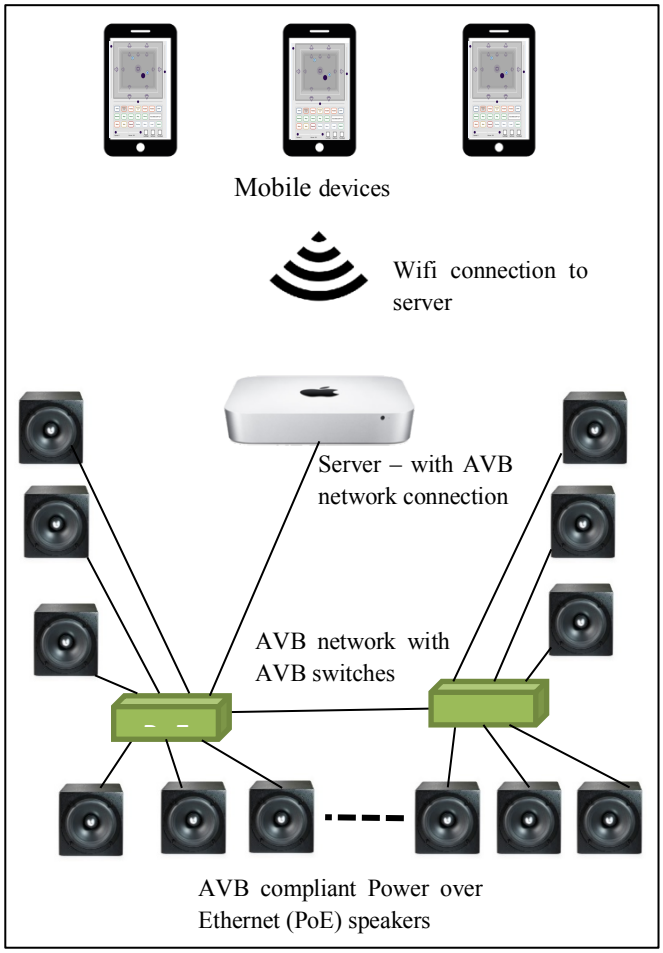


Fig. 17: An immerGo system

User control is via mobile devices, which send 3D coordinates for audio channels to the server. The server uses rendering algorithms to determine mix values to be applied to each channel for each speaker, and sends these mix values to the endpoint speakers using the AVDECC SET_MIXER command. The endpoint speakers each implement a mixer matrix to mix channels to speakers according to requests from the server.

The audio source, typically a DAW, resides on the server and is remotely controllable from the mobile devices. A user will successively select a track on a mobile device, which corresponds to a DAW track, start the DAW and localize the track. These localizations can be saved. The user interface is shown in **Fig. 18** below.

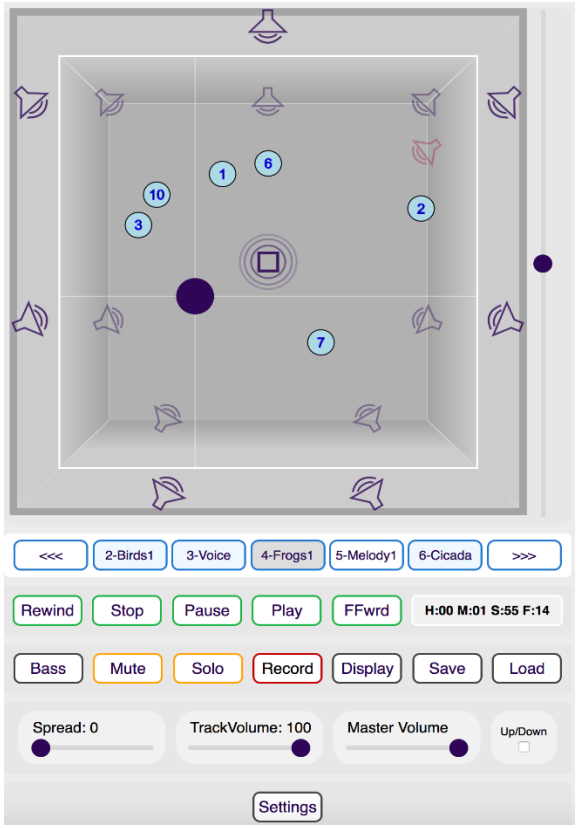


Fig. 19: The immerGo mobile user interface

From a user perspective, the system needs to be responsive to real time localization changes, as well as to stored localization changes. The immerGo server receives MIDI Time Code messages for the synchronized playback of stored localizations. Localizations are updated every quarter frame – at 30 frames per second, this resolves to an update approximately every 8 milliseconds. For a SET_MIXER AVDECC command, a response is received from the endpoint within 200 microseconds.

Apart from the miniDSP devices, ImmerGo has been implemented using AVB endpoint devices with appropriate mixer matrix capabilities, as well as delay and equalization capabilities. However, the control protocol has been either OSC or HTTP. HTTP control has been typically in the order of 100 milliseconds for each message, while OSC messages carried within UDP packets (User Datagram Protocol), have been faster – approximately 8 milliseconds.

For responsive immersive sound applications, it is desirable that the AVDECC protocol be implemented on the endpoints to control the signal processing capabilities within the endpoints, and beyond this for these capabilities to be accessible via appropriate descriptors.

In immerGo, the discovery facilities of AVDECC are used to determine the presence of endpoints, and to use this information for the configuration of speakers. The speaker configuration user interface is shown in **Fig. 20** below.

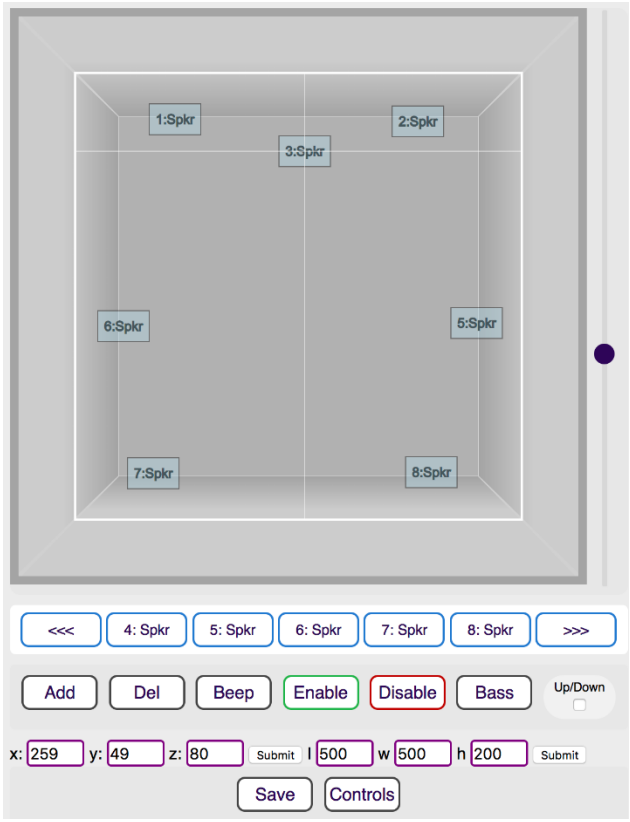


Fig. 20: The immerGo speaker configuration interface

7. Conclusion

Ethernet AVB has a feature set that makes it a good candidate to carry multiple channels of audio for spatial audio applications:

- It is an open standard.
- It provides low latency transmission – a maximum of 2 milliseconds latency.
- There is deterministic audio transmission enabled by stream reservation and flow control mechanisms.
- The clocks of all devices on the network are synchronized to a chosen grandmaster clock.

In particular, all these features are automatically engaged when a network is configured with switches, endpoint devices and controllers. There is no need to configure switches. Above the substrate that ensures transmission, there is a standards-based connection management and control protocol that enables Talker to Listener stream connections, and a range of device controls.

This paper has described how these features have been used in two quite different applications, but which both provide the multiple channel capability required by spatial audio applications. These are the MOTU AVB router and AVBFlex.

The paper has promoted the employment of endpoint processing capabilities to ease the processing resource usage

on workstations. As an example of this, it has described the immerGo spatial audio system, which uses the matrix mixer capabilities of AVB endpoints. In this context, it is important that the control protocol used provide comprehensive discovery and enumeration capabilities, and also that it enables near real time response. The AVDECC protocol has been found to provide these capabilities and response time.

Lastly, the AVNU Alliance is an organization that ensures interoperability between devices that implement Ethernet AVB in their products [12]. The University of New Hampshire InterOperability Laboratory provides an AVNU certification testing process that manufacturers can use to obtain AVNU certification [13].

8. References

[1] J. Evaleens, "Ethernet AVB overview and status", *presented at SMPTE 2014,* October 2014.

[2] IEEE, "IEEE Standard for Device Discovery, Connection Management, and Control Protocol for IEEE 1722 Based Devices", *Document Standard, IEEE Std. 1722.1*, 2013.

[3] Echo Audio, "The Streamware NIC-1", *https://echoavb.com/products/streamware-nic-1*

[4] AudioScience, "The Hono AVB Virtual Sound Card", *http://www.audioscience.com/internet/products/avb/hono_avb_vsc.htm*

[5] MOTU, "Using your MOTU AVB Device as a Mac audio interface over AVB Ethernet", *http://motu.com/avb/using-your-motu-avb-device-as-a-mac-audio-interface-over-avb-ethernet/*

[6] miniDSP, Products for Immersive audio, *https://www.immersivedsp.com/collections/all*

[7] Cycling '74, "Max/MSP", *https://cycling74.com/products/max/*

[8] Cockos Inc. "The Reaper DAW", *https://www.reaper.fm/*

[9] A. Berkhout, D. De Vries, P. Vogel, "Acoustic Control by Wave Field Synthesis", *Journal of the Acoustic Society of America*, vol. 93, May 1993, pp. 2764–2778

[10] T. Carpentier, M. Noisternig, O. Warusfel "Twenty years of Ircam Spat: looking back, looking forward.", *Proceedings of the 141st International Computer Music Conference, Denton, TX, USA, pp270-277, September 2015.*

[11] Foss R., Devonport S. "An Immersive Audio Control System Using Mobile Devices and Ethernet AVB-Capable Speakers", *JAES Volume 66, issue 9, pp. 724-733, September 2018.*

[12] AVNU, "The AVNU Alliance", http://avnu.org/

[13] University of New Hampshire, "AVNU certification testing services", https://www.iol.unh.edu/testing/switching/avnu

AES Reviewed Paper at Tonmeistertagung 2018

Presented* by VDT in cooperation with the

Central European Region of the Audio Engineering Society (AES).

On Human Perceptual Bandwidth and Slow Listening

T. Lund[1], A. Mäkivirta[2]

[1] *Genelec OY, 74100 Iisalmi, Finland, Email: thomas.lund@genelec.com*

[2] *Genelec OY, 74100 Iisalmi, Finland, Email: aki.makivirta@genelec.com*

Abstract

Seventy-four recent physiological and psychological studies revolving around human perception and its bandwidth were reviewed. The brain has ever only learned about the world through our five primary senses. With them, we receive a fraction of the information actually available, while we perceive far less still. A fraction of a fraction: The perceptual bandwidth. Conscious perception is furthermore influenced by long-term experience and learning, to an extent that it might be more accurately understood and studied as primarily a reach-out phenomenon. Considering hearing, time is found to be a determining factor on several planes. It is discussed how such sentient observations could be taken into account in pro audio, for instance when conducting subjective tests; and the term "slow listening" is devised.

1. Introduction

The paper presents a broad review of selected articles and studies that have a bearing on auditory perceptual bandwidth in humans, particularly in the framework of professional audio applications, including listening tests.

In pro audio, double blind ABX testing is regarded as best practice. ITU-R BS.1534, ITU-R BS.1116 and similar standard procedures are generally used as the foundation for extensive statistical analysis, and often able to provide repeatable results. BS.1116 assumes that subjects have the capability to hear all phenomena tested for fully developed prior to participation. However, both guidelines also suggest the use of audio stimuli of 10-25 s duration [1, 2].

User interfaces where quick recognition could be a question of life and death address the limited conscious bandwidth of humans; a horn in a car, flight deck controls (visual design, phasic alerts, haptic alerts), medical equipment etc. A smartphone does every trick in the book to get noticed, a magician knows how not to.

Regardless, subjective audio testing currently puts high faith in that seemingly scarce resource, conscious bandwidth.

Thanks to new non-invasive in vivo experimental techniques, we are getting a better understanding of how *time* can affect sensing. Furthermore, explicit consciousness, the feature considered a hallmark of humans, has lost in importance over recent decades, if physiological and psychological findings have been interpreted correctly [3].

Basing subjective tests solely on explicit conscious and immediate responding may not provide a good enough understanding of long-term effects of a salient experience.

1.1. Interior and exterior

Even thousands of years before antiquity, our forefathers were aware of sensing, its importance and its limitations. The 32.000-year-old painting of an owl in Grotte Chauvet, for instance, recognize ears and eyes as main attributes, see Fig. 1 left. In contrast, we don't really know how music of the great composers sounded just 300 years ago.

Socrates, Plato and later philosophers, with idealism, reason and eventually scientific method, helped us get a firm stand on previous generations' shoulders, and collectively break one step away from the cave. However, every child born today

*: Please note that the AES Reviewed Papers at Tonmeistertagung can be published by both, AES and VDT, in print, online and as PDF download.

ISBN 978-3-9812830-9-9

still spends many years to learn using her or his senses, the same way a long gone baby in Ardèche had to.

Fig. 1: More than 30.000 year old paintings from Grotte Chauvet.

Having noticed how unreliable perception was, René Descartes famously took refuge only in his mind, and declared consciousness to be anchored in our awareness of our own awareness, *je pense, donc je suis*. Enlightenment idealist George Berkeley took the opposite stance, that things only existed by being perceived, an idea that reverberated even into quantum physics, and from there to hypotheses on consciousness and brain topology today [4, 5].

19th century experimental psychologists Ernst Weber and Gustav Fechner discovered our senses to be logarithmic and developed the concept Just Noticeable Difference (JND), still widely relied on in a variety of sensory studies.

However, around 1850, Hermann von Helmholtz pioneered systematic studies on physics and perception. Based on research, he came to the radical conclusion that consciousness does not have access to all data and intermediate results that produce sensation [6]. He called the phenomena "unconscious inference" (in German "unbewusster Schluss"), but he also studied other aspects of involuntary entry into consciousness. Helmholtz's findings were refuted for decades, but they are now at the very heart of recent studies and models, detailed later.

1.2. Individual and collective

In our era, where extensive data about almost any highly specific topic is instantly available, the combination of discoveries from different fields is sometimes what matters more. A modern methodology and encouragement of inter-disciplinary problem-solving is shown in Fig. 2, the quadrant model of Integral or Integrative Perspectivism as proposed by philosopher Ken Wilber [7]. The concept is to consider a question or problem from all quadrants.

Besides from being a methodology under further development at Danish and German universities [8, 9], Integrative Perspectivism has proven a valuable tool for stimulating discussion in international standardisation involving stakeholders from different fields; for instance, engineers, physicians, psychologists, artists, economists and politicians.

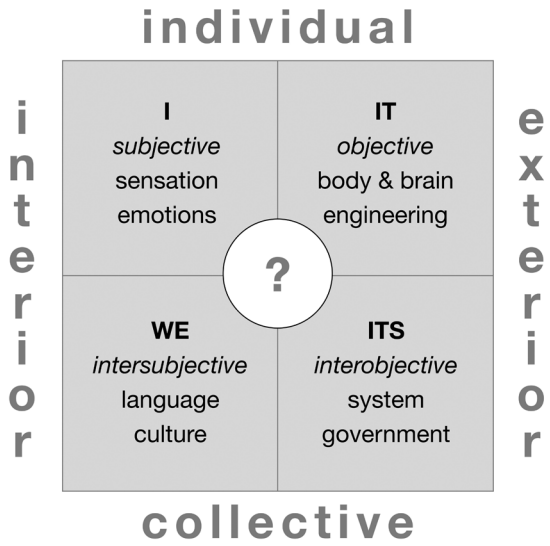


Fig. 2: The quadrant model. Left side subjective, right objective.

Economic utilitarian societies tend to appreciate the exterior (objective) right-hand quadrants more, so the model is also a reminder of more profound human interests, spanning millennia rather than stock cycles.

2. Nomenclature and method

Our sensory system is bombarded with information, leading to a data deluge that cannot be processed in real time. The role of attention is to select information to meet current behavioural goals. Consciousness does not select information but summarizes it, potentially on a longer timescale.

Conscious is used in a narrow, sentient definition: The ability to discriminate, categorize and (to some extent) recall external stimuli that fall within our reception range.

Attention means selective attention. It is the way we can focus on certain aspects of sensory stimuli, mono or multi-modal. The two are therefore entirely different.

The hard problem of consciousness, sentience, the feeling of perception, is important when recognising human art from the past, but outside the scope of this review.

2.1. Reception and perception

A plethora of receptors generate *afferent* nerve impulses about the status of the body, as well as our close and more distant surroundings. Via nerve fibres and a number of synapses, impulses can be relayed to the central nervous system (CNS), allowing the brain to participate in the reception, and occasionally to produce a sensory experience [10].

Reception obviously has a steep funnel associated: We are at a given physical location in space and time, and reception is tuned for conditions that generally matter on planet Earth, for a creature our size and composition, having a certain life-span etc. Our reception apparatus registers only a certain mechanical wave frequency range (hearing, touch), a certain electromagnetic frequency range (seeing, touch), certain smell/taste dimensions etc.

*: Please note that the AES Reviewed Papers at Tonmeistertagung can be published by both, AES and VDT, in print, online and as PDF download.

Considering hearing, the brain indeed is an active participant, not only in the decoding of minute temporal information, but also as the main element of a sense relying heavily on internal tuning. Hearing therefore also features a substantial number of *efferent* nerve fibres. Such fibres constantly send information back to the middle and inner ears, adjusting the reception system itself.

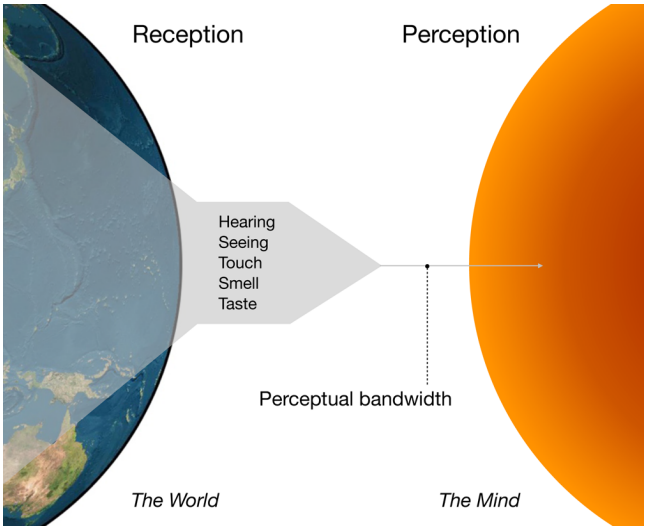


Fig. 3: The two funnels of human apprehension, this review focusing on the second, i.e. perceptual bandwidth.

Perception is distinguished from reception and introduces a second funnel between the exterior world and consciousness, see Fig. 3. Perception is entirely subjective. It is the outcome of sentient brain processing based on experience, expectations, mood, attention and - to some extent - reception.

In this review, *exteroception* is used about stimuli (appearing to originate) from outside the body, and *proprioception* about our sense of movement and relative position of body parts. Because *interoception* in psychology is currently taking on a broader meaning than sensing of internal body status (hunger, pain, body ownership etc.), we shall refrain from further using that term [11].

Perceptual bandwidth is the rate by which we can (consciously at the moment) register sensory stimuli, external or internal.

2.2. Inclusion criteria

A more than 50-year-old finding was inspiration for this review: An estimation of how quickly humans can acquire new information from the exterior. We have searched for results confirming or rejecting Karl Küpfmüller's observation of a modest upper limit for data "input-velocity" (Nachrichtenfluss) [12], and for new findings revolving around the topic. The precise number of bits/s was considered less relevant than verification of the pronounced sensory speed bottleneck itself.

New in-vivo human research and diagnostics tools have been introduced or refined in recent decades, potentially tipping discussions about the perceptual bandwidth to a certain side, or discovering entirely new connections and insights: Non-invasive measurements of various brain event-related potentials (ERP), auditory brainstem responses (AEP) and nuclear magnetic resonance imaging (MRI), especially its temporal version, functional MRI (fMRI); a technique based on hemodynamic response. Neurons require glucose when firing, and the brain has no local energy storage.

Understanding of human exteroception has improved dramatically from early studies based on PET scans 25 years ago, to fMRI today, with better temporal resolution and typically around 1x1x1 mm spatially (voxel 256x256). Animal trials have also been employed, for instance when human experiments were considered unethical.

This review is based on articles and papers listed in PubMed, Embase or PsycINFO, searched first for an AES Convention paper [13] using MeSH terms and keywords between April and June 2017. January to June 2018, a second search was performed, because several trials had unexpectedly reported a pronounced reach-out element to exteroception, leading to updated inclusion criteria. More than 350 articles were found to meet the (new) inclusion criteria. A secondary sorting process was carried out based on title and abstract, limiting the number of studies to 103, thereby adding risk of false negatives and bias.

A tertiary sorting was performed for a condensed presentation of results compatible with this TMT conference paper format, and the number of reported studies limited to 74. Newer studies were prioritized over older. Independent searches between the authors were not systematically performed. Only publications available in English, German, French or a Nordic language were considered for inclusion.

3. Results

Reviews were divided into six categories about perceptual bandwidth, latency of sensation, short duration sound, auditory stress, active sensing and unconscious inference, including long-term learning effects.

3.1. Perceptual bandwidth

Based on physiological evidence, our brain, mostly via cranial nerves, receives sensory data from millions of neurons when we are awake. However, according to Karl Küpfmüller, perceptual bandwidth is limited to 40-50 bits/s in total, and vision and hearing are the only senses that are even close to saturating that bottleneck; while touch, smell and taste can contribute 5 bits/s or less [12].

According to [14], "numerous experimental results indicate that the bandwidth of human perception is severely limited." Consciousness is fairly sparse, and more limited than our introspection would suggest [15].

Several studies also find that during audio-visual stimulation, when participants attend to *just one* modality, the sensory cortical areas serving the attended modality show increased activity and the ignored sensory cortices show a decrease in activity, compared with when attention is directed to the other modality [16, 17, 18, 19]. However, in *divided* attention experiments, audio and visual cortices are not depressed, though performance does decrease. Competition for

*: Please note that the AES Reviewed Papers at Tonmeistertagung can be published by both, AES and VDT, in print, online and as PDF download.

bandwidth appears to happen elsewhere in the brain, possibly where resources are shared, but without the mechanisms being fully understood [19].

Perceptual decoupling (PD), i.e. to disengage attention from perception, is another example where fMRI has provided new insight by charting the meandering between sensory experience and self-generated thoughts uncoupled from perception. PD is shown to happen frequently in daily life, and even called "pervasive" among specialists such as pilots, thus modulating perceptual bandwidth in unforeseeable ways [20]. A high cognitive load can reduce perceptual bandwidth further. Inattentional deafness in aviation was observed in 1/3 of high load cases where subjects overheard alarms they were instructed to report [21].

Finally, studies suggest there could be no fundamental difference between normal perception and hallucinations or pathological conditions, besides from a grading of how much exteroception is used to anchor an experience at all [22].

3.2. Latency of sensation

When physical events happen within our reception range, we cannot experience it instantly. Besides from a delay in the reception system, nerve potentials travel at a finite velocity, and each synapse with its chemical conduction takes time too. The 12 pairs of cranial nerves are categorized as part of the central nervous system (CNS). They can be regarded afferent and efferent highways, for example carrying sensory inputs to the brainstem without introducing extra, time-consuming synapses.

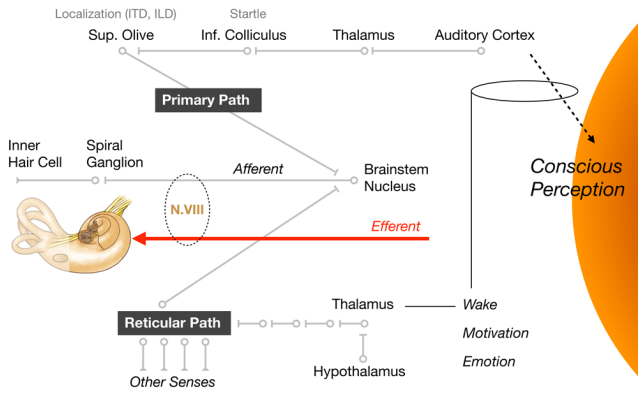


Fig. 4: Auditory nerve (N. VIII), primary and reticular pathways.

Once sensory information arrives at the brainstem, it follows a primary and a reticular path, the latter being used for prioritization between the senses when allocating bandwidth via our perception switchboard, the midbrain's thalamus. Considering the primary auditory pathway, this is where localization and several reflexes origin ahead of conscious awareness, see Fig. 4.

Our conscious mind without attention, however, is associated with a latency of 400 ms where a backwards referral system is hypothesized to make us believe events happen now, in order for a conscious mental image to synchronize with our motor actions [23, 24, 25, 26], see Fig. 5.

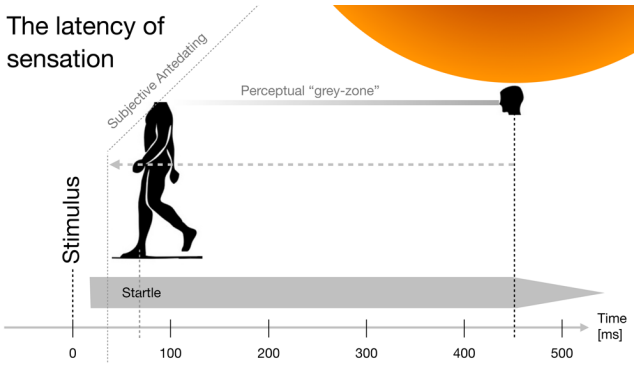


Fig. 5: Latency of sensation, reflexes and perceptual grey-zone.

More recently, in an ERP study based on vision, it has been found that latency depends on whether access is driven by stimulus saliency alone or by a combination of expectations and sensory evidence. Considering the latter, latency of specific neural signatures decreased from 300 ms to 200 ms [27]. Similar effects have been found for hearing [28].

It has long been debated whether perception is time- discrete or continuous, and theories have been proposed that rhythmic brain activity support the periodic sampling of the environment [29, 30]. Cortical theta rhythms (around 1/4 Hz) observed in human scalp EEGs indeed have just been found to be “strong evidence of rhythmic perceptual sampling as an inherent property of the frontoparietal attention network in humans”, even independent of task structure [31]. An underlying perceptual sampling mechanism could potentially also further modulate the timing of sensation.

3.3. Perception of short duration sounds

Basic temporal decoding skills are innate: We are born with the ability to detect differences among all the 500 phonetic distinctions used in the world's languages. Within a few years, however, only the roughly 50 contrasts that are used in our native language are selectively promoted, while we lose the ability to distinguish among the rest [32, 33, 34].

Using Mismatch Negativity (MMN) [35], several studies confirm a perceptual relationship between phonemes in language and rapidly changing acoustic information in music: "By showing similar effects of linguistic and musical expertise, these results argue in favor of common processing of duration in music and speech." [36].

It has also been found that for syllables requiring resolution of temporal cues, musicians are better at identifying unusual phonemes than non-musicians [37]. Subjects that speak multiple languages with pronounced difference between phonemes (Finnish and French) have increased short-term decoding skills in common with musicians [36], confirmed also in [38].

The left auditory cortex is specialized in the processing of acoustic transients as well as short sounds in language [39], musicians listen differently than non-musicians by engaging motor cortex [40], and the style of music played shapes auditory skills [41].

*: Please note that the AES Reviewed Papers at Tonmeistertagung can be published by both, AES and VDT, in print, online and as PDF download.

Music learning in childhood helps language learning [42]; and phonemic templates are formed early in life, so some non-native consonant contrasts are difficult to learn later [43]. Lasting impact of early language learning is also reported in [40], even if exposure does not continue into adulthood.

Finally, a study found that from the early phase of memory encoding, the auditory cortex plays a prominent role in impulse control and emotional learning [45], possibly pointing to it having a more general influence on mental timing in humans.

3.4. Auditory stress and impairment

Some sound stimuli can temporarily or permanently cause reception to degrade, but studies find that not only hair cells in the cochlea are subject to impairment. Recent animal studies have shown that sound exposure can lead to neuronal degeneration, even when hair cells recover and thresholds return to normal. The neurotransmitter, glutamate, becomes toxic with excessive sound exposure (excitotoxic), causing swelling and damage to nerve terminals [46].

Degeneration of cochlear peripheral axons may also be an important component of presbycusis (age-related hearing loss) despite a near-normal hair cell population [47], and remain hidden in (younger) subjects when tested using standard clinical methods [48].

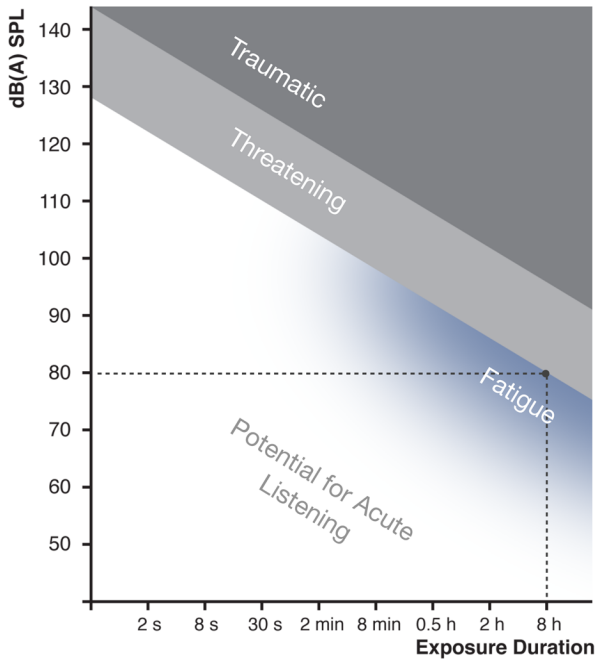


Fig. 6: Safe and potentially harmful sound exposure in adults per day, i.e. sound pressure integrated with time. Adopted from [49].

Besides from overall sound exposure, the quality of sound can also be a factor [49], these variables having the potential to shorten the time it takes before fatigue manifests: Excessive high frequency energy, lack of "quiet transients" and interaural strangeness or unnaturalness. The brain's fastest firing synapses (calyces of Held) are employed in ear comparison, and consequently also prone to fatigue [49, 50], albeit possibly not atrophy.

Recent studies estimating adverse effects of sound exposure per day (24 hours) are summarized in Fig. 6.

3.5. Active sensing

Are we passive receivers of sensory information from our environment, or actively collecting it? Contemporary trials agree on the important roles learning and experience play in perception. However, theories differ in terms of neuronal implementation, largely falling into two categories: A bottom-up accumulation model, and a top-down prediction-delta model [51, 52], with the latter gaining more support recent years [53, 54]; or at least the two complementing each other [55].

Since Alfred Yarbus' seminal work on eye movements was translated into English [56], it has been evident that saccadic eye movements may be modulated top-down by various task demands. A new study finds that implicit learning also guides saccadic movements, but in a qualitative different way than when goal-driven [57].

Different levels of expertise in detecting (pathological) visual features strongly affects saccadic eye movements, to an extent that "expert behaviour is manifested in distinct eye-movement patterns of proactivity, reactivity and suppression." Furthermore, "the fewer fixations needed to detect a visually salient abnormality, the more advanced the level of development." [58].

Saccadic eye movements are also hypothesised to be a more general sign of sensory reach-out whereby we sample the environment in order to test predictions [59, 60]. In such case, reach-out behaviour would likely not relate only to space (vision), but also to time (hearing).

EEG based studies indeed find that the rhythm of amplitude modulated noise and frequency modulated complex tones modulate perception, i.e. immediate explicit detectability is modulated by rhythm [61, 62]. Another auditory-based study concludes that regularity in non-attended items in itself does not capture attention, and that findings can be reconciled as resulting from mechanisms that minimise surprise [63].

It is also reported that real-time sensing is prone to sacrifice optimality in order to ensure self-consistency [64, 65], at times showing preference for a percept that is partially inferred, yet paradoxically considered more reliable than a percept based on external input [66].

3.6. Unconscious inference

This paragraph also includes subsequent learning, cognitive activities carried out unconsciously, and exteroception that produce (more than usual) delayed consciousness.

Unconscious inference in perception is not questioned in recent work. A commonly used mathematical model today, the Bayesian framework, even formalizes Helmholtz's original ideas [6, 55, 67, 68].

Objects can be attended to without consciously being perceived, i.e. attention is able to select invisible objects [69].

*: Please note that the AES Reviewed Papers at Tonmeistertagung can be published by both, AES and VDT, in print, online and as PDF download.

A new study finds that sound can unconsciously, not through attention, disambiguate the contents of visual awareness by facilitating perception of audiovisually congruent stimuli [70]. Other trials have found alpha wave EEG (electro-encephalography) response to hypersonic sound that could not be reported based on explicit consciousness [71, 72].

Positive and negative (ironic) priming affects what involuntarily enters consciousness [73], and problem solving can be performed without consciousness or attention [74, 75].

Various ERPs are under scrutiny as objective prediction error markers, for instance MMN, P300 and contingent negative variation (CNV), where each one's relation to expectation and attention are being examined [76].

Recognition memory, and other types of memory relevant in this context, are being studied specifically [77]. Applying a test protocol that is sensitive to detecting mnemonic mechanisms operating outside of conscious perception, by manipulating subjects' awareness at study and/or at retrieval, it has been found that non-conscious processes can drive accurate recognition memory in either case, or even in both [78].

Responses in auditory cortex are preserved during sleep in rats, and human subjects have been able to learn new sounds during REM or light non-REM sleep [79, 80]. General anesthesia and intensive care units are abundant with reports of patients hearing and learning when they were clearly not conscious [81, 82], and contact has been established to a number of patients in a persistent vegetative state, i.e. having the lowest score, 3, on the Glasgow Coma Scale [83, 84, 85].

4. Discussion

As a natural consequence of our limited perceptual bandwidth, we have to reconsider the notion of sensory *reception*. Could we be systematically underestimating the efferent component of a sensory experience? The studies reported suggest that we generally are, and how Fig. 7 would be a more relevant representation of human probing of the environment than Fig. 3.

Thanks to higher temporal resolution brain imaging techniques, evidence is also mounting that perception is predominantly a reach-out process, and that it could be driven mainly by exteroceptive prediction errors.

Most trials included were based on vision, where brain imaging can be combined with an objective marker, saccadic eye movements, but no studies suggest the fundamental principles to be much different for hearing. On the contrary, our two highest perceptual bandwidth senses are likely to make use of the same basic information gathering and filtering strategies, though specialized in two different dimensions, spatial and temporal.

Aside from temporal neural decoding, reach-out features of hearing would thus include body and head movements.

As organisms, we must be adaptive, but also conserve resources, so reach-out in exteroception would be just enough that we were able to resolve most uncertainty about the environment quickly. Unexpected findings are taken into account by the brain as potential updates of future sentient predictions.

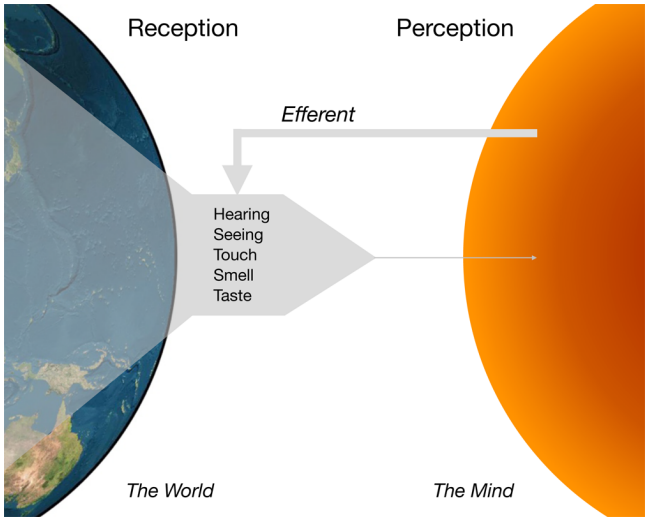


Fig. 7: Exteroception is highly influenced by experience and efferent neural activation.

The active reach-out consensus in general was a surprise finding of this review, though the notion has been held by perceptual psychologists for decades, taking off from the work of J. J. Gibson [86]. Both hypotheses, however, would be readily compatible with a limited perceptual bandwidth in humans.

If sensing primarily serves to correct experience + expectation, the time it takes to assess auditory stimuli close to what one would with unlimited time, depends almost entirely on familiarity with the features and artefacts evaluated.

From the opposite perspective, we gradually become what we sense, without having much notion of what there might be "outside the cave". From generation to generation, our species is getting better at conveying explicit information (symbolic, language), which though can be biased [87], but some of our knowledge is still tacit, and some of that relies on personal sensing. We consequently owe children reasonably unlimited sensory examples so they can fully develop sentient faculties including relevant reach-out behaviour.

Children do not just learn; they reach curiously out to the world and also learn how to learn [88]; but short-duration phonemic contrasts outside our mother-tongue are difficult to learn once we grow up. Another auditory feature, absolute pitch, the ability to identify the pitch of a sound without absolute reference, has also repeatedly been linked to musical training early in life [89, 90], and to learning a native language where pitch is a strong convenor of meaning, possibly supported by genetic disposal [91].

Considering hearing, part of the learning would be familiarization with the language of all humanity - music - performed using discrete acoustical instruments in a fine hall. Under such settings, we are able to combine musical features and the full spectrum of auditory acuity with head movement

*: Please note that the AES Reviewed Papers at Tonmeistertagung can be published by both, AES and VDT, in print, online and as PDF download.

and other reach-out actions: Frequency range, dynamic range, localization, imaging and envelopment.

Since the arrival of composers exploring the potential of abstract instrumental form, for example L. v. Beethoven, music became an even more prominent conveyor of emotions and engagement. Any such musical work, old or new, clearly deserves excellent preservation for future generations to enjoy. Old music scores are subject to today's interpretation, so our experience of such compositions unfortunately are severed, compared to, for instance, visual art of the past. Future generations need not need to have the same kind of regrets about our time.

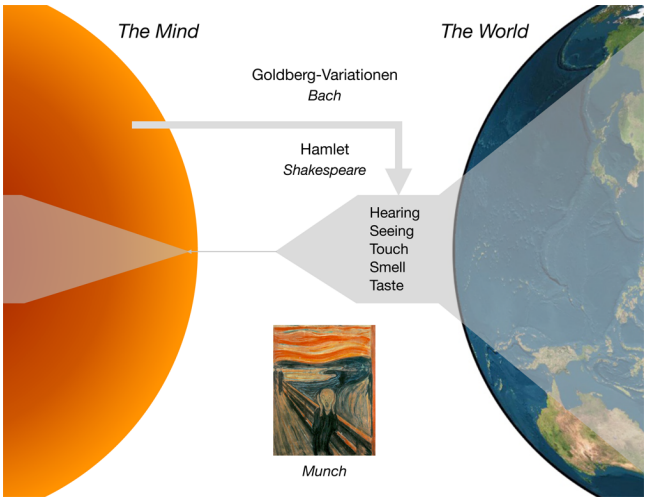


Fig. 8: Perception illustrated as a reach-out, modest return phenomenon; art being the exception.

One way to interpret "art", based on the issues discussed here, would be its unusual ability to make perceptual bandwidth momentarily widen in the receiver, for instance when listening to J. S. Bach, seeing an E. Munch painting or reading W. Shakespeare, Fig. 8. If we look at Fig. 1 right, such a quality may arguably be preserved for 30.000 years; again pointing to aspects of being human that have remained fundamentally unchanged; in essence the upper left-hand quadrant of Fig 2.

4.1. Implications for AES

New technology has the potential to influence all individual and collective aspects of our lives. Tech-robust phonemic contrasts, for instance, could help promote some words and depreciate others, possibly entire languages, like pilot's alphabet has found general use because of poor sound quality in phones, to maintain verbal exchange of meaning at all [92].

Because hearing is not only employed in explicit communication, but also offers a rare, relatively high bandwidth channel capable of carrying tacit information across generations, we must be careful not to over-simplify auditory content and distribution due to current cognitive or technical limitations.

Using traditional subjective testing, it has proven difficult to argue clearly in favor of higher data-rates than 48 kHz/24 bit linear PCM per channel [93]. The same kind of tests, however, have also been used to promote lossy data reduction, where most audio information is discarded, though many interested in sound today notice warbling "space monkey" artefacts and collapsed imaging across platforms, be it broadcast, YouTube, music streaming or phone. That kinds of artefacts might be felt more gravely now as more listeners have become "experts", than when the codecs were originally tested.

Consequently, all three temporal time-scales - the 400 ms grey-zone, fatigue and long-term learning effects - should be taken more into account in future audio standardisation, so our society is not used to rubber-stamp vulgarisation of audio in recording, storage or distribution. From a practical perspective, an automatic date of withdrawal (DOW) e.g. five years later, could be associated with any AES approval of a new audio format or watermarking based on, for instance, less than 48 kHz/24 bit linear per channel. If the new technology is still considered transparent at that time, a continued approval could be issued.

4.2. Subjective testing

Pre-testing should ensure subjects are entirely familiar with artefacts to be detected and reported, when testing is based on conscious responding (BS.1116); but taking into account the mechanisms reviewed whereby perception of auditory stimuli may be paused, delayed or altered (not yet recognized), it is indicated in subjective pro audio testing to generally be highly aware of:
- listener experience,
- listener attention,
- listening duration.

Experience includes intimate familiarity with the features tested as well as the listening environment. Identical or at least standardized listening rooms should be used.

Attention means allocating perceptual bandwidth fully to hearing and focused listening, or less.

Time means enough to satisfy every experience / learning / fatigue-probing criteria, or less. The time required could depend on whether or not the subject is multilingual or a musician, on his or her age, sound pressure level etc., but three practical categories are suggested: Easy listening, trained listening and slow listening.

4.3. Easy listening

Investigation of topics persons with normal hearing should generally be able to evaluate. For instance, if sound is too loud, voice is intelligible, reproduction is flat or immersive. Besides from understanding a language, there is no need to invoke temporal grey-zone skills (Fig. 5) in easy listening.

4.4. Trained listening

Investigation of topics that require conscious listening with attention, for instance relating to the temporal grey-zone (transients), dynamic balance, spectral balance, assessment of imaging etc.

Because experience plays such a fundamental role for our starting point when subjected to sensory stimuli, listeners should either use a room and equipment they know intimately,

*: Please note that the AES Reviewed Papers at Tonmeistertagung can be published by both, AES and VDT, in print, online and as PDF download.

or have plenty of time to get to know an acoustic environment before any tests are performed. Based on a limited perceptual bandwidth and 8 hours of dedicated listening per day, getting to know a room and equipment in any detail would take at least a week, but assuming years would be safer.

This would also be compatible with what we have discussed with many audio professionals, and a story multi award-winning mastering engineer Bob Ludwig told at the 2015 AES convention in New York: He avoids listening at different levels because of the time it takes to mentally re-calibrate. Instead, clients are invited to use headphones in case they wish to listen louder.

Elaborate aspects of trained listening are given in [94], and ITU-R BS.1116 provides more extensive practical recommendations with regard to subjects, screening, training etc. than the casual ITU-R BS.1534. However, we might still underestimate the time required for pre and post listening learning and fatigue assessment, or at least do not emphasize the importance of various temporal elements strongly enough.

4.5. Slow listening

For investigation of audio questions of possible long-term influence, we should consider all four quadrants of Fig. 2, including both interior and exterior aspects over time, but especially when examining any quality which may not be completely familiar.

This review found several mechanisms behind long-term effects of sound exposure from myological or neurological *fatigue*, to various factors causing delayed or modulated entry into consciousness, some clearly via long-term *learning*. Slow listening should therefore at least employ the time an experienced listener needs to potentially quantify fatigue, i.e. typically hours under completely known and controlled conditions, including listening level to keep outside the grey sound exposure areas of Fig. 6.

In case what is tested for is unfamiliar, slow listening could take as long as it would for the subject to learn a new language, maybe more. When reaching out, we first need to know of what to reach for; so subjective tests, even producing repeatable results, may have little long-term relevance if confined in time.

4.6. Research agenda

A variety of recent sensory studies have described exteroception as a mainly experience-based, reach-out driven process. Future research should focus on whether Fig. 7, Fig. 8 or a third principle best reflects a general model of human hearing.

Regardless, both temporal aspects of training (grey-zone, 400 ms) and learning (long-term, years) should be taken into account when testing subjectively for certain audio qualities in content and reproduction. Furthermore, what is broadly categorized as listener fatigue, could possibly be an exhaustion of afferent and efferent pathways, or of one of them, with the outcome not necessarily being the same.

However, objective criteria, not immune to falsification, and repeatable procedures need to be established so "more testing time" does not become a way of defending just any claim.

fMRI sensory studies have typically investigated vision where temporal challenges are few, compared to hearing. New windows into human perception will surely still be opened this way, and "big questions" also verified by nonparametric analysis, where less is assumed a priori [95].

Still, there is a limit to the temporal and spatial resolution of fMRI, and stimuli may need repetition for responses to manifest above the test noisefloor. Furthermore, current scanning methods are typically too noisy for examining topics of interest to pro audio.

A physiological, auditory equivalent of saccadic eye movements, but working in the time domain, could be a stepping-stone from the high spatial resolution of vision to the high temporal resolution of hearing. ERPs such as MMN, P300 or CVN could possibly be tuned for dedicated pro audio applications, for example to investigate early auditory processing and neural exhaustion taking place in the brainstem, while satisfying requirements for repeatability.

Using new ERPs and protocols that mask stimuli from auditory awareness, it might be possible instead to show non-conscious recognition. Some specific trials have already been conducted this way [71, 72], but long-term effects should ideally be tested also to investigate learning with potential efferent consequences, possibly showing less or altered efferent neural activity in expert subjects [58].

5. Conclusion

A broad selection of studies on reception, perception and consciousness, focused on hearing and perceptual bandwidth, were reviewed, divided into categories and summarised.

Previous experience, learning and expectations play the major roles in what is recognized when a human adult probes the environment; and several studies propose perception to be the outcome of primarily an active reach-out process, only sometimes including overt behavior.

Experimental results confirm the bandwidth of human perception to be severely limited, to an extent of it being a scarce resource. Hallucinations or mental illness may be understood as a free-wheeling mind with exteroceptive anchoring further depreciated or lost. Still, our primary senses covering space and time, seeing and hearing, under normal conditions offer higher perceptual bandwidth than touch, smell or taste.

Findings have been related to pro audio listening tests, depending generally on explicit consciousness with attention and short duration stimuli; and complementary methods, compatible with studies reported, have been suggested. The strength of evidence is not only down to p value, especially if we are not asking all the right questions; or if we are not factoring-in time everywhere it is potentially of influence.

*: Please note that the AES Reviewed Papers at Tonmeistertagung can be published by both, AES and VDT, in print, online and as PDF download.

The title of ITU-R recommendation BS.1116 includes the words "small impairment", so we know something is not perfect. What may change over time is the personal and collective evaluation of the significance of the impairment – maybe the "small" is not small enough after all. Because of possible learning and fatiguing effects, "slow listening" is especially important when testing systems or technology where qualities could be unfamiliar.

Science relies on empirical data-gathering, repeatability and verification; but it also relies on theory, and a willingness to strike out for new ones, subject to additional measurements and verification [96]. With all the relevant in vivo research on human perception emerging these years, time is right to consider if a more prominent role should be granted that elusive phenomena – time – also in professional audio.

7. References

[1] ITU-R BS.1534-3, "Method for the subjective assessment of intermediate quality level of audio systems" *ITU-R Recommendation* (2015).

[2] ITU-R BS.1116-3, "Methods for the subjective assessment of small impairments in audio systems" *ITU-R Recommendation* (2015).

[3] Oakley, D. & Halligan, P., "Chasing the Rainbow: The Non-conscious Nature of Being" *Frontiers in Psychology* vol. 14, iss. 8 (2017).

[4] Hameroff, S. & Penrose, R., "Consciousness in the universe. A review of the 'Orch OR' theory" *Physics of Life Reviews*, vol. 11, iss. 1 (2014).

[5] Jedlicka, P. "Revisiting the Quantum Brain Hypothesis: Toward Quantum (Neuro)biology?" *Frontiers in Molecular Neuroscience* vol. 10, iss. 366 (2017).

[6] Helmholtz, H. v., "Treatise of physiological optics: Concerning the perceptions in general" *in T. Shipley, Classics in psychology.* (1925, original published 1856).

[7] Wilber, K., "Sex, ecology and spirituality. The spirit of evolution" *Shambhala Books* (1995).

[8] Tønnesvang, J. et al., "The Four Quadrant Model" *Klim Publishing Books* (2015).

[9] Kleineberg, M., "The Blind Men and the Elephant. Towards an Organization of Epistemic Contexts" *Knowledge Organization 40* vol 5, pp. 340-62 (2013).

[10] Boron, W. F. & Boulpaep, E. L., "Medical Physiology" (2nd ed.), *Elsevier* (2011).

[11] Ceunen, E. et al., "On the Origin of Interoception" *Frontiers in Psychology*, no. 7 (2016).

[12] Küpfmüller, K., "Nachrichtenverarbeitung im Menschen" *Springer Verlag, Taschenbuch der Informatik No. 3*, pp. 429-454 (1974).

[13] Lund, T. & Mäkivirta, A., "On the Bandwidth of Human Perception and its Implications for Pro Audio*" in proceedings of the 143rd AES Convention*, NYC (2017).

[14] Cohen, M. A. et al., "What is the Bandwidth of Perceptual Experience?" *Trends in Cognitive Sciences* vol. 20, no. 5, pp. 324-335 (2016).

[15] Lamme, V. A. F., "How neuroscience will change our view on consciousness" *Cognitive Neuroscience* vol. 1, iss. 3, pp. 204-20 (2010).

[16] Shomstein, S. & Yantis, S., "Control of attention shifts between vision and audition in human cortex" *The Journal of Neuroscience* vol. 24, no. 47, pp 10702-6 (2004).

[17] Johnson, J. A. & Zatorre, R. J., "Neural substrates for dividing and focusing attention between simultaneous auditory and visual events" *NeuroImage* vol. 31, no. 4, pp 1673-81 (2006).

[18] Salo, E. et al., "Brain activity during auditory and visual phonological, spatial and simple discrimination tasks" *Brain Research* vol. 1496, pp 55-69 (2013).

[19] Moisala, M. et al., "Brain activity during divided and selective attention to auditory and visual sentence comprehension tasks" *Frontiers in Human Neuroscience* (2015).

[20] Smallwood, J. & Schooler, J. W., "The Science of Mind Wandering: Empirically Navigating the Stream of Consciousness" *Annual Review of Psychology* vol. 66, pp. 487-518 (2015).

[21] Giraudet, L. et al., "P300 event-related potential as an indicator of inattentional deafness?" *Public Library of Science,* vol. 10, issue 2 (2015).

[22] Seth, A. K., "A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia" *Cognitive Neuroscience* vol. 5, no. 2, pp 97-118 (2014).

[23] Libet, B. et al., "Subjective referral of the timing for a conscious sensory experience: a functional role for the somatosensory specific projection system in man" *Brain* vol. 102, no. 1, pp 193-224 (1979).

[24] Ray, P. G. at al., "Physiology of perception: cortical stimulation and recording in humans" *Neurology* vol. 52, no. 5, pp. 1044-1049 (1999).

[25] Libet, B., "Timing of conscious experience: reply to the 2002 commentaries on Libet's findings" *Consciousness and Cognition* vol. 12, no. 3, pp. 321-331 (2003).

[26] Pockett, S., "The great subjective back-referral debate: do neural responses increase during a train of stimuli?" *Consciousness and Cognition* vol. 15, no. 3, pp. 551-559 (2006).

*: Please note that the AES Reviewed Papers at Tonmeistertagung can be published by both, AES and VDT, in print, online and as PDF download.

[27] Melloni L. et al., "Expectations change the signatures and timing of electrophysiological correlates of perceptual awareness" *The Journal of Neuroscience* vol. 31, no. 4, pp 1386-1396 (2011).

[28] Seibold, V. C. et al., "Temporal attention shortens perceptual latency: a temporal prior entry effect" *Psychophysiology* vol. 48, no. 5, pp. 708-717 (2011).

[29] VanRullen, R. et al., "The blinking spotlight of attention" *PNAS* vol. 104, no. 49, pp 19204-209 (2007).

[30] Bush, N. E. et al., "Whisking mechanics and active sensing" *Current Opinion in Neurobiology* vol. 40, pp. 178-88 (2016).

[31] Helfrich, R. E. et al., "Neural Mechanisms of Sustained Attention are Rhythmic" *Neuron* vol. 99, pp. 854-65 (2018).

[32] Eimes, P. D, et al., "Speech Perception in Infants" *Science*, vol. 171, no. 3968, pp. 301-306 (1971).

[33] Kuhl, P. K. et al, "Linguistic Experience Alters Phonetic Perception in Infants by 6 Months of Age" *Science*, vol. 255, no. 5044, pp. 606-608 (1992).

[34] Martin, A. et al., "Learning phonemes with a proto-lexicon" *Cognitive Science* no. 37, pp. 103-124 (2013).

[35] Näätänen, R. et al. "Early selective-attention effect on evoked potential reinterpreted" *Acta Psychologica* vol. 42, no. 4, pp. 313-329 (1978).

[36] Marie, C. et al. "Musical and linguistic expertise influence pre-attentive and attentive processing of non-speech sounds" *Cortex* vol. 48, no. 4, pg 447-457 (2012).

[37] Zuk, J. et al., "Enhanced Syllable Discrimination Thresholds in Musicians" *Public Library of Science* vol. 8, no. 12 (2013).

[38] Cooper, A. et al., "Thai Rate-Varied Vowel Length Perception and the Impact of Musical Experience" *Language and speech* vol. 60, no. 1, pp. 65-84 (2017).

[39] Johnsrude, I. S. et al., "Left-hemisphere specialization for the processing of acoustic transients" *Neuroreport* vol. 8, no. 7, pp. 1761-1765 (1997).

[40] Alluri, V. et al., "Connectivity patterns during music listening: Evidence for action-based processing in musicians." *Human Brain Mapping*, vol. 38, no. 6, (2017).

[41] Vuust, P. et al., "Practiced musical style shapes auditory skills" *Annals of the New York Academy of Sciences*, vol. 1252, pp. 139-146 (2012).

[42] Zhao, T. C. & Kuhl, P. K., "Musical intervention enhances infants' neural processing of temporal structure in music and speech" *Proceedings of the National Academy of Sciences of the United States* vol 113, no. 19, pp 5212-5217 (2016).

[43] Bomba, M. D. et al, "Phoneme discrimination and mismatch negativity in English and Japanese speakers" *Neuroreport* vol. 22, no. 10, pp. 479-483 (2011).

[44] Bowers, J. S. et al., "Preserved implicit knowledge of a forgotten childhood language" *Psychological Science* vol. 20, iss. 9 (2009).

[45] Grosso, A. et al., "Auditory cortex involvement in emotional learning and memory" *Neuroscience* vol. 299, pp. 45-55 (2015).

[46] Kujawa, S. G & Liberman, M. C., "Adding Insult to Injury: Cochlear Nerve Degeneration after 'Temporary' Noise-Induced Hearing Loss", *The Journal of Neuroscience* vol. 29, no. 45 pp. 14077-85 (2009).

[47] Viana, L. M. et al., "Cochlear neuropathy in human presbycusis: Confocal analysis of hidden hearing loss in post-mortem tissue" *Hearing Research* vol. 327, pp. 78-88 (2015).

[48] Liberman, M. C. et al., "Towards a differential diagnosis of hidden hearing loss in humans" *PLoS One*, vol. 11 (2016).

[49] Gourévitch, B. et al. "Is the din really harmless? Long-term effects of non-traumatic noise on the adult auditory system", *Nature Neuroscience* vol 15, pp. 483-491 (2014).

[50] Hermann, J. et al., "Synaptic Transmission at the Calyx of Held Under In Vivo–Like Activity Levels", *Journal of Neurophysiology* vol. 98, pp. 807-820 (2007).

[51] Luce, R. D., "Response Times: Their Role in Inferring Mental Organization" *Oxford University Press* (1986).

[52] Gregory, R. L., "Perceptions as hypotheses" *Philosophical Transactions of the Royal Society B: Biological Sciences* vol. 290, iss. 1038 (1980).

[53] Hesselmann, G., et al., "Predictive Coding or Evidence Accumulation? False Inference and Neuronal Fluctuations" *PLoS One*, vol. 5 (2010).

[54] Hickok, G., "Predictive coding? Yes, but from what source?" *Behavioral and Brain Sciences*, vol. 36, no. 4 (2013).

[55] Kersten, D. et al., "Object perception as Bayesian inference" *Annual Review of Psychology* vol. 55, pp. 271-304 (2004).

[56] Yarbus, A. L., "Eye Movements and Vision" *Springer* (1967).

[57] Jiang, Y. V. et al., "First saccadic eye movement reveals persistent attentional guidance by implicit learning" *Journal of Experimental Psychology Human Perception & Performance* vol. 4, no. 3 (2014).

*: Please note that the AES Reviewed Papers at Tonmeistertagung can be published by both, AES and VDT, in print, online and as PDF download.

[58] Bertram, R. et al., "The effect of expertise on eye movement behaviour in medical image perception" *PLoS One*, vol. 8 (2013).

[59] Friston, K. et al., "Perceptions as hypotheses: saccades as experiments" *Frontiers in Psychology* no. 3 (2012).

[60] Parr, T. & Friston, K., "The active construction of the visual world" *Neuropsychologia* vol. 104, (2017).

[61] Henry, M. J. & Obleser, J., "Frequency modulation entrains slow neural oscillations and optimizes human listening behavior" *Proceedings of the National Academy of Sciences*, vol. 109, no. 49, (2012).

[62] Hickok, G. et al., "The Rhythm of Perception: Acoustic Rhythmic Entrainment Induces Subsequent Perceptual Oscillation" *Psychological Science*, vol. 26, no. 7 (2015).

[63] Southwell, R. et al., "Is predictability salient? A study of attentional capture by auditory patterns" *Philosophical Transactions of the Royal Society B: Biological Sciences* vol. 372, iss. 1714 (2017).

[64] Stocker, A. A. & Simoncelli, E. P., "A Bayesian model of Conditioned Perception" *Advances in neural information processing systems* no. 20 (2007).

[65] Peters, M. A. K. et al., "Perceptual confidence neglects decision-incongruent evidence in the brain" *Nature Human Behaviour*, vol. 1, iss. 7 (2017).

[66] Ehinger, B. V. et al., "Humans treat unreliable filled-in percepts as more real than veridical ones" *eLife*, vol. 6 (2017).

[67] Luce, R. D., "Detection and recognition" *Handbook of Mathematical Psychology*, pp. 103-189 (1963).

[68] Lorenz, R. et al., "Neuroadaptive Bayesian Optimization and Hypothesis Testing" *Trends in Cognitive Sciences* vol. 21, no. 3 (2017).

[69] Norman, L. J. et al., "Object-based attention without awareness" *Psychological Science* vol. 24, no. 6, pp. 836-843 (2013).

[70] Plass, J. et al., "Automatic auditory disambiguation of visual awareness" *Attention, Perception and Psychophysics* (2017).

[71] Oohashi, T. et al., "Inaudible high-frequency sounds affect brain activity: Hypersonic effect" *Journal of Neurophysiology*, vol. 83, no. 6 (2000).

[72] Fukushima, A. et al., "Frequencies of Inaudible High-Frequency Sounds Differentially Affect Brain Activity: Positive and Negative Hypersonic Effects" *PLoS One*, vol. 9 (2014).

[73] Allen, A. K. et al., "Entry of involuntary conscious contents from ambiguous images" *Psychology of Consciousness* December, pp. 326-337 (2016).

[74] Reverberi, C. el al., "Deduction without awareness" *Acta Psychologica* vol. 139, no. 1, pp. 244-253 (2015).

[75] Hassin, R., "Yes It Can: On the Functional Abilities of the Human Unconscious" *Perspectives on Psychological Science* vol. 8, iss. 2 (2013).

[76] Chennu, S. et al., "Expectation and Attention in Hierarchical Auditory Prediction" *The Journal of Neuroscience*, vol. 3, pp. 11194-205 (2013).

[77] Barak, O. & Tsodyks, M., "Working models of working memory" *Current Opinion in Neurobiology* vol. 24, pp. 20-24 (2014).

[78] Chong, T. et al., "Recognizing the unconscious" *Current Biology* vol. 24, iss. 21 (2014).

[79] Sela, Y. et al., "Responses in Rat Core Auditory Cortex are Preserved during Sleep Spindle Oscillations" *Sleep* vol. 39, no. 5 (2016).

[80] Andrillon, T. et al., "Formation and suppression of acoustic memories during human sleep" *Nature Communications* vol. 8, no. 179 (2017).

[81] Lawrence, M., "The unconscious experience" *American Journal of Critical Care*, vol. 4, no. 3, pp. 227-232 (1995).

[82] Tapson, K., "'It's the hearing that is last to go': a case of traumatic head injury" *Brain Journal of Nursing* vol. 24, no. 5, pp. 277-281 (2015).

[83] Owen, A. M. et al., "Detecting awareness in the vegetative state" *Science* vol. 313, no. 5792 (2006).

[84] Hampshire, A. et al., "Assessing residual reasoning ability in overtly non-communicative patients using fMRI" *Neuroimage Clinical* vol. 2, pp. 174-183 (2012).

[85] Owen, A. M. & Coleman, M., "Functional neuroimaging of the vegetative state" *Nature Reviews Neuroscience* vol. 9, pp. 235-43 (2008).

[86] Gibson, J. J., "The senses considered as perceptual systems" *Boston: Houghton Mifflin* (1966).

[87] Caliskan, A. et al., "Semantics derived automatically from language corpora contain human-like biases" *Science* vol. 356 (2017).

[88] Siegler, R. S., "Emerging minds: the process of change in children's thinking" *Oxford University Press* (1996).

[89] Moulton, C. "Perfect pitch reconsidered" *Clinical medicine* vol. 14, no. 5, pp. 517-9 (2014).

[90] Leite, R. B. et al., "Music Proficiency and Quantification of Absolute Pitch: A Large-scale Study among Brazilian Musicians" *Frontiers in neuroscience* vol. 10 (2016).

[91] Deutsch, D. et al., "Absolute pitch among American and Chinese conservatory students: prevalence differences and evidence for a speech-related critical period" *Journal of the Acoustical Society of America* vol 119, no. 2, pp 719-22 (2006).

*: Please note that the AES Reviewed Papers at Tonmeistertagung can be published by both, AES and VDT, in print, online and as PDF download.

[92] http://www.spelltool.com/en/index.php

[93] Reiss, J. D., "A Meta-Analysis of High Resolution Audio Perceptual Evaluation" *Journal of Audio Engineering Society* vol. 64, no. 6 (2016).

[94] Bech, S. & Zacharov, N., "Perceptual Audio Evaluation – Theory, Method and Application" *Wiley publications* (2006).

[95] Nichols, T. E. & Holmes, A. P., "Nonparametric permutation tests for functional neuroimaging: A primer with examples" *Human Brain Mapping* vol. 15, no. 1 (2002).

[96] Koopmans, T. C., "Measurement without theory" *The Review of Economic Statistics*, vol. 29, no. 3, pp. 161-72 (1947).

Illustrations

Figures 3, 4, 5, 7, 8 original artwork by the authors. Citation granted with declaration of source.

Ear Signals are Bodily Signals
Binaural Audio as a Form of Engineering at the Human Body

M. Rumori[1]
[1] *Institute of Electronic Music and Acoustics Graz, Austria, Email: rumori@iem.at*

Abstract

Binaural audio technology is based on ear signals, that is, acoustic signals as they occur at our outer hearing apparatus such as the ear canals or the eardrums. Such signals are sought to be captured or synthesised for reproducing them at the same location, usually by means of headphones. If the reproduction is sufficiently accurate with respect to the reference situation, the same auditory impression shall be evoked by listening. In practice, several factors impede the realisation of this idealised assumption, among them technical, psychoacoustic and aesthetic ones.

From a bodily perspective, ear signals represent an intermediate step in the processing chain of human hearing: they result from the outer hearing apparatus, which, by its physical properties, transforms the acoustic surrounding into a signal pair that contains spatial information to be evaluated by the brain. Capturing and reproducing ear signals therefore means to interfere with the bodily function of listening, to decouple the outer ear temporally and spatially, or to replace it, e.g., by a dummy head or by the pinnae of another person. I regard this interference as a form of engineering at the human body, even if it is subtle and non-invasive.

The dummy head microphone is not a pictorial representation of the human outer hearing apparatus but it is designed to take over its function: it is supposed to actually perform a partial act of hearing. Like other areas of engineering, binaural audio is deeply connected to mimesis, which, with reference to Walter Benjamin, may be regarded as a fundamental human faculty of convergence to nature. The paper will relate aspects of binaural technology to concepts of mimesis, leading to aesthetically effective categories such as the magic and the uncanny.

Introductory Remark

This paper contains the literal text of my talk at Tonmeistertagung on November 15th, 2018, with only a few minor corrections. Deviating from my usual habits, the text has been written out in full prior to the presentation, because I decided to strictly adhere to a certain structure (as explained in the second part of the talk). In order to retain this structure in the proceedings of Tonmeistertagung, I refrain from providing a more scholarly version of the paper. However, aforementioned structure did not allow to cover all aspects announced in the abstract, at least not in the probably expected level of detail. They will be reflected in a future publication.

Owing to the performative character of the talk, not all of the text parts marked as quotations here were clearly presented as such at the conference, especially those by John Cage.

1.

Good morning, ladies and gentlemen.

I assume that most of you are familiar with binaural audio, so I would like to start with a thought experiment. We learned that binaural audio deals with so-called ear signals. Where at your body occur ear signals that correspond to the present situation, now while I am talking and you are listening to my voice? Try to imagine the signals “taking place” at their location in this very moment. Is that an experiment, is it an experience? Do you consider your body being involved?

I regard ear signals as something the human body produces from the sound around us. More precisely, our acoustic surrounding is transformed by the outer hearing apparatus, which involves our head, the layout of our ears and shoulders, the shape of our pinnae. The result is a signal pair that contains spatial information known as cues, such as time and level differences, or spectral distortion. Ear signals then proceed through our middle ears into the inner ears, where they get “data-reduced” in order to be received by the brain as an “mp3 file,” as we learned from Jürg Jecklin in yesterday’s keynote.

When we carry out binaural recordings, we seek to pick up ear signals after they have been provided by our outer hearing apparatus. To this end, we place microphones anywhere into our ear canals or at our ear drums, in between the outer and middle ears.

In a binaural recording situation using our own head, do we

ISBN
978-3-9812830-9-9
The authors hold the copyrights of all of content in this manuscript. They have granted the VDT the right to publish it in print, online and as PDF download on *https://tonmeistertagung.com*

still need our outer hearing apparatus for something else than the sake of recording? Probably we still use it for orientation in case we carry out binaural field recordings, or it is active for monitoring as a kind of body's own in-ear monitoring system. But in many situations we do not. Most often, our outer hearing apparatus is split off from ourselves. Actually it could be removed from the rest of our signal processing chain, and sometimes it even is, for example, when the ear canals are closed while recording.

We then literally become a dummy head, because we are only interested in exploiting its physical properties, like a well-shaped piece of wood. The rest of ourselves may even become an obstacle whose influence we seek to eliminate, or *to minimise,* as we say in engineering. When measuring binaural impulse responses in a laboratory situation, our head is fixated such that it really behaves like a dummy head, unlike a living body, which wants to move.

When listening to binaural audio, we aim at reproducing the ear signals at the same location they were recorded. We try to inject them using the same tap in the signal processing chain of hearing we created for recording, but this time into the other direction, towards the middle ears. That's why we wear headphones or attach earbuds to our ear canals. We want to bypass our outer hearing apparatus. Again, a portion of our body becomes an obstacle whose influence we seek to eliminate, pardon, to minimise, this time the parts of the outer hearing apparatus.

It is now time to digest what I just said in a short break.

2.

> This is a composed talk for I am making it just as I make a piece of music. [. . .]
>
> That music is simple to make comes from one's willingness to accept the limitations of structure. Structure is simple because it can be thought out, figured out, measured. It is a discipline which, accepted, in return accepts whatever [. . .].
>
> How could I better tell what structure is than simply to tell about this, this talk which is contained within a space of time [exactly twenty] minutes long?
>
> That [twenty] minutes has been divided into [four] [. . .] parts, and each unit is divided likewise. [. . .] [4, 110–12]

After each four minutes of talking, a general pause of one minute follows. Each block of talking is divided into one minute of presenting a thought or an aspect, two minutes of reflecting and deconstructing it, and one minute of drawing further speculations from it. Right now, we are already in the second unit of the second part, that is, the reflection on the idea of giving a composed, pre-structured talk, and its deconstruction.

I restrict myself to this pre-determined structure today because I became aware of the fact that I never did this before. Instead, when giving a talk or a lecture, I did whatever I thought I have to do for this occasion. I copied from what I saw or heard others doing. I rarely reflected on what I experienced although I very much noticed that some presentations are more pleasant, inspiring, or moving to me than others.

Recently I was wondering: What if others did copy what I do, in turn? Would I want to leave them with my unconsidered habits based on limited abilities to copy from others? Wouldn't I want to exemplify different ways of acting on a situation like this one, of using it as an opportunity to experiment, to take some risk, ideally to provoke a rewarding experience both for myself and for the audience?

You may suspect that my attempt to perform a composed talk is also copied from somebody else. This assumption is actually true. The idea is stolen from American composer John Cage, and even parts of my text are literally taken over from his *Lecture on Nothing,* first published in 1959. The difference is that I did not copy ideas by John Cage before, at least not consciously, and that nowadays I almost never see somebody copying John Cage.

I am now at the end of the second part of my talk. What will follow is another break of approximately one minute. If you have no idea how to spend this minute of your life, simply listen. Should you find yourself thinking of something, try to drop the thought and return to listening.

3.

Earlier I said that we split off our outer hearing apparatus for recording and for playing back binaural audio. Like this, we are able to reconstruct a physical hearing constellation we are actually not in: temporally, spatially, or both. We may hear a concert that takes place in a concert hall when we are actually not there but *as if* we were there, or we may hear it later, when it actually does not take place, *as if* it would now. The key is to introduce such a split and to use the separated parts independently: the outer for recording, the inner for reproducing.

What we cannot reproduce, though, is actually *listening,* because it involves the perception of auditory events, as Jens Blauert calls it, which in turn involves cognition [3]. And again, Jürg Jecklin reminded us yesterday that perception does not take place in our hearing apparatus.

But let us stick with hearing and come back to the aforementioned split somewhere slightly beyond the boundaries of our head. The same mechanism is exploited in other techniques of binaural audio. Measuring binaural impulse responses of our head, as mentioned, is merely a special case of recording. What we measure is the acoustic properties of our isolated outer hearing apparatus.

Likewise, when we carry out recordings or measurements with a dummy head, we use a split-off copy of our outer hearing apparatus at its place. It is most probably not an exact copy of *our* hearing apparatus. Likewise, when we interpolate or synthesise impulse responses from measurements or models. I will come back to the dummy head in a minute.

On the reproduction side, we may also use loudspeakers involving crosstalk cancellation, and we need this exactly for the purpose of bypassing the outer hearing apparatus. In the case of headphones, a lot of effort is put into making these transparent. Headphone equalisation, the compensation of the coupling between the headphone and the head, sometimes even individualised, are all performed in order to make our outer hearing apparatus disappear, sorry, to minimise its interference.

Back to the dummy head microphone. To me, there are at least three interesting things to it:

1. It actually performs the act of hearing partially, "it hears,"
2. it is a human-made tool modelled after a part of the human body in order to take over its function,
3. there is something sublime to it because it personifies an idealised, synthesised human ability.

As we cut off our outer hearing apparatus when listening to binaural audio, reproducing a recorded situation would simply not work if the dummy head did not produce ear signals in our place, that is, if it did not do what usually our outer hearing apparatus does: a part of spatial hearing.

A specialised, engineered tool that takes over body functions is usually called a *prosthesis* if it re-enables abilities that were lost. As we usually did not lose our outer hearing capabilities when engaging with binaural audio, I would probably not go as far as calling the dummy head a part of a prosthesis. However, I would attribute some *prosthetic qualities* to it, because we use binaural audio to enhance our hearing abilities: listening to synthesised virtual spaces that do not exist or that we would never be able to visit in person is one example.

For the third break, if you like, stuff your fingers into your ears and close your eyes.

4.

> Slowly, as the talk goes on, we are getting nowhere and that is a pleasure. It is not irritating to be where one is. It is only irritating to think one would like to be somewhere else. [4, 118–19]

Binaural audio enables us to listen with somebody else's ears or with ears that do not exist in reality. In fact, something like this happens regularly when we hear binaural signals recorded with binaural microphones by somebody else or with a dummy head, or synthesised with generalised or other people's indivual impulse responses. It might even be questioned whether we actually hear with our own ears when listening to our own binaural recordings. Surprisingly enough, we can quite well adapt our spatial listening to foreign ears.

The dummy head has something sublime to it: it aims at averaging the acoustic properties of many people's outer

hearing apparatus such that as many as possible people can adjust their listening to this average. Although *not everything that limps is a comparison,*[1] this reminds me at the legend of painter Zeuxis of Herakleia in Greek Antiquity, who lived around 430 B. C. E. For the city of Kroton, Zeuxis was commissioned to paint an image of Helena to be exhibited in the Temple of Hera. As Helena was admired as the most beautiful woman ever, he sought a technique to picture such immense beauty.

Elizabeth Mansfield analysed:

> [...] Zeuxis cannot work without a model. [...] But he realizes that a real model cannot serve as a bridge to the ideal. [...] The painter cannot *conceive* of ideal form. If he could, no model would be necessary. At the same time, he searches but cannot confirm the existence of the ideal in the world he perceives. This leads to a crisis: How can the ideal exist if it can neither be seen nor imagined? And, as Plato explains, the non existence of the ideal would eliminate the possibility of transcendent experience. [5, 28–9, original emphasis]

Zeuxis solved the problem by selecting the five most beautiful women of the city and by creating a synthesis of each model's most significant elements of beauty. As I said, not everything that limps is a comparison, and the greatest beauty ever is probably not the same as the best working average. But what is shared here in my opinion is a notion of *the ideal* that is sought to be reached.

The chair of this session, David Griesinger, told me two years ago about an experience he made when he wanted to suspend a dummy head microphone above the head of the conductor of an orchestra he was recording. The conductor refused to have somebody else hearing in such a prominent position. Or, probably even more disturbing, to have a *thing* that appears like *somebody* who is listening. That the thing actually hears we already learned. The observation that we tend to attribute the ability of listening to the dummy head reveals that it carries something uncanny to us.

Ear signals are body signals: they are created by the body and they belong to the body, and without engineering we cannot get hold of them. Actually getting hold of them is therefore engineering at the human body. That, in turn, is part of our human fundamental faculty of mimesis, our major productive force, as Walter Benjamin puts it [2]. Like any other engineering, it exemplifies a convergence to nature, which, according to Adorno and Horkheimer, must never and can never be fully reached if barbarism is to be avoided [1].

[1] This is a play on words hinting at a literally translated saying in German, which does not work in English. A "limping comparison" is one that does not really fit.

5. References

[1] Adorno, T. W. and Horkheimer, M.: Dialectic of Enlightenment, Stanford University Press, Stanford, 2002

[2] Benjamin, W.: On the Mimetic Faculty. In: Selected Writings 1926–1934, Belknap Press, Cambridge/Mass., 1999, 721–22

[3] Blauert, J.: Spatial Hearing. MIT Press, Cambridge Mass./London, 1997

[4] Cage, J.: Lecture on Nothing. In: Silence. Lectures and Writings, Wesleyan University Press, Middletown/Connecticut, 1961, 109–26

[5] Mansfield, E. C.: Too Beautiful to Picture. Zeuxis, Myth, and Mimesis, University of Minnesota Press, Minneapolis/London, 2007

On the localization of auditory objects created by directional sound sources in a virtual room

F. Wendt and M. Frank
Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Austria
Email: {wendt, frank}@iem.at

Abstract

A directional sound source can create auditory objects at distinct locations in a room by varying its directivity pattern and orientation. Recent perceptual studies could show that the direct-to-reverberant energy ratio is the most relevant cue for their distance from the listening position. The lateralization, on the other hand, is affected by the precedence effect. Thus, the azimuthal angle of auditory objects depends on signal type and time/strength of single wall reflections compared to the direct sound. This contribution evaluates the perception of auditory objects in a virtual room regarding both distance and azimuth angle relative to the listening position. In the evaluation, listeners were asked to aurally match the position of an omnidirectional sound source with the position of the auditory objects created by the directional sound source in the same virtual room. The evaluation examines the influence of directivity pattern and beam orientation of the directional source, as well as signal type and source position. The results are modeled by the extended energy vector and the inter-aural cross correlation coefficient and they are compared to previous studies.

1. Introduction

The localization (direction and distance) of sound in a room is not only depending on the position of the sound source and the listener, but also on the directivity of the source and the reflection properties of the room. Loudspeakers with controllable directivity, such as the IKO [1], make use of the reflection paths to create static and moving auditory objects [2, 3] inside the room. The auditory distance between the auditory object and the listener is mainly determined by the direct-to-reverberant energy ratio, which can effectively be controlled by the spherical harmonic order of the beam pattern [4]. The recent study revealed that changing the order of a listener-facing beam yields significantly increasing auditory distance for each integer decrease of the order from 3 to 0. The lateralization of the auditory object, i.e. the lateral distance between the auditory object and the sound source, is influenced by the precedence effect. Thus, more transient sounds are perceived closer to the sound source in comparison to stationary sounds [3]. However, masking effects can be employed to suppress the precedence effect and to increase the lateralization of transient sounds [5].

In previous studies, modeling of the perceived direction has been done using either impulse responses measured in a real environment with a directional sound source in a room [6] or simplified simulations using the image-source method for a shoebox room and a generic directional sound source [2]. Both kinds employed an extended version of the energy vector [7, 8]. This model was originally developed for the prediction of localization in surrounding loudspeaker arrays at off-center listening positions, where sound is not arriving simultaneously at the listener's ears.

This contribution extends the existing studies by sound sources with 5^{th}-order beam patterns and employs a simplified auralization of a shoebox room using a combination of an image-source model and a feedback delay network.

In contrast to previous experiments, the evaluation applies a novel method, where listeners are not asked to indicate the Cartesian coordinates of the auditory object, but to aurally match the position of the auditory object with a freely movable omnidirectional source inside the same auralization.
The paper is structured as follows: Section 2 describes the experimental setup, including the room simulation and the conditions. The following section presents the results of the listening experiments focusing on the influence of the directivity order, the signal, and the source position. Subsequently, the results are modeled using the extended energy vector for direction and the direct-to-reverberant energy ratio and the inter-aural cross correlation coefficient for auditory distance. Finally, the results are compared to previous experiments that used a real 3^{rd}-order loudspeaker in a room and the measurement-based virtualization thereof.

2. Listening experiment

The listening experiment was carried out in the shoebox-shaped IEM lecture room ($7.0\,\mathrm{m} \times 8.3\,\mathrm{m} \times 3.0\,\mathrm{m}$, $T_{60} = 0.58\,\mathrm{s}$). To obtain maximum controllability the perception of directional sound sources in this room is studied using a simulated sound field played back over headphones. The simulation consists of direct sound, early specular reflections, and diffuse reverberation binaurally rendered in 7^{th}- order Ambisonics using the freely available IEM plug-in suite[1] with Reaper as DAW. The positioning of the omnidirectional source in the virtual room to match the localization of the directional source is done with PureData and control data of source position and audio playback is send to Reaper via Open Sound Control (OSC). The playback employs an RME Babyface Pro sound card with AKG K702 headphones. Figure 1 shows the processing scheme of the listening experiment.

[1] plugins.iem.at

ISBN
978-3-9812830-9-9
The authors hold the copyrights of all of content in this manuscript. They have granted the VDT the right to publish it in print, online and as PDF download on https://tonmeistertagung.com

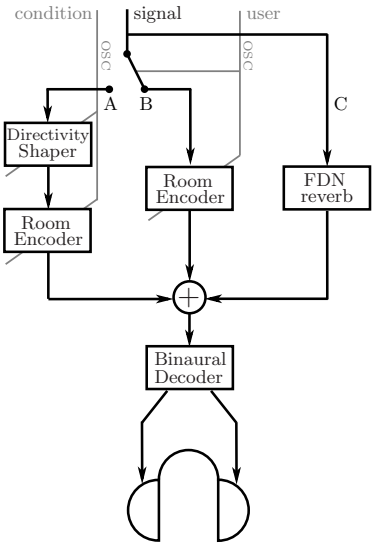


Fig. 1: Processing scheme for the simulated sound field using Reaper and the IEM plug-in suite. Gray lines indicate the adjustments by the condition and listener transmitted by PureData via OSC.

2.1. Room simulation

The sound field simulation is processed in three parallel effect chains A – C, cf. Fig.1. Direct sound and early reflections are generated separately for the directional (A) and omnidirectional sound source (B) using the *RoomEncoder* plug-in. Using Sabine's reverberation equation, we estimate an overall frequency-independent reflection attenuation r from volume V, surface S, and reverberation time T_{60} of the room:

$$r = 1 - \frac{0.161\,V}{T_{60}\,S}, \tag{1}$$

which is used for the 127 early reflections in the *RoomEncoder*. The directivity of the directional sound source is considered by the *DirectivityShaper* plug-in. It encodes orientation θ and order n of a frequency-independent directivity pattern using the so-called max-$\boldsymbol{r}_{\mathrm{E}}$ weights [9] into the Ambisonics domain which is processed by the *RoomEncoder*.

In the third chain C, a feedback delay network is used for simulating the directivity-independent diffuse reverberation. The parameters of the *FDNreverb* are derived from measured room impulse responses (RIR) of the IEM lecture room. Parameters *room size* and *reverberation time* are adjusted to match the temporal decay of the RIR, cf. Fig. 2. The *filter* is adjusted in way that it sounded convincingly enough in comparison to the real room and the *fade in time* of the reverb is set to blend between specular and diffuse part of the simulation to guarantee a similar energy progress of simulated and the measured RIR.

After summing up direct sound and early specular reflections with the diffuse reverberation, the *BinauralDecoder* is used for binaural synthesis. It convolves the simulated RIR with head-related impulse response (HRIR) measurements of the Neumann KU100 [10] and includes headphone equalizations for the used headphones.

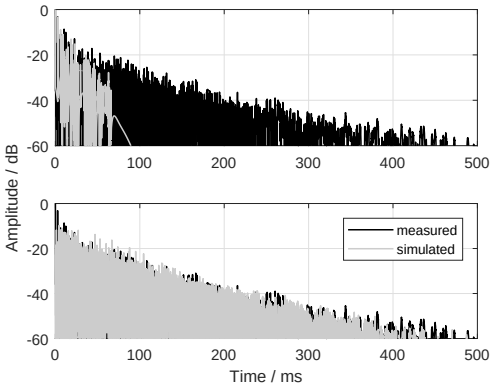


Fig. 2: Upper: simulated direct sound and specular reflections using the RoomEncoder plug-in (gray) together with measured RIR (black) at position P_1. Lower: simulated diffuse reverberation using the FDNreverb plug-in (gray) together with measured RIR (black) at P_1.

2.2. Conditions

Two source positions P_1 and P_2 of the directional source are tested, both with a virtual distance of $4.2\,\mathrm{m}$ and azimuth angles of $-7°$ and $40°$ with respect to the listener. The sound source is tested with four beam orientations for each position. Three of them are normal to the sidewalls, whereas the orientation facing towards the lateral adjacent wall corresponds to the respective specular direction. This yields beam orientations $\theta_1 = (0°, 90°, 180°, 235°)$ for source position P_1 and $\theta_2 = (0°, 115°, 180°, 270°)$ for P_2. Tested directivity orders for position P_1 are $n = (1, 3, 5)$ and $n = (1, 3)$ for P_2. Additionally, a reference condition is included in the test consisting of P_1 with order $n = 0$.

During the experiment, the listener was sitting in the IEM lecture room and her/his physical position corresponded to the position within the simulation. The schematic layout of the room was shown on the computer screen and the listener was not aware of the position of the directional sound source. A marker within the schematic room layout could be moved around and its position data controlled the position of the omnidirectional source in the *RoomEncoder* (B, see Fig. 1). The listeners task was to match position of this omnidirectional source in the virtual IEM lecture room with the position of auditory objects created by the fixed directional sound source in the same room. The binaural playback did not include dynamic rendering incorporating the head rotation of the listener. Instead, the rendering was done in a way that the listener was always facing the directional source. Numbers at the walls indicated the viewing direction for both source positions. In order to match the vision with the simulated sound field, listeners had to look at the corresponding number. In this way they could roughly adjust the position of the marker (drag and drop) with the fixed auditory object. For fine adjustments, an A/B switch allowed to directly compare the auditory object with the omnidirectional source. Moreover, the position of the latter could be adjusted by two sliders controlling its azimuth and distance.

Tested excitation signals were played back in loop until listeners were satisfied with the match between omnidirectional source and the auditory object of the directional source.

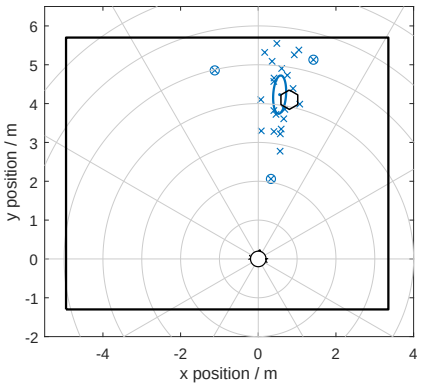


Fig. 3: Individual answers collected by the reference condition $n = 0$ for P_1 ans S_1. Mean value and corresponding 95%-confidence ellipse are based on Hotelling's T^2 distribution. Answers not taken into account in the calculation are encircled.

Signal S_1 was a 750-ms-long pink noise burst, linearly faded in and out by $250\,\mathrm{ms}$, followed by a 500-ms-long pause. Additionally, source position P_1 with directivity order $n = 3$ was tested with male speech (S_2).
This yielded 25 conditions, which were tested twice in an individual random permutation. Twelve listeners participated in the experiment that lasted 60 minutes on average. All of them were experienced listeners and reported normal hearing acuity.

3. Results

Obtained results for the omnidirectional reference condition at P_1 are shown in Fig. 3. Hexagon and head symbols indicate positions of directional sound source and listener, respectively. Ideally, for the reference condition answers represented with an $\times$ should coincide with the position of the directional sound source.

For a compact representation of individual data points, we compute the 2-dimensional mean value and the corresponding 95%-confidence area. Therefore outliers $\otimes$ are removed and the remaining data is shifted to have zero mean. A singular value decomposition is applied to get the unbiased standard deviation estimator. With this we can calculate half-axes of the ellipse representing the 95%-confidence area with the mean value at its center using the Hotelling's T^2 distribution. Rotation and size of the obtained ellipse are indicators of the data distribution. Answers spread more along the distance compared to the angle which agrees with literature according to which we are better in estimating the lateralization compared to the distance [11]. Nevertheless, mean value and sound source almost coincide with a slightly bigger deviation in direction compared to distance.

3.1. Influence of directivity order

Fig. 4 shows means and corresponding 95% confidence areas for all conditions with signal S_1 at P_1. Beam orientations θ_1 are color coded and schematically depicted as solid lines around the directional sound source. It can be seen that with increasing order n the distance of confidence ellipses to the directional sound source tends to increase.

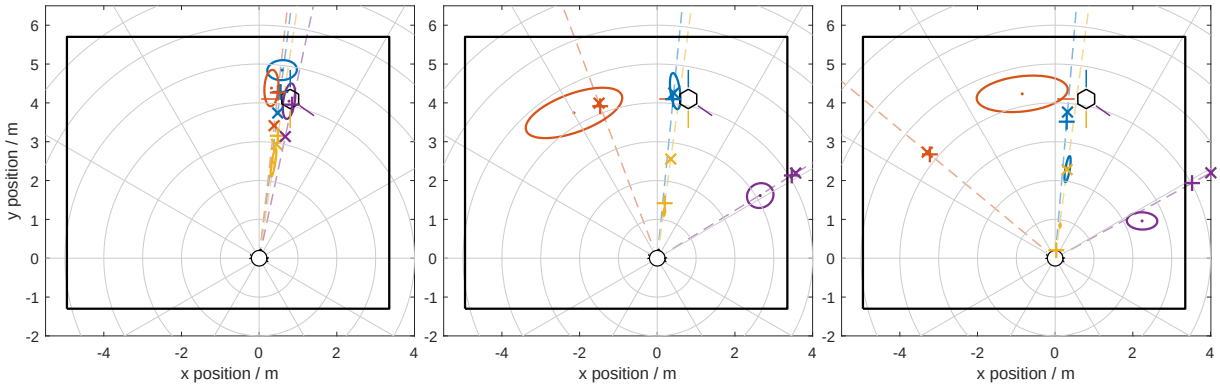
This shift away from the directional sound source agrees with the beam orientation and – from the listeners perspective – lateral orientations $\theta = (90°, 235°)$ mainly yields lateral shifts, whereas medial orientations $\theta = (0°, 180°)$ shift the smaller ellipses along the distance.

The latter effect is explained by the variation of the direct-to-reverberant energy ratio (DRR), a cue for distance perception [11], which is influenced by the source directivity. For $\theta = 0°$ the auditory distance decreases with increasing the order n and a minimum distance of less then $1\,\mathrm{m}$ is achieved with $n = 5$. Agreeing with [4] for the back-facing condition $\theta = 180°$ the auditory distance increases by increasing $n = 0 \rightarrow 1$. Interestingly, for $n = 1 \rightarrow (3, 5)$ auditory distances do not increase further as the sound reflected from the frontal wall prevails the direct sound.
For lateral source orientations the lateralization mainly increases with the order. This is not true for the $\theta = 90°$ and $n = 5$. The reason therefore is source splitting, which can be seen in the bimodal distribution of individual angles for this condition, cf. Fig. 5. The delay between direct sound and the strong lateral reflection $\tau_1 = 21\,\mathrm{ms}$ exceeds the individual echo threshold and the perception of a single auditory object breaks apart. In this case the precedence effect is no longer active and two auditory objects are perceived at distinct locations in the room.
The decreased distance of $\theta = 235°$ for $n = 5$ compared to $n = 3$ indicates that similarly to $\theta = 0°$ the reflection takes over the distance perception.

3.2. Influence of signal and source position

The influence of the signal $S_{1/2}$ is examined in Fig. 6(a) – (b) showing means and corresponding 95% confidence areas for both signals with order $n = 3$ at source position P_1. For medial orientations $\theta = (0°, 180°)$ there is almost no influence, neither on the distance nor on the lateralization. Contrastingly, for lateral orientations $\theta = (90°, 235°)$ we see that auditory objects are less lateralized for the speech signal S_2. This agrees with [3, 12] and is explained by a higher proportion of transient signal components within S_2 and supports findings on the specific properties of the precedence effect for transient signals [13, 14].

Fig. 6(c) – (d) show the results for source position P_2 with orders $n = (1, 3)$. In contrast to P_1, the distances of ellipses to the directional sound source do not always increase with increasing order. Moreover, the direction of auditory objects does not always correspond with the beam orientation and in some cases, it also depends on the directivity order.
The confidence area of $\theta = 270°$ with $n = 1$ is shifted along the medial direction. Contrastingly, for $n = 3$ a more lateral shift is obtained. A reason therefore could be the fact that except for specular orientation $115°$, beam orientations are no longer medial or lateral. Thus, depending on the directivity order, either the distance or direction cues dominate the localization.



(a) $n = 1$ (b) $n = 3$ (c) $n = 5$

Fig. 4: Order dependent means and corresponding 95% confidence ellipses for beam directions 0° (front, blue), 90° (left, red), 180° (yellow), −90° (violet) for P_1 ans S_1. The $\boldsymbol{r}_E$-model results (dashed lines) use $w_\tau = -0.8\,\mathrm{dB/ms}$. Distances models are shown as markers with $\times \ldots f(\mathrm{DRR})$ and $+ \ldots f(1 - \mathrm{IACC})$.

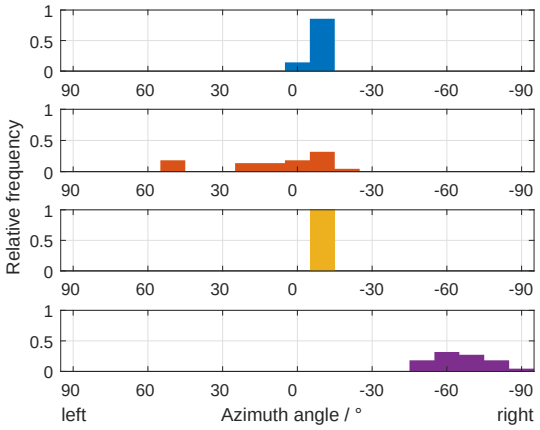


Fig. 5: Histogram of obtained directions from the listeners perspective for directivity order $n = 5$ and signal S_1 evaluated at source position P_1. See Fig. 4 for respective color coding.

4. Modeling the localization

In the previous section experimental results were mainly explained by dividing the localization task in direction and distance. Accordingly, we use the same division for the modeling of the experimental data.

4.1. Direction

Obtained directions of auditory objects are predicted using a so-called vector-based model. In contrast to more complex binaural models, the evaluated weighted energy vector works with information of the coherent sound arriving from multiple sound paths based on delay and direction information of a image-source model.

The extended energy vector is defined as

$$\boldsymbol{r}_\mathrm{E} = \frac{\sum_{i=1}^{I} (w_{\tau,i}\, w_{d,i}\, g_i)^2\, \boldsymbol{\theta}_i}{\sum_{i=1}^{I} (w_{\tau,i}\, w_{d,i}\, g_i)^2} \tag{2}$$

and employs amplitude gain g_i, dissipation weight $w_{d,i}$, and unity vector $\boldsymbol{\theta}_i$ for each sound path i.

Moreover, to incorporate the strength of the precedence effect, the relative importance of the sound path is considered by a time-dependent weight $w_{\tau,i} = w_\tau \tau_i$, with the relative delay τ_i compared to the direct sound. The signal-dependent factor w_τ is found by tuning it to get the best matching results of the experimental data with our model. For calculation, a simple third-order image-source model is used which includes the room geometry and its reflection properties calculated with Eq. (1). Orientations of obtained energy vectors are represented as dashed lines in Fig. 4 for S_1 with $w_\tau = -0.8\ \mathrm{dB/ms}$. For all orders $n = (1, 3, 5)$, the model achieves good matching results and almost all vectors hit the corresponding confidence ellipse. Only for $n = 5$ and $\theta = 90°$ the model deviates from the experimental results. This explained by the bimodal answer distribution for this condition and the modeled direction corresponds to the direction of the auditory object reported in only 17% of the answers, cf. Fig. 5.

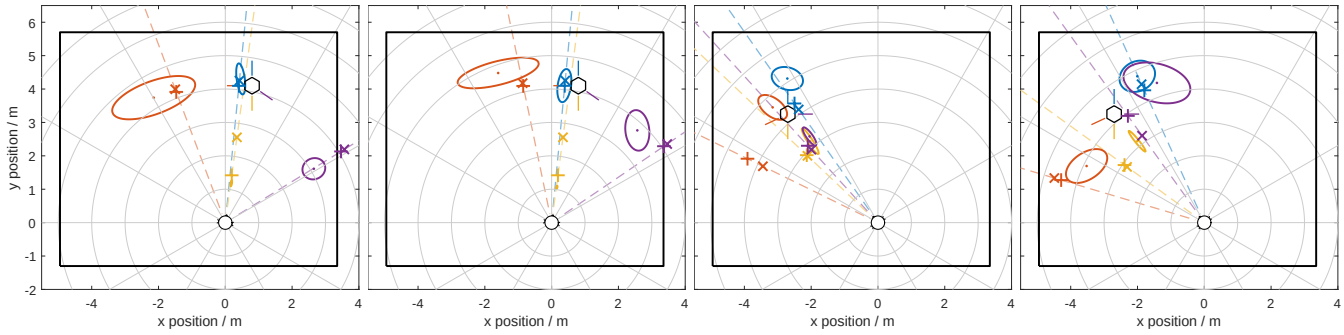
In Fig. 6(a) – (b) the signal-dependence of the $\boldsymbol{r}_\mathrm{E}$ model is examined with $w_\tau = -0.9\ \mathrm{dB/ms}$ for S_2. Except for $\theta = 115°$ the predicted lateralizations hits corresponding ellipses. The modeling of source position P_2 shown in Fig. 6(c) – (d) with S_1 and $w_\tau = -0.8\ \mathrm{dB/ms}$ is not as accurate as for P_1.

4.2. Distance

For modeling the auditory distance, two metrics of the spatial sound field and their regression to the experimental data are studied. The most obvious distance predictor is the direct-to-reverberant energy ratio (DRR). It is widely accepted for prediction of auditory distance [11] and is defined as

$$\mathrm{DRR} = 10 \log_{10} \frac{\int_{0\mathrm{ms}}^{T} h_W^2(t) dt}{\int_{T}^{\infty} h_W^2(t) dt}. \tag{3}$$

The time constant T regards only the direct sound and h_W is the omnidirectional RIR (W channel) for source position $P_{1/2}$. Regression analysis of all answers fits a linear regression function $f(\mathrm{DRR}) = k\mathrm{DRR} + d$ depending on the DRR and yields $k = -0.085$ and $d = 2.88$, cf. Fig. 7. The correlation of the DRR with the experimental data is poor ($R^2 = 0.34$) and the modeled auditory distance for $n = 1$ is mainly underestimated, whereas for $n = (3, 5)$ it is mainly overestimated, cf. Fig. 4 and Fig. 6.



(a) S_1 with $n = 3$ at P_1 (b) S_2 with $n = 3$ at P_2 (c) S_1 with $n = 1$ at P_2 (d) S_1 with $n = 3$ at P_2

Fig. 6: Means and corresponding 95% confidence ellipses for different beam directions. The $\boldsymbol{r}_E$-model results (dashed lines) use signal-dependent temporal weights w_τ. Distances models are shown as markers with $\times \ldots f(\mathrm{DRR})$ and $+ \ldots f(1-\mathrm{IACC})$.

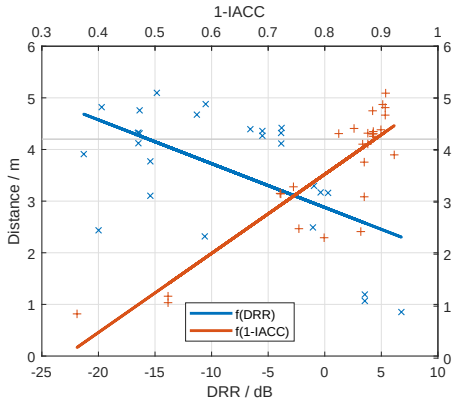


Fig. 7: Regression model DRR: $R^2 = 0.34$, 1-IACC: $R^2 = 0.77$.

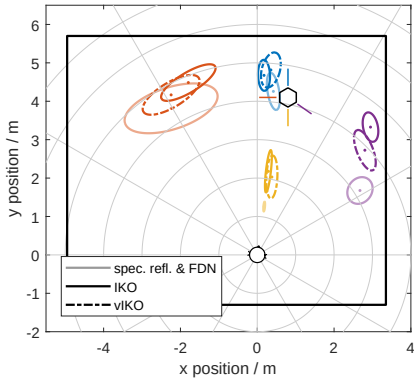


Fig. 8: 95% confidence ellipses obtained by the simulated sound field (light) in comparison to the a real directional sound source in the IEM lecture room (solid) and a binaural rendering thereof (dotted).

In [3] the distance of auditory objects created by a directional sound sources was successfully modeled by the inter-aural cross correlation coefficient (IACC). It is based on the inter-aural cross correlation function (IACF):

$$\mathrm{IACF}(\tau) = \frac{\int_{0\mathrm{ms}}^{80\mathrm{ms}} h_l(t) h_r(t+\tau) dt}{\sqrt{\int_{0\mathrm{ms}}^{80\mathrm{ms}} h_l^2(t) dt \int_{0\mathrm{ms}}^{80\mathrm{ms}} h_r^2(t) dt}}, \qquad (4)$$

with the binaural room impulse response $h_{l,r}$. The IACC is defined as the maximum absolute value within $\tau = \pm 1\,\mathrm{ms}$:

$$\mathrm{IACC} = \max_{\forall \tau \in [-1\mathrm{ms};1\mathrm{ms}]} |\mathrm{IAFC}(\tau)|. \qquad (5)$$

A linear regression yields $f(1-\mathrm{IACC}) = 7.61(1-\mathrm{IACC}) - 2.55$ a better correlation with the experimental data ($R^2 = 0.76$, cf. Fig. 7). The $1-\mathrm{IACC}$ model works qualitatively well and markers are within or – if the direction of $\boldsymbol{r}_\mathrm{E}$ did not hit – at the same distance as the corresponding confidence ellipse for most conditions, cf. Fig. 4 and Fig. 6.

5. Comparison to previous results

Generally, listeners reported a natural sounding acoustic scenery. Although head movements were not taken into account in the rendering, the auditory feedback of the listeners-controlled omnidirectional sound source was convincing. Most listeners reported to not need any visual feedback of the surrounding room for the marker placement.

Zaunschirm et al. in [12] studied the perception of auditory objects created in a real setup using the icosahedral loudspeaker array (IKO), which is able to create directivities up to $n = 3$. In the experiment, listeners indicated the positions of the auditory objects directly as Cartesian coordinates with a marker on a screen. Moreover, a virtualization of the setup was studied and compared to the real IKO. This virtual IKO (vIKO) employs multiple input multiple output convolutions comprising measured RIRs and dynamic binaural Ambisonic rendering. The positioning of listener and directional sound source corresponded to P_1 with a similar excitation signal (1.5-s-long pink noise burst linearly faded in and out by 0.5 s). Fig. 8 compares the obtained 95%-confidence areas of the IKO and vIKO with the results of our experiment with $n = 3$ and S_1. Although ellipses do not coincide for all beam orientations, the match is astonishing. Especially if one considers that compared to our simple sound field rendering, the measured RIR used within the vIKO include the real directivity pattern of the IKO, as well as all frequency-dependent reflection paths of the room.

6. Conclusion

We presented a listening experiment studying how directivity and orientation of a virtual directional sound sources influences the localization of auditory objects.

The auralization is simplistic and includes frequency-independent directivity patterns, an image-source model for direct sound and early reflections, and a feedback delay network for the diffuse reverberation. To evaluate the localization of auditory objects created by the fixed directional source, the listeners had to match its position with a freely movable omnidirectional source inside the same auralization.

The obtained results indicate that for source position P_1 auditory objects are shifted along the direction indicated by the beam orientation. Distinct signal-dependent lateral shifts are obtained for lateral orientations up to the 3^{rd} order. For 5^{th}-order beam patterns, this lateral shift saturates and the auditory objects tend to split because the individual echo threshold is reached.
Medial orientations influence the auditory distance by controlling the direct-to-reverberant energy ratio which is known to be a cue for distance perception. Agreeing with [4] for a listener-facing orientation $\theta = 180°$, the increase of the directivity order yields a decrease of the auditory distance and for directivities facing away from the listener $\theta = 0°$ the auditory distance increases. If the reflection of the adjacent wall is strong enough, it can prevail the direct sound and dominate the distance perception. This is seen in the results for beam orders $n = (3, 5)$ at $\theta = 0°$.

Successful modeling of the experimental results was presented. For modeling the direction we used the extended energy vector. It is calculated with a simple specular model of the auralized room and includes a signal-dependent weighting of the sound paths. The auditory distance is modeled with the inter-aural cross correlation coefficient which is typically used to quantify the apparent source width and includes the binaural RIR.

Finally, the obtained results for the 3^{rd}-order beam patterns are compared to the perception of a real directional sound source and a measurement-based simulation thereof which was studied in a previous experiment. Although our virtualization is simplistic the matching results indicate that it sounded convincing and our assessment method worked well.

Acknowledgments

We thank all listeners taking part in the experiment. Special thanks go to Franz Zotter, who came up with the matching idea. Our research was partly funded by the Austrian Science Fund (FWF) project nr. AR 328-G21, Orchestrating Space by Icosahedral Loudspeaker (OSIL).

References

[1] F. Zotter, M. Zaunschirm, M. Frank, and M. Kronlachner, "A beamformer to play with wall reflections: The icosahedral loudspeaker," *Computer Music Journal*, vol. 41, no. 3, pp. 50–68, 2017.

[2] F. Zotter and M. Frank, "Investigation of auditory objects caused by directional sound sources in rooms," *Acta Physica Polonica*, vol. 128, no. 1-A, 2015.

[3] F. Wendt, G. K. Sharma, M. Frank, F. Zotter, and R. Höldrich, "Perception of Spatial Sound Phenomena Created by the Icosahedral Loudspeaker," *Computer Music Journal*, vol. 41, pp. 76–88, mar 2017.

[4] F. Wendt, F. Zotter, M. Frank, and R. Höldrich, "Auditory Distance Control Using a Variable-Directivity Loudspeaker," *Applied Sciences*, vol. 7, no. 7, p. 666, 2017.

[5] J. Linke, F. Wendt, F. Zotter, and M. Frank, "How masking affects auditory objects of beamformed sounds," in *DAGA 2018*, 2018.

[6] F. Zotter, M. Frank, F. Wendt, and M. Zaunschirm, "Auditory object localization of the variable-directivity icosahedral loudspeaker," in *Fortschritte der Akustik*, vol. 2, pp. 1–3, 2016.

[7] P. Stitt, S. Bertet, and M. van Walstijn, "Extended energy vector prediction of ambisonically reproduced image direction at off-center listening positions," *J. Audio Eng. Soc.*, vol. 64, no. 5, pp. 299–310, 2016.

[8] E. Kurz and M. Frank, "Prediction of the listening area based on the energy vector," in *Proceedings of the 4th ICSA, Graz*, (Graz), pp. 52–58.

[9] J. Daniel, *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*. PhD thesis, Université Paris 6, 2001.

[10] B. Bernschütz, "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100," *Fortschritte der Akustik – AIA-DAGA 2013*, pp. 592—-595, 2013.

[11] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *Acta Acustica united with Acustica*, vol. 91, no. 3, pp. 409–420, 2005.

[12] M. Zaunschirm, M. Frank, and F. Zotter, "An Interactive Virtual Icosahedral Loudspeaker Array," in *Fortschritte der Akusitk*, (Aachen), 2016.

[13] W. M. Hartmann, "Localization of sound in rooms," *The Journal of the Acoustical Society of America*, vol. 74, no. 5, pp. 1380–1391, 1983.

[14] B. Rakerd and W. M. Hartmann, "Localization of sound in rooms, II: The effects of a single reflecting surface.," *The Journal of the Acoustical Society of America*, vol. 78, pp. 524–533, 1985.

How the perception of moving sound beams is influenced by masking and reflector setup

J. Linke[1], F. Wendt[2], M. Frank[2], and F. Zotter[2]
1 *University of Music and Performing Arts Graz, Austria, Email: julian-michael.linke@student.kug.ac.at*
2 *Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Austria*
Email: {wendt, frank, zotter}@iem.at

Abstract

The icosahedral loudspeaker array (IKO) uses strongly focused sound beams to "orchestrate" reflecting surfaces resulting in the perception of auditory objects at distinct locations in the room. Previous studies revealed that stationary sounds are localized more distant from the IKO in comparison to transient sounds due to the precedence effect. Nevertheless, we could show in a direct comparison using static sound beams that the application of maskers can partly suppress the precedence effect resulting in an increased distance for transient sounds. This contribution extends the study towards trajectories of moving sound beams and also investigates the influence of a concave setup of reflectors around the IKO that is typically used in practice.

1. Introduction

In electroacoustic music, the icosahedral loudspeaker array (IKO) is used to create auditory objects at distinct locations in a room by varying its directivity pattern and orientation [1]. An earlier study [2] could show that the perception of auditory objects strongly depends on the signal, how sound beams are configured, and on the room situation. The perception of auditory objects can be divided into two categories:
(1) static objects created by a fixed sound beam and
(2) auditory movements caused by beam trajectories such as turns. To enforce the percepts and increase the number of discretely localizable directions, the use of reflectors in a concave setup behind the IKO has been proven as artistic practice, affecting the sound-propagation paths, cf. Fig. 1.

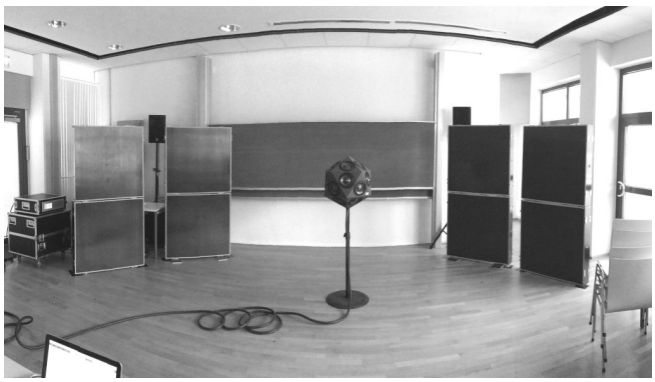


Fig. 1: Experimental setup with the IKO and two reflector pairs in the IEM lecture room.

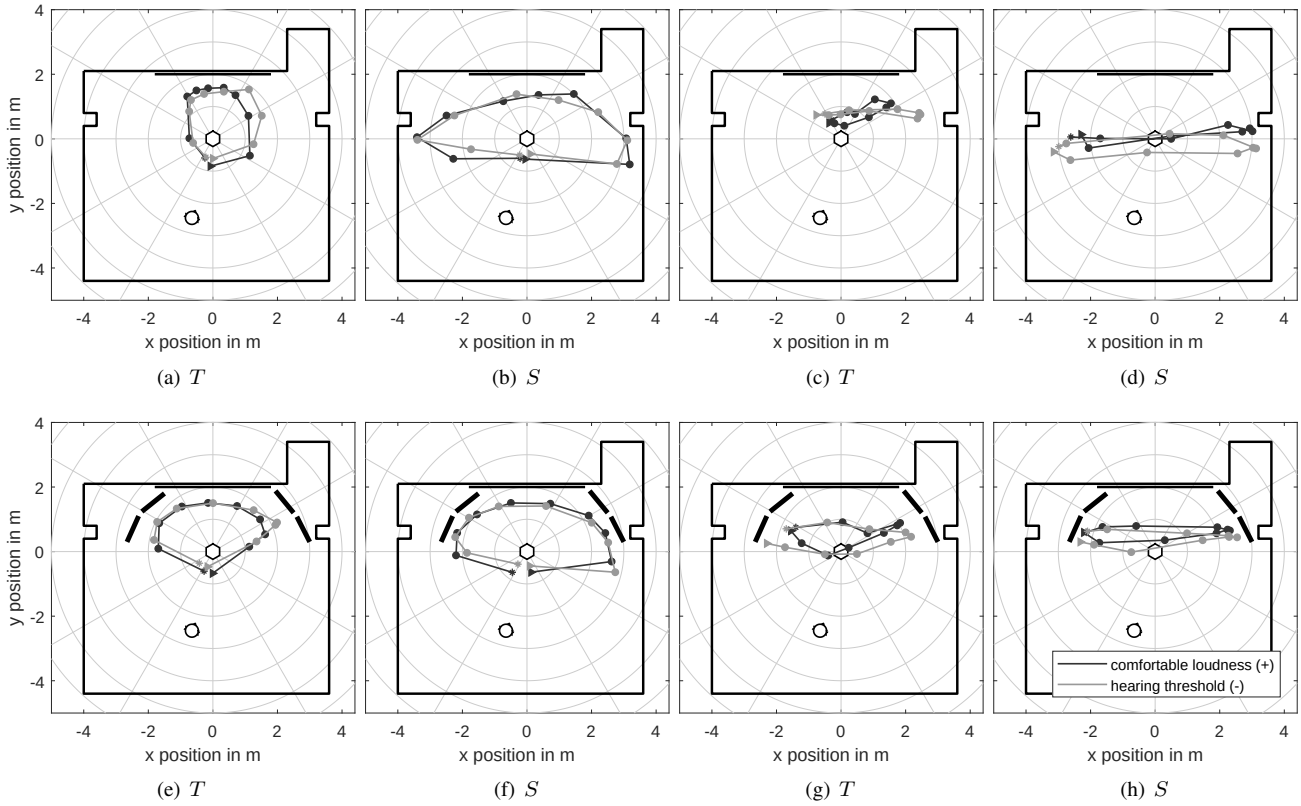
The perceived direction of auditory objects of both categories is affected by the precedence effect. Depending on the signal type and time/strength of single wall reflections compared to the direct sound, they can appear more or less lateralized from the IKO. Studies revealed that the addition of maskers weaken the precedence effect [3]. In a preceding study [4] we could apply this effect to static beamformed sounds with reflectors. By adding an omnidirectional masker, transient auditory objects appeared to be more lateralized and distant from the IKO.

This contribution extends the previous study by investigating moving sound beams and the auditory objects they cause. We present a listening experiment studying the influence of: (1) two different masking levels, (2) stationary and transient sounds, (3) two typical setups for music performances, and (4) two types of beam trajectories.

2. Experimental Setup

The listening experiment was carried out in the IEM lecture room ($6.8\,\mathrm{m} \times 7.6\,\mathrm{m} \times 3\,\mathrm{m}$ with $T_{60} = 0.56\,\mathrm{s}$). The IKO was placed near the blackboard of the lecture room with the listener in front of it, slightly offset at a distance of $2.6\,\mathrm{m}$.
For the transient sounds T, a sequence of regular bursts was used as signal. In [4] different masking types had been studied and transient sounds had been found to be lateralized most if their level had been close to the room noise in quite. Thus, T was played back either just above hearing threshold representing the masked transient sound T_- or at comfortable level representing the non-masked transient sound T_+.

ISBN
978-3-9812830-9-9
The authors hold the copyrights of all of content in this manuscript. They have granted the VDT the right to publish it in print, online and as PDF download on https://tonmeistertagung.com



(a) T (b) S (c) T (d) S

(e) T (f) S (g) T (h) S

Fig. 2: Mean values for each time interval coded in grayscale for both playback levels for stationary sounds S and transient sounds T. Upper row: without reflectors. Lower row: with reflectors. Columns 1 and 3: single beam rotating counterclockwise by 360° (R). Columns 2 and 4: linear fade between 2 opposing beams (F).

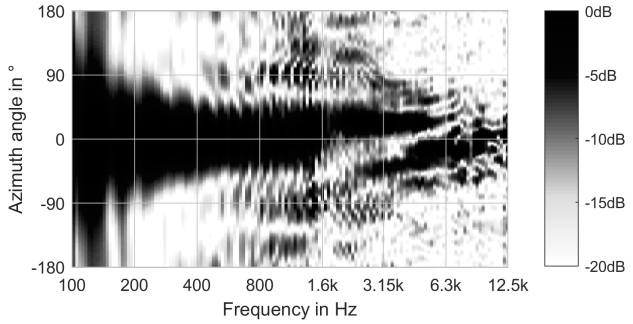


Fig. 3: Horizontal cross-section through measured 3^{rd}-order max-r_{E} beampattern [5, 6] of the IKO normalized to its maximum.

For stationary sounds S pink noise was used. Similarly than for the transient sound, the stationary sound was evaluated both at comfortable loudness (S_+) and just above hearing threshold (S_-).

All sounds were tested with two beam trajectories using 3^{rd}-order beams at IKO's horizontal plane, cf. Fig. 3. The trajectory R consisted of a single beam rotating counterclockwise by 360° in the horizontal plane and starting at $\theta = 180°$ (0° axis was facing the blackboard behind the IKO). Trajectory F represented a pendulum and consisted of a linear fade between two opposing beams at $\theta = \pm 90°$. The trajectory started with the beam at $\theta = 90°$ (left), faded to the beam at $\theta = -90°$ (right), and then back to $\theta = 90°$.

The continuous movement (rotating/fading) of the trajectory lasted 5 s and was played back in a loop. The listeners were asked to localize the auditory objects and adjust ten markers in 2D Cartesian coordinate system to the perceived location in successive half-second intervals during playback. Markers were flashing successively at the associated playback time, and they could be moved using a mouse on a graphical interface showing the floor plan of the test setup. Playback was repeated until listeners are satisfied with the match between marker placement and what they heard.

Both signals, their levels, and both beam trajectories were tested in full permutation yielding 8 conditions. To test the influence of reflectors on the perception, the experiment was performed once with and once without reflectors consecutively by each listener. Each time, all 8 conditions were tested twice in random order. Half of the listeners started the experiment with reflectors and the other half without. For reflector placement, two reflector pairs were arranged at a distance of approximately 2.5 m at both sides of the IKO, cf. Fig. 1. Ten experienced and normal-hearing listeners participated in the test, which lasted about 1 hour.

3. Results

For representation of the experiential data, we calculate two-dimensional means of the obtained positions for each time interval of a condition. In Fig. 2 the 10 mean values are shown for both playback levels (black: comfortable loudness; gray: hearing threshold) of the signal type S/T for the beam trajectory R/F with/without reflectors. Markers ▷ and ∗ indicate starting and ending position of the trajectory, respectively. The hexagon and the head symbol indicate position of the IKO and the listener, respectively and the black bars indicate the positioning of the reflectors.

A measure of lateralization is found by calculating the maximum lateralization range $\Delta x = |x_{\max}| - |x_{\min}|$ of Cartesian values for each listener.
Let us start with conditions without reflectors shown in Fig. 2 (a-d). An analysis of variance (ANOVA) of the lateralization range Δx reveals the *trajectory type* not to be a significant parameter ($p = 0.99$). By reducing the level of the excitation signals, the direct sound is masked by the room noise and according to [4] we expect higher lateralizations for the soft transient sound (gray) compared to the sounds at comfortable level (black). This masking effect is obtained only for beam trajectory F and lateralizations of T_- are just slightly more lateralized then the signal without masker (T_+), cf. Fig. 2 (e). For stationary sounds, no masking effect is obtained and we can conclude that *masking* does not necessarily increase the lateralization of transient auditory objects created by moving sounds.
Obviously, a more prominent parameter is the *signal type*. Auditory objects of the stationary sound S are lateralized all the way to the wall, whereas transient sounds T stick at the IKO. For both trajectories, the mean lateralization range Δx of the stationary sound is twice as large as for the transient sound, cf. Fig. 4.

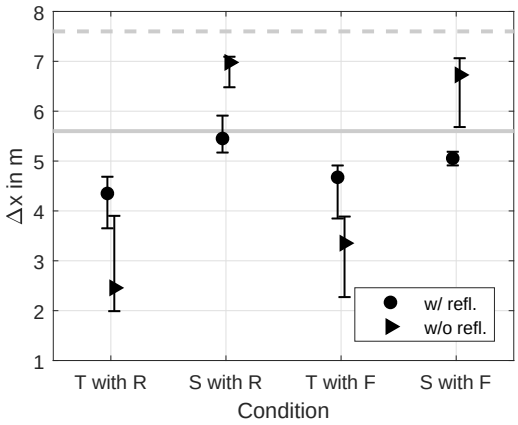


Fig. 4: Means and 95% confidence intervals of the lateralization range Δx combined for both signal levels for the four signal types give a comparison of the setups with and without added reflectors. Horizontal lines indicate the maximum distances between both reflector pairs (solid) and both lateral room walls (dashed).

The results obtained with additional reflectors are shown in Fig. 2(e-f). Similarly than without reflectors, the trajectory type is not significant and the masking effect for transient sounds is negligible. Interestingly, however, the influence of the *signal type* is not as pronounced as for the setup without reflectors. Although still significant, the differences of corresponding means are reduced and the sizes of 95% confidence intervals indicate a lower scattering for conditions with additional reflectors, cf. Fig. 4. This is because auditory location of stationary sounds is laterally limited by the reflectors, while the lateralization for transient ones is extended. Moreover, transient sounds are distributed more uniformly at both sides of the IKO with reflectors, cf. Fig. 2 (c) vs. 2 (g).

4. Conclusion

We studied the influence of maskers and reflectors on the perception of transient and stationary auditory objects created by moving soundbeams. In contrast to [4] for moving sound beams, the use of maskers did not yield the desired effect as lateralization of transient auditory objects barely increased. The use of reflectors on the other hand had a significant impact on the lateralization for all conditions and yielded a more uniform and less signal-dependent perception. This is explained by an increase of the reflected energy for transient sounds which weakens the precedence effect. Moreover, auditory objects of stationary sounds are laterally limited by the reflector positions. In this way, we found scientifically evident reason in the artistic practice that has evolved for the IKO over the recent years.

Acknowledgments

We thank all listeners taking part in the experiment. Our research was partly funded by the Austrian Science Fund (FWF) project nr. AR 328-G21, Orchestrating Space by Icosahedral Loudspeaker (OSIL).

5. References

[1] G. K. Sharma, F. Zotter, and M. Frank: Orchestrating wall reflections in space by icosahedral loudspeaker: findings from first artistic research exploration. In Proc. ICMC/SMC, Athens, pp. 830-835, 2014.

[2] F. Wendt, G. K. Sharma, M. Frank, F. Zotter and R. Höldrich: Perception of Spatial Sound Phenomena Created by the Icosahedral Loudspeaker. Computer Music Journal, pp. 76-88, 2017.

[3] Y. Chiang and R. L. Freyman: The influence of broadband noise on the precedence effect. The Journal of the Acoustical Society of America, vol. 104, pp. 3039-3047, 1998.

[4] J. Linke, F. Wendt, F. Zotter and M. Frank: How Masking Affects Auditory Objects Of Beamformed Sounds. DAGA - 44. Jahrestagung für Akustik, 2018.

[5] J. Daniel: Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia. PhD thesis, Université Paris 6, 2001.

[6] F. Zotter and M. Frank: All-round ambisonic panning and decoding. AES: Journal of the Audio Engineering Society, vol. 60, no. 10, pp. 807-820, 2012.

Hörversuche zur Entwicklung eines neuartigen Mehrkapsel-Mikrofons

Hendrik Paukert[1], Jonathan Ziegler[1, 2], Andreas Koch[1]
[1]*Hochschule der Medien Stuttgart, Germany, Email: paukert@hdm-stuttgart.de*
[2]*Eberhard Karls Universität Tübingen, Germany Email: zieglerj@hdm-stuttgart.de*
[1]*Hochschule der Medien Stuttgart, Germany, Email: kocha@hdm-stuttgart.de*

Abstract

Für die Entwicklung eines neuartigen digital prozessierten Mikrofonarrays, haben wir im Vorfeld einen Hörversuch zur Untersuchung verschiedener Störgeräusche vorgestellt [1]. Diese dienten dem besseren Verständnis, welchen Faktoren bei der Entwicklung der Algorithmen die größte Beachtung geschenkt werden muss. In diesem Teil werden die Hörversuche, deren Audiodatengenerierung sowie die Ergebnisse zur Einschätzung und Abgrenzung verschiedener Fremd- und Eigenalgorithmen zur Entwicklung des Mehrkapsel-Mikrofonarrays im Haupteinsatzgebiet der Sprache vorgestellt. Hierzu wird das Open Source Tool "WAET" [2] für die menschliche-, sowie die STOI-Methode [3] für die algorithmische Bewertung genutzt.

1. Vorüberlegungen

Der vorausgegangene Hörversuch zur Untersuchung von Störgeräuschen wurde komplett in Max/MSP [2] programmiert und von den Teilnehmern offline an einem Laptop bearbeitet. Um den aktuellen Test nun schneller aufbauen und auch online einer größeren Anzahl an Teilnehmern anbieten zu können, sollte das frei zugängliche „Web Audio Evaluation Tool" - kurz „WAET" [3] genutzt werden. Dieses Tool bietet u.a. die Möglichkeit, einen den ITU-R BS.1534-Empfehlungen [4] entsprechenden Test erstellen zu können. Im Detail kann damit eine zufällige Prüfdatenausgabe, unterschiedliche Skalenbeschriftungen automatische Lautstärkennormalisierung, die Einbindung von Kommentarfeldern, versteckter Anker- und Referenzdatei und das Prüfen auf versehentlich doppelt eingebundene Dateien umgesetzt werden. Auch können weitere Daten wie z.B. Bearbeitungsdauer, Anzahl der Abspielwiederholungen gespeichert und der Hörversuch den Teilnehmern von einem php-fähigen Server browserbasiert angeboten werden.

Auf Grund der zu erwartenden großen Datenmengen, sollte außerdem eine zeitsparende und weitgehend automatisierte Auswerteroutine möglich sein und, um Überbeanspruchung der Probanden zu vermeiden, die Bearbeitungsdauer von 30 min. im Mittel nicht überschritten werden.

1.1 Vor- und Nachteile von Online-Hörversuchen

Durch online zugängliche Hörversuche kann eine große Anzahl an möglichen Teilnehmern erreicht werden. Diese können den Bearbeitungszeitpunkt außerdem frei wählen und den Test bequem an einem beliebigen Ort bearbeiten. Auch für den Versuchsdurchführenden reduziert sich der Aufwand für Auf- und Abbau der Testumgebungen an einem spezifischen Ort, ebenso reduziert sich der Organisationsaufwand.

Nachteile sind jedoch die indirekten Hilfe- bzw. Assistenzmöglichkeiten bei auftretenden Fragen und Problemen, sowie die schlechte Kontrollmöglichkeit des verwendeten Equipments. Ein Online-Versuch sollte also möglichst übersichtlich und selbsterklärend aufgebaut und die Aufgabenstellung dem entsprechend geeignet sein. Durch die in diesem Zusammenhang teilweise geringen Unterschiede unserer Testdaten in Bezug auf Nachhallzeit, Klangveränderung und Bildung von Artefakten, bat sich generell die Nutzung von Kopfhörern an. Auch erwies sich uns die Nutzung einer eingemessenen Abhöre in akustischen optimierter Umgebung als wenig relevant, da in diesem Hörversuch keine frequenz-vollumfänglichen Musik-bearbeitungen, sondern Sprache abgefragt wird. Weiterhin kann davon ausgegangen werden, dass der jeweilige Proband für die Bearbeitung des Hörversuches nur einen Kopfhörer nutzt und somit die Bewertung der Audiodaten zueinander stimmig ist. Die Typenbezeichnung der Kopfhörer fragten wir aus Interesse zusätzlich ab.

2. Zu untersuchende Algorithmen

Haupteinsatzgebiet des neu entwickelten Mikrofonarrays werden Konferenzen sein. Hierzu wurde an der Hochschule der Medien ein Tracking-Algorithmus zur Detektion und Verfolgung jeweiliger Sprecher entwickelt [5], [6]. Eine aus drei Mikrofonkapseln synthetisierte Nieren- oder Superniercharakteristik, kann in Echtzeit ein aktuelles Sprachereignis erfassen, verfolgen und so von unerwünschten Schallereignissen besser freistellen.

Um diesen Freistellungseffekt weiter erhöhen zu können, soll zusätzlich ein Beamformer- bzw. Dereverb-Algorithmus implementiert werden. Hierzu stand prozessiertes Audiomaterial des Beamformers von Bernfried Runow [7], [8], zwei Dereverb-Plug-Ins und ein Spaced-Array-Konferenz-Komplett-System zur Verfügung. Durch die fixe Architektur der verschiedenen Algorithmen, konnte jedoch nur Runows Beamformer alle drei Mikrofonsignale des neuen Mehrkapselmikrofons nutzen und verfügte somit theoretisch über das größte Potential. Ein Plug-In konnte so nur das Summen-Trackingsignal, das andere nur zwei Signale des Mikrofons nutzen. Eine weiterhin geprüfte Spaced-Array-Konferenzanlage, verfügt über ihr eigenes räumliches Arraymikrofon mit 24 Kapseln und dient dem direkten Vergleich zu einem auf dem Markt bereits erhältlichen System.

ISBN
978-3-9812830-9-9
Der Autor hält die Urheberrechte für den gesamten Inhalt seines Manuskripts. Er überlässt dem VDT das Recht auf Veröffentlichung in gedruckter Form, online oder als PDF-Download von https://tonmeistertagung.com

Nach Möglichkeit wurden die verschiedenen Algorithmen auch in verschiedenen Stärkeeinstellungen abgeprüft. Inklusive der Ankerdatei, dem bewusst verschlechterten Signal, und der Referenzdatei, dem optimalen Signal, umfasste der Hörversuch 9, jeweils in zwei Sprachen zu bewertende Audiofiles. Angesichts der gesetzten Bearbeitungsobergrenze von 30 Minuten, erwies sich diese Anzahl bei mehreren Vorversuchen bereits als ausreichend.

Bezeichnung	Beschreibung
Anker	Ankerdatei mit 3,5kHz-Lowpass und einer der Kategorie entsprechenden Degradierung durch hohe Nebengeräusche, Rauschen oder hohen Hallanteil. Dieses Signal muss vom Probanden als „schlecht“ bewertet werden.
Referenz	Nach Möglichkeit optimales Signal (trockene Studiosprachaufnahme), ohne Hall, Nebengeräusche oder Rauschen. Dieses Signal muss als „gut“ bewertet werden.
Kugel	Kugelsignal des Prototypenmikrofons
Runow 50%	B. Runows-Beamforming-Algorithmus in mittlerer Stärke, gespeist mit den drei einzelnen Kapselsignalen und der vom Tracker ermittelten Richtungsinformation.
Runow 100%	B. Runows-Beamforming-Algorithmus in voller Stärke, gespeist mit den drei einzelnen Kapselsignalen und der vom Tracker ermittelten Richtungsinformation.
Plug-In Nr.1 50%	DAW-Plug-In in mittlerer Stärke, gespeist mit zwei geführten Mono-Signalen.
Spaced-Array	Gesamtsystem mit eigenem räumlichen Mikrofonarray und schwacher Einstellung des produkteigenen Algorithmus. Stärkere Einstellungswerte verschlechterten das Signal enorm und mussten daher vom Test ausgeschlossen werden.
Plug-In Nr.2 50%	DAW-Plug-In in mittlerer Stärke, gespeist mit geführtem Mono-Signal.
Plug-In Nr.2 stark	DAW-Plug-In in starker Einstellung, gespeist mit geführtem Mono-Signal. Noch stärkere bzw. maximale Einstellung verschlechterten in unserem Fall die Signalqualität sehr stark.

Tab. 1: Übersicht der genutzten Signale

3. Generierung der Audiodaten

Zur Generierung der Audiodaten nutzten wir den neu entwickelten Aufbau einer reproduzierbaren Konferenzumgebung: über mehrere Iterationen hinweg konnten wir hier, auch durch die detaillierte BA von Robin Hirt [9], eine „virtuelle Konferenz“ gestalten. Mit diesem Mikrofonteststand kann über ein Lautsprechersetup mit definierten Abständen und Zuspielszenarien die Wirkungsweise der verschiedenen Algorithmen bzw. Algorithmen-Revisionen geprüft werden.

Insgesamt werden 10 Lautsprecher in zwei Ringen bzw. Radienabständen vom zu prüfenden Mikrofon aufgebaut. Über die vier Lautsprecher des inneren Rings werden die nahezu reflexionsfreien Sprachdateien mehrerer Personen getrennt und auch parallel ausgegeben. Die Lautsprecher des äußeren Rings erzeugen ein- und mehrkanalige Atmogeräusche wie z.B. Straßenlärm oder das Öffnen und Schließen einer Tür. Weiterhin befindet sich, nahe am Mikrofon, ein einzelner Lautsprecher, welcher z.B. Papier-, Tassen- und Stiftgeräusche ausgibt, sowie in 4 Metern Abstand ein weiterer Lautsprecher zur Generierung von sehr indirekten und räumlichen Signalen. Zur Simulation von Körperschall, z.B. verursacht durch Vibrationen eines Laptops oder Smartphones, kann zur Simulation optional ein kleiner E-Motor am Tisch befestigt werden. Durch eine elastische Aufhängung des Mikrofons ist hier allerdings schon von einer guten Körperschallentkopplung auszugehen.

Zur Nutzung im Hörversuch wurde letztendlich ein 3s kurzer Zeitabschnitt der Aufnahmen in deutscher und englischer Sprache ausgesucht. Längere Abschnitte machten aus Gründen der Vergleichbarkeitsschwäche der menschlichen Klangwahrnehmung und der dadurch folgenden Erhöhung der Versuchsdauer keinen Sinn.

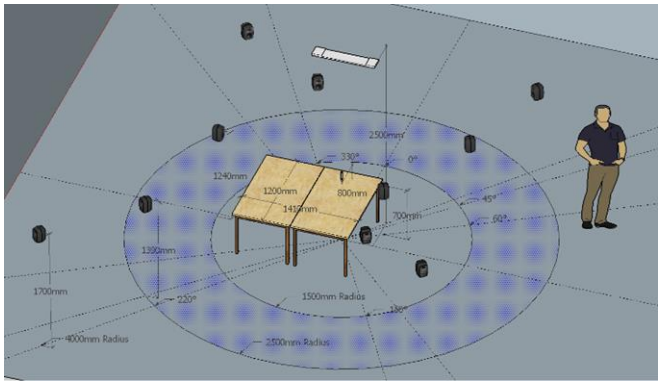


Abb. 1: Aufbauskizze der virtuellen Konferenz [10]: Zu sehen sind die beiden konzentrischen Lautsprecherringe, ergänzende Nah- und Fernlautsprecher, das kleine Prototypenmikrofon in der Mitte sowie das Konferenzsystems in Deckenmontage.

Abb. 2: 360°-Aufnahme des Genelec 1029A [11]-Lausprechersetups der virtuellen Konferenz: Die Raummaßen betragen 8,3x8,2x3,8m, die RT60 2,31s im Bereich von 500-1000Hz 2,31s (leerer und akustisch unbehandelter Raum).

4. Versuchsinterface

Das WAET stellt verschiedene Interface-Arten wie z.B. AB-, ABX- und Checkbox-Verfahren zur Verfügung. Anstatt des klassischen Mehrfachschieberegler-Interface der klassischen

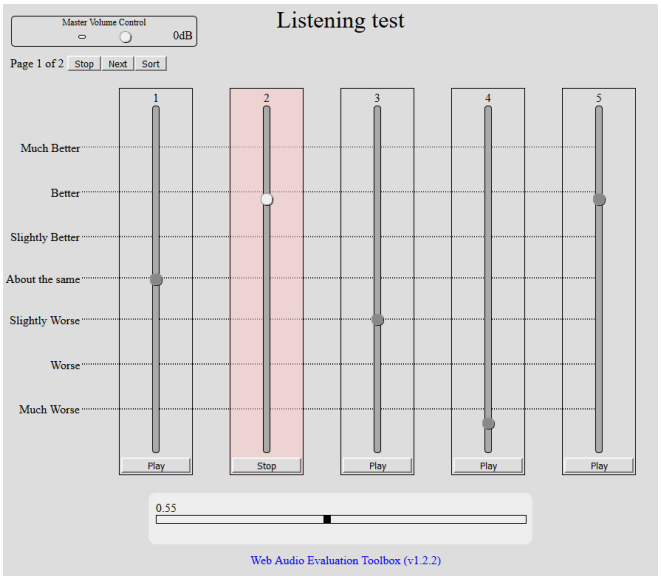


Abb. 3: Mehrfachschieberegler-Interface des Web Audio Evaluation Tool: jedes der fünf angebotenen Signal verfügt über einen eigenen Schieberegler

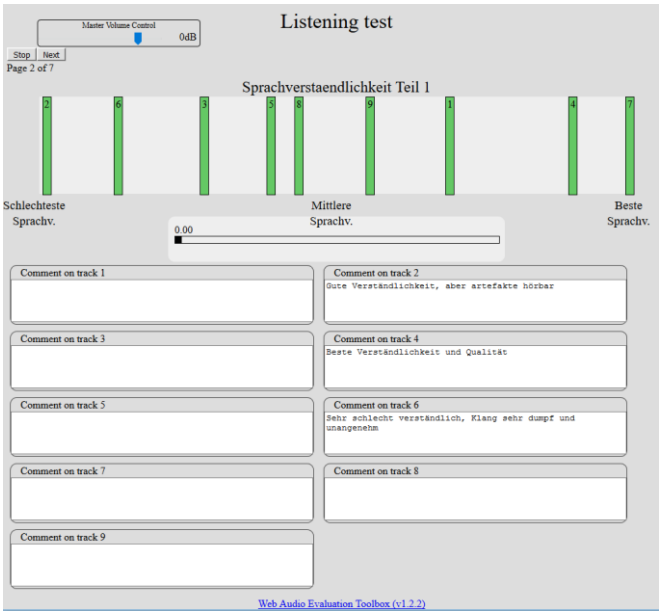


Abb. 4: einachsiges Hörversuchsinterface mit dem Web Audio Evaluation Tool: Jeder der grünen Balken repräsentiert eine auf der Skala zu positionierende Sprachdatei. Weitere Auffälligkeiten können anhand der Kommentarfelder weitergegeben werden.

MUSHRA-Methode (**Abb. 3**), fiel die Entscheidung zu Gunsten der APE-Variante (Audio Perceptual Evaluation) [12]. Wie in **Abb. 4** ersichtlich, wird hierbei nur eine Achse genutzt, auf der alle Soundfiles als Schiebebalken repräsentiert sortiert werden müssen. Das Interface ist dadurch kompakter und lässt pro Prüfabschnitt mehr Platz für Kommentarfelder oder andere Zusatzoptionen. Außerdem schien uns der einachsige Aufbau die Relation der Soundfiles zueinander mehr in den Vordergrund zu stellen, als dies bei getrennten Achsen der Fall wäre [Vgl. 13]. Durch die Textfelder konnten die Probanden auch möglicherweise gar nicht abgefragte, ihnen aber als wichtig erscheinende Informationen weitergeben. Um ein mögliches Biasing bzw. bewussten oder unterbewussten Manipulationen entgegenzuwirken, wurden die Audiodaten pro Abschnitt außerdem zufallsverteilt ausgeben.

Auf Grund des weiten Wertebereichs des einachsigen Aufbaus jedoch, kann bei nicht vollständiger Nutzung der Skala eine Verzerrung der Standardabweichung σ auftreten: ein Proband könnte seine Daten z.B. nur in der unteren Hälfe der Skala anordnen, ein anderer jedoch in der oberen Hälfte. Durch diesen Offset würde die σ aller Daten nun fälschlicherweise ansteigen. Das WAET bietet hier an, gewünschte Soundfiles in einstellbaren (Toleranz-)Bereichen an den Skalenenden anordnen zu müssen. Das nachträgliche Normalisieren der vielen Ergebniswerte kann so vermeiden werden. In unserem Falle musste also die Ankerdatei, welche vom Probanden definitiv als schlechteste Datei erkannt werden muss, am linken Rand, und die Referenzdatei, am rechten Rand angeordnet werden. Diese Maßnahme stellt auch sicher, dass jeder Teilnehmer den Test aktiv und aufmerksam durchführt und ihn bis zum Erkennen und dem korrekten Einordnen der Dateien nicht fortführen kann. Die restlichen Sprachdateien können nun zwischen beiden Werten verteilt werden. Leider konnte die angezeigte Fehlermeldung bei Nichterkennen nicht zu einem verständlichen Hinweis umformuliert werden.

Weiterhin erlaubt die WAET-Testumgebung eine Überprüfung ob jedes Audiodatei komplett abgespielt und bewegt worden ist. Das ist wichtig, da die Dateien, abgesehen von Anker- und Referenzdatei, sonst an ihren zufallsgenerierten Startpositionen stehen bleiben können.

5. Beschreibung der Prüfabschnitte und Ergebnisse

5.1. Einleitende Fragen

Zu Beginn des Hörversuches wurde eine einleitende Beschreibung angezeigt und die verwendeten Kopfhörer, das Alter der Teilnehmer sowie die audiospezifischen Erfahrungsbereiche abgefragt. Wie in **Abb. 5** erkennbar, gaben die meisten Teilnehmer an, ein oder mehrere Instrumente zu spielen und sich mit dem Aufnehmen und Bearbeiten von Musik, Foleys oder Sprache zu befassen. Ein weiterer großer Teil gab an, Musikliebhaber und Konsument von höherwertiger Audiotechnik zu sein. Da auch Mehrfachnennungen möglich waren, wurden oft zwei oder drei Kategorien parallel genannt, siehe **Abb. 6**.

Bei den genutzten Kopfhörern wurden die Marke Beyerdynamic [14], gefolgt von Sennheiser [15] und AKG [16], am meisten genannt (**Abb. 7**). Die vielen Einzelmeldungen unterschiedlicher Marken wurde unter „sonstige Nennungen“ zusammengefasst. Das Durchschnittsalter aller 59 Teilnehmer betrug 39 Jahre, allerdings mit einer hohen Streuung. Jüngster war 19 und ältester Teilnehmer 63 Jahre jung. Auch die durchschnittliche Bearbeitungsdauer schwankte stark, lag im Mittel allerdings bei knapp 22 Minuten.

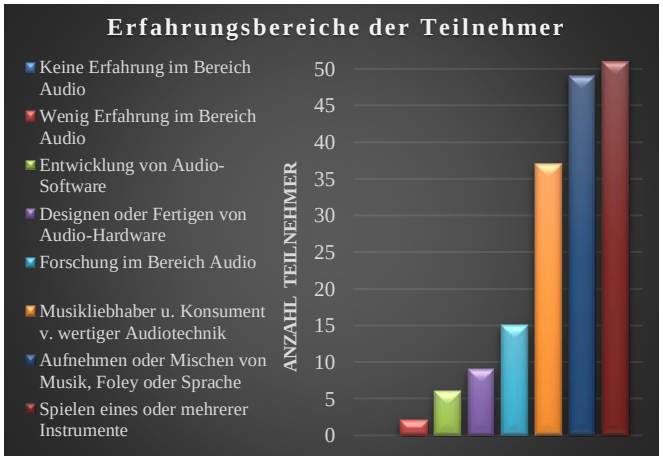


Abb. 5: Erfahrungsbereiche der Teilnehmer: am meisten kennen sich die Versuchsteilnehmer bei Audioproduktionen aus und spielen ein oder mehreren Instrumente.

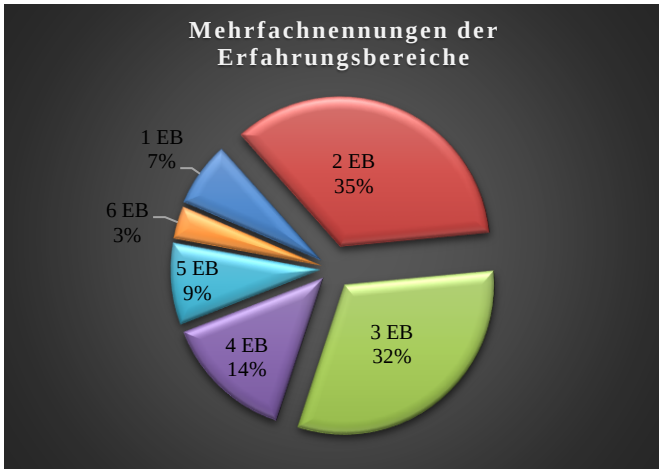


Abb. 6: Mehrfachnennungen der Teilnehmer: 67% nannten max. 3 Erfahrungsbereiche parallel.

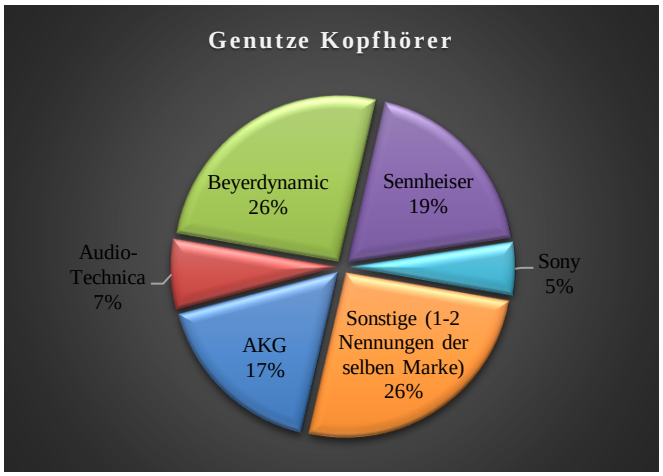


Abb. 7: Genutzte Kopfhörer der Teilnehmer: die größte Blöcke bilden mit insg. 62% Beyerdynamic, Sennheiser und AKG.

5.2 Beschreibung der Balkendiagramme

Die nachfolgenden Diagramme setzen sich aus den Bewertungen der abgeprüften Sprachdateien anhand von blauen Balken, sowie deren Standardabweichung σ anhand von roten Balken zusammen. Der Wert 0,0 steht für eine sehr schlechte und der Wert 1.0 für eine sehr gute Bewertung. Niedrige Werte der roten Balken zeigen eine niedrige σ auf. In diesem Zusammenhang liegen die σ-Werte der Anker- und Referenzdatei generell und trotz des Toleranzbereichs für die Zwangspositionierung an den Skalenenden auf einem sehr niedrigen Niveau. Für den ersten Prüfabschnitt „Training“ steht die 1,0 für eine als sehr hallig empfundenen Bewertung.

5.3 Training

Damit sich die Teilnehmer an die Testumgebung gewöhnen konnten, wurde ein vom Prüfumfang reduzierter Trainingsmodus implementiert. Anstatt 9 wurden nur 6 Soundfiles in einer Sprache, Deutsch oder Englisch, abgefragt. Auch wenn das Training dadurch nicht voll bewertet werden kann, können die Daten auf Grund der Menge der Teilnehmer und der abgefragten Eigenschaft „empfundene Halligkeit“ zumindest für eine erste Abschätzung genutzt werden. Grundsätzlich stellt das Kugelsignal das unbearbeitete und räumlichste Realsignal dar und wird nur von der künstlich verhallten Ankerdatei übertroffen (3s Nachhallzeit). Nachfolgend zur nachhalllosen Referenzdatei, wurde B. Runows Beamformer als sehr trocken bewertet.

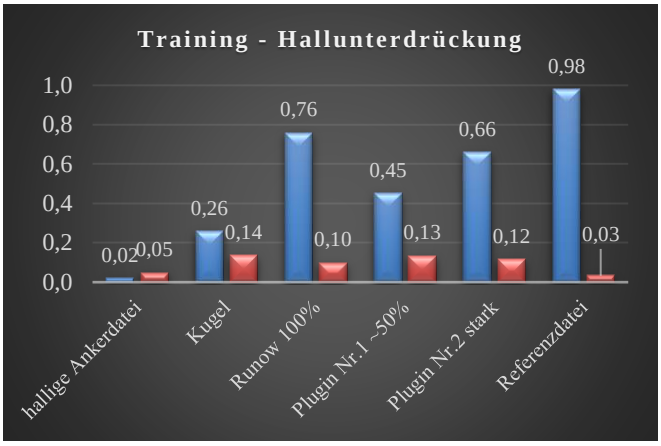


Abb. 8: Ergebnis „empfundene Halligkeit“ im Trainingsmodus: Runows Beamformer wird mit geringster σ und Halligkeit bewertet.

5.4 Sprachverständlichkeit

Eine Haupteigenschaft für (Konferenz-)Mikrofone ist die Verständlichkeit der Sprache. Hierzu wurden nun alle im Versuch vorkommenden Sprachdateien in deutscher und englischer Sprache abgefragt.

Bei beiden Durchläufen wurde B. Runows Beamformer im Maximaleinstellung als bestes bewertet, dicht gefolgt von Plug-In Nr.2. Deutlicher Verlierer ist hier das Spaced-Array-Konferenzsystem, dessen Sprachverständlichkeit deutlich unter einer übermäßigen Signalverschlechterung durch das interne Processing leiden musste.

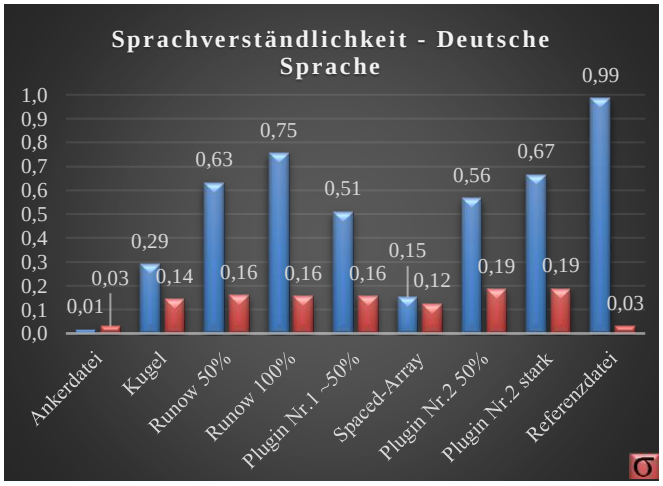


Abb. 9: Ergebnis „Sprachverständlichkeit“ in deutscher Sprache

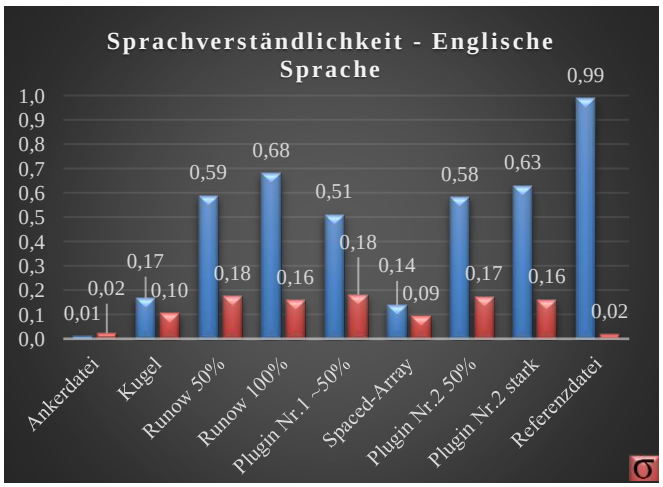


Abb. 10: Ergebnis „Sprachverständlichkeit“ in englischer Sprache

5.5 Unterdrückung der Störgeräusche

In diesem Abschnitt fragten wir die Sprachdateien mit Fokus auf die Störgeräuschunterdrückung ab. Das vorher weit abgeschlagene Konferenzsystem konnte hier deutlich aufholen. Sein starkes Processing versteht es, Nebengeräusche wirksam zu unterdrücken, im Vergleich zur Konkurrenz litt dadurch aber auch erheblich die Qualität, womit die erhöhte σ zusammenhängen könnte. Abgesehen von der störgeräuschsfreien Referenzdatei, führt B. Runows Algorithmus in Maximaleinstellung erneut die Auswertung an.

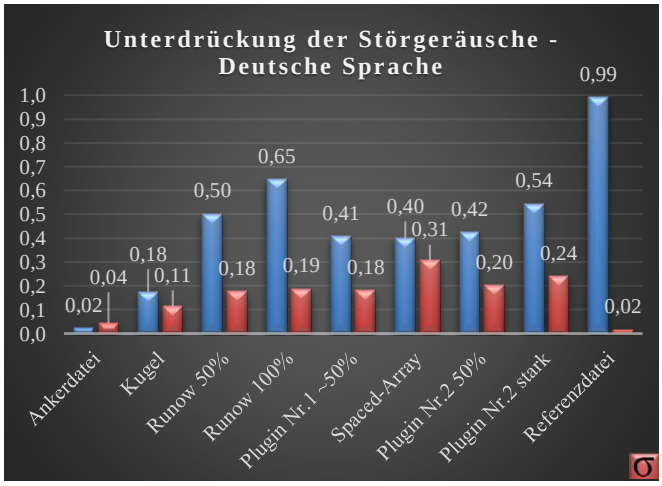


Abb. 11: Ergebnis „Unterdrückung der Störgeräusche“ in deutscher Sprache

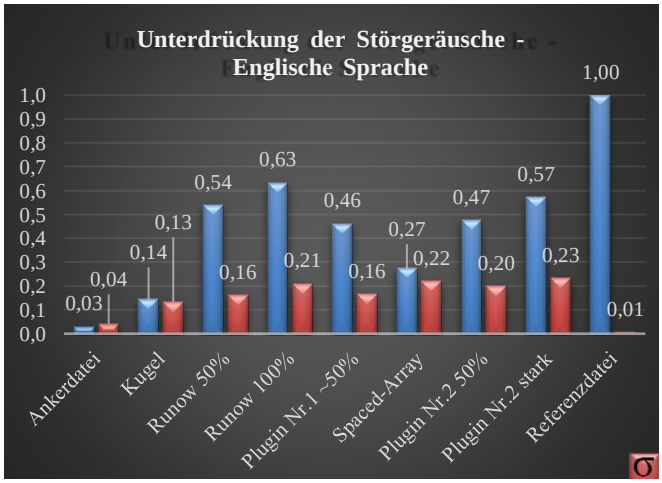


Abb. 12: Ergebnis „Unterdrückung der Störgeräusche“ in englischer Sprache

5.6 Empfundene Qualität

Das Gehör reagiert sehr empfindlich auf Veränderungen des Klangs der Sprache und schon geringe Nuancen werden wahrgenommen. Als weiterer wichtige abzuprüfende Eigenschaft galt daher die empfundene Qualität der bearbeiteten Sprachdateien. Hierbei liegt das Spaced-Array nur knapp über der bewusst mit Rauschen und Bandpassfilterung degradierten Ankerdatei. Trotz hoher Sprachverständlichkeit und Störgeräuschunterdrückung kann B. Runows Beamformer auch hier den ersten Platz behaupten. Da gute Qualität ein individuelles Maß ist, vermieden das Einbinden einer vermeintlich hochwertigen Referenzdatei. Auch in diesem Prüfabschnitt wichen die Ergebnisse in deutscher und englischer Sprache nur gering voneinander ab.

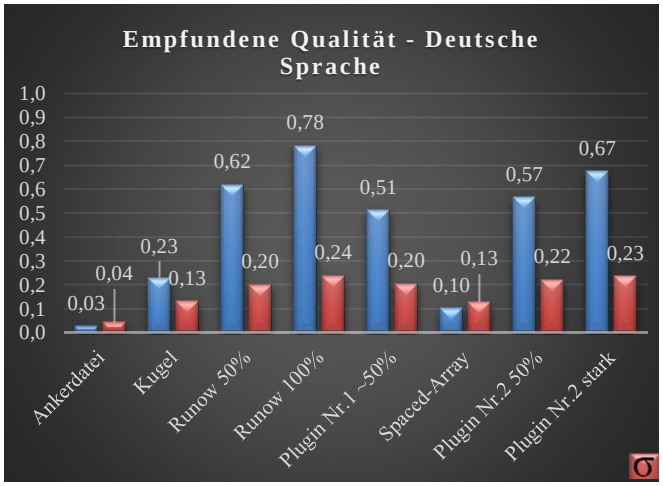


Abb. 13: Ergebnis „empfundene Qualität“ in deutscher Sprache

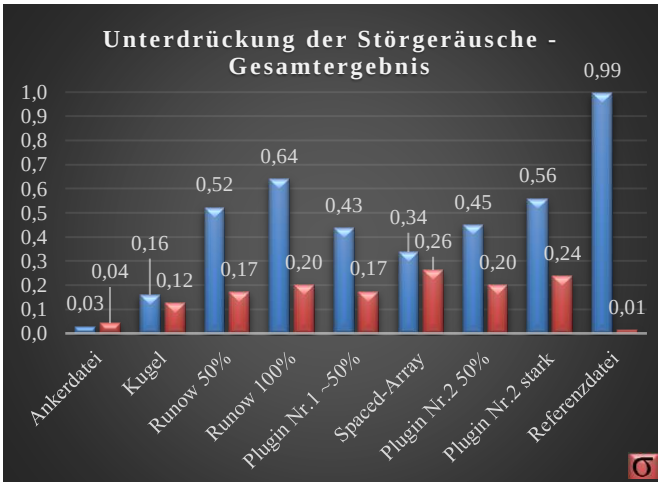


Abb. 16: Gesamtergebnis „Unterdrückung der Störgeräusche“

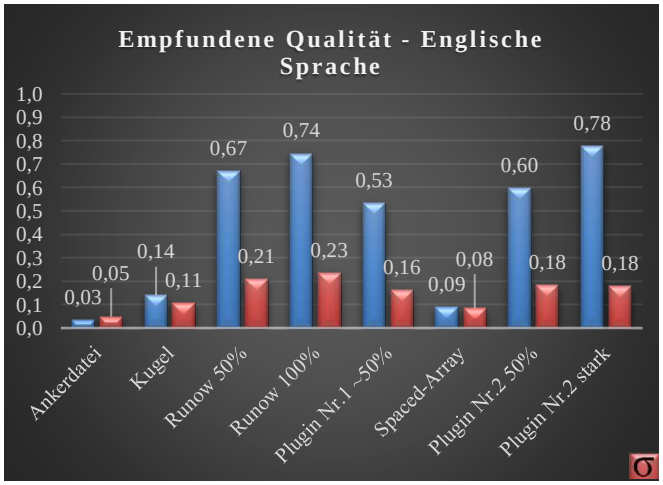


Abb. 14: Ergebnis „empfundene Qualität“ in englischer Sprache

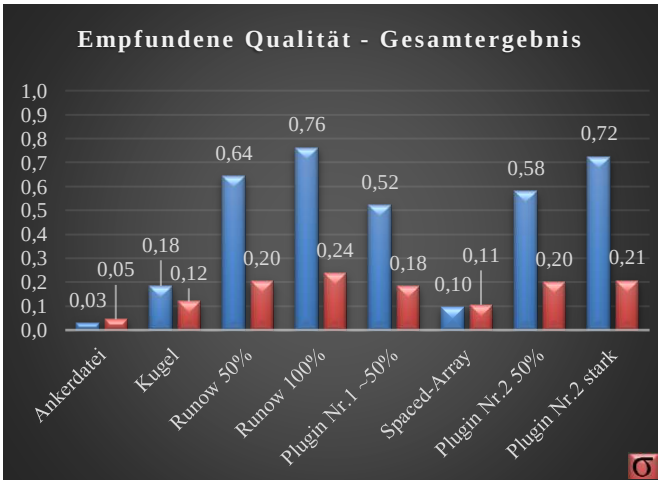


Abb. 17: Gesamtergebnis „empfundene Qualität“

5.7 Zusammenfassende Ergebnisse beider Sprachen

Zu besseren Übersicht werden hier nochmal die zusammenfassenden Ergebnisse beider Sprachen in Summe dargestellt.

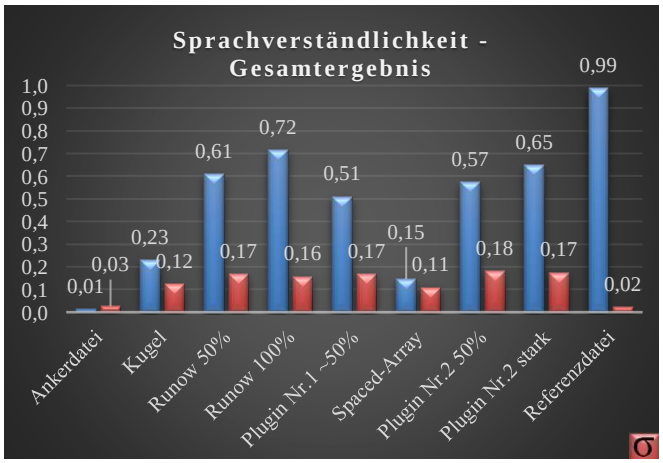


Abb. 15: Gesamtergebnis „Sprachverständlichkeit“

5.8 Sprachverständlichkeitsindex STOI

Zum automatischen Bewerten der Sprachverständlichkeit bietet sich das frei zugängliche Matlabscript „STOI“ [17] an. „STOI“ steht für Short Time Objective Intelligibility Messurement und wurde von C.H. Taal et al an der Delft University of Technology in Holland entwickelt. Das Skript vergleicht in unserem Fall die trockene Studioaufnahme (Referenzdatei) mit denen, über das in der virtuellen Konferenz mit dem Prototypenmikrofon aufgenommen und Dateien. Hohe Ergebniswerte bedeuten eine hohe Sprachverständlichkeit, so dass auch beim Vergleich zweier gleicher Dateien stets eine 1.0 herauskommt. Wir konnten so unsere Sprachdateien erneut abprüfen, gespannt darauf, in wie fern sich die Ergebnisse des Prüfalgorithmus mit unseren bisherigen Ergebnissen decken würden.

Grundsätzlich liegen die Ergebnisse näher beieinander, als bei der durch den Menschen durchgeführten Bewertung. Dadurch sollte den Nachkommastellen mehr Beachtung geschenkt werden. Interessant ist, dass beide Sprachdateien des Spaced-Arrays schlechter als die Ankerdatei bewertet wurde. Wie bei der Ankerdatei sind die Daten des Spaced-Arrays im

Frequenzband stark beschnitten - für den Bewertungs-Algorithmus wahrscheinlich zu stark und in einem als wichtig bewertetem Bereich. Auch das indirekte Kugelsignal wird in beiden Fällen erkannt und in Summe nicht viel besser bewertet als das räumliche Kugelsignal. Interessanterweise schneidet das Tracking Signal bei der STOI-Bewertung auf ähnlich hohem Niveau ab, wie mit kombiniertem Beamformer bzw. Dereverb. Grund dafür könnte die interne Korrelationsmessung bis nur 5000Hz sein, welche a) das Processing der höheren Frequenzen unbewertet lässt und b) die Artefaktbildung korrelationsbedingt höher bewertet als der Mensch, der hier teilweise deutliche Unterschiede hören kann.

Die weiteren Sprachdateien bzw. Algorithmen liegen relativ nahe beieinander, allerdings zeichnet sich auch in diesen geringen Bereich das Führen des Beamformer von B. Runow ab. Deutlich erkennbar ist auch, dass das Plug-In Nr. 1 in dieser Bewertungsrunde stark aufholen kann. Zur Ergänzung wurden zwei weitere Einstellungen des Plug-Ins ausprobiert.

Bezeichnung	Beschreibung
Ankerdatei	Ankerdatei mit 3,5kHz-Lowpass und einer Degradierung durch hohe Nebengeräusche und Rauschen. Dieses Signal muss als „schlecht“ bewertet werden.
Referenz	Sie diente im STOI-Versuch als Vergleichssignal (Wert 1.0). Aus Übersichtsgründen wird dieses Signal daher im Chart nicht dargestellt.
Kugel	Kugelsignal des Prototypenmikrofons
Tracking	Sprachverfolgungs-Algorithmus Hochschule der Medien Stuttgart mit synthetisierter Niere bis Superniere.
Runow 50%	B. Runows-Beamforming-Algorithmus in mittlerer Stärke, gespeist mit den drei einzelnen Kapselsignalen und der vom Tracker ermittelten Richtungsinformation.
Runow 100%	B. Runows-Beamforming-Algorithmus in maximaler Stärke, gespeist mit den drei einzelnen Kapselsignalen und der vom Tracker ermittelten Richtungsinformation.
Plug-In Nr.1 ~25%	DAW-Plug-In in schwacher Einstellung, gespeist mit zwei geführten Mono-Signalen.
Plug-In Nr.1 ~50%	DAW-Plug-In in mittlerer Einstellung, gespeist mit zwei geführten Mono-Signalen.
Plug-In Nr.1 100%	DAW-Plug-In in maximaler Stärke, gespeist mit zwei geführten Mono-Signalen.
Spaced-Array	Gesamtsystem mit eigenem räumlichen Mikrofonarray und schwacher Einstellung des produkteigenen Algorithmus.
Plug-In Nr.1 50%	DAW-Plug-In in mittlerer Stärke, gespeist mit geführtem Mono-Signal.
Plug-In Nr.2 stark	DAW-Plug-In in starker Einstellung, gespeist mit geführtem Mono-Signal.

Tab. 2: Übersicht der ergänzten Signale für die Sprachverständlichkeitsanalyse „STOI“

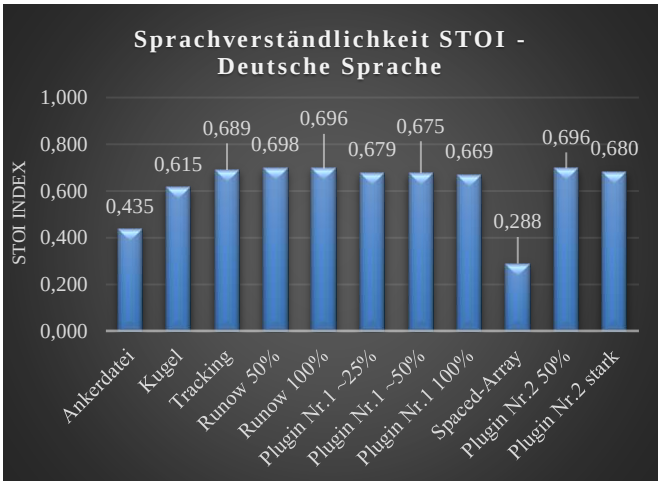


Abb. 18: Ergebnis Sprachverständlichkeitsanalyse „STOI“ in deutscher Sprache

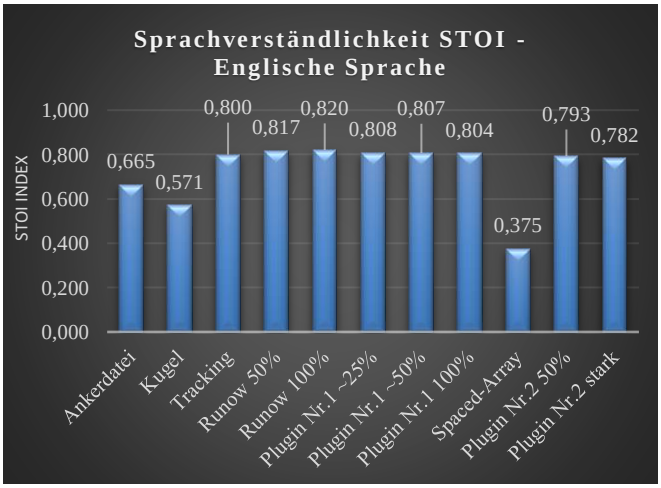


Abb. 19: Ergebnis Sprachverständlichkeitsanalyse „STOI“ in englischer Sprache

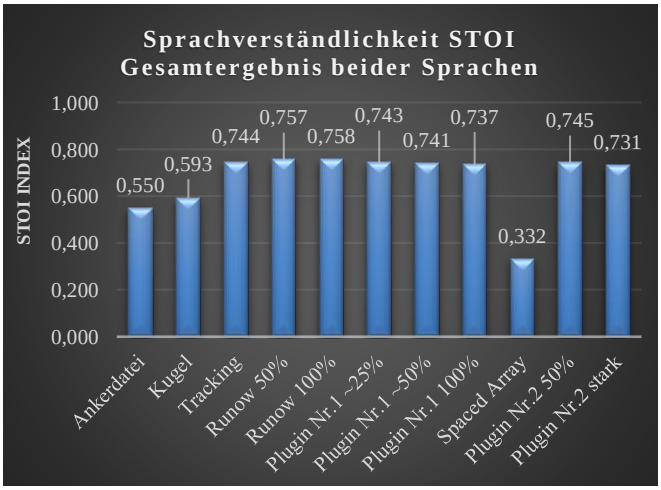


Abb. 20: Gesamtergebnis Sprachverständlichkeitsanalyse „STOI“ beider Sprachen

6. Datenauswertung

Pro Teilnehmer standen über 70, insgesamt über 4000 Werte zur Auswertung bereit. Zur zügigen Verarbeitung der Daten musste beim Erstellen des Hörversuches jedoch beachtet werden, dass WAET die zufällige Reihenfolge der ausgegeben Sprachdateien genauso in der Ergebnisdatei abspeichert. Daher müssen die jeweiligen Sprachdateien bei der Programmierung des Hörversuchs mit Indizes versehen werden um sie nachträglich sortieren zu können. Anschließend konnten die sortierten Datenblöcke in eine Sammelliste übertragen und ausgewertet werden.

7. Zusammenfassung und Ausblick

Durch das Web Audio Evaluation Tool konnte ein individuell programmierbarer und den ITU-Empfehlungen folgender Hörversuch einem großen Teilnehmerkreis online zugänglich gemacht werden.

Insgesamt wurde die Kombination des Trackingalgorithmus mit B. Runows Beamformer in Maximaleinstellung als bestes bewertet. Dieser Beamformer konnte allerdings auch alle drei Kapselsignale nutzen, so dass auch die anderen Algorithmen theoretisch Optimierungspotential besitzen. Plug-In Nr.1 lag mit seinen Ergebnissen oft knapp hinter denen des Plug-Ins Nr.2 und konnte vor allem bei den Ergebnissen der automatischen Sprachverständlichkeitsanalyse STOI deutlich aufholen. Das eigenständige Konferenzsystem wurde von allen prozessierten Dateien in allen Kategorien als schlechtestes bewertet und konnte nur im Bereich der Störgeräuschunterdrückung einigermaßen überzeugen.

Weiterhin können die gesammelten Daten zur fortführenden Mikrofonentwicklung genutzt werden und die Grundlage zur Implementierungen eines Beamformers und dessen Stärkeeistellungen bilden.

8. Quellenangaben

[1] H. Paukert, J. D. Ziegler: “Listening Tests in the Process of Microphone Development”: Tagungsbericht der 29. internationalen Tonmeistertagung des Verbands deutscher Tonmeister e.V., Nov. 2016, Seite 273-280, ISBN 978-3-9812830-7-5, URL: https://www.tonmeister.de/index.php?p=tonmeistertagung/2016/downloads

[2] Cycling ’74, Max 7 perpetual licence, URL: https://www.cycling74.com

[3] N. Jillings, B. De Man, D. Moffat, J. D. Reiss: “Web Audio Evaluation Tool: a browser-based Listening Test Environment”, Centre for Digital Music, Queen Mary University of London, URL: https://github.com/BrechtDeMan/WebAudioEvaluation Tool

[4] “Method for the subjective assessment of intermediate quality level of coding systems”, International Telecommunication Union (ITU) Empfehlung BS.1534-3, 2015, URL: https://www.itu.int/rec/R-REC-BS.1534-3-201510-I/en

[5] J. D. Ziegler, A. Koch, A. Schilling: „Speech classification for acoustic source localization and tracking applications using convolutional neural networks”, Audio Engineering Society convention 145, October 2018

[6] J. D. Ziegler, H. Paukert, A. Koch, A. Schilling: “Speaker Tracking with Coincident Microphone Arrays and Convolutional Neural Networks”, IEEE Journal, aktuell ausstehende Bewilligung

[7] B. Runow, O. Curdt: “Mikrofonarrays in der professionellen Audioproduktion”, Tagungsbericht der 28. Internationalen Tonmeistertagung Verband deutscher Tonmeister e.V., Nov. 2014, Seite 263-269, ISBN 978-3-9812830-5-1 URL: https://www.tonmeister.de/index.php?p=tonmeistertagung/2014/downloads

[8] B. Runow, O. Curdt, A. Schilling: „Richtrohrmikrofone versus Mikrofonarrays“, Tagungsbericht der 29. internationalen Tonmeistertagung des Verbands deutscher Tonmeister e.V., Nov. 2016, Seite 221-228, ISBN 978-3-9812830-7-5, URL: https://www.tonmeister.de/index.php?p=tonmeistertagung/2016/downloads

[9] R. Hirt: „Entwicklung einer virtuellen Konferenz unter besonderer Berücksichtigung der Reproduktion von zuvor aufgenommener Sprache“, Bachelorthesis, Hochschule der Medien Stuttgart, 2017

[10] Timbre Inc., URL: https://www.sketchup.com/

[11] Genelec Inc. URL: https://www.genelec.com/support-technology/previous-models/1029a-studio-monitor

[12] B. D. Man, J. D. Reiss, “APE: Audio Perceptual Evaluation toolbox for Matlab”, 136th AES Convention 2014, Berlin, Germany

[13] B. D. Man, J. D. Reiss, “APE: Audio Perceptual Evaluation toolbox for Matlab”, 136th AES Convention 2014, Berlin, Germany, Seite 1

[14] Beyerdynamic GmbH & Co. KG, URL: http://www.beyerdynamic.de

[15] Sennheiser electronic GmbH & Co. KG, URL: https://de-de.sennheiser.com

[16] Harman Deutschland GmbH, URL: https://de.akg.com/

[17] C.H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen: A short-time objective intelligibility measure for time-frequency weighted noisy speech”, Delft University of Technology, Niederlande

realHDaudio for freeDSP: Ein DIY-AD/DA-System für professionelle Audioanwendungen

C. Epe[1], D. Leckschat[1], B. Rieß[1], Y. Proch[1]

[1] *Hochschule Düsseldorf, Germany, Email: christian.epe@hs-duesseldorf.de*

Abstract

freeDSP ist ein Projekt der Technischen Universität Dresden und bietet preisgünstige und einfach umsetzbare DSP-Technik für DIY-Anwendungen. Verwendet wird ein System-on-a-Chip DSP, dessen integrierte AD- und DA-Wandler von der Performance her jedoch nicht an heutige Studioqualität heranreichen. Zu diesem Zweck wurde ein realHDaudio-Board entwickelt, welches trotz geringer Baugröße eine Audio-Performance erreicht, die bisher nur großen und vor allem hochpreisigen Studiosystemen (z.B. Signal-to-Noise-Ratio größer 118 dB) vorbehalten war. Das Board (Abbildung 1) lässt sich über einen Erweiterungsslot mit dem freeDSP-Board, aber auch mit jedem anderen DSP/FPGA-System verbinden, sofern entsprechende Schnittstellen zur Verfügung stehen.

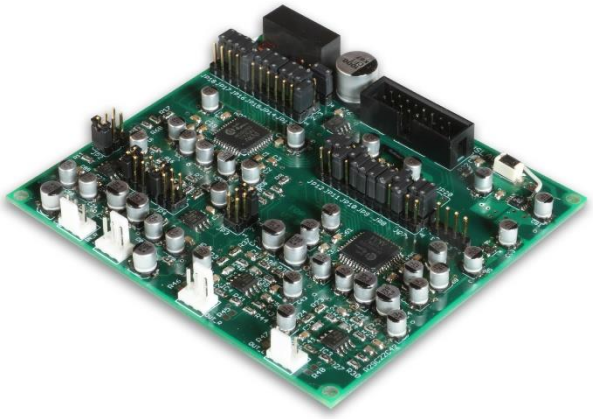


Abb. 1: realHDaudio-Board Version 2.0

1. Konzept

Abbildung 2 zeigt ein Blockschaltbild des entwickelten Erweiterungsboards. Es kann über einen „Expansion Slot“ mit dem freeDSP oder jedem anderen System, welches über eine I2S-Schnittstelle verfügt, verbunden werden. Die Basis des Boards bilden ein AKM AK5397EQ AD- und ein AKM AK4495 DA-Konverter. Beide ICs können mit einer maximalen Samplerate von 768 kHz und einer Bittiefe von bis zu 32 Bit betrieben werden. Sie zeichnen sich zudem durch sehr gute Audiomesswerte aus. Der AK5397EQ zählt mit bis zu 127 dB Signalrauschabstand (SNR) im Stereobetreib und 130 dB SNR im Monobetrieb momentan zu den rauschärmsten Wandlern auf dem Markt. Die Konfiguration der ICs erfolgt über auf dem Board befindliche Jumper (Abb. 2, Hardware Control) oder beim AK4495 wahlweise über I^2C. Die Anbindung der Wandler an ein Hauptsystem (DSP/FPGA usw.) erfolgt je nach ihrer Konfiguration über I^2S (BCK, LRCK, SDATA), Left-Justified oder TDM. Die Masterclock (MCK) muss vom Hauptsystem zur Verfügung gestellt werden und beträgt im Idealfall 24,576 MHz, um alle verfügbaren Sampleraten der Wandler nutzen zu können. Sie wird über einen Clock-Buffer an die Wandler weitergeben. Im Fall des freeDSP gelangt über den Expansion Slot eine Spannung von 12V auf das Board, aus der die nötigen Spannungen für die ICs (3,3V, 5V) sowie für die symmetrisch ausgelegten, analogen Eingangs- und Ausgangstufen (+9V/-9V) generiert werden. An einen freeDSP CLASSIC lassen sich bis zu 4 der realHDaudio-Boards anschließen.

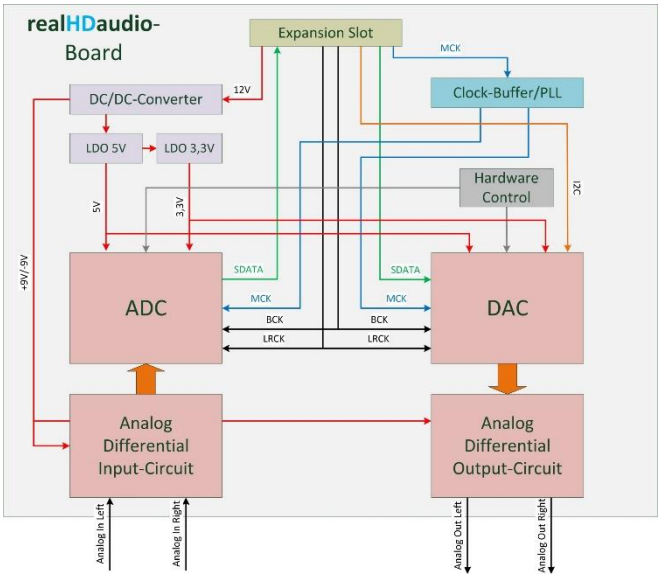


Abb. 2: Blockschaltbild realHDaudio-Board Version 2.2

2. Analoge Eingangsstufe/Gainstruktur

Die äußeren Rahmenbedingungen legen in erster Linie den erlaubten Pegel am Eingang des AD-Wandlers fest. Die Verstärkung der Eingangsstufe (Abbildung 3) muss anschließend dem gewünschten maximalen Eingangspegel angepasst werden, um den Wandler entsprechend korrekt auszusteuern. Der AK5397EQ erlaubt an einem seiner Eingangs-Ports (positiver oder negativer Eingang) maximal

ISBN
978-3-9812830-9-9

Der Autor hält die Urheberrechte für den gesamten Inhalt seines Manuskripts. Er überlässt dem VDT das Recht auf Veröffentlichung in gedruckter Form, Online oder als PDF-Download von https://tonmeistertagung.com

2,8 Vpp (2,13 dBu). Die Spannung muss zudem mit einem Bias von 1,9 V angelegt werden, so dass sich die resultierende Spannung an einem Port zwischen 0,5 V und 3,3 V bewegt. Um nun eine Verstärkung für die OP-Stufe festlegen zu können, müssen mögliche Eingangspegel definiert werden. Die zu definierenden Pegel beziehen sich jeweils auf die, auf beide Eingangs-Ports zusammen gegebene Spannung. Diese teilen sich dann jeweils auf den positiven und negative Eingangs-Port auf. Als mögliche Eingangspegel sollen drei Fälle betrachtet und die entsprechende Verstärkung definiert werden:

- „Rundfunkpegel“: Der Rundfunknormpegel ist ein Standard für professionelle Audiogeräte und liegt bei +6 dBu (1,44 VRMS).
- „Consumer-Pegel“: Der Consumer-Pegel wird häufig mit -10 dBv (0,316 VRMS) definiert.
- „Mikrofon“: Die Empfindlichkeit eines dynamischen Standardmikrofons (z.B. Shure SM58) liegt bei 2 mV/Pa. Definiert man 115-120 dBSPL als maximal geplanten Schalldruck vor dem Mikrofon (Schlagzeug in einem Meter Entfernung) führt das zu einem Ausgangssignal von 32 mV (-27 dBu). Eine Verstärkung von 35 dB wäre nötig, um den Wandler richtig anzusteuern.

Um die Verstärkung für alle drei Fälle anpassen zu können, wird R1 aus Abbildung 3 mit Hilfe eines Jumpers variabel gestaltet. Die Werte für R2, R3 und R4 werden aus dem Datenblatt des AK5397EQ übernommen. Für jeden Fall wird zudem eine Aussteuerungsreserve (7 dB bzw. 10 dB) eingeplant. Mit Hilfe von Formel 1 wurden für R1 folgende Werte (Standard E-Reihe) berechnet.

- „Rundfunkpegel“: R1=430 Ohm
- „Consumer-Pegel“: R1=2,7kOhm
- „Mikrofon“: R1=39kOhm

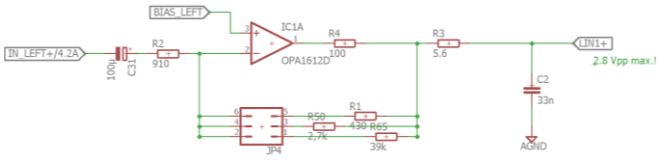


Abb. 3: Analoge Eingangsstufe (hier nur eine Phase)

$$L_{dB} = 20 * \log\left(\frac{R1}{R2} * \frac{R1}{(R1 + R4)}\right) \quad [\text{dB}] \quad (1)$$

Mit den gewählten Widerstandswerten für R1 führen folgende Pegel zur Vollaussteuerung des Wandlers:

- Max. Input „Rundfunkpegel“: +14 dBu
- Max. Input „Consumer-Pegel“: -4 dBv
- Max. Input „Mikrofon“: 122 dBSPL

Die resultierende Gain-Struktur des Audiopfades je nach Jumper-Konfiguration in der Eingangsstufe ist noch einmal in Abbildung 4 dargestellt.

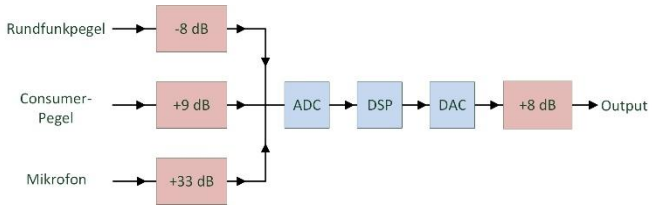


Abb. 4: *Gainstruktur des Systems*

3. Platine und mechanischer Aufbau

Aus dem fertigen Schaltplan wurde ein Layout für eine 4-lagige Platine erstellt. Die Oberseite der Platine bildet Layer 1. Auf ihm wurden alle SMD-Bauteile untergebracht. Außerdem wird hier das primäre Routing der Stromversorgung, Audiosignal- und teilweise digitalen Signalleitungen vorgenommen. Layer 2 stellt den Masse-Layer (GND) für die gesamte Schaltung dar. Um Störungen der analogen Schaltungsteile durch die digitalen Schaltungsteile zu vermeiden, werden die zugehörigen Massebereiche voneinander getrennt und im Bereich der zentralen Versorgungsspannung zusammengeführt. Mit Hilfe von Vias wird die Masse direkt in der Nähe der jeweils angeschlossenen Bauteile an die Oberfläche geführt. Auf Layer 3 werden die Versorgungsspannungen zu den jeweiligen Bauteilen verteilt. Auch hier musste darauf geachtet werden analoge und digitale Schaltungsteile zu trennen. Auf Layer 4 verlaufen unter anderem die I²S-Signale mit seriell angelegten 0 Ohm-Widerständen, welche im Falle von zu starken „Overshoots“ der Clock-Signale bestückt werden können. Abbildung 5 zeigt einen Lautsprechercontroller bestehend aus einem freeDSP-CLASSIC und drei realHDaudio-Boards in einem 1HE-Gehäuse.

Abb. 5: *realHDaudio LS-C2, bestehend auf einem freeDSP CLASSIC und drei realHDaudio-Boards*

4. Messwerte und max. Ratings

Bei der Messung der audio-charakteristischen Eigenschaften ist die Signalverarbeitung im DSP-System abgeschaltet und es findet somit lediglich eine AD- und DA-Wandlung bzw.

Weiterleitung durch den DSP statt. Charakteristische Messwerte sind der Tabelle 1 zu entnehmen.

Analog In ”Rundfunkpegel" bei 0,0022% THD-N	14,9	dBu
Noise bei +80dB digital gain (A-w.) Mono Mode	-33	dBu
SNR Mono Mode	127,9	dB
Noise bei +80dB digital gain (A-w.) Stereo Mode	-30	dBu
SNR Stereo Mode	124,9	dB

Tab. 1: Maximum Ratings ADC

Analog Out bei 0 dBFS	16,2	dBu
Noise DAC (A-weighted)	-102,8	dBu
SNR (+3dB im Mono Mode)	119,0	dB

Tab. 2: Maximum Ratings DAC

SNR	117,9	dB
THD+N bei 7 dBu Input	-103,3	dB
THD bei 7 dBu Input	-110,6	dB
Crosstalk channel 1 bei 14 dBu	-123,2	dB
Crosstalk channel 2 bei 14 dBu	-128,1	dB

Tab. 3: Maximum Ratings DAC

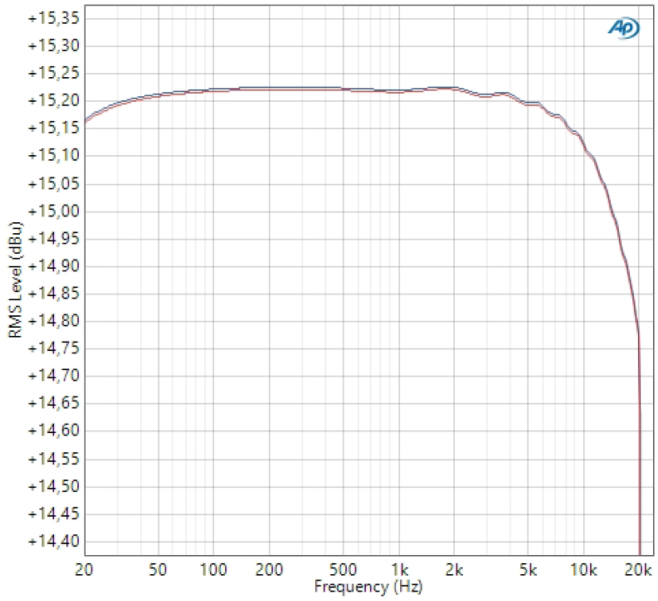


Abb. 7: *Frequenzgang bei 14 dBu Eingangssignal*

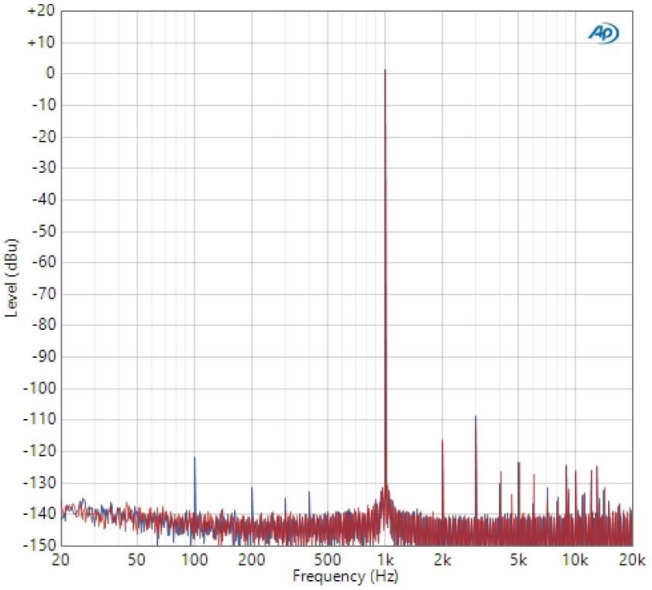


Abb. 6: *Spektrum bei 1kHz, 0 dBu Eingangssignal*

5. Zusammenfassung und Ausblick

Ergebnis dieser Arbeit ist ein kompaktes Erweiterungsboard mit professionellen Audioeigenschaften. Ursprünglich gedacht für den freeDSP lässt es sich darüber hinaus mit jeden anderen DSP/FPGA-System verbinden, sofern die nötigen Schnittstellen vorhanden sind. Das Board wird Anfang 2019 auf der Webseite des freeDSPs als Erweiterungsplatine gelistet und steht dann als Open-Source-Variante zur Verfügung. Stand Nov. 2018 existiert neben der freeDSP CLASSIC-Version eine CLASSIC SMD B-Variante, die zwar über vier passende Expansion Slot verfügt, aber aufgrund zweier nicht erreichbarer DSP-Pins momentan noch nicht mit dem realHDaudio-Board nutzbar ist. Soweit diese in einer Revision des freeDSP CLASSIC SMD B erreichbar sind, können auch hier bis zu vier realHDaudio-Boards verwendet werden.

7. Quellen

[1] www.realHDaudio.de

[2] www.freedsp.cc

The Fundamental Problem of the Spectral Subtraction

B. Runow[1], J. D. Ziegler[2], H. Paukert[3], A. Schilling[4], O. Curdt[5]

[1] *Wilhelm-Schickard Institut, Eberhard Karls University, Tübingen, Email: bernfried@runow.info*
[2] *Wilhelm-Schickard Institut, Eberhard Karls University, Tübingen, Email: zieglerj@hdm-stuttgart.de*
[3] *Hochschule der Medien Stuttgart, Email: paukert@hdm-stuttgart.de*
[4] *Wilhelm-Schickard Institut, Eberhard Karls University, Tübingen, Email: schilling@uni-tuebingen.de*
[5] *Hochschule der Medien Stuttgart, Email: curdt@hdm-stuttgart.de*

Abstract

Spectral Subtraction is often used for noise reduction and speech enhancement. It is an important tool of digital audio signal processing. Since its introduction in 1979, several problems like Phase Errors, Cross-time Errors and Magnitude Errors cause rather disappointing results. Beyond these errors, there is a fundamental problem within the basic principles of Spectral Subtraction, which is documented in this publication.

1. Introduction

Spectral Subtraction is a widespread method to dynamically process the spectrum of a digital audio signal. It gives you the possibility to edit a signal in a specific spectral range. The basis for this procedure is the discrete Fourier transform (DFT), which converts a time-series signal into the frequency domain and makes frequency analysis possible. In the spectral domain it is possible to edit individual spectral components, the so-called spectral coefficients. This makes it possible to subtract information from a specific frequency component. Finally, the processed signal can be resynthesized by means of an inverse discrete Fourier transform (iDFT). Therefore, the edited signal is available in the time domain once again.

The crucial advantage of the Spectral Subtraction is given by the short-time Fourier transform (STFT). With the STFT, it is possible to decompose a continuous stochastic signal and transform each time segment into the spectral domain. There, the time segments can be edited one after another. After the inverse transformation, the time segments can be recomposed into a continuous signal.

Because of the segmental processing, it is possible to edit each segment individually. This means, we can create an adaptive, real-time signal processing algorithm with a short latency. This is the reason for the importance of the Spectral Subtraction in the last decades. A multitude of applications use this technique, like noise reduction and speech enhancement.

2. Fundamentals

2.1. Windowing of a Signal

The segmentation of a continuous input signal $x(n)$ can be achieved with a window function $w(n)$, as we can see in Fig. 1.

Each segment is multiplied with the window function $w(n)$:

$$x_{win}(n) = x(\eta_{win} + n) \cdot w(n)\,, \qquad (1)$$

where $n = 0,1,2,\dots,N-1$ is the discrete time index and N the length of the segments. The variable η_{win} defines the first sample of the current segment.

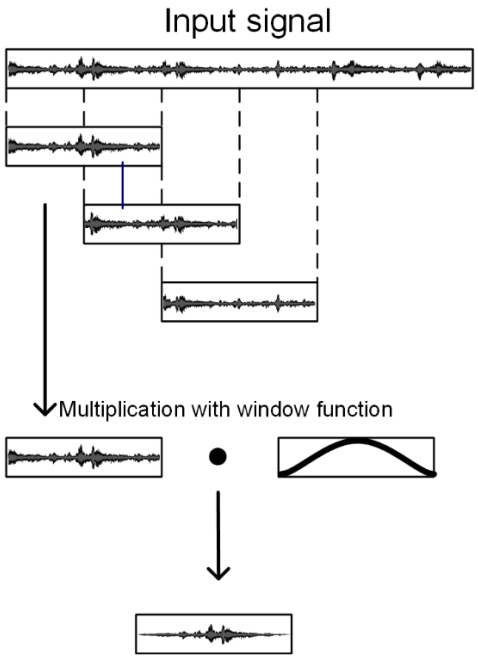


Fig. 1: Windowing of a continuous input signal using the von-Hann window function with an overlap of 50%.

An overlap of the segments is possible. Depending on the length of overlap, a compatible window function has to be chosen. The sum of the successive window functions always has to be one. This restriction is given in order to prevent a distortion of the signal within the resynthesis process, more precisely through the multiplication with the window

ISBN
978-3-9812830-9-9
The authors hold the copyrights of all of content in this manuscript. They have granted the VDT the right to publish it in print, online and as PDF download on *https://tonmeistertagung.com*

function. This means that the windowing must result in a constant amplification of 1.

If we don't want an overlap of segments, we can choose the rectangular window:

$$w_{rect}(n) = 1\,, \tag{2}$$

with $n = 0,1,2,\dots,N-1$.

If we want an overlap of 50%, we can, for example, choose the von-Hann window function:

$$w_{hann}(n) = \frac{1}{2} - \frac{1}{2}\cos\left(2\pi\frac{n}{N-1}\right), \tag{3}$$

with $n = 0,1,2,\dots,N-1$.

In Fig. 2 you can see the von-Hann window functions for the segmentation of an input signal. Any window function can be used as long as the constraint of constant amplification is met.

Because of the use of a window function, a segment is also called window.

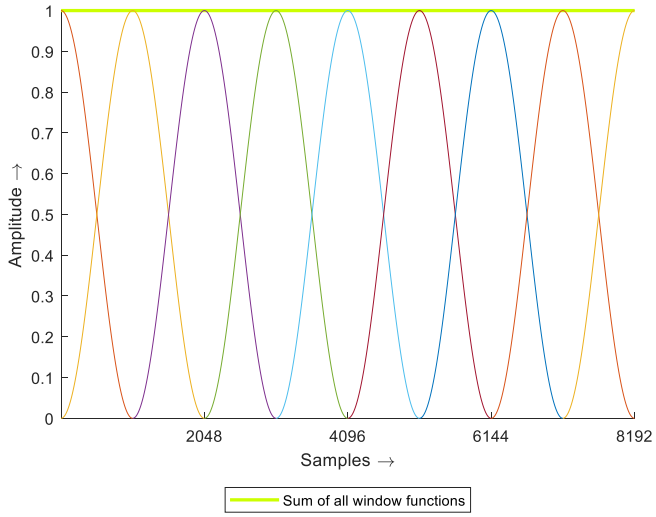


Fig. 2: Von-Hann window functions with a length of 2048 samples and their sum.

2.2. Short-Time Fourier Transform

After the segmentation of the input signal, the short-time Fourier transform (STFT) uses the Discrete Fourier transform to transport each window into the frequency domain. We obtain the DFT-coefficients using [4][8]:

$$X_{win}(k) = \sum_{n=0}^{N-1} x_{win}(n)\cdot e^{-j2\pi k\frac{n}{N}}, \tag{4}$$

where $k = 0,1,2,\dots,N-1$ is the discrete frequency index.

Each DFT coefficient represents a constant oscillation with the dedicated frequency f_k:

$$f_k = f_s\cdot\frac{k}{N}, \tag{5}$$

where f_s represents the sampling frequency which was used for the sampling during the digitalisation of the input signal. The absolute value of the DFT coefficient is the amplitude $|X_{win}(k)|$ of the oscillation and $\angle X_{win}(k)$ describes the corresponding phase angle.

By means of the inverse discrete Fourier transform we can transport the spectral signal $X_{win}(k)$ back into the time domain [4][8]:

$$x_{win}(n) = \frac{1}{N}\sum_{k=0}^{N-1} X_{win}(k)\cdot e^{j2\pi n\frac{k}{N}}, \tag{6}$$

with $n = 0,1,2,\dots,N-1$. Thus, the two signal sequences $x_{win}(n)$ and $X_{win}(k)$ are a transform pair.

Finally, the processed signal segments can be recombined according to the defined overlap.

2.3. Characteristics of the STFT

The Short-time Fourier transform has a number of characteristics which are accurately described in the relevant literature [2][4][8][9]. Two of these characteristics are especially important for Spectral Subtraction: the periodicity and the resolution of time and frequency.

2.3.1. Periodicity

The exponential function $e^{-j2\pi kn/N}$ behaves in a periodic fashion depending on N. This results the periodicity of the DFT and consequently of the STFT [4][8]:

$$X_{win}(k) = X_{win}(k+N) \tag{7}$$

and

$$x_{win}(n) = x_{win}(n+N)\,. \tag{8}$$

2.3.2. Time Resolution and Frequency Resolution

By using a clever analogy to the Heisenberg uncertainty principle, Küpfmüller points out that it is not possible to simultaneously achieve both a high resolution in time and in frequency within the spectral domain [7].

The background of this principle is the identical length N of the transform pair consisting of the time-domain signal $x_{win}(n)$ and the signal in the frequency domain $X_{win}(k)$. To get a high frequency resolution, we need a preferably long signal length. Contrarily we achieve a high time resolution using a short window in the time domain as this enables us to compute an individual spectrum for each short time segment.

Fig. 3-5 make this uncertainty principle clear. We can see several spectra over time. The test signal, which is a sine wave changing its frequency every second, was transformed into the spectral domain by means of STFT. The charts differ in the window lengths which were used for the STFT.

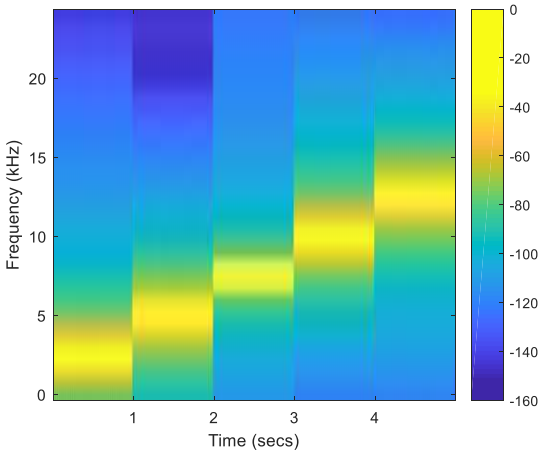


Fig. 3: Spectrogram of a sine wave changing its frequency every second. Analysed using STFT with a window length of 64 samples and a sampling frequency of 48kHz.

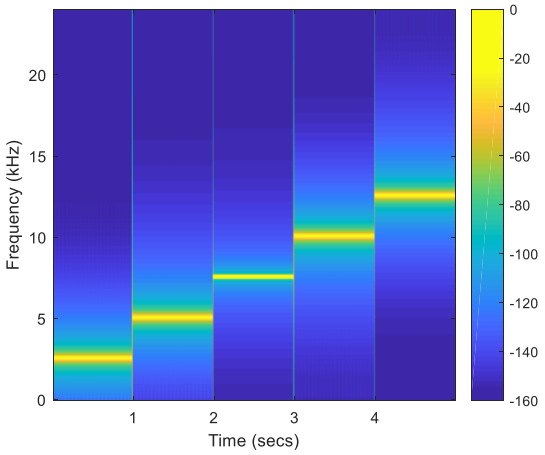


Fig. 4: Spectrogram of a sine wave changing its frequency every second. Analysed using STFT with a window length of 512 samples and a sampling frequency of 48kHz.

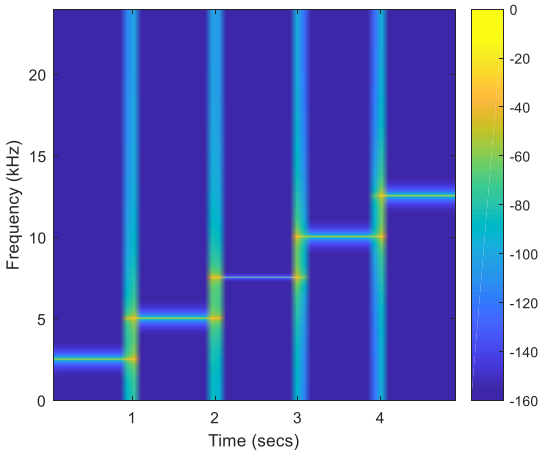


Fig. 5: Spectrogram of a sine wave changing its frequency every second. Analysed using STFT with a window length of 8192 samples and a sampling frequency of 48kHz.

We can solve this conflict with the help of a process called 'Zero Padding'. To get a high frequency resolution for a short time segment we can add a number of zeros at the end of the windowed time signal:

$$\tilde{x}_{win}(n) = \begin{cases} x_{win}(n) & for\ 0 \le n \le N-1 \\ 0 & for\ N \le n \le N+L-1\,, \end{cases} \quad (9)$$

where $n = 0, 1, 2, \dots, N+L-1$ and L represents the number of the added zeros. Thus, it is possible to simultaneously achieve a high time resolution and a high frequency resolution within the STFT.

2.4. Spectral Subtraction

During Spectral Subtraction the amplitudes of two spectral signals are subtracted from each other. If $|X_{win}(k)|$ is the minuend and $|U_{win}(k)|$ is the subtrahend, we obtain the difference [3]:

$$|Y_{win}(k)| = |X_{win}(k)| - |U_{win}(k)| \cdot \upsilon(k, p=1)\,, \quad (10)$$

where $\upsilon(k,p)$ is a real weighting factor which regulates the subtrahend $|U_{win}(k)|$, so that $|Y_{win}(k)|$ cannot assume negative values:

$$\upsilon(k,p) = \begin{cases} 1 \cdot \iota & for\ |U_{win}(k)| \le |X_{win}(k)| \\ \dfrac{|X_{win}(k)|^p}{|U_{win}(k)|^p} \cdot \iota & for\ |U_{win}(k)| > |X_{win}(k)| \end{cases}. \quad (11)$$

The real factor $0 \le \iota \le 1$ defines the intensity of the Spectral Subtraction. If $\iota = 0$, there is no subtraction. If $\iota = 1$, the subtraction is maximal. The quotient of $|X_{win}(k)|$ and $|U_{win}(k)|$ prevents that $|Y_{win}(k)|$ can become negative if the absolute value of $U_{win}(k)$ is larger than the absolute value of $X_{win}(k)$.

If we don't want to subtract the amplitudes, but the power, equation (10) is modified to produce $|Y_{win}(k)|$:

$$|Y_{win}(k)| = \sqrt{|X_{win}(k)|^2 - |U_{win}(k)|^2 \cdot \upsilon(k, p=2)}\,. \quad (12)$$

A more general form can be written as:

$$|Y_{win}(k)| = \left(|X_{win}(k)|^p - |U_{win}(k)|^p \cdot \upsilon(k,p)\right)^{\frac{1}{p}}. \quad (13)$$

This is often named parametric spectral subtraction [5] and sets a variable exponent. With $p = 1$ we obtain the spectral subtraction from (10) and with $p = 2$ we obtain the spectral subtraction of the power from (12).

Combined with the phase $\angle X_{win}(k)$ of the input signal $X_{win}(k)$, the output signal can be computed with:

$$Y_{win}(k) = |Y_{win}(k)| \cdot e^{j\angle X_{win}(k)}\,. \quad (14)$$

To an extent, this operating sequence is a makeshift method. It is to be expected that after the subtraction, the correct phase of $Y_{win}(k)$ is not identical to the phase of the input signal $X_{win}(k)$. Jens Groh asserts that the correct phase often cannot be derived [6]. Thus, in many cases, the correct phase of the output signal is simply unknown. Studies have shown, that phase corruption in the spectral domain is considerably less perceptible than a corruption of the amplitude in this domain [10].

Finally, the output signal can be transformed back into the time domain by using the iDFT:

$$y_{win}(n) = \frac{1}{N}\sum_{k=0}^{N-1} Y_{win}(k) \cdot e^{j2\pi n\frac{k}{N}}. \tag{15}$$

Thereby the output signal is as long as the input signal and consists of N samples.

3. Spectral Subtraction as a Time-Variant System

The Spectral Subtraction can be considered as a time-variant system with a varying processing and parameters that can change from window to window.

Hence, we are able to write the subtraction in the spectral domain from (13) as a multiplication:

$$\begin{aligned} |Y_{win}(k)| &= \left(|X_{win}(k)|^p - |U_{win}(k)|^p \cdot \upsilon(k,p)\right)^{\frac{1}{p}} \\ &= |X_{win}(k)| \cdot \left(1 - \frac{|U_{win}(k)|^p}{|X_{win}(k)|^p} \cdot \upsilon(k,p)\right)^{\frac{1}{p}}. \end{aligned} \tag{16}$$

Then the amplitude response of this system is:

$$H_0(win,k) = \left(1 - \frac{|U_{win}(k)|^p}{|X_{win}(k)|^p} \cdot \upsilon(k,p)\right)^{\frac{1}{p}}, \tag{17}$$

and the frequency response of each window is:

$$H_{win}(k) = H_0(win,k) \cdot e^{j\angle H_{win}(k)}. \tag{18}$$

Assuming $\angle H_{win}(k) = \angle X_{win}(k)$, the equations (17) and (18) leads us to:

$$H_{win}(k) = \left(1 - \frac{|U_{win}(k)|^p}{|X_{win}(k)|^p} \cdot \upsilon(k,p)\right)^{\frac{1}{p}} \cdot e^{j\angle X_{win}(k)}. \tag{19}$$

Like the input signal $X_{win}(k)$, the frequency response consists of N DFT-coefficients. Thus, the spectral output signal $Y_{win}(k)$ can be computed as a product of the spectral input signal and the frequency response $H_{win}(k)$:

$$Y_{win}(k) = X_{win}(k) \cdot H_{win}(k) \tag{20}$$

A multiplication in the spectral domain corresponds to a convolution of the equivalent signals in the time domain [2]:

$$\begin{aligned} y_{win}(n) &= x_{win}(n) * h_{win}(n) \\ &= \sum_{m=0}^{N_{IR}-1} x_{win}(n) \cdot h_{win}(n-m), \end{aligned} \tag{21}$$

where $h_{win}(n)$ describes the impulse response of the system and N_{IR} is the length of this impulse response.

4. The Fundamental Problem

4.1. The Length of the Output Signal

The length of the output signal of a convolution is [2]:

$$N_{conv} = N_{input} + N_{IR} - 1, \tag{22}$$

where N_{input} is the length of the input signal, N_{IR} is the length of the impulse response and N_{conv} is the length of the convolved signal.

Considering the convolution in (21), both the input signal and the impulse response are of length N. Therefore, the output signal consists of $2N-1$ samples.

This means, that the output signal computed using convolution in the time domain is nearly twice as long as the output signal which is computed using Spectral Subtraction in the spectral domain and which has N samples. Thus, the output signal $y_{win}(n)$ in (21) cannot be the same as the output signal in (15) with (13) and (14), as we can see in Fig. 6.

The reason for this is the static signal length in the spectral domain and the periodicity of the DFT. The periodicity presupposes a continuous repetition of the finite output signal. The modifications of the DFT coefficients cause an extension of the signal when transformed back into the time domain. The part of the processed signal after the Nth sample will be continued at the beginning of the window. Since the STFT does not take this repetition at the recombination of the windows into account, an error inevitably occurs. We receive an overlap with a signal part, which is inserted at the wrong time position. This error becomes apparent when the signal is compared directly with the output signal, which is computed by convolution in the time domain. In Fig. 6 we can see the differences between the output signal of the Spectral Subtraction and the output signal of the convolution.

4.2. Zero Padding is no Solution

By using zero padding, we can reduce the effective length of the input signal in relation to the length of the window $N_{input} + L$. Consequently, the length of the frequency response $H_{win}(k)$ increases and for this reason the length of the impulse response $h_{win}(n)$ will increase up to the extended window length of $N_{input} + L$ samples.

The constraint that the output signal fits into the window without an overlap is only fulfilled in the case of $N_{input} = 1$:

$$\begin{aligned} N_{input} + N_{IR} - 1 &\leq N_{input} + L \\ 2N_{input} + L - 1 &\leq N_{input} + L \\ N_{input} &\leq 1. \end{aligned} \tag{23}$$

This case is unusable for Fourier analysis.

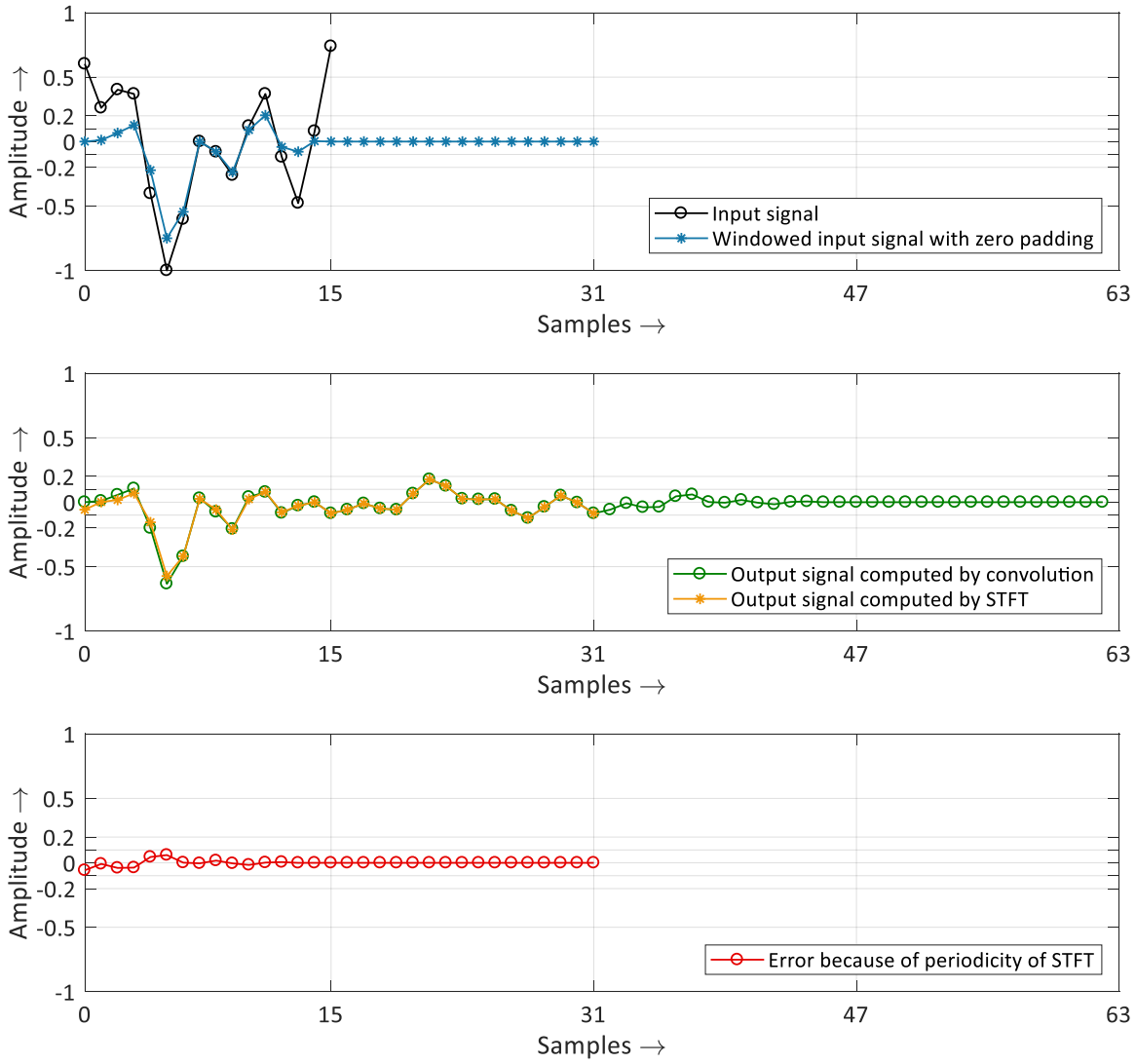


Fig. 6: Comparison of Spectral Subtraction using STFT and the equivalent processing with a convolution in the time domain. A windowed test signal of 16 samples is processed with Spectral Subtraction in the spectral domain and with a convolution in the time domain. The last diagram shows the signal part which is at the wrong position in the output signal when processed with the Spectral Subtraction.

4.3. An Example to Illustrate

To illustrate the behaviour of the DFT in combination with Spectral Subtraction we generate a window of a synthetic input signal:

$$x(n) = \cos\left(2\pi\frac{n}{N}\right) + \sin\left(4\pi\frac{n}{N}\right) + \cos\left(8\pi\frac{n}{N}\right), \quad (24)$$

with $N = 16$ and $n = 0,1,2,\dots,N-1$. The result is the black graph in Fig. 6. We multiply this input signal with the von-Hann window function from (3):

$$x_{win}(n) = x(n) \cdot w_{hann}(n)\,. \quad (25)$$

To get a better frequency resolution we add 16 zeros:

$$\tilde{x}_{win}(n) = \begin{cases} x_{win}(n) & for\ 0 \le n \le 15 \\ 0 & for\ 16 \le n \le 31\,. \end{cases} \quad (26)$$

We receive the windowed input signal with zero padding, as illustrated by the blue graph of Fig. 6.
As an example, we reduce the third, fifth and ninth DFT coefficients by about 70%, using Spectral Subtraction and (4), (10) and (14). The result is the output signal of the Spectral Subtraction, shown as the orange graph. Now we compare this result with the equivalent processing using convolution in the time domain. By means of (19) with $p = 1$ and (21), we receive the green graph. The difference of these two output signals (red graph) shows the wrongly inserted part of the signal, occurring due to the periodicity of the DFT.

5. Analysis of the Impulse Response

If we look to the impulse response $h_{win}(n)$ of the Spectral Subtraction, which is the inverse Fourier transform of $H_{win}(k)$:

$$h_{win}(n) = \frac{1}{N}\sum_{k=0}^{N-1} H_{win}(k) \cdot e^{j2\pi n\frac{k}{N}}\,, \quad (27)$$

it becomes apparent, that the maximum of the impulse response is located at the first sample $n = 0$, as we can see in Fig. 7.

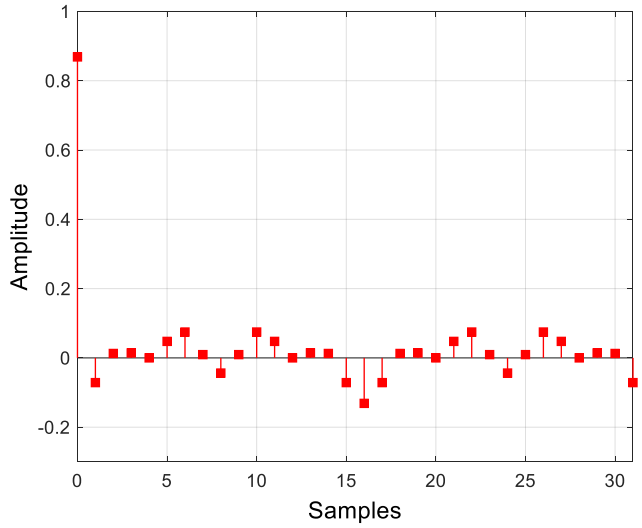


Fig.7: Impulse response of the Spectral Subtraction, computed with (10) and (14). The third, fifth and ninth DFT coefficients are reduced by ~70%.

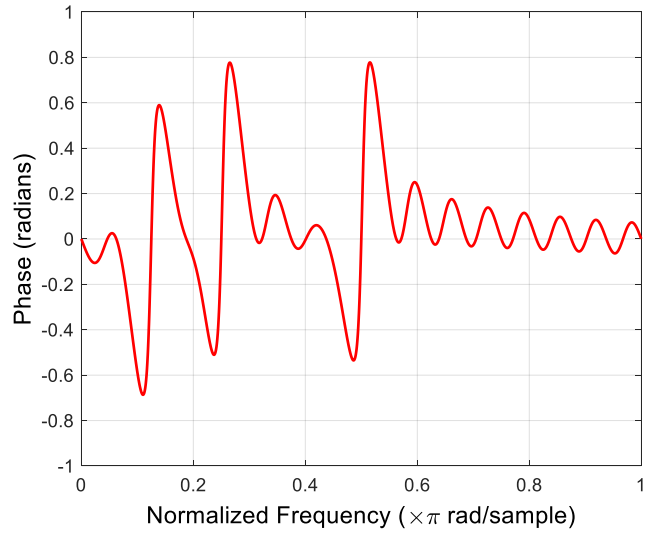


Fig. 8: Phase response of the Spectral Subtraction, computed with (10) and (14). The third, fifth and ninth DFT coefficients are reduced by ~70%.

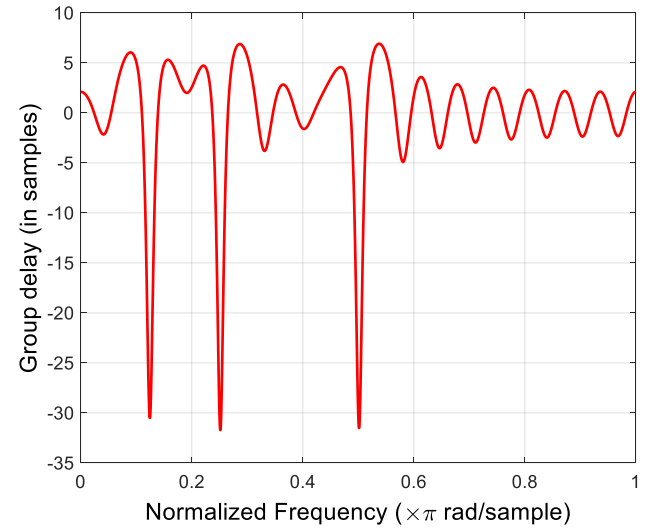


Fig. 9: Group delay response of the Spectral Subtraction, computed with (10) and (14). The third, fifth and ninth DFT coefficients are reduced by ~70%.

Furthermore, the samples $n = 1$ to $n = 31$ are axis-symmetric to $n = 16$. This impulse response behaves as if multiplied with the Heaviside step function:

$$\mathfrak{H}(n) = \begin{cases} 0 & for\ n < 0 \\ 1 & for\ n \geq 0\,, \end{cases} \tag{28}$$

and shows a nonlinear phase shift, as we can see in Fig. 8 and a strong varying group delay depending on frequency, as we can see in Fig. 9. We obtain the strongest group delay at the three processed DFT coefficients.

To prevent nonlinear phase shifting and an inconstant group delay, we must shift the phase within the processing in the spectral domain, depending on frequency. The phase of the DFT coefficients representing high frequencies with a short wavelength have to be shifted more than the phase of DFT coefficients representing low frequencies. For an impulse response with an even length and an even symmetry we obtain the phase difference [2][9]:

$$\theta(k) = -\frac{N-1}{2}\Omega\,, \tag{29}$$

where $\Omega = 2\pi k/N$ is the normalized complex angular frequency. If we include this phase difference in (14), we receive:

$$Y_{win}(k,\theta) = |Y_{win}(k)| \cdot e^{j\angle X_{win}(k)} \cdot e^{j\theta}\,. \tag{30}$$

We can call this enhanced algorithm ‘Advanced’ Spectral Subtraction.

In Fig. 10–12 we can see the symmetric impulse response, the linear phase response and the constant group delay of the Advanced Spectral Subtraction using (10) and (30).

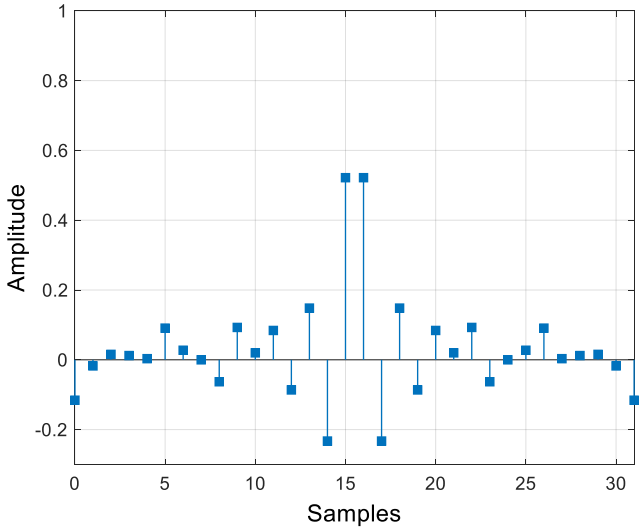


Fig. 10: Impulse response of the Spectral Subtraction, computed with (10) and (30). The third, fifth and ninth DFT coefficients are reduced by ~70%.

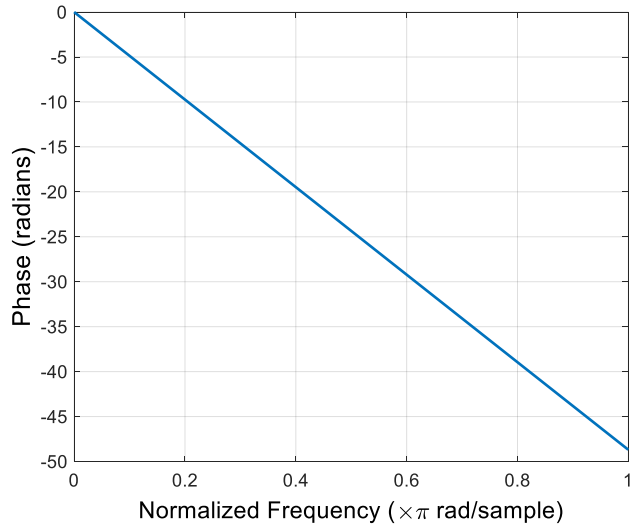


Fig. 11: Phase response of the Spectral Subtraction, computed with (10) and (30). The third, fifth and ninth DFT coefficients are reduced by ~70%.

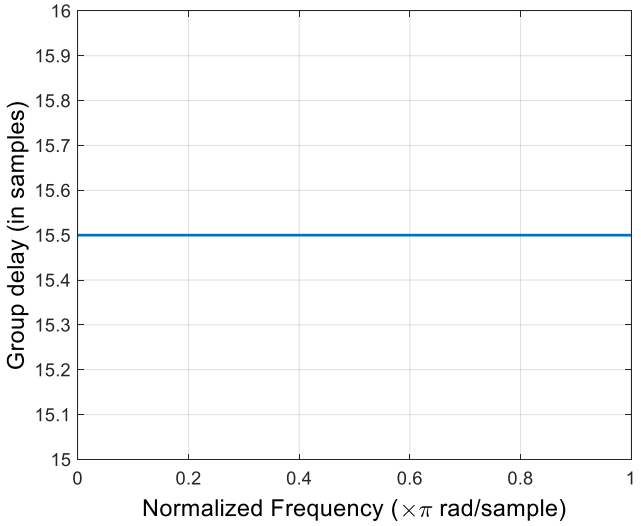


Fig. 12: Group delay response of the Spectral Subtraction, computed with (10) and (30). The third, fifth and ninth DFT coefficients are reduced by ~70%.

As we can see in Fig. 12, the Advanced Spectral Subtraction results in a constant group delay, which also means that the processing has a latency of one half window length.

Finally, we can take a look at the two magnitude responses, computed by Spectral Subtraction using (10) and (14) and by the Advanced Spectral Subtraction using (10) and (30).

The two magnitude responses show strong similarity. We can see the three attenuations, with the red one providing a slightly narrower band width. It also becomes apparent that the Spectral Subtraction with linear phase has a low-pass behaviour at very high frequencies. This is the result of an impulse response with an even length and an even symmetry [2][9]. In the vast majority of cases, this behaviour is of little to no consequence. For example, in digital audio signal processing with a sampling frequency of $f_s = 48kHz$, the cut off is located above the upper limit of human perception.

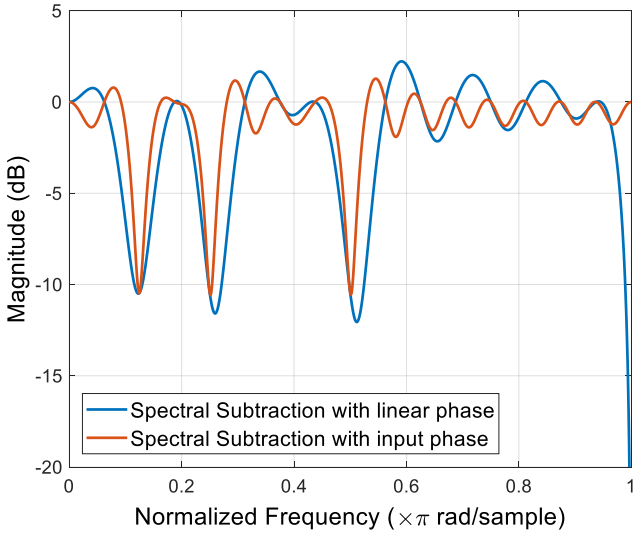


Fig. 13: Magnitude response of the Spectral Subtraction. The blue graph is computed with (10) and (30) and the red graph is computed with (10) and (14). The third, fifth and ninth DFT coefficients are reduced by ~70% within the processing.

6. Conclusion

We can state that processing in the frequency domain makes the signal longer. The signal part by which the output signal is longer than the input signal corresponds to the transient effect and decay process of the impulse response. The crucial point is to arrange the transient and decay parts at the correct time position in the output signal.
If we do the processing in the spectral domain via STFT, because of the periodicity, we receive an overlap in the output signal during resynthesis. This means, that we have a signal part at the wrong time position. Since the STFT does not take this repetition into account, an error inevitably occurs.

The subjective perception of this error is relatively small. Furthermore, it is not the reason of the artefact called ‘musical noise’. Presumably, the resulting error is covered by stronger artefacts like the aforementioned ‘musical noise’, which can occur because of a dynamic processing in the spectral domain, too.

Irrespective of this, it is recommended to work around this error. For example, the resulting amplitude response can be smoothed. This approach minimizes the error, but it does not completely prevent it. To obtain the correct output signal, the frequency response can be generated. By means of the iFFT, we receive the impulse response of the processing. Now it is possible to compute the output signal with convolution of the windowed input signal and the impulse response in the time domain. This means, that the algorithm has more calculating steps and needs more time for the processing. However, with the fast convolution we have a fast-acting tool, which uses the fast Fourier Transform FFT.

The question arises as to why the fast convolution can compute the output signal without an error while still using the DFT. When we use the fast convolution, we have the

windowed input signal and the complete processing information within the impulse response. We don't have to generate the frequency response in the spectral domain. The fast convolution fills up the windowed input signal and the impulse response with enough zeros to fit the entire output signal into the window.
This is still not possible if we generate the frequency response of the Fourier transformed window with the input signal in the spectral domain, like the Spectral Subtraction does. In this case, the frequency response and for this reason the impulse response are always as long as the transformed window. Therefore, the output signal does never fit into the window.

We can conclude that Spectral Subtraction has a fundamental problem within its approach. But it is possible to work around this weak spot and prevent the occurring error. Furthermore, we can use a phase shift within the processing, so that the 'Advanced' Spectral Subtraction does not have any nonlinear phase response or inconstant group delay.

7. References

[1] Benesty, J.; Chen, J.; Habets, E.A.P.: Speech Enhancement in the STFT Domain, Springer Briefs in Electrical and Computer Engineering. Springer, Berlin, 2011.

[2] Bellanger, M.G.: Digital Processing of Signals - Theory and Practice. John Wiley and Sons Ltd, Chichester, 2000.

[3] Boll, S. F.: Suppression of Acoustic Noise in Speech Using Spectral Subtraction. IEEE Tran. on Acoustics, Speech and Signal Processing ASSP-27, 2, 1979, S. 113-120.

[4] Briggs, W. L.; Van Emden, H.: The DFT - An Owners' Manual for the Discrete Fourier Transform. Society for Industrial and Applied Mathematics, Philadelphia, 1995.

[5] Etter, W., Moschytz, G. S.: Noise reduction by noise-adaptive spectral magnitude expansion. Journal of the Audio Engineering Society 42 (1994), S. 341-349.

[6] Groh, J.: Verringerung von Kammfilterverzerrungen bei Multimikrofonaufnahmen. Tagungsbericht 26. Tonmeistertagung (2010), S. 616-625.

[7] Küpfmüller, K.; Kohn, G.: Theoretische Elektrotechnik und Elektronik. Springer-Verlag, Berlin, Heidelberg, 2000.

[8] Neubauer, A.: DFT – Diskrete Fourier-Transformation. Springer Vieweg, Wiesbaden, 2012.

[9] Oppenheim, A.V.; Schafer, R.W.: Digital Signalprocessing. Prentice Hall, Englewood Cliffs, 1975.

[10] Vary, P.: Noise suppression by spectral magnitude estimation-mechanism and theoretical limits. Signal Processing 8 (1985), S. 387-400.

A position-dynamic binaural synthesis of a multi-channel loudspeaker setup as an example of an auditory augmented reality application

Stephan Werner[1], Mina Voigt[2], Florian Klein[1], Annika Neidhardt[1],
[1] *Technische Universität Ilmenau, Electronic Media Technology Group, Ilmenau, Germany,*
Correspondence Email: stephan.werner@tu-ilmenau.de
[2] *MinaLord, Berlin, Germany, Email: post@minalord.berlin*

Abstract

This contribution presents an insight into a position-dynamic binaural synthesis system used in an auditory augmented reality (AR) application. Audio signals from six loudspeakers are reproduced binaurally in a real room. The listener can move freely within an area of approx. 4 m x 4 m surrounded by a virtual loudspeaker setup. The head pose and the listener position are tracked and the corresponding binaural room impulse responses are continuously provided. A setup of five loudspeakers corresponds to a surround setup. One additional loudspeaker is located within the walkable area. A self-created music installation is reproduced with the virtual loudspeaker setup. This study investigates techniques and concepts for creating an interactive musical experience with such a system. Technically, panned audio signals between the surround speakers and discrete audio objects at the several speakers are used to create an acoustic scene. The listener's movements are incorporated as a design element.

1. Introduction

Existing audio systems can reproduce spatial audio in a way that artificial and real audio objects are perceived as plausible audible events in a virtual and/or augmented environment. An auditory illusion of a spatial acoustic environment can be created with the help of existing position-dynamic binaural synthesis systems [1]. The occurrence of such a plausible auditory illusion depends on an adequate technical realization and on several context dependent quality parameters like congruence between synthesized scene and the listening environment or individualization of the technical system for example.

This contribution presents a binaural synthesis system, which adds six virtual sound sources to a real room. The realization is based on BRIRs measured with a sparse distribution in that room. To provide a certain area of action for the listener, a dense grid of BRIRs is generated using energy- and time-based BRIRs synthesis approaches. The system is used to auralize a spatial audio mix.

2. Auditory Augmented Reality

The so-called augmented reality refers to a subset of the so-called reality-virtuality continuum described by Milgram et al. [2]. Figure 1 classifies this subset into the mixed reality approaches. A real environment is supplemented by perceptual events caused by the synthesis of virtual objects. The auditory augmented reality is a technically mediated reality that adds virtual audio objects at defined positions in a real listening room. The virtual objects are simulated with the help of a binaural synthesis system in this case. The aim of the system is to create a plausible auditory illusion for the moving listener.

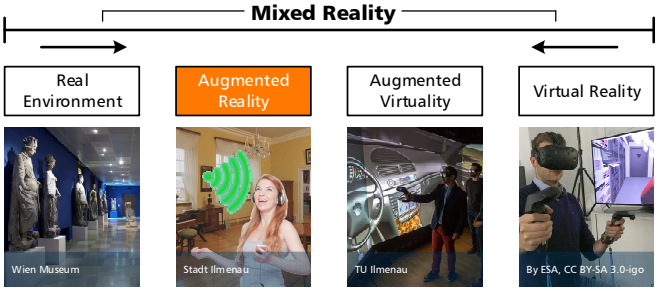


Fig. 1: The reality-virtuality continuum after Milgram et al. [2] with applications as examples.

The technical system must meet the user's expectations for the presented and extended environment. The challenge is to minimize perceptual conflicts resulting from deviations between the perceived and the desired nature of the constructed reality in our cognitive system. These deviations can have a negative impact on the experienced quality in terms of plausibility. Own evaluations of such a system are shown in [1,3] for the quality features perceived externalization and localization of the audio objects as well as for overall impression.

3. Binaural Synthesis System

A technical system which addresses AAR should be able to add virtual audio objects to a real room, should allow one or more listeners to freely move within the augmented real room, and should create a plausible auditory illusion. The additional virtual audio objects should be perceptually fused with the real room. Audio systems which allow a listener to freely move within a room already exist. These systems usually use either simulated BRIRs [4,5], or measurements of BRIRs in the actual listening room. The last approach can (also) provide

ISBN
978-3-9812830-9-9
The authors hold the copyrights of all of content in this manuscript. They have granted the VDT the right to publish it in print, online and as PDF download on *https://tonmeistertagung.com*

a very convincing auditory perception [6]. The drawback is the amount of measurements that must be done for such a system.

Approaches which try to minimize the amount of measurements are for example described by Savioja et al. They merge calculations of head-related transfer functions, acoustic room simulations, apply reverberation, or source directivities [5]. Algazi et al. developed a technique called Motion-Tracked Binaural (MTB) sound for capturing, recording, and reproducing spatial sound using a circular array of microphones on a sphere [7]. When playing back the sound via headphones the headphone signals are interpolated from the microphone recordings. Pörschmann et al. developed an approach which creates BRIRs by using HRTFs and omni-directional representation of the room reverberation [8].

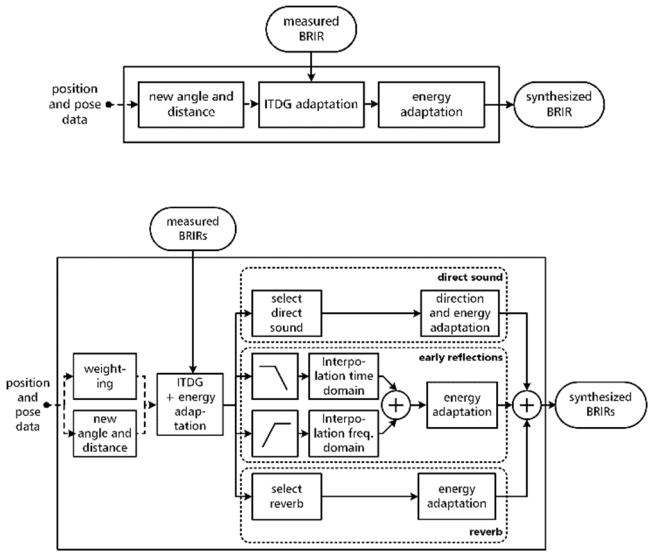


Fig. 2: Block diagrams of the two BRIR synthesis approaches; top: BRIR synthesis using one measured BRIR and adaptation of the initial time delay gap (ITDG) and the overall energy; bottom: approach using two or three measurements and interpolation like the MTB approach.

The approaches used in this paper are based on work from Füg [9] and Mittag [10] which synthesize new BRIRs at different positions in a real room from BRIRs measured at other positions. The methods generate numerous BRIRs from spatially sparse measurements and thus from a few BRIRs. The BRIRs are created by adaptation of time and frequency-based distance cues and applying spatial interpolation (see figure 2). The principle applicability of these or similar approaches is shown in [1,3] for example.

4. Walkable Audio Installation

The binaural synthesis system is used to synthesize six loudspeakers in a room. A centre speaker (C) is positioned at 0° and 3.5 m from the midpoint of an assumed circle. A left and a right speaker (L, R) is positioned at +/-30°, another left and right speaker (SL, SR) is placed at +/- 120°. An additional speaker (S) is placed at 135° (clock-wise orientation) and at 1.4 m from the assumed midpoint.

The listener can move within an area of 4 m x 4 m. The resolution of the binaural synthesis for translation is 0.25 m using a grid of squares. The resolution for horizontal head movements is limited to 5°. This means that the five speakers are outside the walkable area, but the sixth speaker is within. The real-time binaural rendering is realized by the pyBinSim tool [11]. Special tailored audio content was created to make use of this technical system. The musical piece provides and enveloping acoustical surrounding with multiple audio objects which can be discovered best by walking around and exploring the whole area.

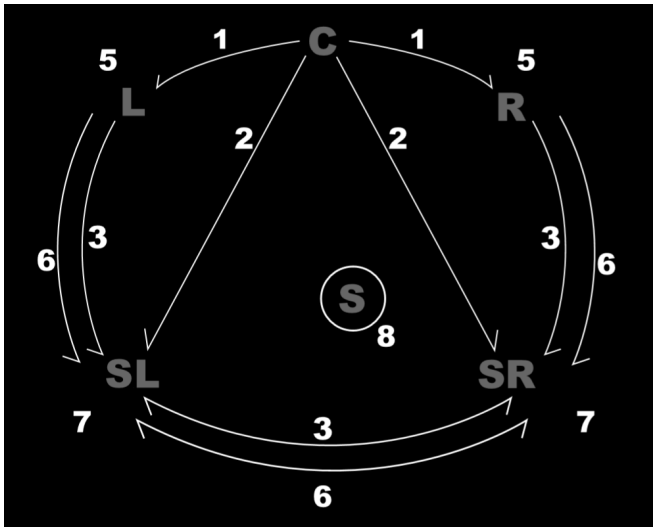


Fig. 3: Graphical overview of the auralized virtual loudspeaker positions; numbers indicate the audio mix (see text).

Figure 3 shows a sketch of the realized audio installation. The letters indicate the six speaker positions while the numbers indicate the development of the audio.

1. The music starts with three different bird samples that move around before they settle at the front stereo pair (L, R), the rear (SL, SR), and center (C).
2. The soundscape is being disturbed by thunder followed by the bass line which focuses on the rear and the centre front.
3. A wind like sound starts moving around the speakers.
4. The specials start kicking in. You can hear a dog at the special (S) speaker and voices coming from all four corners.
5. The drum kit focuses on the front stereo pair with an extra hihat stereo channel at the rear.
6. A ride cymbal played with a brush is moving around the space.
7. The percussion first stays at the rear before later in the song they also start moving around the whole space.
8. A monologue kicks in at the special speaker.
9. As the drums and the bass stop playing, the soundscape of birds, wind, percussion and specials start moving around again more, before they fade out.

A stereo version (down mix) can be found here: https://soundcloud.com/lord_mina/1015-zirkkel-one .

5. Conclusion

The contribution presents an application of our position-dynamic binaural synthesis system. It uses synthesized BRIRs from few real measurements to auralize six sound source positions and a 4 m x 4 m walkable area.

An audio installation is developed for this setup. The audio mix realizes a soundscape which can be interactively explored by the listener. We use such setups to investigate the influence of an exploring listener on the perception of spatial audio in virtual and augmented reality scenarios.

We are convinced that such systems are an exciting and challenging field of activity for engineers and artists. We always keep our eyes open regarding cooperation.

6. References

[1] Brandenburg, K., Cano, E., Klein, F., Köllmer, T., Lukashevich, H., Neidhardt, A., Sloma, U., and Werner, S., "Plausible Augmentation of Auditory Scenes using Dynamic Binaural Synthesis for Personalized Auditory Realities", Audio Engineering Society Conference on Audio for Virtual and Augmented Reality, Redmond, USA, 2018.

[2] Milgram P. et al., "Augmented Reality - A class of displays on the reality-virtuality continuum". In: SPIE-Telemanipulation and Telepresence Technologies 2351, 1994.

[3] Werner, S., Neidhardt, A., Klein, F., and Brandenburg, K., "Comparison of Different Methods to Create an Interactive Augmented Auditory Reality Scenario Using Sparse Binaural Room Impulse Response Measurements," in 44th Annual Meeting for Acoustics, DAGA, Munich, Germany, 2018.

[4] Zimmermann, A. and Lorenz, A., "LISTEN: a user-adaptive audio-augmented museum guide," User Modeling and User-Adapted Interaction, 18(5), pp. 389–416, 2008.

[5] Savioja, L., Huopaniemi, J., Lokki, T., and Väänänen, R., "Creating Interactive Virtual Acoustic Environment," J. Audio Eng. Soc., 47(9), pp. 675–705, 1999.

[6] Brinkmann, F., Lindau, A., and Weinzierl, S., "On the authenticity of individual dynamic binaural synthesis." J. Acoust. Soc. Am., 142(4), pp. 1784–1795, 2017.

[7] Algazi, V. R., Duda, R. O., and Thompson, D. M., "Motion-Tracked Binaural Sound," in 116th Audio Engineering Society Convention, Berlin, Germany, 2004.

[8] Pörschmann, C., Stade, P., and Arend, J., "Binauralization of omnidirectional room impulse responses-algorithm and technical evaluation," in 20th Int. Conf. on Digital Audio Effects (DAFx), UK, pp. 345–352, 2017.

[9] Füg, S., Werner, S., and Brandenburg, K., "Controlled Auditory Distance Perception using Binaural Headphone Reproduction - Algorithms and Evaluation," in VDT Int. Conv., 27. Tonmeistertagung, pp. 614–620, 2012.

[10] Mittag, C., Werner, S., and Klein, F., "Development and Evaluation of Methods for the Synthesis

of Binaural Room Impulse Responses based on Spatially Sparse Measurements in Real Rooms," in 43rd Annual Meeting for Acoustics, DAGA, Kiel, Deutschland, 2017.

[11] Neidhardt, A., Klein, K., Knoop, N., and Köllmer, T., "Flexible Python Tool for Dynamic Binaural Synthesis Applications", 142nd Convention of the Audio Engineering Society, Berlin, Germany, 2017.

Traveler's portable plate absorber for field applications

R. Mores
University of Applied Sciences Hamburg, Hamburg, Email: robert.mores@haw-hamburg.de

Abstract

Fieldworkers in musical acoustics often conduct recordings of ethnological sound sources. Cataloguing musical instruments by means of impulse responses, for instance, demands independence from individual acoustical rooms in the field. A free field would be perfect, however, not while travelling. To achieve at least some neutrality, a simple plate absorber is suggested to suppress floor reflections. Wall or ceiling reflections might be late anyway when rooms are large, so reflection-free impulse responses can then be captured. Conventional molton curtains, otherwise not very absorbing, can be arranged to function as plate resonator. After some analytical suggestion and some impedance tube pretests, the absorption of an applicable absorber reaches down to 100 Hz as confirmed by ISO 365 measurements. Stripes of corrugated paper - arranged meander-like, upright on the floor - are strong enough to support some layers of molton curtain. The total weight is less than 1 kg and the volume is roughly 10 l for every m² of absorber, fitting into every travelling bag. But the absorber can also be build ad-hoc anywhere due to the simplicity of material and construction.

1. Introduction

Field recording of musical instruments ask for defined acoustical properties of the envelopping room, especially for purposes such as cataloguing or vibroacoustical analyses. A free field is usually not around in remote places, and it is impractical to move instruments, or to bring instruments and laboratory conditions together otherwise. A free field outside buildings is an option but regularly ambient sound sources are too loud. While it seems also impractical to carry loads of equipment, we search for a pragmatic way to build necessary absorbers ad-hoc, or carry such absorbers in a traveller's bag. Molton curtain, or at least cloth or dust sheets can be expected to be available in remote places, as well as corrugated paper, or at least cardboard.

2. Analytical Suggestion

Molton curtain, while lying directly on the floor, or hanging directly at a wall has only little damping effect, and only forhigh frequencies. However, when arranged at a specific distance to a reflecting floor, or wall, the cloth acts as plate resonator. The resonator's frequency is determined by the distance and the mass per area according to

$$f = \frac{c}{2\pi} \cdot \sqrt{\frac{\rho}{m_A \cdot d}} \quad (1)$$

with the density, ρ, and the speed of sound, c, for the given air, the mass per area m_A, and distance d between cloth and reflecting plane. Using 1.2 kg/m³ (sea level at 20 °C), and one or two layers of conventional Molton curtain with mA =300 g/m², the frequency varies by distance as shown in Fig. 1.

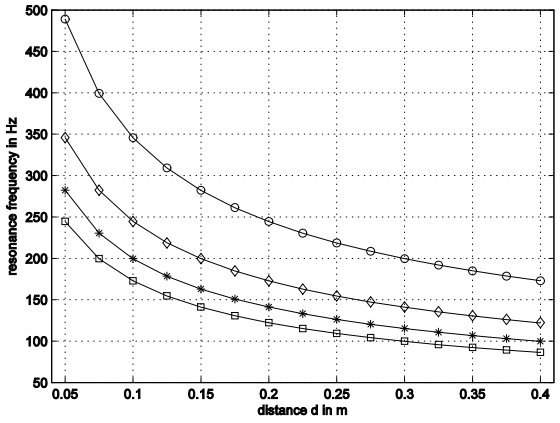


Fig. 1: Resonance frequency (analytical prediction) for one (o), two (◊), three (*), and four (□) layers of cloth with 300g/m² arranged at distance *d* from a reflecting plane.

3. Impedance Tube Measurements

Fig. 2 shows the results of impedance tube measurements. One or two layers of mentioned Molton curtain show little absorption, when directly arranged at the reflecting surface, $d = 0$. When brought to distance relative to a reflector in the tube, the absorption increases, especially for low frequencies.

ISBN
978-3-9812830-9-9
The authors hold the copyrights of all of content in this manuscript. They have granted the VDT the right to publish it in print, online and as PDF download on *https://tonmeistertagung.com*

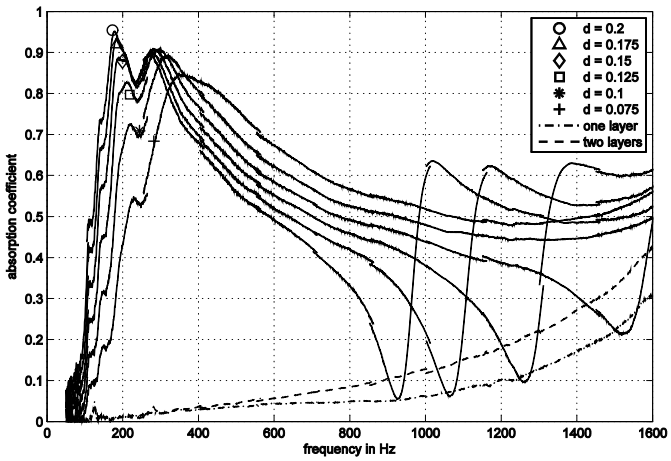


Fig. 2: Absorption measured in an impedance tube for one (-·) and two (--) layers of cloth with 300 g/m² arranged at distance d = 0 m from a reflective plane, as well as two layers arranged at distance d as indicated in the legend. Position of symbols indicate the predicted resonance frequency of the related plate absorber.

Predicted resonance frequencies are indicated in the figure where the respective legend symbol is placed on indiviual traces. The prediction is perfect for d = 0.2 m. For other frequencies it is moderate due to edge effects of the probe in the tube. Another impairing effect is the step at 125 Hz caused in the given impedance tube under all probe conditions. However, the general agreement of tube measurement results with prediction is encouraging, in particular the considerably high absorption at considerably low frequencies, below 200 Hz. These findings suggest to forward to the next step of application.

4. Absorption Measurement in Echo Chamber acc. to ISO 354

While heading towards applications two things change: (i) the sound is likely to be diffuse in real applications while it was directional in the impedance tube, (ii) the layers of cloth must be brought to a defined distance relative to a reflecting plane, by any means, which will introduce additional material. For reasons of a pragmatic approach, stripes of corrugated paper are used for keeping the space between plates. These are just strong enough to support some layers of cloth. The stribes are arranged meander-like on the floor before the layers of cloth are gently let down on the construction. Four layers of somewhat a thinner Molton curtain (160 g/m²) are used which amount to the same weight per area as above. The absorption quality is verified by measurements in a echo chamber. ISO 354 [1] is followed apart from a few parameters: the absorption area is not 10 m² to 12 m² but only some 4 m², and six rather than twelve speaker-microphone transfer-paths are used for averaging.

Fig. 3 shows the results for T30 measured in a echo chamber of volume V = 152.2 m³, at temperature T = 27 °C and at 30% relative humidity.

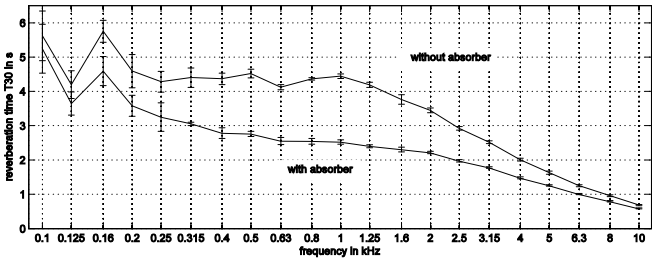
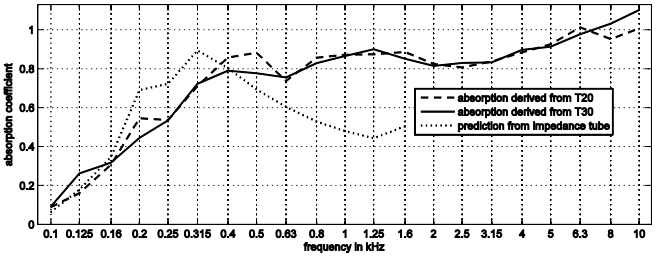


Fig. 3: RT30 measured according to ISO 354 in an echo chamber of volume V = 152.2 m³, without, and with an absorber of size 4 m², consisting of 10 cm heigh corrugated paper arranged on the floor, and four layers of Molton curtain, m_A = 160 g/m² (nominal), on top.

The absorption coefficient, as derived from the RT30 times, respects given absorber edge effects [2], as measured during setup. Fig. 4 shows the results for absorption coefficients derived from T30 but also those derived from T20, in



comparison with the prediction from the impedance tube.

Fig. 4: Absorption coefficient derived from ISO 354 measurements for the absorber of size 4 m², consisting of 10 cm heigh corrugated paper arranged on the floor, and four layers of Molton curtain, m_A = 160 g/m² (nominal), on top.

In the application, the real absorber does fairly well reach into the low frequency range as predicted. An absorption of 0.5 was expected at about 180 Hz. The real absorber reaches this value at about 190 Hz (with reference to RT20), or 230 Hz (with reference to RT30). The other positive outcome is that the drop back at 1.25 kHz does not appear. The absorption rather continues to grow up to 100% due to the diffusive sound field.

Fig. 5 shows the implementation of a 17 cm high absorber, for which all measurements have been repeated. The higher absorber reaches further down in terms of frequency. An absorption coefficient of 0.5 was expected at 125 Hz and also confirmed at 125 Hz for the real absorber by both references, RT20 and RT30. An absorption of 0.8 is reached at 250 Hz. Unlike the 10 cm high absorber, the 17 cm high absorber shows a drop down to 0.75 at 1.25 kHz, predicted were 0.2 at 1.05 kHz, see Fig. 2. The higher the absorber, the more likely it seems that the predicted drop appears in the real absorber. However, the drop is much weaker than expected.

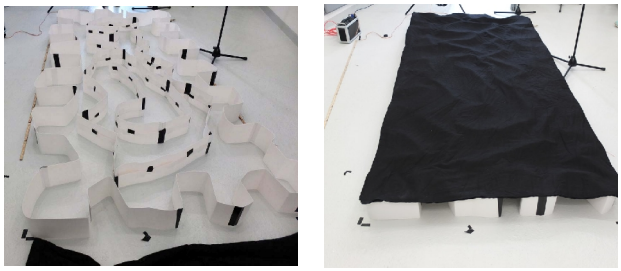


Fig. 5: Floor absorber consisting of (left) a supporting structure made of corrugated paper with (right) four layers of Molton, m_A = 160 g/m², on top of the 17 cm high absorber.

To conclude, the results from the impedance tube can be used as a predictor for the real absorber in a diffuse sound field. The real absorber is likely to maitain a high absorption, without a strong drop, throughout the frequency range, beginning with a lower bound frequency somewhat higher than predicted.

5. Application

An absorber can now be realized at low weight. For instance, a 1 m² absorber arranged on the floor should be sufficient to record guitar impulse responses. The total mass is about 600 g for the Molton and 200 g for the corrugated paper. The volume is less than 10 l, sufficiently little to fit into a traveller's bag. Using the absorber at the floor, advantageously in the middle of a larger room, will allow to take early reflections into account only from walls or the ceiling. Taking the speed of sound propagation, 5 m distance to either wall should allow for almost clean impulse responses for up to 30 ms.

The absorber can also be useful when recording sound samples during performance, for instance when playing in front of a wall. Sound engineers might wish to apply such a simple absorber to get rid of disturbing reflections, for instance when applying close-ups.

7. References

[1] DIN EN ISO 354:2003-12, Akustik - Messung der Schallabsorption in Hallräume (ISO 354:2003); Deutsche Fassung EN ISO 354:2003, Beuth Verlag.

[2] Sauro, R., Vargas, M., & Mange, G. (2009). Absorption coefficients-part 2: is "edge effect" more important than expected?. *Internoise, Ottawa, Canada, August 2009*.

Towards the Development of an Automatic 2D- to 3D-Audio Upmix Algorithm

J. Mäsel[12]
[1] *Film University Babelsberg, Email: info@tuneandfocus.com*
[2]*This work was performed as a Master Thesis at SONY Europe Ltd., EuTec SSG, Stuttgart in 2016*

Abstract

Different approaches to the blind upmixing of 5.1-stereophonic material to 3D-audio systems are presented and a perceptually effective upmixing algorithm is proposed. The suggested method focuses on the generation of additional lateral content (using vector decomposition for the extraction of correlated material) as well as on the creation of additional content for the upper listening hemisphere (using blind reverberation extraction based on signal analysis with recurrent neural networks and inverse impulse response estimation for spectral subtraction processes). An extensive set of stimuli (classical music / AV content) has been created for the formal expert listening test which achieved promising results for the method proposed, showing significant increase in width and depth perception as well as listener envelopment. Key words: Blind Upmixing, 3D-Audio,Monopole Synthesis, Reverberation Extraction, 2-to-3 Upmix, Primary Component Analysis

1. Introduction

The amount of audio-visual content mixed especially for 3D-audio formats has been constantly growing over the last years. Still, most of the consumer content is only available in stereophonic formats. This also holds true for existing market formats like radio broadcasts or digital concert hall subscriptions.
It is therefore desirable to design a blindly working, 'one-size-fits-all' algorithm capable of upmixing content in order to provide material for novel consumer playback formats featuring height loudspeakers.
However, to protect a work's integrity, the artistic intention of the mix must not be affected by the automatic upmix. Any harmful changes regarding the basic audio quality, tone colour, source localisation, or other attributes affecting spatial perception have to be avoided.

The main premise for the algorithm developed in the context of this work is always that, in the best case scenario, it improves the examined parameters, while in the worst it does not perform worse than the original mix.
As there are already many well studied techniques for upmixing 2-channel stereophonic content to 5.1, this work concentrates solely on 5.1 source material which was created from our own classical music recordings in Berlin Philharmony as well as relevant AV content. Furthermore, 3D mixes from the according recordings were created to serve as high quality anchor in the later expert listening test.

For the creation of a spatial sound field, the 'Monopole Synthesis' (MPS) method, whose underlying theoretical principles are derived from Wave Field Synthesis, is used [1]. The methods under test that are used to generate material for the use of upmixing a 2-dimensional source to a 3-dimensional sound field are

- the extraction of correlated material from the two adjacent side channels L/L_s and R/R_s to feed the new virtual speakers L_m/R_m. This approach uses Principal Component Analysis (PCA) with certain modifications.
- the extraction and redistribution of the source's original reverberation to the upper listening hemisphere. We suggest the use of concepts from machine learning as well as feature creation derived from human reverberation perception later used for re-levelling and reverberation time estimation.
- a virtual 5.1 system which expands the sources' radius, allowing the leaking of sound into the height and additional side speakers.

2. Methods

2.1. Primary-Ambient-Separation

One approach to upmixing two channels to three is to take advantage of common stereophonic mixing techniques. Sounds that are to appear in between two speakers are placed there as phantom sources. Phantom sources that appear in the exact centre between two speakers are those sounds that are fully correlated and, level-wise, equally distributed onto each of the two stereo speakers so there are neither Interchannel Level nor Time Differences (ILTD).
Extracting the correlated material from two channels now provides material to feed a third channel. An according loudspeaker could then be placed in between the two stereo speakers, depicting the physical equivalent of the former phantom source.
Considering 2-channel stereo mixes, signal decomposition into decorrelated and correlated material can be achieved by using an approach called Principle Component Analysis (PCA), developed by Pearson [2]. PCA is a geometrical method used for vector decomposition. Following Michael Goodwin et al. [3], a multichannel audio signal can be seen as a conglomerate of vectors $\vec{X}_{im}$ for each of the m channels. These vectors can be divided into correlated and decorrelated signal parts considering further assumptions which is often referred to as Primary-Ambient-Separation.

ISBN
978-3-9812830-9-9
The authors hold the copyrights of all of content in this manuscript. They have granted the VDT the right to publish it in print, online and as PDF download on https://tonmeistertagung.com

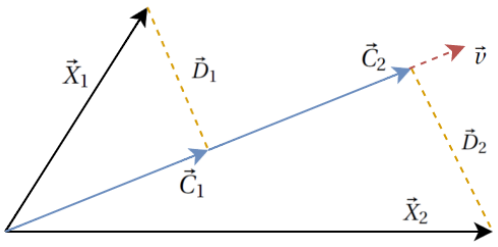


Fig. 1: *Orthogonal vector decomposition of a stereo signal* $\{\vec{X}_1, \vec{X}_2\}$ *using PCA. Note that* $\vec{C}_1$ *and* $\vec{C}_2$ *do not fit Eq. 4 here. After Goodwin (2007) [4]*

They make several assumptions for the input matrix:

For $X = [\vec{X}_1, \vec{X}_2, ...\vec{X}_m,]$

$$X = C + D = \vec{v}(\rho_1, ...\rho_m) + [\vec{A}_1, ...\vec{A}_m] \quad (1)$$

C represents correlated material with correlation index $r = 1$ and D all signal components but C. $\vec{v}$ is the unit vector which best describes the conglomerate of channel vectors and ρ the corresponding coefficients for the different $\vec{C}_m$ in matrix C.

Further assumptions are that $\vec{C}_1$ and $\vec{C}_2$ are *fully* correlated

$$\vec{C}_1 = a.\vec{C}_2 \quad (2)$$

with a representing the complex difference in amplitude. Moreover, that all $\vec{C}$ and their corresponding $\vec{D}$ are orthogonal and thus fully uncorrelated:

$$\vec{C}_m \vec{D}_m = 0 \quad (3)$$

His assumptions were modified for this work's purpose so that all *correlated* components have the same magnitude

$$|\vec{C}_m| = |\vec{C}_{m+1}| \quad (4)$$

For the PCA-based vector decomposition, the first thing to do is to compute the inter-channel correlation r_{12} and the auto-correlation r_{mm} with

$$r_{ij} = \vec{X}_i \vec{X}_j \quad (5)$$

Next, the principal unit vector $\vec{v}$ (Eq. 1), that is the one that results in the highest energy for the correlated component C and fulfils above assumptions 2-3, has to be found. Put simply, this is a maximisation problem for the correlation between $\vec{v}$ and the channel signals $\vec{X}_1$ and $\vec{X}_2$.

The correlated component is now calculated for each channel, projecting $\vec{X}_m$ onto $\vec{v}$ as seen in Fig. 1. It becomes obvious here that $\vec{C}_1$ and $\vec{C}_2$ are correlated as they are multiples of $\vec{v}$. Further, one can see that $\vec{D}_m$ and $\vec{C}_m$ are orthogonal to each other.
To fulfil Eq. 4, only those $\vec{C}_m$ with the same absolute value shall be taken into consideration.

For the test material processing, we used an improved PCA algorithm by Ibrahim [5], modified for this work in order to hand back only those correlated vectors that have the same absolute value. Thus, no ILTDs exist and only the phantom sources from the very middle between the front and corresponding rear speakers are extracted. L / L_s and R / R_s are then fed with the decorrelated ('*Ambient*') components while L_m and R_m receive the correlated ('*Primary*') material.

2.2. Reverberation Extraction

From each of the five original monophonic channels, reverberation was extracted and distributed in the listening sphere using MPS. Subjective experiments with many examples of this work's stimuli corpus were carried out in order to find the best parametrisation. It was attempted to compare the largest amount of possible positions and spread variations and the best quality of extracted reverberation.

We suggest an algorithm using a combination of different approaches within the context of machine learning, that is reverberation time estimation using recurrent neural networks (RNN) as can be seen in [6]. The extraction process follows the principles presented by Soulodre [7], assuming

1. the imperceptibility of the human auditory system towards short-time phase variations
2. the stability in amplitude for nearby IRs in a room, compared to their strong variations in phase
3. the derivation of the phase of the reverberant signal from the dry signal (the reverberant signal is a 'mapping' of the dry signal at all times naturally) so the input signal's phase can be taken as an estimate for the dry and the reverberated component's phase

Basically, the temporal progression of an impulse response (IR) corresponds to a finite impulse response (FIR) filter. Under the circumstance that all the FIR filter stages representing the room IR are precisely known, and all their poles lie inside the unit circle, that is, a minimum / linear phase filter, a correspondent infinite impulse response (IIR) filter can compensate the FIR.
The spectral subtraction is done regarding only the *magnitude* of the input signal, not the *phase*, in consistency with above explained point (3), stating that the input's phase and the dry / reverberant component's phase are equal, and also with point (1) and (2), stating that our hearing is insensitive to short-term phase variations. That is, the same room impression goes along with strongly varying phase responses and relatively constant magnitude responses. These are the underlying assumptions that justify neglecting the phase information from the IR and choosing a linear phase FIR filter for it.
Based on the RNN's RT estimation, an approximate room IR can be guessed. Put simply, this is everything needed to compute the linear phase FIR, the compensating IIR, and with their help, the direct and reverberant component estimates.

Finally, to obtain an adequate measure for the reverberation 'intensity', that is, how much of the extracted material should be heard (comparable to a dynamic dry/wet controller), it is proposed to estimate the direct to reverberation ratio (DRR) from the material's modulation spectrum (MS), calculating the ratio of low to high frequency MS (For further information on MS, see [6]).
The reason for an intensity control for the extracted reverberation based on DRR shall be illustrated with an example from the preceding subjective experiments. During this time it was found that e.g. in intimate parts of a film with mainly dialogue from the center channel, that is a rather

monophonic impression, the dereverberation algorithm used had to be turned down in volume, distributing too much of the extracted reverberation in the upper hemisphere and thus watering the artistic intention of very directional speech with very little envelopment. Thus, in the case of already very little reverberation, the danger of 'spatial over-enhancement' might be reduced by attenuating the extracted material. On the other hand, for the very reverberant orchestral examples, amplification of reverberation seemed to improve envelopment and the material withstood higher reverberation levels without being severely affected.

3. Experiment Setup

The listening test is based on ITU.BT.500 recommendation for the "MUlti Stimulus test with Hidden Reference and Anchor" (MUSHRA). In order to compare more than just one single system to the reference, subjects could switch at will between any of the systems. This allowed subjects to more easily detect differences between the systems, and grade them with a high degree of resolution. They were to compare the methods explained below for different stimuli regarding the parameters 'Basic Audio Quality', 'Tone Colour', 'Localisation Plausibility', '3D Listener Envelopment', 'Width' and 'Depth'.

The listening room's acoustics were aimed at simulating living room conditions. Hardwood floor, wooden walls and a diffuse ceiling plastering provide a vivid room impression with measured $RT_{60} \approx 0.25s$.
The loudspeakers for the horizontal 7.0 setup were seven Genelec 8030A with linear frequency response (no dips activated). For the height speakers, 4 Tennokoe speakers were mounted on the Genelec speakers' backs, facing towards a mobile gypsum plasterboard intermediate ceiling construction in order to create precise reflections. Furthermore, two Genelec 7050B active subwoofers were placed in the front and back of the studio.

4. Listening Test Evaluation

	PCA	RevEx	v5.1
Basic Audio Quality	**59,6**	57,53	34,33
Tone Colour	55,53	**56,07**	42,33
Localisation Plausibility	57,67	**58,44**	38
3D Envelopment	64,4	**72,8**	45,87
Width	63,87	**69,67**	47,67
Depth	64,33	**73,6**	41,6

Tab. 1: *Average mean results for the ratings over all stimuli for the three systems to be compared. Best results are highlighted*

Test subjects (all of them are long year professionals in the AV industry, namely film mixers, recording producers and engineers) were told to evaluate the criteria as depicted in tab. 1 for a reference stimulus (further referred to as "6-ch"), the upmixing methods described above (further referred to as "PCA" for the generation of lateral content, "RevEx" for the extraction and redistribution of reverberation and "v5.1" for the virtual 5.1 system based on MPS) and a low quality anchor (further referred to as "LoQual"). Clear tendencies in the evaluation of the three systems under test could be shown. Subjects generally rated the PCA and RevEx systems highly positively regarding the spatial attributes '3D LEV, 'width', and 'depth' (see Fig. 5, 6 and 7). No severe changes in tone colour could be observed except for those arising from the extensive use of height speakers introducing a different spectral behaviour compared to the main speakers (see Fig. 3). The observed localisation shifts of sound objects were rated positive for these systems as well (see Fig. 4). Furthermore, basic audio quality seems to be improved by them (see Fig. 2).
The v5.1 system is rated much poorer than all the other systems (except for the LoQual system of course) for all criteria, even though it uses more loudspeakers than the reference version.

Post-test discussions with the expert listeners helped to support the results shown and specify the reasons for their ratings. Apparently, positive impressions from each of the systems were highly dependent on the program material. As a result, the desired improvements for spatial parameters were best perceivable for silent, intimate material with only very little energy portions in low MS frequency bands. Loud, chaotic material with many discrete sound objects panned to all directions was considered particularly hard to evaluate.

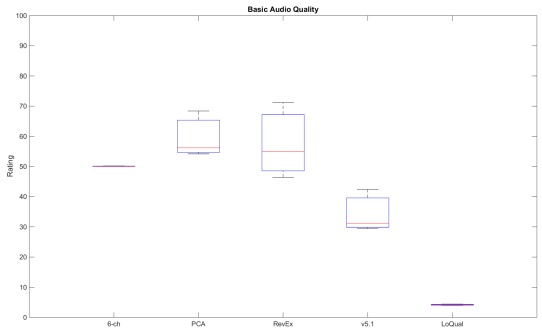


Fig. 2: *Results for the average ratings (y-axis) on basic audio quality over all stimuli for the systems under test (x-axis)*

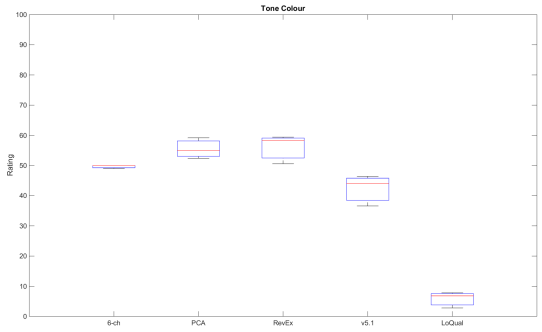


Fig. 3: *Results for the average ratings (y-axis) on tone colour over all stimuli for the systems under test (x-axis)*

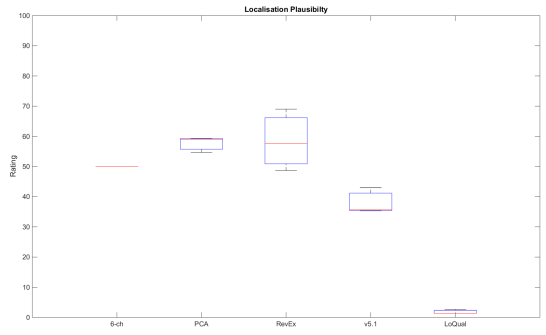


Fig. 4: *Results for the average ratings (y-axis) on localisation plausibility over all stimuli for the systems under test (x-axis)*

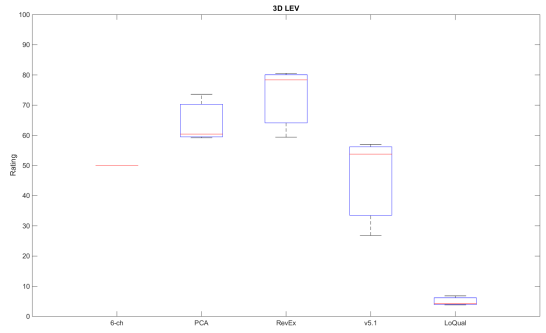


Fig. 5: *Results for the average ratings (y-axis) on 3D listener envelopment over all stimuli for the systems under test (x-axis)*

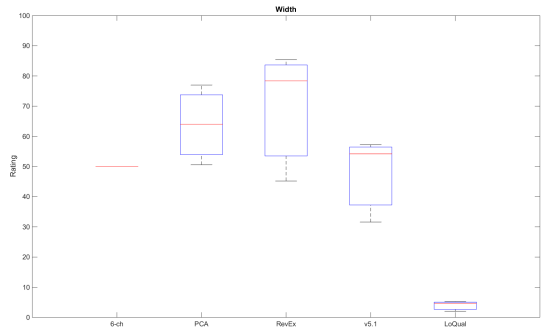


Fig. 6: *Results for the average ratings (y-axis) on width over all stimuli for the systems under test (x-axis)*

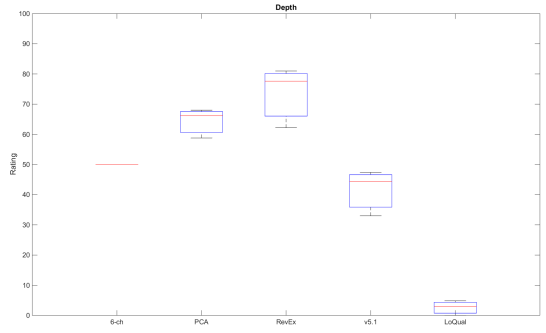


Fig. 7: *Results for the average ratings (y-axis) on depth over all stimuli for the systems under test (x-axis)*

5. Proposed Model

Based on the listening test results, the proposed model consists of a combination of the PCA and RevEx systems as depicted in Fig. 8. It is considered to particularly increase depth, width, and 3D LEV, whilst assuring only minimal degradation of tone colour and basic audio quality in the worst cases, if at all.

The proposed model is a generative model. This means that it tries to modify the original 5.1 tracks as little as possible and generate new material for the upmix out of them instead. It is generally divided into four stages:

- The reverberation stage, where RT_{60} gets estimated, DRR computed and original reverberation extracted from each monophonic input channel. It creates 5 new channels which get re-levelled based on the measured DRR.
- The lateral signal generator, where correlated material gets extracted from the L/L_s and R/R_s speaker pair to generate two new channels, L_m/R_m. This is the only stage where parts of the original mix get modified. They are then adjusted in level against the unmodified centre speaker.
- The spatialisation stage, where, partly based on the user dependent loudspeaker setup, the newly generated channels are positioned in the listening hemisphere.
- The output section, where spectral shortcomings of the height speakers are compensated before the final distribution to the hardware output. The LFE channel is sent directly to the hardware output without further modification.

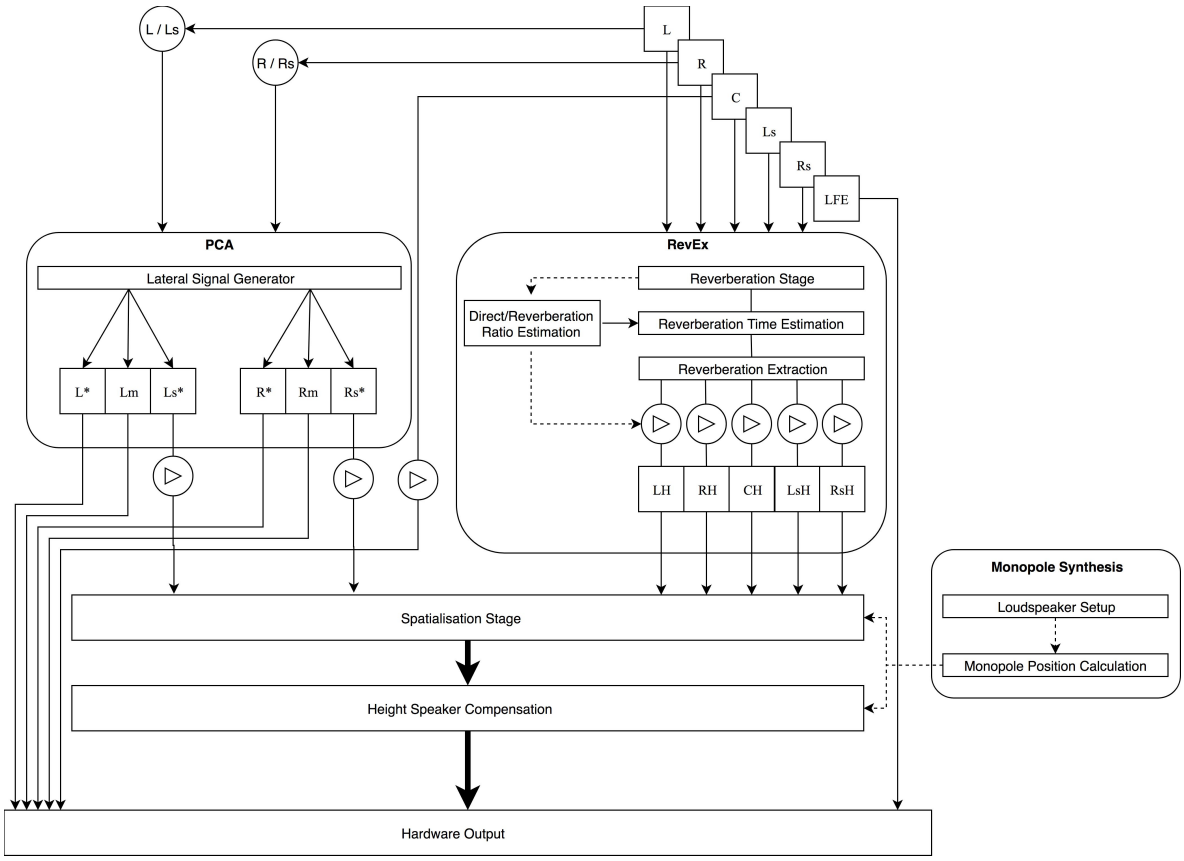


Fig. 8: *Proposed Model*

6. References

[1] Franck Giron. Method, Device and System for Approximating the Synthesis of a Target Sound Field, 2015.

[2] Karl Pearson. On Lines and Planes of Closest Fit to Systems of Points in Space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(1):559–572, 1901.

[3] Michael M . Goodwin and Jean-Marc Jot Valley. Primary-Ambient Signal Decomposition and Vector-Based Localization for Spatial Audio Coding and Enhancement. 2007.

[4] Michael M. Goodwin. Geometric Signal Decompositions for Spatial Audio Enhancement. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 409–412, 2008.

[5] Karim M. Ibrahim. Primary-Ambient Separation of Audio Signals. 2016.

[6] João F Santos and Tiago H Falk. Blind Room Acoustics Characterization Using Recurrent Neural Networks and Modulation Spectrum Dynamics. 2016.

[7] Gilbert A Soulodre. About this Dereverberation Business: A Method for Extracting Reverberation from Audio Signals. 2010.

Software-defined decoder and encoder for data cassette formats of historical audio systems

K. M. Indlekofer[1]

[1] *Hochschule RheinMain, University of Applied Sciences, Wiesbaden, Germany,*
Email: michael.indlekofer@hs-rm.de

Abstract

Data cassette tape formats based on standard audio Compact Cassettes have been commonly employed for non-volatile data storage for historical electronic musical instruments and other audio production and processing systems. Within this article, examples of software-defined decoder and encoder tools for such data cassette formats are presented, providing a means of data recovery, analysis, and conversion. This approach thus enables the digital preservation and re-usage of data recordings of historical instruments and audio systems.

1. Introduction

Historically, audio Compact Cassettes [1,2] have been used as a low-lost medium for data storage for sound generators and audio equipment such as drum machines, synthesizers, and organs. On such "data cassettes", a pseudo-audio signal is employed to encode digital information, for example drum patterns, arrangements, sound programs, patch data, and system configurations.

Reliably usable data rates on a data cassette are on the order of a few 100-1000 bits per second, requiring a recording time of a few seconds up to minutes in order to save a typical payload data amount of a few kilobytes. Typical digital modulation methods [3] that have been used on data cassettes are for example FSK (frequency shift keying), Manchester code, Bi-Phase-Mark code, and general bit group coding techniques. Such schemes enable simple clock recovery, and obey pulse duration constraints as well as spectral limitations. In addition, parity bits or checksums typically provide a rudimentary means of error detection.

Within this article, examples of software-defined decoder and encoder tools for such data cassette formats are presented, enabling data recovery and analysis as well as data conversion. The main goal of this approach is to enable digital preservation, archiving, and re-usage of musically valuable compositions, arrangements, and sound designs of historical drum machines, organs, and synthesizers. The employed multi-platform, open source approach provides an implied documentation of the particular historical storage formats.

2. Data Storage on Compact Cassettes

Certain vintage audio devices such as drum machines, synthesizers, and organs are still of great interest, representing electronic musical instruments with unique sound quality and being part of the cultural heritage of the 20th century. From a technical perspective, such historical audio equipment typically employed early microprocessors (CPUs, typically with 8bit or 16bit data busses only), thus being specialized embedded computer systems. Typically, these CPUs were responsible for system and memory control, user interface and display control, input/output interfaces, and real-time functionality (e.g. in rhythm pattern control and sequencing, sound generator control etc.). Typical "essence" data amounts in vintage audio devices were on the order of 1-10KB for example for drum patterns, sequences, arrangements, sound programs, short audio waveforms, patch data, and system configurations. In addition, the usage of longer audio samples required additional, larger memories (e.g. in sample EPROMs or RAM [22]). Typical historical means of non-volatile storage of such "essence" data (or even complete memory "dumps") are for example:

- memory cartridges (typically with EPROMs, EEPROMs, or NVRAM [22], internally or externally)
- "data cassette" with magnetic tape based on the standard audio Compact Cassette [1,2] format (low-cost, low CPU demands, commonly available)
- other cassette formats based on magnetic tape (specialized cassette and drive formats)
- floppy disks (large data amount but expensive in the 1970s-1980s)
- RS232 packets saved by a host computer (expensive in the early 1980s, mostly for experts)
- MIDI System Exclusive messages [6] saved by a MIDI sequencer/tool (commonly available since the mid-1980s)

Fig. 1: Picture of a typical "data cassette" based on the standard audio Compact Cassette (of type "C 10" in the shown example).

In this contribution, the focus is on data cassette formats (see e.g. Fig.1) which are based on the audio Compact Cassette

ISBN
978-3-9812830-9-9

The authors hold the copyrights of all of content in this manuscript. They have granted the VDT the right to publish it in print, online and as PDF download on https://tonmeistertagung.com

[1,2]. Here, the permanent magnetic recording of a digitally encoded pseudo-audio signal on a standard audio Compact Cassette tape is employed for non-volatile data storage purposes, employing suitable digital modulation and coding methods as described below. As for the audio Compact Cassette recording format, the corresponding IEC standard [1] defines the underlying technical (mechanical, magnetic etc.) parameters and properties. As a major advantage of such a data cassette format, the physical media and the tape drives were standardized and commonly available as mass products, providing a low-cost data storage solution. Consequently, such data cassettes were widely established in the context of digital storage purposes for early home- or personal-computer systems throughout the late 1970s and the early 1980s (see for example [4]). Similarly, data cassette-based approaches have been used for microprocessor-controlled audio systems as well.

From the perspective of data input/output for such systems, data cassettes were sometimes the only means of external data storage and provided sort of a data interface to the external world. In order to enable restoration, recovery, and conversion of historical compositions and arrangements which were created on such vintage audio systems, it is important to have a "bridge" to modern technology via software-defined data cassette decoder/encoder tools in combination with (still) readily available Compact Cassette recorders/players, common audio interfaces for computers, and software for digital audio recording (e.g. [20]). In particular, the conversion of the recorded data content into modern formats such as WAV-files [5] or SMF-files (MIDI) [6] is very useful in order to enable a re-usage of patterns or audio waveforms etc. on modern systems. Furthermore, such decoder/encoder tools provide a means for improved archiving and for a re-usage of original historical equipment.

A huge variety of historical data cassette storage formats were used by various audio equipment manufacturers. Quite often, proprietary digital modulation variants and coding formats as well as data structures and protocols were employed, some of them neither being officially documented nor being standardized. As a result, data cassettes quite often are not interchangeable between device models and, in the worst case, the usability is constrained to a particular audio device with a particular firmware version. From the perspective of digital preservation and data recovery, consequently, it is necessary to analyze such data cassettes individually, for example by inspection of cassettes with known content or with the help of plausibility considerations. If possible, original vintage equipment can be used for testing and analysis purposes of old or new recordings. Once a historical data cassette format has been de-ciphered successfully, the implementation of corresponding open source software-defined decoder and encoder tools, as presented in this contribution, yield an implied de-facto documentation of the historical formats ("documentation by open source implementation"). One has to note that these implementations are "best-effort" approaches, often based on analysis results, possibly with some assumptions if information is missing.

Due to the usage of the audio Compact Cassette format, such data cassettes exhibit characteristic signal limitations and degradations which are possible causes for errors, resulting from the physical properties of the underlying magnetic recording method and the chosen parameters (e.g. tape speed and track width etc.):

- linear distortion: sub-optimal frequency response (amplitude and phase characteristics), in particular at the lower and upper frequency limits
- non-linear distortion (Note: a square-wave-like signal is quite immune)
- noise and hum
- dropouts
- mechanical, chemical, and magnetic degradation, magnetic tape print-through effects, aging effects
- speed deviation, wow and flutter

As a consequence, more or less severe signal degradation and, in turn, possible data loss results, depending on the storage and usage conditions.

Recording of a pseudo-audio signal on a Compact Cassette audio tape thus imposes certain requirements on the employed modulation scheme and pseudo-audio signal waveforms:

- polarity independence
- pulse-width constraints and bandwidth-limitations (e.g. 100Hz...10kHz), signal must be DC-free
- tolerance against noise and hum within certain limits (for the given signal-to-noise ratio (SNR))
- tolerance against speed-deviation and wow/flutter within certain limits
- self-synchronizing modulation scheme (for clock recovery)
- simple algorithm (for low CPU demands and low-cost implementation)

Considering the given limitations and constraints, certain digital modulation schemes and line codes [3] and variants thereof were commonly chosen for data cassette formats in the past, for example (see also Fig.2 for illustration):

- FSK = frequency shift keying (independent of polarity, simple analog FM demodulator can be used, low data rates typically)
- Manchester code = phase encoding (polarity dependent, however, polarity detection/correction via suitable sync pattern possible)
- Bi-Phase-Mark code (equivalent to 1/2-clock shifted inverted differential Manchester code, independent of polarity)
- general bit group coding techniques

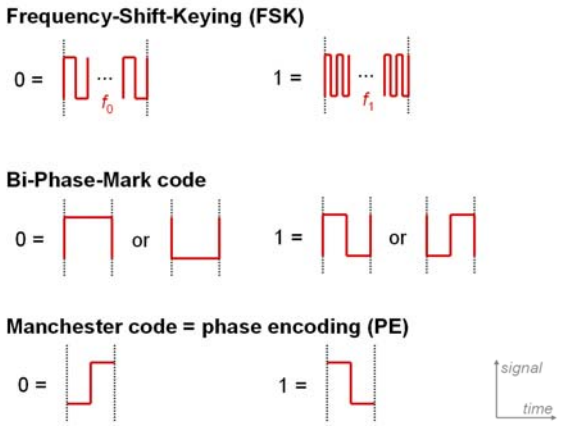


Fig. 2: Examples of typical digital modulation schemes for data storage using pseudo-audio signals on data cassettes. (Note: Manchester code variant with inverted bits also possible.)

As a realistic example, Fig.3 shows the audio signal from a data cassette recording of a "LinnDrum" [7,8] drum machine, using a Manchester code variant with a line bitrate of approximately 1.3kbit/s. (Here and in the following, "Audacity" [20] has been used for recording and visualization.)

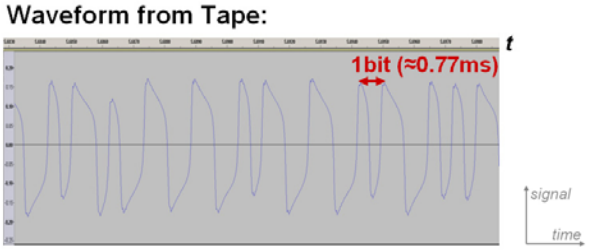


Fig. 3: Typical audio signal from a data cassette, using Manchester code with a line bitrate of 1.3kbit/s.

In Fig.4, a larger time scale of the same example is shown, exhibiting two dropouts (as indicated by arrows).

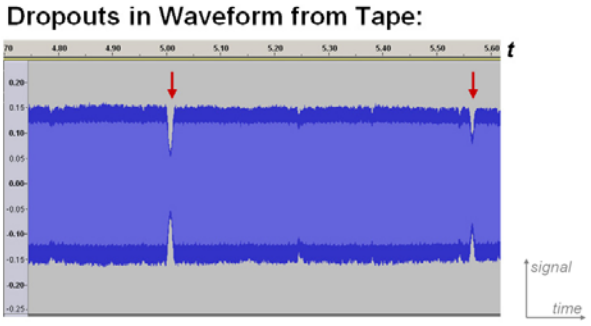


Fig. 4: Typical audio signal from a data cassette, exhibiting dropouts.

Fig.5 shows the corresponding long-time-analysis audio frequency spectrum. The minimum required audio bandwidth in this example would be approximately 5.2kHz in order to include at least the sidebands of the 1st and 3rd harmonic of the modulated square-wave-like signal.

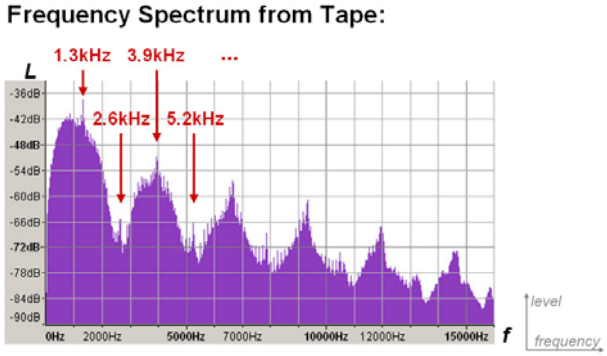


Fig. 5: Typical audio frequency spectrum from a data cassette, using Manchester code with a line bitrate of 1.3kbit/s.

In the following chapters, a software-defined decoder and encoder concept is presented, enabling data recovery and analysis as well as data conversion of typical historical data cassette formats of certain vintage audio equipment. Subsequently, examples of real audio devices and corresponding open source tools are discussed.

3. Software-Defined Decoder

In this chapter, the concept of a software-defined decoder scheme is presented, focussing on data cassette formats which employ the Manchester code or the Bi-Phase-Mark code. In principle, such a decoder is based on a hierarchical demodulation/decoding scheme which is visualized in a simplified form in Fig.6.

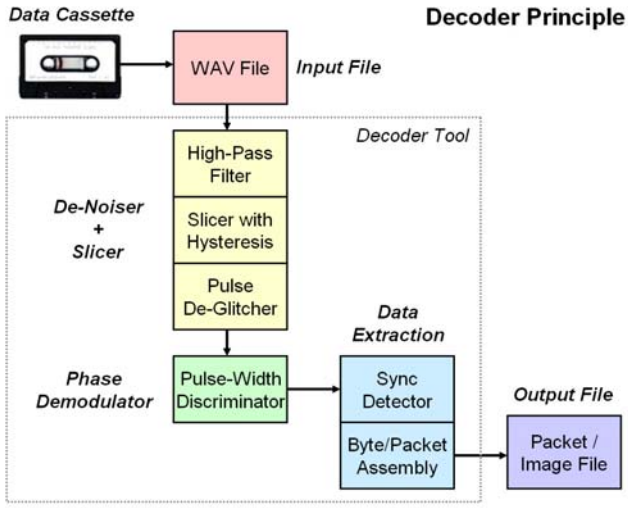


Fig. 6: Simplified principle of a software-defined decoder for data cassettes using Manchester code.

As an input signal, the decoder software tool requires a digital audio recording of the data cassette in a suitable file-format such as a WAV-file [5], which is obtained by usage of a Compact Cassette tape recorder/player attached to a digital audio interface and a recording software (e.g. [20]) with suitable parameters (e.g. 44.1kHz and 16bit PCM).

In the first signal processing step of the decoder, low-frequency noise and hum reduction is achieved by means of suitable digital high-pass filtering as visualized in Fig.7.

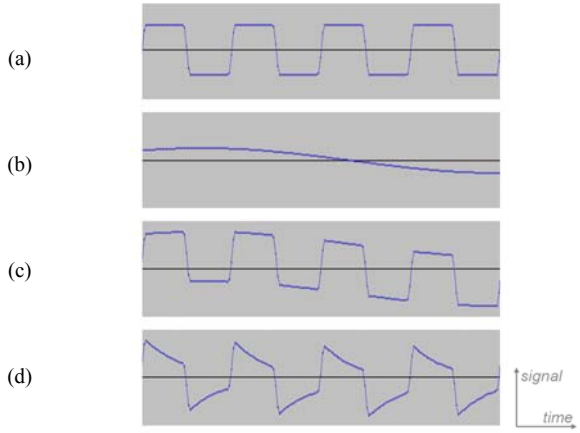


Fig. 7: Illustration of audio signals as a function of time. Using a high-pass filter to remove low-frequency (LF) noise and hum, and to emphasize slopes. (a) Original signal, (b) LF-noise and hum, (c) Resulting signal, (d) Filtered signal.

In the next step, general noise suppression is achieved with the help of a bit slicer with a hysteresis behaviour as shown in Fig.8.

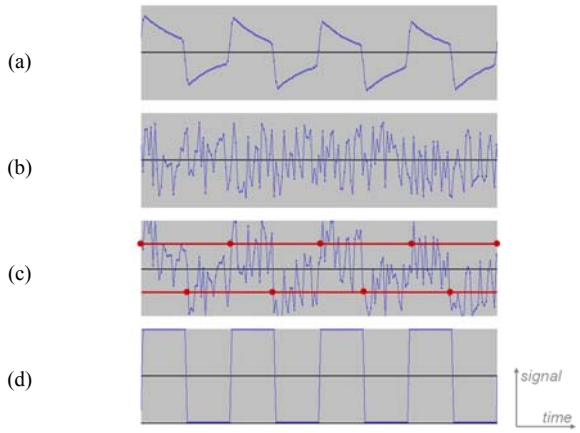


Fig. 8: Illustration of audio signals as a function of time, using a slicer with hysteresis against noise. (a) Original signal, (c) Noise, (c) Resulting signal for slicer with hysteresis, (d) Output signal.

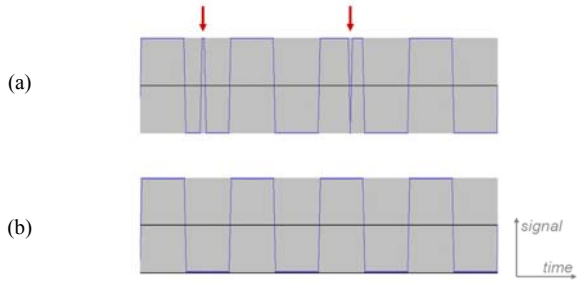


Fig. 9: Illustration of signals as a function of time, using a de-glitcher. (a) Signal with glitches, (b) Output signal where pulses are removed which are too short.

The resulting signal is supplied into a pulse-de-glitcher processing module in order to removes "spikes" which are short pulses that have a pulse width below a given threshold. Fig.9. visualizes the de-glitcher behaviour.

In the following processing step, a pulse-width discriminator is used in order to derive 0 and 1 bits, depending on the employed modulation scheme. In the case of a Manchester code, the pulse-width discriminator is based on the following algorithm as shown in Fig.10:

- 2 "short" pulse-widths → bit remains
- 1 "long" pulse-width → toggle bit

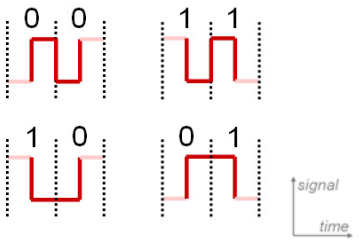


Fig. 10: Manchester code: Possible signal forms and resulting bits using a pulse-width discriminator for demodulation.

Here, the duration threshold for "short" and "long" pulses depends on the actual line bitrate. Pulse durations are measured with the help of sample clock counters which are synchronized by signal slopes, providing a self-synchronizing demodulation scheme. Immunity against a certain degree of wow/flutter results from a duration margin between "short" and "long". (Note that within the described algorithm, "long" pulses are also employed for a correct bit clock alignment.)

In the case of the Bi-Phase-Mark code, the pulse-width demodulator has to be modified as:

- 2 "short" pulse-widths → bit 1
- 1 "long" pulse-width → bit 0

(Alternatively, the algorithm for the Manchester code could be employed as well in certain cases, however, requiring a look-up table for byte translation.)

Next, the bit stream is supplied into the data extraction module. Depending on the usage of start/stop/parity line bits per data byte, the bits are assembled via a shift-register into symbols and, finally, payload bytes are obtained which are typically organized within data packets/blocks. For synchronization purposes (in particular for the definition of data packet/block starts), suitable sync bit-patterns are employed on data cassettes. Within the software-defined decoder, the sync pattern detector is realized using a shift-register (e.g. 64 bit). Unique sync patterns are also mandatory for the detection of signal polarity in the case of the Manchester code. Here, the detection of the sync pattern or the inverted sync pattern provides a means to determine polarity inversion and, consequently, for polarity correction. As a result, the scheme becomes polarity-independent and self-synchronizing. Depending on the data cassette coding

format, recovery (with re-synchronization) after a dropout is possible if the sync pattern occurs frequently enough, for example with every data packet. In this respect, a coding format with many small packets (where at least a subset can be restored) is advantageous as compared to a format with one large memory dump packet. For error detection and to enable consistency checks, many data cassette formats use packet checksums, which can be employed within the software-defined decoder in order to identify errors or inconsistencies.

Finally, the resulting data packets/blocks or a memory image are saved into suitable output files. These output files can be further processed with the help of a corresponding data packet parser tool which enables the analysis and listing of the contained "essence" data. Typically, data packets may contain a packet header which, for example, specifies a block address, block size, and content type information. Depending on the format, the data packet parser furthermore may provide conversion functionalities, for example the export of sequencer or rhythm patterns to SMF-files (MIDI) [6] and of audio samples ("waveforms") to WAV-files [5].

4. Software-Defined Encoder

Complementary to the software-defined decoder described in the previous chapter, the current chapter presents the concept of a software-defined encoder for data cassettes, again focussing on data cassette formats which employ the Manchester code or the Bi-Phase-Mark code. The simplified principle of such an encoder is visualized in Fig.11.

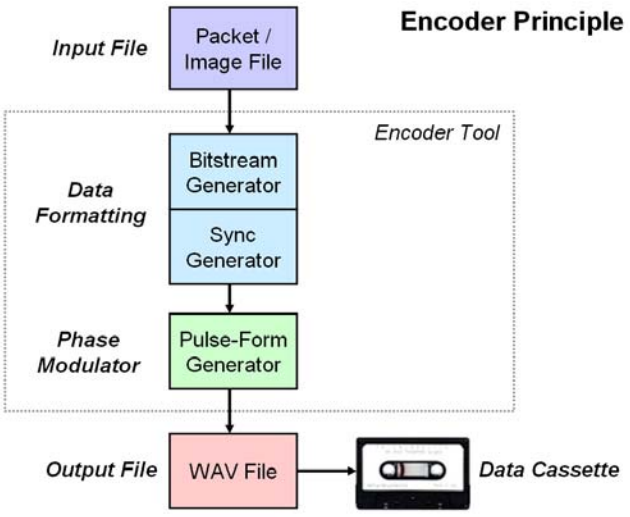


Fig. 11: Simplified principle of a software-defined encoder for data cassettes using Manchester code.

At its data input, the encoder accepts data packets/blocks or a memory image, supplied in form of an input file which has the same format as the output file which is generated by the decoder.

Within the first processing step of the encoder, the input data is formatted within a bitstream generator and all required sync patterns and start/stop/parity bits are added, according to the specific data cassette format. In the next step, the corresponding pseudo-audio signal is generated by means of a modulator. Here, the required digital modulation scheme is employed, for example a Manchester code or a Bi-Phase-Mark code, preferably using a signal look-up table. At the encoder output, the pseudo-audio signal is written to a WAV-file.

The resulting pseudo-audio WAV-file, in turn, can be recorded to a real data cassette. Alternatively, such a pseudo-audio WAV-file (from the encoder output or a digitized recording of an original data cassette) can be supplied directly to the vintage audio system by using a modern computer with audio interface and WAV-player/recorder software (e.g. [20]), thus bypassing the usage of a real cassette. Vice versa, the pseudo-audio output signal of the vintage audio equipment can be recorded directly by a modern computer in a similar way, without the usage of a real cassette. This offers the possibility to establish a de-facto data communication interface between the vintage audio equipment and a modern computer, based on a standard audio interface, audio player/recorder software, and the described software-defined decoder/encoder tools. Furthermore, it is possible to re-encode old data cassette recordings for signal re-conditioning with corrected encoding and data formatting, employing the decoder and the encoder tools in series.

In order to avoid aliasing artefacts of the generated pseudo-audio signal, the employed audio sample rate f_S is preferably chosen as an integer multiple of twice the line bitrate. For example, f_S = 25kHz can be chosen for a line bitrate of 2500bit/s, as illustrated in Fig.12 for the square-wave-like pseudo-audio signal which results from a sequence of bits 1 using a Manchester code.

Fig. 12: Sampled square-wave-like signal as a function of time. Shown example: $f_S = 10\,f$.

One has to note that the construction of the encoder tool is usually much simpler than the construction of the corresponding decoder tool. In particular, the decoder has to deal with noise and possible errors.

5. Examples

In this chapter, three examples of Compact Cassette-based [1,2] data cassette formats of historical drum machines and associated software-defined decoder/encoder tools are presented. These tools were released as open source software [16-18] and enable data extraction and analysis, and furthermore, provide export functionality of sequencer and rhythm patterns to SMF-files (MIDI) [6] and audio samples to WAV-files [5]. Playback and analysis of the resulting WAV- or SMF-files, in turn, is possible for example via applications such as "Audacity" [20] or "Sound Font

Midiplayer" [21], respectively. In the following sections, data cassette and data structure formats of three historical drum machines are described, based on information published in product documentations [7-15] and based on analysis results that were obtained with the help of the described software-defined decoder/encoder tools [16-18], employing real data cassette recordings or signals from original equipment.

5.1. Linn Electronics LinnDrum

The "LinnDrum" [7,8] drum machine (see Fig.13) was manufactured by Linn Electronics, Inc..

Fig. 13: Picture of a LinnDrum drum machine.

The LinnDrum drum sounds are produced from digital drum sound samples which are in 8bit µ-law format and are stored in dedicated internal EPROMs. This drum machine provides memory for 2x7x7 "patterns" and 7x7 "songs", internally saved in non-volatile RAM ("NVRAM"). The timing resolution is 48PPQ fixed.

The data cassette format of the LinnDrum employs a Manchester code variant with 10 line bits per data byte (8 data bits + 2 start/stop bits) with a line bitrate of approximately 1305.5bit/s and a packet checksum for error detection. Two types of LinnDrum data cassette recordings are possible: pattern + song memory dump for type "EVERYTHING", or song + used patterns for type "SONG". Fig.14 shows the spectrogram of a typical data cassette recording of the LinnDrum for these two recording types.

Fig. 14: LinnDrum example: Typical spectrogram of data cassette recordings for cassette types "EVERYTHING" (approx. 9 sec duration in the example) and "SONG" (approx. 9 sec duration in the example), respectively.

Furthermore, Fig.15 shows the corresponding long-time-analysis audio spectrum with typical frequency blocks which are centered at odd multiples of 1.3kHz (stemming from the underlying line bitrate).

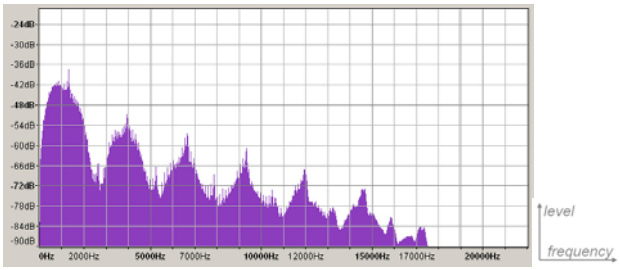


Fig. 15: LinnDrum example: Typical spectrum (playback signal from tape). Line bitrate: 1.3kbit/s.

The open source tool set "linndrumcass" [16] contains the software-defined decoder, encoder, and data parser applications for the LinnDrum data cassette format. Export functionality of patterns to SMF-files (MIDI) is provided.

The LinnDrum patterns employ a variable memory allocation scheme. Here, 3 byte event codes are used, containing information about an additional wait time-interval (to next event) and a drum trigger code. 12 different drum sounds can be triggered (partially exclusive), some of them having level/pitch/decay variants. Further details of the LinnDrum data cassette format and the employed data structures can be found in [16].

5.2. E-mu Systems Drumulator

The "Drumulator" [9,10] drum machine (see Fig.16) was manufactured by E-mu Systems, Inc..

Fig. 16: Picture of a Drumulator drum machine.

The Drumulator drum sounds are produced from digital drum sound samples which are in 8bit µ-law format and are stored in dedicated internal (EP)ROMs. This drum machine provides memory for 6x6 "segments" and 8x8 "songs" (8 with older firmware), internally saved in non-volatile RAM ("NVRAM"). The timing resolution is 24PPQ fixed.

The data cassette format of the Drumulator employs a Bi-Phase-Mark code variant with 9 line bits per data byte (8 data bits + 1 odd-parity bit) with a line bitrate of 2500bit/s. A Drumulator data cassette recording contains a memory dump of the NVRAM. Fig.17 shows the spectrogram of a typical data cassette recording of the Drumulator.

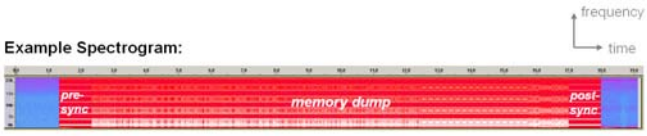


Fig. 17: Drumulator example: Typical spectrogram of a data cassette recording (approx. 17 sec duration).

Furthermore, Fig.18 shows the corresponding long-time-analysis audio spectrum with typical frequency blocks which are centered at odd multiples of 2.5kHz (stemming from the underlying line bitrate).

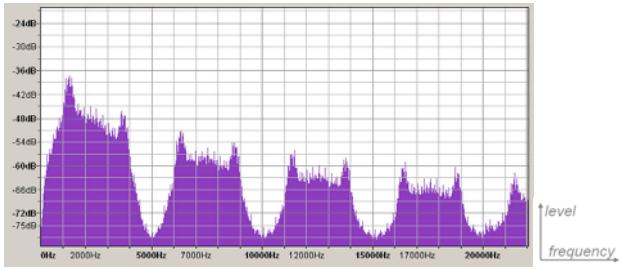


Fig. 18: Drumulator example: Typical spectrum (drum machine output). Line bitrate: 2.5kbit/s.

The open source tool set "drumucass" [17] contains the software-defined decoder, encoder, and data parser applications for the Drumulator data cassette format. Export functionality of patterns to SMF-files (MIDI) is provided. Due to the odd-parity bit per data byte, the Bi-Phase-Mark code (de)modulator within the current version of the decoder/encoder of [17] is implemented as a Manchester code (de)modulator with 1/2-clock shift and 1 start bit per data byte, employing a byte lookup table for data byte translation.

The Drumulator segments employ a variable memory allocation scheme. Here, 1 byte and 3 byte event codes are used, containing information about time-slot position and a drum trigger code. 12 different drum sounds can be triggered (partially exclusive) with an individual "accent" flag. Further details of the Drumulator data cassette format and the employed data structures can be found in [17].

5.3. WERSIMATIC CX5 DrumComposer

The "WERSIMATIC CX5 Drum Composer" [11] drum machine (see Fig.19) was manufactured by WERSI-electronic GmbH & Co.KG.

Fig. 19: Picture of a WERSIMATIC CX5 Drum Composer drum machine.

The CX5 drum sounds are produced from digital drum sound samples which are in 8bit linear PCM format and are stored in dedicated internal (EP)ROMs. This drum machine provides memory for 2x16 user-programmable "rhythms" and 16 "sequences", internally saved in non-volatile RAM ("NVRAM"). In addition, 2x2x16 fixed "rhythms" are stored within an internal EPROM. The timing resolution is variable ("Times" = 1...16 steps per "Beat"). With this drum machine, external data storage is possible on proprietary cartridges (NVRAM-based, for 2x16 "rhythms" and 16 "sequences") and on data cassettes.

The data cassette formats of the CX5 and some WERSI DX-series organs [12-14] are very similar and are based on a Manchester code variant with 10 line bits per data byte (8 data bits + 2 start/stop bits) with a line bitrate of 2500bit/s and a packet checksum for error detection. Within such a data cassette recording, multiple small packets with individual address and type code are used, corresponding to data packets of the WERSI RS232 protocol [15]. The CX5 in particular uses a data subset of the rhythm-data packet format of the "WERSIMATIC CX4" subsystem of the WERSI DX400/DX500 [14] organs.

Fig.20 shows the spectrogram of a typical data cassette recording of the CX5. Periodic features corresponding to 64 rhythm-data packets (to be assembled into 2x16 rhythms) can be seen.

Fig. 20: CX5 example: Typical spectrogram of a data cassette recording (approx. 76 sec duration).

Furthermore, Fig.21 shows the corresponding long-time-analysis audio spectrum with typical frequency blocks which are centered at odd multiples of 2.5kHz (stemming from the underlying line bitrate).

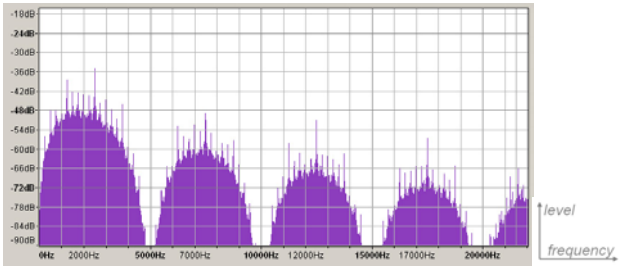


Fig. 21: CX5 example: Typical spectrum (drum machine output). Line bitrate: 2.5kbit/s.

The CX5 rhythms employ a static memory allocation scheme. Here, a fixed space for 2 measures with max. 16 steps per measure is provided per rhythm. 2 drum trigger codes with 3 bytes each are allocated per step, corresponding to 2 measure variants ("normal" and "intro/break"). 24 drum sounds can be triggered, organized as 8 slots with 3 mutually exclusive sounds and a "soft/loud" flag each.

The open source tool set "wersidatencassette" [18] contains the software-defined decoder, encoder, and data parser applications for the data cassette format of the CX5 and some WERSI DX-series organs. Export functionality of patterns to SMF-files (MIDI) is provided. In addition, WERSI RS232 packet import/export is possible, as well as the export of "Waves" of some WERSI DX-series organs to

WAV-files. Further details of the data cassette format and the employed data structures can be found in [18].

One has to note that some older WERSI systems, for example the "WERSIMATIC CX2", use an FSK-based modulation scheme (see for example the tool set "wersicx2datencassette" [19]), whereas the CX5 and some WERSI DX-series organs, which were discussed in this section, employ Manchester code (see above).

6. Summary

In this contribution, the digital preservation of the content of Compact Cassette-based data cassettes for historical audio systems such as drum machines, synthesizers, and organs was discussed. The concept of software-defined decoder and encoder tools was presented. Due to the suggested open source implementation, a de-facto documentation of the proprietary modulation and coding formats is provided. Based on the developed open source tool sets, implementation examples and analysis results for data cassette formats of three vintage drum machines were discussed.

As a possible future perspective, the described approach could be extended to further historical data cassettes formats of vintage audio systems and musical instruments of interest. In addition, the features and the usability of the software-defined decoder/encoder tools could be expanded and, furthermore, the software could be ported to mobile device platforms, for example for tablet computers or smart phones.

7. References

[1] IEC 60094 standard, parts 1-7: "Magnetic sound recording and reproducing systems"

[2] see e.g. Wikipedia Free Encyclopedia: "Compact Cassette", *https://en.wikipedia.org/wiki/Compact_Cassette*, accessed 03-Nov-2018

[3] see e.g. R.W.Middlestead: "Digital Communications with Emphasis on Data Modems", Wiley

[4] see e.g. webpage by H. Johnson: "Computer Data Cassette Formats", *http://www.retrotechnology.com/restore/cass_data.html*, accessed 07-Nov-2018

[5] Microsoft: "Resource Interchange File Format (RIFF)" file format, "WAVE" file format

[6] MIDI Manufacturers Association: "Complete MIDI 1.0 Detailed Specification", "SMF" file format

[7] Linn Electronics, Inc.: LinnDrum Operating Instructions

[8] Linn Electronics, Inc.: LinnDrum Technical Information

[9] E-mu Systems, Inc.: Drumulator Owners Manual, Version 1.0, 1981

[10] E-mu Systems, Inc.: Drumulator Service Manual, Version 1.0, 1983

[11] WERSI: "Drum Composer CX5 Bedienungsanleitung", BA 392

[12] WERSI: "ALPHA DX350 Bedienungsanleitung", BA 328

[13] WERSI: "ALPHA DX300 / ALPHA DX350 / CONDOR DX100 Technische Unterlagen", BA 329

[14] WERSI: "ALPHA DX400/500 Bedienungsanleitung", BA 333

[15] WERSI RS232 protocol: WERSI "ALPHA DX350 Bedienungsanleitung", BA 328, pp.77-82

[16] K.M.Indlekofer: "linndrumcass", *https://sourceforge.net/projects/linndrumcass/*, accessed 03-Nov-2018

[17] K.M.Indlekofer: "drumucass", *https://sourceforge.net/projects/drumucass/*, accessed 03-Nov-2018

[18] K.M.Indlekofer: "wersidatencassette", *https://sourceforge.net/projects/wersidatencassette/*, accessed 03-Nov-2018

[19] K.M.Indlekofer: "wersicx2datencassette", *https://sourceforge.net/projects/wersicx2datencassette/*, accessed 11-Nov-2018

[20] Audacity Team: "Audacity", *https://www.audacityteam.org/*, accessed 03-Nov-2018

[21] Z.Bacskó (Falcosoft): "Soundfont Midi Player", *http://falcosoft.hu/softwares.html#midiplayer*, accessed 03-Nov-2018

[22] "RAM" = random access memory, "SRAM" = static RAM, "NVRAM" = non-volatile RAM (e.g. based on SRAM with backup battery), "ROM" = read-only memory, "EPROM" = erasable programmable ROM, "EEPROM" = electrically erasable ROM

Disclaimer

We are not affiliated in any way with companies/persons mentioned in this document. All brand names and trademarks are property of their respective owners. The contents of this document are intended for educational/academic/non-commercial purposes only. The author does not take responsibility for the contents of any links referred to. Product parameters, formats etc. are described for illustration or academic documentation purposes and are not guaranteed to be correct.

When I looked at a shellac record, I saw the sound! Reality or illusion?

Stefano S. Cavaglieri

Swiss National Sound Archives, Lugano, Switzerland, Email: SergioStefano.Cavaglieri@nb.admin.ch

Abstract

Nowadays, in the digital age, with a high degree of miniaturization and advanced techniques at our disposal, we think that we can easily transform just everything into an accurate series of bits and bytes. For example, looking at the surface of a shellac disc through a microscope, we can clearly see the groove excursions, which correspond to the waveform of the audio signal they represent. So, there must be a simple way to digitize them... or is it less obvious than one may think? Let's expose the measurable reality, in an objective way, analysing and comparing (electro)mechanical vs. optical reproduction systems.

1. Introduction

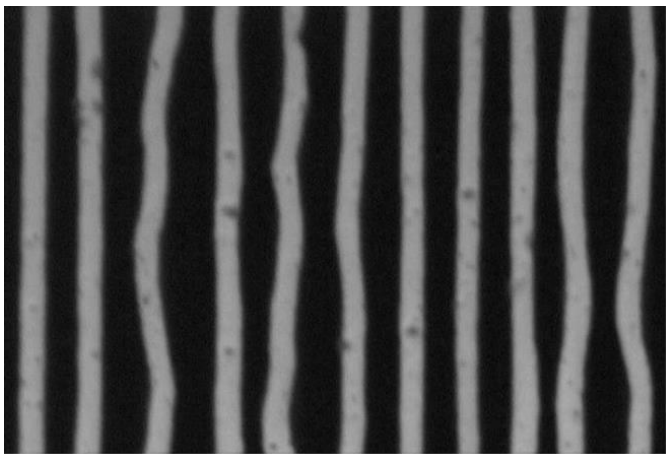


Fig. 1: A section of a shellac record, as seen from the top.

1.1. When I looked at a shellac record, I saw the sound!

By observing a disc using a magnifier or a microscope, we can literally "see" the sound: it is represented by the lateral excursions of the groove.

1.2. Reality or illusion?

If it's visible, it means that the audio information is contained in the image of the surface.

With imaging, we can capture a vast amount of time-related information in a snapshot.

Some say, that imaging is an objective means for transferring the physical representation of a sound recording (i.e. the groove) to another media, while retaining full authenticity.

2. VisualAudio

So, why not just trying to put all this into practice? The following steps may do the job:

- The disc is photographed (high resolution copy)
- The film is scanned (image digitization)
- The raw image is processed (groove reconstruction)
- The final image is translated to sound (sound extraction)

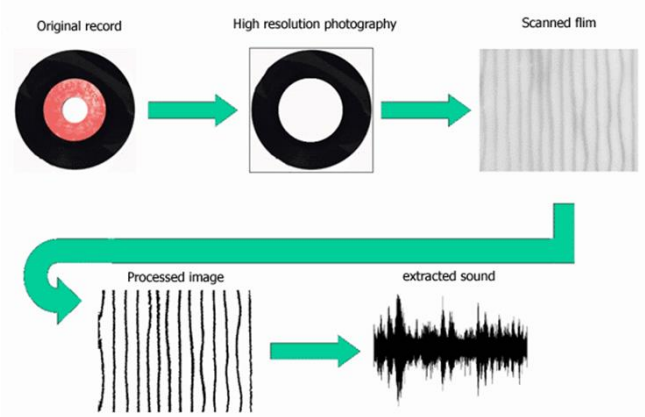


Fig. 2: Schematic overview of the VisualAudio concept.

3. Re-recording

3.1. Principles

Thou shalt ensure the authenticity of the recording!!!

- In order to do so: We should extract the best possible signal from the original recording.
- Which translates into: We are required to carefully adjust all mechanical as well as electronic components to match the specs of the recording.
- Because: On the copy, corrections for misalignments are technically limited.

ISBN
978-3-9812830-9-9

The authors hold the copyrights of all of content in this manuscript. They have granted the VDT the right to publish it in print, online and as PDF download on *https://tonmeistertagung.com*

This lead to develop some best practices (i.e. the art and science of re-recording), which meant dealing with:

- Groove geometry (stylus shape and size, height adjustment, lateral balance, cartridge azimuth, tracking adjustment, bias adjustment)
- Electrical adjustments (velocity vs. amplitude sensitive retrieval, level, de-emphasis)
- Mechanical adjustments (speed, centring, de-warping)

Because of the complexity of the above procedures, as well as the presence of a number of grey zones, we aimed for a broader view, asking ourselves a couple of questions:

- How accurate is mechanical playback?
- How does it compare to optical playback?
- Several scientists, engineers, audiophiles, and students gave it a try, but most projects failed. Why?

In 2016, the IASA established an Emerging Technologies subcommittee (of its Technical Committee) to eventually shed light on it.

3.2. Workflows

3.2.1. Mechanical

A typical professional mechanical re-recording workflow is defined in the following steps:

- Clean the record
- Replace inner- and outer sleeves
- Visually inspect the grooves
- Choose the right stylus
- Adjust the tonearm geometry
- Set the equalization, if any
- Centre the record
- Determine and set the right speed
- Adjust the playback level
- Listen carefully (pre-A/D)
- *Make a high-resolution copy*
- Listen carefully (**post A/D**)

3.2.2. Optical

A typical VisualAudio re-recording workflow is defined in the following steps:

- Clean the record
- Replace inner- and outer sleeves
- Shoot an analogue picture
- *Scan the negative film*
- *Process the digital image*
- *Make a high-resolution copy*
- *Determine and set the right speed*
- *Set the equalization, if any*
- *Adjust the playback level*
- Listen carefully (*post A/D*)

Italics = it happens in the digital domain
Regular = it happens in the analogue domain

3.3. Average timings

3.3.1. Mechanical

The time required for a typical professional mechanical re-recording of 1 side of a shellac record is around 16 minutes (stylus choice, mechanical and electrical adjustments)

3.3.2. Optical

The time required for a typical VisualAudio re-recording of 1 side of a shellac record is outlined in the following steps:

- Photography: 10' (disc positioning, picture shooting, film development)
- Film scanning: 30' (film positioning, ring-by-ring scanning and merge)
- Image processing and image-to-sound translation: 2h 30' (groove alignment and reconstruction, image-to-sound translation, de-emphasis)

4. Converters

When it comes to produce a digital audio copy, we inevitably must go through some sort of a converter, which may be one of the following.

4.1. Analog-to-digital converters (ADC)

An ADC converts an analogue electric signal (usually voltage) into a digital representation (numbers).

All ADCs work under the same principle: they need to convert a signal to a certain number of bits, typically 16 or 24 in audio applications. Just as a remainder:

$16\ \text{bit} = 2^{16} = 65'536\ \text{values}$
$24\ \text{bit} = 2^{24} = 16'777'216\ \text{values}$

The choice of the ADC for a given application is defined by the requirements you have: if you need speed, use a fast ADC; if you need precision, use an accurate ADC; if you are constraint in space, use a compact ADC.

The main characteristics of any ADC can be defined as:

- Coding scheme
- Offset, gain error, non-linearity
- Resolution
- Conversion speed

4.2. Bitmap-to-wave converters (BWC)

2D imaging technology is straightforward: a raster (or bitmap) image representing a signal (i.e. a waveform) is subdivided into pixels, each with its planar coordinates that tell you about the instant amplitude related to time.

The main characteristics of any ADC can be defined as:

- Coding scheme
- Resolution
- Conversion speed

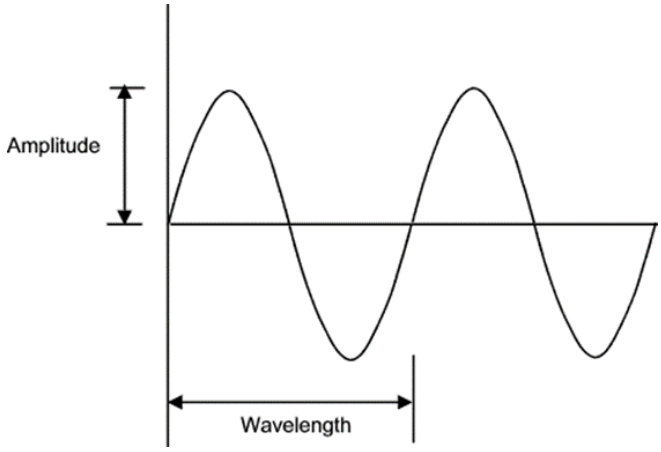


Fig. 3: A very simple audio signal.

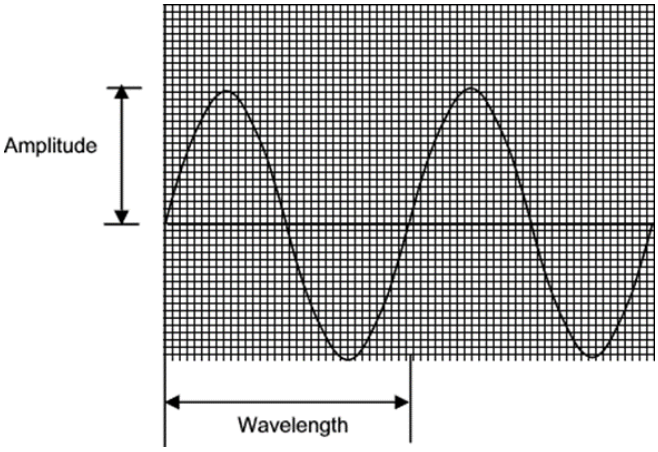


Fig. 4: The same audio signal represented as a pixel array.

4.3. The coding scheme

The BWC coding scheme can be represented by a staircase, similar as for an ADC, although it contains both amplitude and time information.

Even in the ideal case, there is always a minimum error of ±1/2 LSB, called the quantization error, as well as of ±1/2 sample.

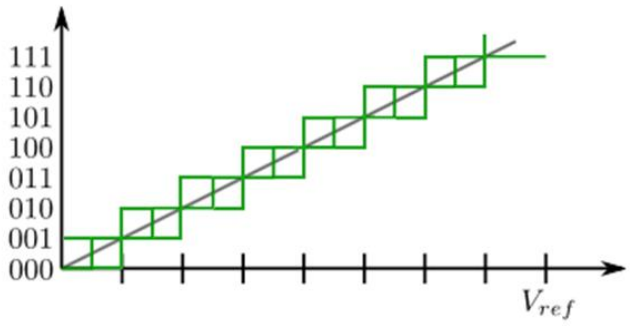


Fig. 5: An ideal coding scheme.

4.4. ADC accuracy

An essential component that you will find in all analogue-to-digital converters is the comparator. The comparator requirements become very demanding for high-resolution ADCs. The specifications of the comparator are of paramount importance, because:

- The input-referred offset voltage must be below 1 LSB
- The input-referred noise must be below 1 LSB
- It must resolve a comparison for differential inputs below 1 LSB, where:
    - 16bit ADC with a reference voltage of 5V: 1 LSB = 5V / 216 = 76μV
    - 24bit ADC with a reference voltage of 5V: 1 LSB = 5V / 224 = 298nV

But this is just the tip of the iceberg. Each and every aspect (i.e. architecture and characteristics), but also the manufacturing, as well as the environment in which the ADC is used, leads to a different interpretation of the signal, therefore produce different numbers.

When we read these numbers back and convert them to a signal, the digital-to-analogue converter (DAC) has no clue on what happened in the opposite stage, so the sound will simply change.

4.5. BWC accuracy

The key element in all bitmap-to-wave converters is the pixel size relative to the maximum excursion of the groove. The pixel size requirements become very demanding for high-resolution imaging systems. The dimensions of a pixel are of paramount importance, because:

- They tell you how much of a groove is captured within a pixel
- The pixel height must be below 1 LSB
- The pixel width must be below 1 sample

Imaging a coarse-groove record at 16bit/44.1kHz, assuming a maximum groove excursion of 75μm, at 100mm from the centre of the disc, and a rotation speed of 78rpm, produces the following figures:

- 1 LSB = 75μm / 216 = 1.14nm
- 1 sample = 628mm / 60*78rpm / 44’100 = 18.5μm

While imaging at 24bit/96kHz to the same specs, produces the following figures:

- 1 LSB = 75μm / 224 = 4.47pm
- 1 sample = 628mm / 60*78rpm / 96’000 = 8.5μm

5. Optical retrieval systems

In order to do really understand where we are aiming to, it is necessary to expose the specifications of 2 production-ready optical retrieval systems.

5.1. IRENE 2D

- Uses a line sensor with 4096 pixels of 7µm square size each.
- Uses optics of roughly x10, so it images 0.7µm on the disc surface.
- Uses high contrast illumination in order to find edges, for example the groove bottom. To find the location of an edge it does edge detection across a set of pixels which appear with varying greyscale. Therefore, the edge resolution is better than one-pixel width. There is single number to quote here but clearly, in this mode, resolution can be 0.25µm or better.
- In the time direction, the square pixels are effectively elongated since it images a disc in rotating motion. Depending upon the exposure time and rotational velocity, the pixels correspond to (in the case of 78rpm media) 104kHz, 208kHz, or 312kHz time sampling.

5.2. VisualAudio

- Shoots a 1:1 analogue picture on a mid-size (35x45cm) Agfa Alliance CE film, which has a silver-halide crystal granularity of 0.2µm, to form dye-clouds of approx. 10µm each.
- Scans the film using a line sensor with 2048 pixels of 10µm square size each.
- Uses optics of 4x, so it images 2.5µm on the analogue picture surface.
- Coarse edge detection is first applied to detect the presence of all the edges over the image, even in case of damages or non-homogeneous groove's wall illumination. Finer edge detection then locates accurately the edge candidates to subpixel precision. The groove reconstruction step selects the candidates that best match the existing traces. With this procedure, resolution can be said to be close to 0.25µm.
- In the time direction, the square pixels are elongated since it images a disc in rotating motion. Depending upon rotational velocity, the pixels correspond to (in the case of 78rpm media) 85kHz or 170kHz time sampling.

The constraints encountered on both of these optical retrieval systems can be summarized as follows:

- Sampling (i.e. the subdivision in the time domain) does not represent an issue. Capture speed can be set so to have at least 1 pixel/audio sample.
- Capture can only happen at the edges, or the bottom of the groove, or combined. Cannot trace the walls the same way as a stylus does.
- Quantization is the real issue. Lack of information leads to processing limitations, such as for equalization (de-emphasis) and restoration.

6. Quantization issues

The figures marked as "best case" are only reachable, assuming that the systems are capable to add some sort of a luminance rating to the planar coordinates, in the conversion algorithm.

Maximum full-scale groove excursion @ around 70Hz:

- At worst – VisualAudio: 75µm / 2.5µm = 30 = 5bit
  " " – IRENE 2D: 75µm / 0.7µm = 107 = 7bit
- At best – Both: 75µm / 0.25µm = 300 = 9bit

Average full-scale groove excursion @ around 1kHz:

- At worst – VisualAudio: 30µm / 2.5µm = 12 = 4bit
  " " – IRENE 2D: 30µm / 0.7µm = 43 = 6bit
- At best – Both: 30µm / 0.25µm = 120 = 7bit

Average full-scale groove excursion @ around 10kHz:

- At worst – VisualAudio: 15µm / 2.5µm = 6 = 3bit
  " " – IRENE 2D: 15µm / 0.7µm = 21 = 5bit
- At best – Both: 15µm / 0.25µm = 60 = 6bit

7. Physical limits

The figures below are valid for true 16-bit quantization. The situation gets worse above 1kHz, because de-emphasis cannot be applied prior to image contour detection.

Maximum full-scale groove excursion @ around 70Hz, 16bit:

- (Sub)Pixel size is 0.25µm
- Should be 1.14nm
- Ideal (sub)pixels should be 219 times smaller

These following figures do not make sense. As a comparison, in a water molecule, the distance from the centre of the oxygen atom to the centre of one of the hydrogen atoms is about 95.84pm.

Maximum full-scale groove excursion @ around 70Hz, 24bit:

- (Sub)Pixel size is 0.25µm
- Should be 4.47pm
- Ideal (sub)pixels should be 55'928 times smaller

8. Conclusions

This might sound as the calm after the storm, imaging has still some benefits anyway:

- The image per se is an accurate copy of the original signal. It is also a quick and reliable way to freeze the degradation of the surface of a disc.
- A certain depth of field may be granted by well-known photographic principles.
- The image is captured without interfering with the surface of the disc. There is no need to manipulate the disc, except for positioning.

- Discs in virtually any condition (even delaminated, broken, deformed, etc.) can be read and the sound restored.
- Each and every disc format (size, speed, cutting, etc.) can be read using the same equipment.
- Image processing is well established. It is relatively easy to make corrections to the physical incoherencies of the disc.

And there is room for improvement:

- Image capture - Film vs. linear scanner vs. confocal probe vs. nanotechnology; 2-dimensional vs. 3-dimensional approach (such as IRENE 3D)
- Image processing - Raster vs. vector graphics vs. holography; Chromatic detection in addition to luminance
- Signal processing - Modelling of the tonearm pivoting and inertia + stylus tracking, in order to determine some sort of a cross-emphasis curve; Define equivalent de-emphasis curves, considering the amplitude (of optical systems) vs. velocity approaches (of electromechanical systems)

9. References

[1] John Elison: Stylus-Tip Velocity relative to Groove Amplitude and RIAA Equalization

[2] Patrick Feaster: Calculating Velocities: another step in “playing” waveform images

[3] Kevin Gray: Inverse RIAA Revisited

[4] Carl Haber: IRENE 2D specs

[5] Keith Howard: Cut and Thrust: RIAA LP Equalization

[6] Jim Lesurf: Keep in Contact

[7] EnjoyTheMusic.com: Cartridge History

[8] OnMyPhD.com: Analog-Digital Converters

Webbasierte Hörversuche unter unkontrollierten Abhörbedingungen am Beispiel einer Audioqualitäts-Evaluierung von Lautheitstools

T. Rauscher[1], M. Meier[2]

[1] *IRT GmbH, München, Deutschland, Email: thomas.rauscher@irt.de*

[2] *IRT GmbH, München, Deutschland, Email: michael.meier@irt.de*

Abstract

Anhand eines Hörversuchs – gemäß einer erweiterten Methode nach ITU-R BS.1116 – zur subjektiven bzw. perzeptiven Qualitätsbeurteilung verschiedener Lautheitsalgorithmen werden die Ergebnisse zweier verschiedener Teilnehmer-Gruppen (Panels) gegenübergestellt.

Der browser-basierte Hörversuch wurde einerseits als Laborexperiment mit definierten Versuchsbedingungen wie akustisch optimiertem Raum nebst hochwertigem Audio-Interface und Elektrostat-Kopfhörern und andererseits als Feldexperiment mit entsprechend variablen Abhörbedingungen und Online-Version des Versuchs durchgeführt. Die Probanden, bestehend aus hörerfahrenen Versuchspersonen, sollten hierbei verschiedene Stimuli, angepasst auf bestimmte vorgegebene Lautheitszielwerte, zu einem unbearbeiteten Referenzsignal unter Zuhilfenahme von ausgewählten Qualitäts-Merkmalen (Attributen) bewerten. Die Erweiterung des Laborexperiments erfolgte unter dem Gesichtspunkt einer Steigerung der Reichweite bzw. der Probandenanzahl. Im Vordergrund stand daher die direkte Vergleichbarkeit der Ergebnisse aus diesen beiden Versuchsgruppen-Szenarien, um in Ergänzung Aussagen über Reliabilität und ökologischer Validität gewinnen zu können.

1. Einführung

Hörversuche haben für die Ermittlung von subjektiver Wahrnehmung aber auch von qualitativen Eindrücken einen hohen Stellenwert. In den meisten Fällen werden hierfür unter kontrollierten Bedingungen zu bewertende Stimuli einer Probandengruppe (Panel) dargeboten. Die Schwierigkeit der Generalisierbarkeit von Laborversuchen kann durch eine Versuchsdurchführung als Feldexperiment prinzipiell kompensiert werden. In dieser Untersuchung werden mit Hilfe einer Browser-basierten Hörversuchs-Software beide Szenarien miteinander verbunden unter der Prämisse einer Erweiterung des Laborexperiments durch ein Probanden-Panel eines Feldversuches. Anhand eines im IRT durchgeführten Hörversuchs zur Bestimmung der Audioqualität von Lautheitsprozessoren (Systeme) soll diese Möglichkeit überprüft und im Folgenden unterstützt durch grafische und statistische Ergebnisse validiert werden. Ziel ist es, die grundlegende Fragestellung zu beantworten, ob die Bewertungen und Ergebnisse des Laborexperiments unabhängig von äußeren Einflüssen im Feldexperiment reproduzierbar sind. Das heißt: sind die Bewertungsmittelwerte der jeweiligen Stimuli in der direkten Gegenüberstellung der beiden Szenarien gleich?

2. Hintergrund

Hintergrund dieser Evaluierung ist die Lautheitsbearbeitung für die Online-Distribution innerhalb des öffentlich-rechtlichen Rundfunks, vergleichbar der bereits 2012 umgesetzten Lautheitsaussteuerung gemäß EBU R 128 [1] im Fernsehbereich. Die Bereitstellung von Medieninhalten über die Mediatheken und ähnlichen Plattformen wird auch für die Rundfunkanstalten ein immer wichtiger werdender Verbreitungsweg. Es ist hierfür entsprechend den bestehenden Distributionswegen ebenfalls eine adäquate Audioqualität sicherzustellen. Die Nutzung von Online-Inhalten erfolgt in großem Maße durch mobile Endgeräte, welche eine Begrenzung der maximalen Wiedergabelautstärke aus regulatorischen Gründen zum Schutz vor Hörschäden aufweisen. Es ist daher erforderlich, sendeseitig eine entsprechende Anpassung der Lautheitswerte vorzunehmen, da derzeit eine Wiedergabe von auf -23 LUFS produziertem Content auf mobilen Endgeräten mit ausreichendem Pegel nicht möglich ist. Eine Anpassung dieser Rundfunkinhalte sollte nun möglichst unkompliziert, automatisiert und mit hoher Qualität erfolgen. Es wurde deshalb eine Evaluierung von Lautheitsprozessoren ins Leben gerufen, um deren qualitative Möglichkeiten herauszuarbeiten. In dieser Untersuchung wurden vordergründig fünf Algorithmen zur filebasierten Lautheitsanpassung miteinander verglichen. Unter der Vorgabe einer Einbindung dieser Algorithmen in die Open Source-Plattform *FFmpeg* – deren Verwendung für die Online-Distribution einen immer größer werdenden Stellenwert einnimmt – sollte ein Szenario eines prozessoptimierten Workflows nachgebildet werden. Durch eine im Vorfeld durchgeführte objektive messtechnische Untersuchung der Algorithmen konnte eine bestmögliche Parametrisierung mittels eines fixen Parametersets für jeden Algorithmus bereitgestellt werden. Diese Vorgehensweise

ISBN 978-3-9812830-9-9
Der Autor hält die Urheberrechte für den gesamten Inhalt seines Manuskripts. Er überlässt dem VDT das Recht auf Veröffentlichung in gedruckter Form, online oder als PDF-Download von https://tonmeistertagung.com

verfolgte hiermit auch weitestgehend einen praktischen Ansatz. Ein wichtiger Aspekt bei der Prozessierung der Audiobeispiele ist in der Wahl der Lautheitszielwerte begründet. Basierend auf bereits existierenden Vorschlägen für einen einheitlichen Zielwert für die Programmlautheit (IL_k) wurden für diese Evaluierung alle Audiobeispiele auf einen Zielwert von -16 LUFS angepasst. Um eine hohe Audioqualität zu gewährleisten, sollte mit einer Justierung der Programmlautheit ebenfalls eine entsprechende Anpassung der Loudness Range (LRA) einhergehen. Aus empirischer Erfahrung vorangehender IRT-interner Untersuchungen und Entwicklungen wurde ein LRA-Zielwert von 10 LU gewählt.

3. Hörversuchsmethode

Basierend auf einem ABX-Vergleich nach ITU-R BS.1116 [2] sollten die teilnehmenden Probanden die Stimuli jeweils zu einem unbearbeiteten Referenzsignal vergleichen und beurteilen. Zusätzlich wurde den Probanden die Möglichkeit geboten, die qualitativen Unterschiede zwischen Stimulus und Referenzsignal mittels spezieller Attribute näher zu bestimmen. Die Wahl der Attribute wurde an typische Artefakte angelehnt, welche im Grenzfall bei unsachgemäßer Verwendung eines Dynamikalgorithmus entstehen. Somit kann das Versuchsdesign als mehrfaktoriell mit den Attributen als sogenannte unabhängige Variablen und der Audioqualität als abhängige Variable angesehen werden. Als Messwiederholungsplan ausgelegt bietet dieses Versuchsdesign den Vorteil einer geringen Probandenzahl, um valide Ergebnisse zu erzielen. Die für diese Evaluierung gewählte Methode zielt auf die Beurteilung kleiner Unterschiede ab. Als Nachteil erweist sich jedoch die fehlende Möglichkeit eines direkten Vergleiches der Ergebnisse oder einer absoluten Bewertung.

In dieser Evaluierung wurde ein Referenzsignal als zwingend angesehen, da eine Beurteilung dynamischer Verhältnisse ohne einen Bezug als schwierig erachtet wurde und bestimmte dynamische Effekte in manchen Bereichen als Stilmittel verwendet werden. In diesem Sinne käme als alternative Versuchsmethode ein MUSHRA-Test in Frage, mit dem Vorteil der direkten Gegenüberstellung der Testobjekte. Ein wesentlicher Nachteil wären die sehr diffizilen Unterscheidungsmerkmale der prozessierten Signale, deren separate Beurteilung als wesentlich einfacher erachtet wurde. Der Hörversuch wurde mit Hilfe einer im IRT entwickelten Hörversuchssoftware umgesetzt. Ausschlaggebend für diese Entwicklung stellt die Einbindung verschiedener Methoden und Standards, die Umsetzung komplexerer Szenarien sowie die Anpassung an jeweilige Bedürfnisse dar. Als Client-Server-Architektur ausgelegt besteht der größte Vorteil der Software in einer Browser-basierten Anwendungsmöglichkeit und der individuellen Gestaltung der Rahmenbedingungen. Die Oberfläche der Software ist in Abb. 1 dargestellt.

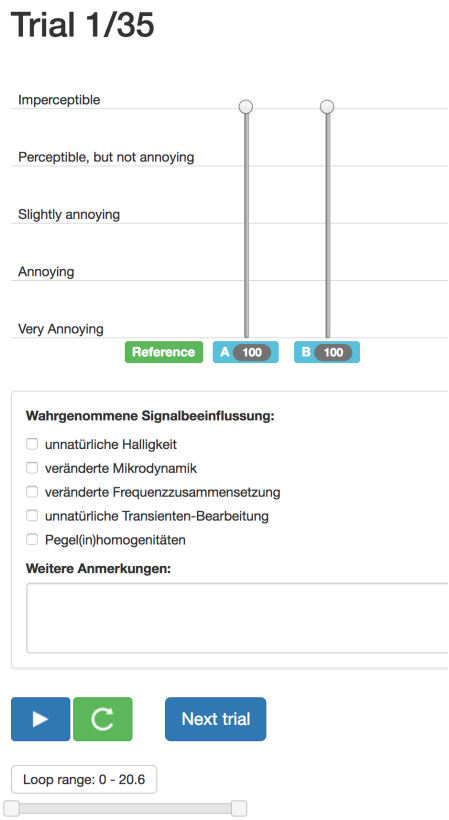


Abb. 1: Oberfläche der IRT-internen Hörversuchssoftware in Anlehnung an ITU-R BS.1116 mit zusätzlicher Möglichkeit von Attributangaben zur Spezifizierung der wahrgenommenen Signalbeeinflussung.

Erkennbar ist neben den Attributen auch ein zusätzliches Kommentar-Feld für Anmerkungen oder zur Angabe individueller Attribute.

Als Audiosignale wurden 7 verschiedene Beispiele (siehe Tab.1) mit je einer maximalen Dauer von ca. 20 sec. gewählt.

Programm/ Soundfile	IL_k [LUFS]	LRA [LU]	TP_{max} [dBTP]	Genre
Annunciation	-19,0	25,9	-0,1	Chor
Mahler 5	-19,5	23,2	0,0	E-Musik
Musik Mod	-21,6	18,0	-0,2	Radio
Sanssouci	-22,5	9,9	0,0	TV/Doku
SOKO	-23,4	11,4	-0,8	TV/Serie
Tricycle	-26,3	17,3	-0,1	U-Musik
Zarathustra	-14,4	18,5	0,0	E-Musik

Tab. 1: Übersicht der im Hörversuch verwendeten Audiobeispiele (Programm) unter Angabe der Lautheitswerte der Originaldateien und dem jeweiligen Genre. (IL_k = Programmlautheit, LRA = Loudness Range, TP_{max} = maximaler TruePeak-Wert).

Aufgrund dieser zeitlichen Beschränkung war es in dieser Evaluierung nicht möglich, länger andauernde dynamische

Verhältnisse (Makrodynamik) zu untersuchen. Hierfür wären Beispiele in der Größenordnung von Minuten nötig gewesen, jedoch steht hier generell die Schwierigkeit des auditiven Kurzzeitgedächtnisses einer adäquaten Evaluierung im Wege. Die Beispiele wurden aus unterschiedlichen Genres wie E- und U-Musik, Radio- und TV-Inhalten und einem speziellen Chor-Stück ausgewählt, um eine möglichst große Bandbreite an senderelevanten Inhalten zur Verfügung zu stellen. Im Wesentlichen wurden Signale gewählt, bei deren Bearbeitung durch einen Lautheitsprozessor sowohl eine Programmlautheits- als auch Loudness Range- bzw. eine Spitzenpegelanpassung erfolgen sollte. Es wurden deshalb keine Signale ausgewählt, bei der lediglich eine Absenkung des Materials nötig wäre, da hier keine Artefakte oder Qualitätsbeeinflussungen zu erwarten sind.

Der Versuchsaufbau des Laborexperiments wurde in einem akustisch optimierten Raum mit definierten Abhörbedingungen (Fremdgeräuschpegel = 22 dB(A), Nachhallzeit T_{60} = 0,25 s @ 1 kHz) umgesetzt. An zwei mittels akustischer Trennwände voneinander separierten Versuchsplätzen, ausgestattet mit hochwertigen Kopfhörern (Stax Lambda Pro) und professionellem Interface (RME Babyface Pro II), konnten Probanden den Versuch zeitlich parallel absolvieren (vgl. Abb. 2).

Abb. 2: Hörversuchs-„Kabine“ in einem akustisch optimierten Raum des IRT mit Elektrostat-Kopfhörern und professionellem Audiointerface.

Nach der Durchführung dieses Laborexperiments mit 20 Probanden, bestehend aus Experten und hörunerfahreneren Personen (sog. Laien oder Consumer) wurde der Versuch als Feldexperiment umgesetzt. Dabei diente eine erweiterte Fassung der Versuchsanleitung den Feld-Probanden zur selbständigen Durchführung des Versuches. Zur Erhöhung der Validität wurden zusätzlich zur Randomisierung des Versuchsablaufs weitere Vorkehrungen getroffen, um äußere Einflüsse zu minimieren. Auf der einen Seite wurde der Versuch zur Vermeidung von Einflüssen durch das äußere Schallfeld wie im Laborversuch auf reine Kopfhörerwiedergabe beschränkt. Hierzu war es ebenfalls nötig, den Probanden vergleichbare Wiedergabepegel bereitzustellen. Dies wurde anhand eines Einpegelsignals – einem Sprachsignal, das als akustisches Referenzsignal für Messungen zur Sprachverständlichkeit Verwendung findet – ermöglicht, welches die Probanden auch nach Unterbrechung des Versuches verwenden konnten, um eine adäquate Wiedergabelautstärke einzustellen. Den Probanden wurden auf der anderen Seite Hörbeispiele für jedes Attribut sowie einem Trainings-Durchlauf zur Gewöhnung an das Versuchsszenario dargeboten. Zudem wurde ein Fragebogen (anonymisierte Abfrage persönlicher Daten wie Alter und Hörerfahrung und technischer Daten wie Interface und Kopfhörer) und eine detaillierte Anleitung inklusive dem Hinweis auf Einhaltung von Pausen präsentiert. Da für die Feld-Evaluierung hauptsächlich Tonschaffende der Rundfunkanstalten adressiert waren, bestand dieses Panel vorwiegend aus Experten.

4. Auswertung

4.1. Probandendaten

Während die Probanden-Panels beider Szenarien altersmäßig einer Normalverteilung folgen (mit tendenziell älteren Personen im Feldexperiment), ist eine eher ungleichmäßige Verteilung der Hörerfahrung zu beobachten. Die Grafiken in Abb. 3 verdeutlichen die Aufteilung des Labor-Panels in Experten und Consumer einerseits und der höhere Anteil an Experten im Feld-Panel andererseits. Für letzteres Panel konnte aus den Fragebögen die Verwendung von größtenteils professionellen Studio-Kopfhörern (geschlossen und halboffen) nebst der vorrangigen Nutzung von On-Board-Soundkarten mit eher geringerer Audioqualität gewonnen werden. Über weitere Abhörbedingungen können entsprechend der Versuchsanordnung keinerlei Aussagen getroffen werden.

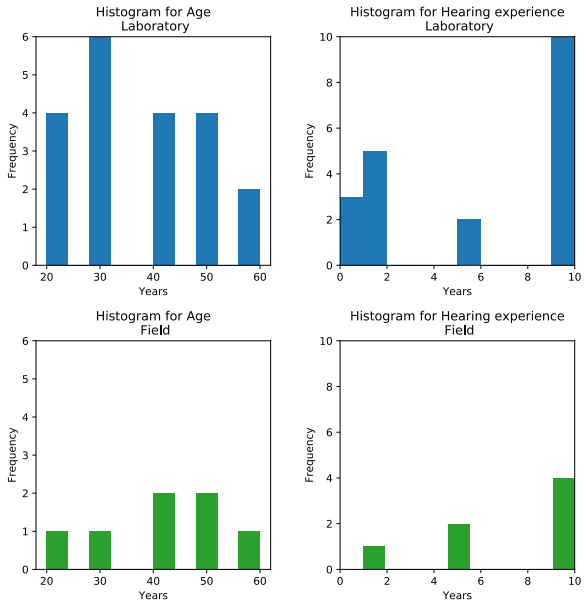


Abb. 3: Auswertung der Probanden-Daten der beiden Versuchs-Szenarien. Oben: Labor-Versuch, unten: Feld-Versuch. Links: Histogramm der Altersverteilung, rechts: Histogramm der Hörerfahrung.

4.2. Ergebnisse

Abb. 4 verdeutlicht die Bewertungen der nach dem Post-Screening verbleibenden 13 Labor-Probanden und 8 Feld-Probanden, aufgeschlüsselt nach den jeweiligen bearbeiteten Audiobeispielen (Programmen) gegenübergestellt für drei Systeme. Die Skala stellt gemäß der statistischen Beurteilungsmethode laut BS.1116 die Differenz zwischen dem zu bewertenden Stimulus und der versteckten Referenz dar. Positive Bewertungen bedeuten demnach eine irrtümliche Bewertung dieser versteckten Referenz durch den Probanden. Es wird zunächst deutlich, dass vor allem das Chor-Signal („Annunciation“) außerhalb der Bewertungskriterien dieser Methode liegt, da keiner der Algorithmen in der Lage ist, dieses Audiobeispiel im Rahmen der geforderten Zielwerte mit guter Audioqualität zu prozessieren. Grundlegend kann und muss hier die Verwendung der Zielwerte in Frage gestellt werden. Demgegenüber werden die beiden TV-Signale („SOKO“ und „Sanssouci“) größtenteils mit guter Audioqualität, d.h. ohne erkennbare Artefakte bearbeitet. Die übrigen Signale weisen je nach Algorithmus und Programmart meist erkennbare (perceptible) bis störende (annoying) Artefakte auf. In der direkten Gegenüberstellung der beiden Versuchsszenarien fällt eine tendenziell strengere Bewertung durch die Feld-Probanden auf, was auf den prozentual höheren Anteil an Experten zurückführbar sein könnte. Basierend auf den Angaben der Probanden bezüglich ihrer Hörerfahrung konnte dies statistisch jedoch nicht nachgewiesen werden. Die Tendenz der programmabhängigen Unterschiede ist in beiden Szenarien dennoch in vergleichbarem Maße ausgeprägt, ebenso wie Unterschiede in der Prozessierungsqualität der einzelnen Systeme. Hier ist jedoch ein direkter Vergleich der Systeme im Sinne einer absoluten Bewertung nicht möglich, da die Stimuli für sich jeweils zu einer Referenz bewertet wurden, nicht jedoch im direkten Vergleich. Tendenziell scheint Algorithmus 2 in der Beurteilung sowohl durch die Labor- als auch durch die Feld-Probanden mit geringerer Qualität zu prozessieren als die beiden anderen Algorithmen.

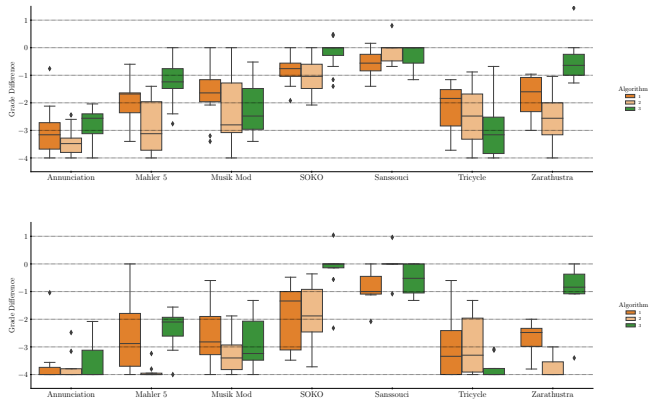


Abb. 4: Explorative Statistik: grafische Darstellung der Bewertungen der Algorithmen 1, 2 und 3 mittels Box-Plot aufgeschlüsselt nach dem jeweiligen Programm aus beiden Hörversuchs-Szenarien. Oben: Labor-Versuch, unten: Feld-Versuch. Die Skala entspricht der Differenz zwischen zu beurteilendem Stimulus und der versteckten Referenz.

In der Gesamtbeurteilung fehlen zwei der fünf getesteten Algorithmen. Eine Verwendung dieser beiden Algorithmen in einem professionellen Umfeld ist nicht zielführend, da die Bearbeitungsergebnisse die vorgegebenen Zielwerte meist nicht erreichten. Gerade diese Abweichungen jedoch führen in vielen Fällen zu einer besseren Audioqualität, was in vergleichbarem Maße aus den Bewertungen beider Probanden-Panels gefolgert werden kann. Das Gesamtergebnis der Evaluierung der drei verbleibenden Systeme ist in Abb. 5 aufgeführt mit insgesamt eher mäßigem Qualitäts-Ergebnis, aber tendenziell vergleichbaren Einstufungen der Systeme, wobei Algorithmus 2 im Vergleich zu Algorithmus 3 aufgrund der fehlenden Überlappung der Konfidenzintervalle signifikant niedriger bewertet wurde. Ebenfalls ersichtlich ist ein größeres Konfidenzintervall bei den Bewertungen des Feld-Szenarios aufgrund der geringeren Probandenanzahl.

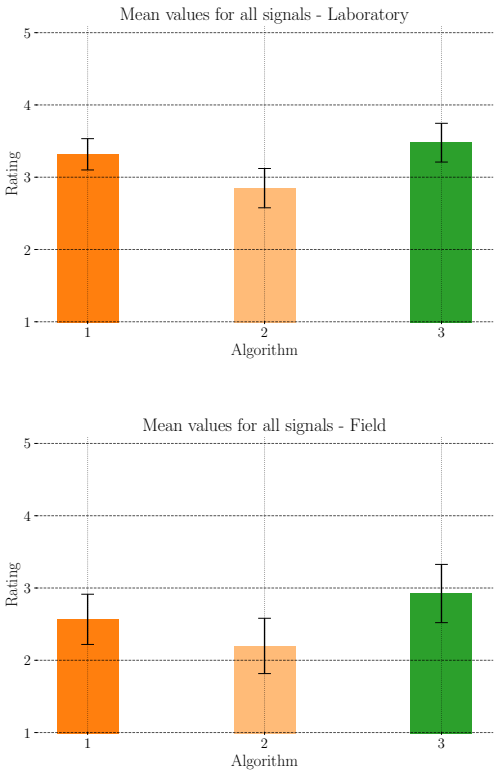


Abb. 5: Gegenüberstellung der Mittelwerte der Gesamtbewertungen für die Algorithmen 1, 2 und 3 mit 95%-Konfidenzintervall aus beiden Hörversuchs-Szenarien. Oben: Labor-Versuch, unten: Feld-Versuch.

Der Vollständigkeit halber ist in Abb. 6 das zusammengeführte Ergebnis aus beiden Versuch-Szenarien dargestellt, welches die Tendenzen bezüglich der Qualität der Systeme entsprechend wiedergibt.

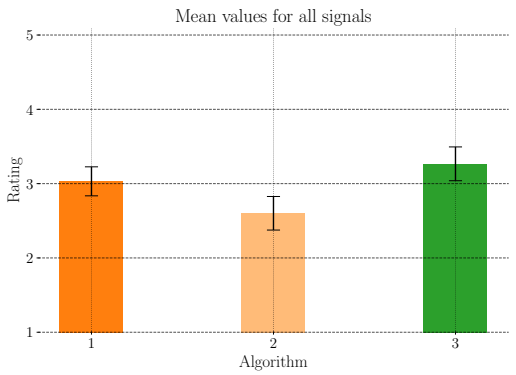


Abb. 6: Gegenüberstellung der Mittelwerte der Gesamtbewertungen für die Algorithmen 1, 2 und 3 mit 95%-Konfidenzintervall aus beiden Hörversuchs-Szenarien zusammengenommen.

4.3. Statistische Auswertung

Die statistischen Auswertungen der beiden Szenarien zeigten in Bezug auf eine Gegenüberstellung der Mittelwerte und Varianzen ein vergleichbares Resultat. Die Nullhypothese gleicher Mittelwerte der Bewertungen der Systeme muss für beide Szenarien verworfen werden. Ebenso muss die Nullhypothese der gleichen Mittelwerte der Bewertungen der Programme innerhalb eines Systems für alle Systeme verworfen werden und zwar in gleichem Maße für beide Szenarien. Ausschlaggebend für eine Erweiterung des Probanden-Panels durch das Feldexperiment ist jedoch die Nullhypothese, dass für beide Szenarien die Mittelwerte der Bewertungen der jeweiligen Programme innerhalb eines Systems gleichwertig sind. In 80% der Fälle ist diese Forderung gegeben (vgl. Tab. 2), was – bis auf eine Ausnahme – aufgrund der nicht vollständig gegebenen Normalverteilung durch einen Kruskal-Wallis-Test als nicht-parametrische Alternative zu einer ANOVA bestätigt wurde (vgl. [3]).

Programm/ Soundfile	System/Algorithmus				
	1	2	3	4	5
Annunciation	1,54	1,43	4,35	3,31	2,98
Mahler 5	2,30	8,52	10,25	0,57	4,59
Musik Mod	3,85	3,36	2,29	1,53	3,02
Sanssouci	1,54	0,60	1,49	0,43	0,97
SOKO	7,52	4,57	0,00	1,44	0,88
Tricycle	4,12	1,01	3,48	7,55	0,57
Zarathustra	8,82	9,91	1,63	0,33	3,04

Tab. 2: Vergleich der Mittelwerte aus beiden Versuchs-Szenarien durch Varianzanalyse (ANOVA): F-Statistik mit oberem F-Wert von 4,38; Signifikanzniveau 0,05. H_0: $\mu_1=\mu_2$. Die rot markierten F-Werte kennzeichnen ein Verwerfen der Nullhypothese.

Eine Erweiterung des Probanden-Panels scheint für dieses Experiment geglückt zu sein. Ein Einfluss der äußeren Bedingungen auf die Versuchs-Ergebnisse wie verwendetes Equipment, Abhörumgebung oder Probanden-bezogene Umstände ist nicht nachweisbar. Für eine tiefergehende und weiterführende statistische Analyse wurde als Post-hoc-Test ein multipler Paarvergleich nach Tukey (vgl. [4]) vorgenommen mit dem Ergebnis wie in Abb. 7 dargestellt. Das Chor-Signal (ganz unten) wird im Schnitt als qualitativ eher niedriger bewertet. Die Gegenüberstellung zeigt eine vermeintlich homogenere Bearbeitung durch Algorithmus 1 über alle Signale hinweg, während Algorithmus 3 teils sehr gute Ergebnisse erzielt, allerdings mit größeren Diskrepanzen in Bezug auf die unterschiedlichen Programme. In der Tendenz ist diese Erkenntnis in beiden Szenarien identifizierbar mit sichtbar größeren (Gruppen-) Konfidenzintervallen und tendenziell strengeren Bewertungen im Feldversuch.

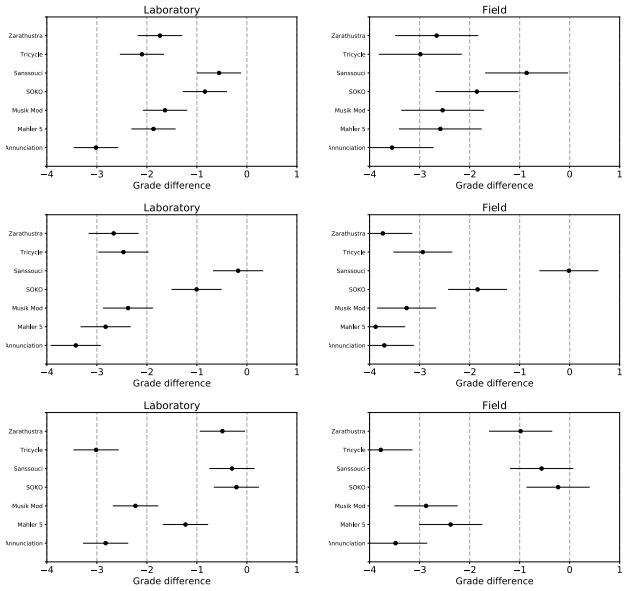


Abb. 7: Grafische Darstellung der Post-hoc-Analyse mittels multiplem Paarvergleich nach Tukey durch Gegenüberstellung der Bewertungs-Mittelwerte und Gruppenkonfidenzintervalle der jeweiligen Programme innerhalb der Systeme 1, 2 und 3. Links: Labor-Versuch, rechts: Feld-Versuch.

Eine abschließende Auswertung der Attribute unter Berücksichtigung aller Programme ergab keinen signifikanten Unterschied im direkten Vergleich der Algorithmen. Im Detail jedoch sind einige Schwachstellen der Algorithmen bei Betrachtung der einzelnen Programme zu beobachten. Unter Anwendung eines exakten Tests nach Fisher (vgl. [5]) wurden die Häufigkeiten jedes einzelnen Attributs innerhalb eines Systems in Abhängigkeit der Programme verglichen unter der Nullhypothese der Unabhängigkeit der Attribut-Angaben vom jeweiligen Szenario. Wie in Abb. 8 erkennbar ist die Wahrscheinlichkeit einer zufälligen Unabhängigkeit in den meisten Fällen größer dem Signifikanzniveau von 5%, und die Nullhypothese muss somit größtenteils verworfen werden. Es ist davon auszugehen, dass die Attribut-Angaben der Probanden aus den beiden Versuchs-Szenarien korrelieren.

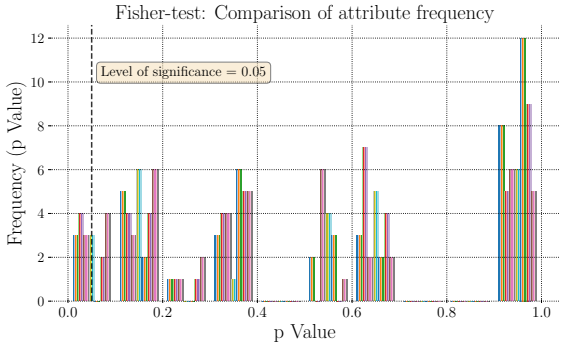


Abb. 8: Vergleich der Attribut-Häufigkeiten der beiden Hörversuchs-Szenarien: p-Wert und dessen Häufigkeit ausgewertet für jedes Attribut innerhalb eines Systems in Abhängigkeit des Programms gem. einem exakten Test nach Fisher unter der Annahme unabhängiger Häufigkeiten (Nullhypothese).

5. Schlussfolgerungen

Die statistischen Resultate lassen die Schlussfolgerung zu, dass sich aufgrund der hohen Übereinstimmung der Bewertungen die Ergebnisse aus beiden Versuchs-Szenarien gegenseitig ergänzen, eine Erweiterung des Probanden-Panels somit erfolgreich vorgenommen werden kann.

Als übergeordnetes Ziel dieser Evaluierung könnte im Anschluss dieser Experten-Befragung mit Hilfe der vorgestellten Browser-basierten Software-Lösung eine Consumer-Evaluierung zu dieser Thematik unternommen werden. Hierfür müsste allerdings im Vorfeld ein geeignetes Versuchsdesign beispielsweise als Zufallsgruppenplan eruiert werden. Des Weiteren ist es unabdingbar, ausreichend valide Vorversuche durchzuführen, um verschiedene Einflüsse wie Abhörumgebung, Art des Endgerätes oder Nutzerpräferenzen abschätzen zu können.

6. Literatur

[1] EBU: Technical Recommendation R128. Loudness normalisation and permitted maximum level of audio signals. Genf, 2014

[2] ITU: Recommendation ITU-R BS.1116-3. Methods for the subjective assessment of small impairments in audio systems. Genf, 2015

[3] Bech, S., Zacharov, N.: Perceptual Audio Evaluation. Wiley & Sons, Chichester, 2006

[4] Borz, J., Schuster C.: Statistik für Human- und Sozialwissenschaftler. Springer-Verlag, Berlin Heidelberg, 2010

[5] Sheskin, D.°J.: Handbook of Parametric and Nonparametric Statistical Procedures, Fifth Edition. Chapman & Hall, Boca Raton, 2011

Concept of Earth Concert Hall

J. Schulz[1]
[1] *Potsdam, Germany, Email: joerg.schulz@tonmeister.de*

Abstract

The idea behind the Earth Concert Hall (ECH) is to create a globally distributed portal system for announcing and broadcasting live audiovisual events over the Internet: artists or their management will use the portal to announce concerts to be broadcast across the globe. They stream their live music, available to consumers in various quality levels at different financial conditions. The lowest level would be to offer banner advertising without paying for music, which implies an advertising effect for the Earth Concert Hall itself, that would promote its dissemination. By contrast, registered viewers consume high-resolution ad-free qualities by a billing model. Artists receive a share of revenue to be defined. For this system to operate across borders, well-established cooperation partners are needed to represent artists' rights in the respective countries. These old partners developed by the proliferation of mechanical recordings and have a reliable international distribution of dividend, for example when third party rights are affected. Authors and composers, too, should receive their share of the proceeds of the Earth Concert Hall performances due too existing network of cooperating traditional regional collecting societies worldwide. The high-performance infrastructure of global communication achieves a status that could enable the comprehensive implementation of the abovementioned solutions.

1. Introduction

The goal of the Earth Concert Hall (ECH) is to create a global portal system for announcing and broadcasting live audiovisual events over the Internet. Thus, artists or their management can proclaim and present concerts worldwide through such a portal by streaming their live music to consumers with varying levels of quality and cost. The lowest resolution level can be funded by banner advertising without payment. As a side effect, these free performances promote ECH and its spread on our planet. In contrast, high-resolution sound and image recordings are offered only to registered viewers. The quality follows a billing model. Artists receive a share of the turnover.

The tasks to be solved to create a global distributed portal system for the announcement and transmission of live audiovisual events over the Internet are cross-border. The representation of artists' rights in all countries requires well-established cooperation partners. Existing collecting societies, founded in an era of mechanical recordings, have a reliable international dividend payout for creatives. Authors, composers and third parties may receive their share of the proceeds of the ECH performances as the network of cooperating traditional regional collecting societies exists worldwide. An essential part of the required infrastructure already exists: high performance data networks across the globe.

The user interface of the Earth Concert Hall would show the image of our rotating planet Earth. If you zoom in, the picture changes to a map. When a location of a live concert appears on the screen, its music is played through real-time broadcast, seemingly loudest from the center of the displayed earth view.

2. Project Planning

First, a suitable project planning method has to be defined.

2.1. Choice of Method

An experience in computer architecture teaches that complex distributed computer systems are designed like large building projects in civil engineering or construction. Therefore one can speak of computer architecture. The chosen synergetic planning model now used in this article is SIA 112 [1]. SIA stands for the Swiss Engineering and Architects Association.

2.2. Strategic Planning

This section 2.2 outlines the first step in SIA 112.

2.2.1. Formulation of Needs

So far, there is nothing comparable in the world that allows anyone to watch each concert in real time via a single portal on the Internet. The global high performance communication infrastructure enables such a solution.

2.2.2. Approaches and Strategies for Solutions

Even commercial vendors are increasingly using open source software to reduce development costs. An open source project development team saves money because its members use components for free, but may also need to provide their work - depending on the licensing model - to the society. Lengthy negotiation phases between contract partners are thus avoided.

Volunteers, commercial, governmental and international supporters are welcome to create such a supraregional cultural network that presents live music across borders, offering artists new revenue opportunities.

ISBN
978-3-9812830-9-9

The authors hold the copyrights of all of content in this manuscript. They have granted the VDT the right to publish it in print, online and as PDF download on https://tonmeistertagung.com

Computer architecture: Repeater servers with low data transmission costs and spread across the earth, allow connections to regional end users. A unified Debian Linux server distribution with pre-configured settings ensures safe operation of the ECH nodes. Payment services are handled by external specialists, reducing the security risk for ECH servers.

2.3. Pre-studies

2.3.1. Definition of the Project

The Earth Concert Hall aims to create a global distributed portal system for announcing and broadcasting live audio visual events over the Internet.

2.3.2. Feasibility Study

The demand of the music users, which depends on the popularity of the transmitted performances, determines the load of the system. Forecasts of the number of users to reach have to be created according to marketing aspects. This avoids server resource bottlenecks.

2.3.3. Selection Process

The self-optimization process of open source projects eliminates the need to select different bidders, as would be required for commercial product development.

2.4. Project Work

First of all, the project goal must be fixed and should not be changed significantly due to the complex dependencies of the large system.

2.4.1. Pre-project by means of prototyping

A working prototype has to be programmed and installed on a few servers. Already in the prototyping is to pay attention to scalability, in order to reuse written code and to develop a truly expandable system. Using a simple prototype menu design, this subsubsection outlines the necessary data collection to register artists and their performances, as well as listeners who want premium broadcast quality.

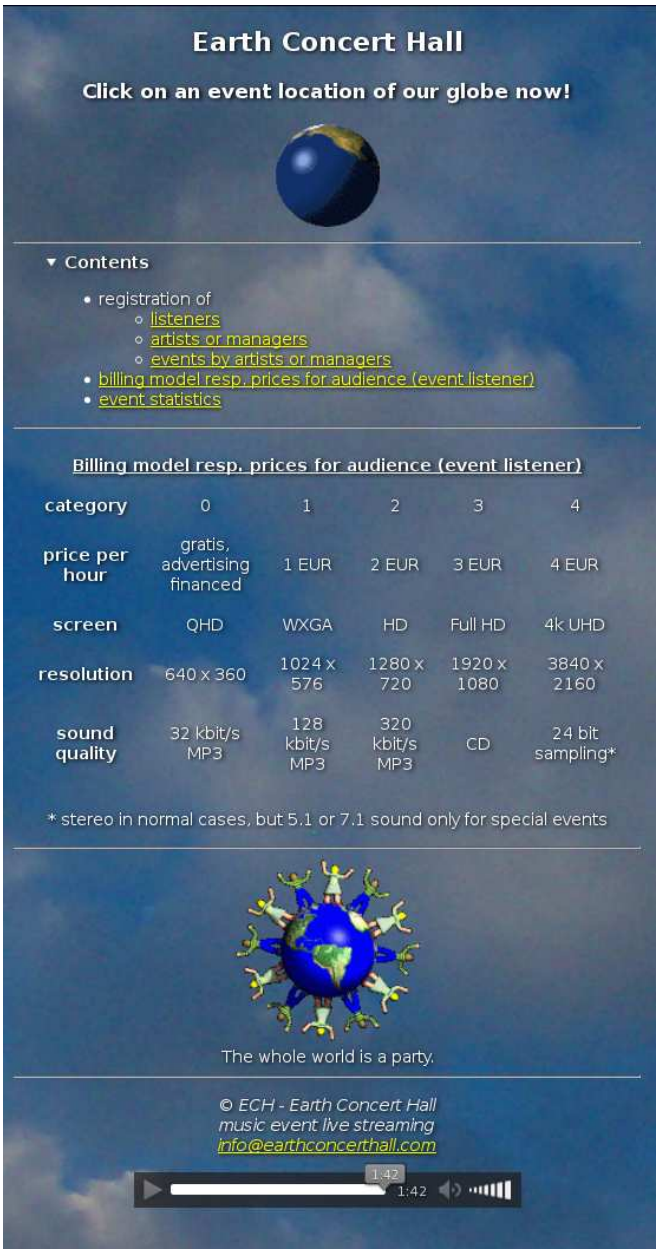


Fig. 1: Menu of some prototype features for the Earth Concert Hall, implemented in the multimedia-enabled HTML 5 standard, compatible with many web clients on different operating systems

ECH Main Menu: Figure 1 outlines the goal of implementing the Earth Concert Hall. By clicking on a location or venue of our globe you get to a music event, which is transmitted via live stream. In addition, you can choose in a menu whether you want to be registered as a listener, artist or manager or announce planned events. Furthermore, the billing model respectively prices for the audience as well as event statistics should be displayed. For example, the billing model might be differentiated by price for the target group: Of the five image transfer quality categories, the lowest is free and funded by advertising, while the higher are available with rising prices, starting with QHD and ending with UHD, ie 4k resolutions. The sound quality varies between 32 kbit/s with data reduction and high-resolution lossless 24-bit sampling quality. The HTML 5 standard used ensures high compatibility with various web browsers and operating systems when transferring multimedia content.

Artists or managers register for the Earth Concert Hall by completing a form sheet (see picture 2). Required are salutation, first and last name, e-mail address, desired password and its repetition, telephone, street, postal code, town, country (whereby the system automatically inserts a 3-letter code) as well as the payment connection.

Events are registered in the Earth Concert Hall by artists or their management (cmp. figure 3). The applicant clicks on the virtual venue on the globe or on a map in an enlarged view. The music event form asks for name, date, time as Greenwich Mean Time (GMT), location, street, postal code, city, country (also here the system automatically inserts a 3-letter code). In addition, salutation, first and surname of the artist or manager of this event must be stated as well as e-mail address and password. A brief overview explains rules

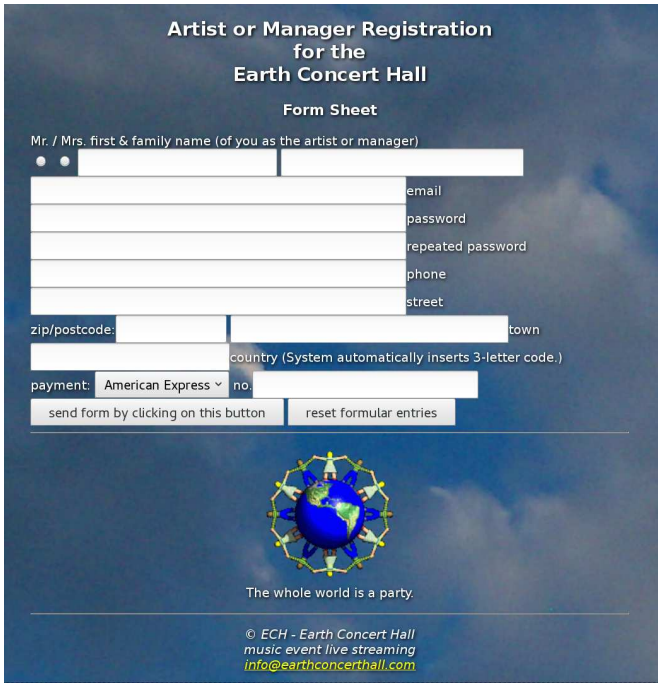


Fig. 2: Input field in which artists or their managers can register

for the payout to artists or event managers in the previously mentioned five broadcast quality categories.

The listener registration in the Earth Concert Hall is realized by a form sheet that includes the salutation (see fig. 4), first and last name, email address, password and its repetition, house number, street, postal code, town, country, and here too the server automatically generates a 3-letter code for the state. The desired payment method can also be specified.

2.4.2. Final Building Project

If the prototype works, the next expansion stages can be defined in order to then distribute the system across the globe.

2.4.3. Licensing and Management Procedure

In order to ensure a wide distribution of the entire system in the future, the GNU Public License (GPL) is recommended as an open source license. This guarantees that anyone who develops code for the project will also make it permanently available and free.

2.5. Tendering Process

2.5.1. Tendering

A call for tenders does not take place because open source software is used. Motivation: All of the TOP 500 supercomputers work with the Linux open source operating system [2].

2.5.2. Compare Offer

Even open-source systems are subject to the competition of detail solutions. Similar to the Linux distribution Debian four to three development phases over several years are used: (experimental), unstable, test, stable.

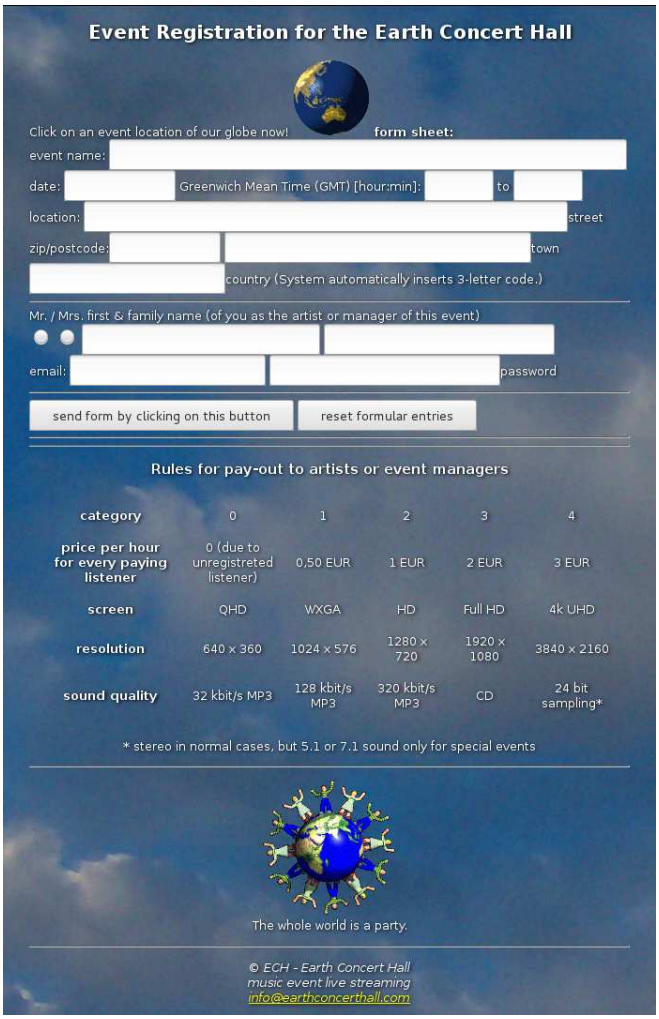


Fig. 3: Here the planned performances are reported.

2.5.3. Application Proposal

If the new version is stable, it will be transferred to the production system.

2.6. Realization

2.6.1. Detailed Design

A project group has to be established and can be organized under the reserved domain: http://earthconcerthall.net/

Within the project special working groups are to be formed covering legal, organizational and technical areas.

2.6.2. Implementation

The software developer working group within the project would program the complex system according to the tasks defined by the organizers.

An implementation can be organized through a developer platform such as GitLab [3].

2.6.3. Commissioning and Completion

After a successful test phase, the Earth Concert Hall can be put into operation for all users around the globe, with a prior estimate of the need to avoid overloading when opening.

For the striking look and feel of the user interface, the

Fig. 4: At this point, listeners register to receive performances in high-definition transmission quality.

rotating earth is important, in whose regions the listener immerses. Such a solution is provided by Google Earth and now Google Maps, which integrates very well with web design [4]. An alternative would be OpenStreetMap, which does not require the transfer of data to third parties when requesting geo-data by the ECH user, if the geo-information service OpenStreetMap works locally, which is privacy-friendly [5].

The domain earthconcerthall.com has been reserved for productive use by the listeners of the performances.

3. Conclusion and Outlook

This short paper outlined the conception of a virtual earth concert hall to be planned including the manifold tasks in such a complex project.

Even before planning the productive operation, ”Privacy by Design and by Default” must be one of ECH’s development goals in order to meet the world’s leading requirements of the European Data Protection Regulation.

One possible project add-on would be the Earth Concert Archive (ECA) as an entry to past performances with a similar billing model as for the ECH.

4. References

[1] Röthlisberger, H.: Das neue Leistungsmodell SIA 112. CH-3645 Gwatt-Thun, 2018, URL: http://www.roeplaner.ch/buch_neuerungen/text/k_2.htm

[2] List Statistics — TOP500 Supercomputer Sites. URL: https://www.top500.org/statistics/list/

[3] The first single application for the entire DevOps lifecycle - GitLab. URL: https://about.gitlab.com/

[4] Google Maps – Earth View. URL: https://www.google.de/maps/@51.1758062,10.4541192,2.33z

[5] OpenStreetMap. URL: https://www.openstreetmap.org/

MEMS-Loudspeaker – A Novel Class of Electroacoustic Transducers for Mobile Audio Applications

L. Ehrig[1], B. Kaiser[1], H. Conrad[1], He. Schenk[1], D. Schuffenhauer[1], M. Stolz[1], M. Gaudet[1,2], Ha. Schenk[1,2]

[1] *Fraunhofer Institute for Photonic Microsystems, Germany, Email: lutz.ehrig@ipms.fraunhofer.de*

[2] *Brandenburgische Technische Universität Cottbus-Senftenberg, Lehrstuhl Mikro- und Nanosysteme, Germany*

Abstract

Wireless earphones require a tremendous degree of integration: In addition to the electroacoustic transducer at least the power supply and a radio module have to be accommodated, supplemented by several microphones for hands-free applications or for noise reduction, as well as position and bio sensors, memory and CPU. With increasing integration, the pressure on the individual components increases in terms of size, energy efficiency and costs. In the field of microphones, MEMS technology has enjoyed great popularity for almost 10 years, and due to their benefits (size, cost, integration) MEMS microphones have almost completely supersede other microphone technologies in mobile audio devices and opened up new fields of applications.

For MEMS speakers, a similar development is expected as for the MEMS microphones. However, the development of MEMS-based loudspeakers is much more challenging – even for earphones – due to the large mechanical deflection required to provide sufficient volume displacement and sound pressure level. In this article the work of the Fraunhofer IPMS in Dresden on a novel actuator for MEMS speakers is presented. The actuator works according to the electrostatic principle and uses a bimorph-like mechanism to transfer an electrostatic force acting perpendicular within a beam into a bending movement. Such actuators can be grouped and optimized for specific frequency ranges according to their application. The use of this technology is particularly suitable for in-ear audio applications like hearables, earphones, IEMs and hearing aids due to its small size and high efficiency.

1. Introduction and Motivation

The acronym "MEMS" refers to micro electro-mechanical systems, i.e. devices in the µm-scale. Their details are only visible with the aid of microscopes and they are often so complex that they are referred to as micro-machines. The manufacturing of MEMS involves methods used for the fabrication of integrated circuits along with the selective removal and the addition of material layers.

The scale of MEMS devices and the techniques used for fabrication lead to an unique set of features and advantages: predominantly that batch fabrication is possible and that common CMOS processes can be used resulting in low costs per unit in mass production. Further advantages are obviously the small size and light weight of the devices, the high energy efficiency at least regarding electrostatic and piezoelectric devices, the enhanced performance and reliability due to high resolution devices (e.g. for optical applications) and the possibility to design arrays of devices and to obtain tight tolerances is given.

Adopting the same fabrication processes and materials as for integrated circuits the integration of MEMS sensors and actuators with electronic circuits necessary to read out or to process signals is obvious. Therefor the application of MEMS suggests itself in fields where this integration leads to better performance. An example is the development and rise of MEMS microphones and the integration of interface circuits. They have been evolved from simple pre-amplifiers to complex mixed-signal circuits including A/D-converters and basic voice recognition capabilities for wake-up functionalities [1]. With this integration of functionalities combined with superior performance and lower costs MEMS microphones superseded electret condenser microphones in many applications and lead to new ones.

The question arises why MEMS speakers have not been successful already [2]. Even for small closed volumes as for in-ear applications one main challenge is to displace enough air volume to attain sufficient sound pressure levels. As a matter of course the operation needs to be dynamic in a wide frequency range (e.g. 20 Hz up to 20 kHz) with good linearity, i.e. low distortion. Further the established competing loudspeaker technologies like moving coil, balanced armature and electrostatic have been evolved and optimized over many decades – not only in terms of technology, materials and manufacturing processes, but also in terms of supplementary system components like amplifiers, signal processing etc.

2. "Nanoscopic Electrostatic Drive"-Technology

Conventional electrostatic speakers need high voltages due to the large gap between the electrodes. The voltage ranges from several hundred Volts for in-ear systems to over 1 kV for floor speakers. The voltage can be reduced by reducing

ISBN 978-3-9812830-9-9

The authors hold the copyrights of all of content in this manuscript. They have granted the VDT the right to publish it in print, online and as PDF download on *https://tonmeistertagung.com*

the gap, but this would limit the maximum excitation of a diaphragm and consequently the maximum sound pressure level.

To overcome this drawback a bimorph bending actuator has been developed at Fraunhofer IPMS, which uses the ability of lithographic processes and sacrificial layers to fabricate very small electrode gaps. The large electrostatic forces in these gaps can be used with an appropriate micromechanical design to be transformed to bending actuation.

Such an actuator is depicted in Fig. 1. It consists of two electrodes with insulating spacers between them. In case of an electrical potential *V* between the top (green) and the bottom (gray) electrodes an electrostatic field *E* is generated, which leads to the attraction of the electrodes due to the Coulomb force. Because of the roof-like topography of the top electrode the attraction results in a lateral strain and a force *F* leading to a downward bending of the cantilever tip. Shifting the spacers to the top of the roof-like-structured electrodes an upward bending can be achieved. The actuator mechanism is described in detail in [3].

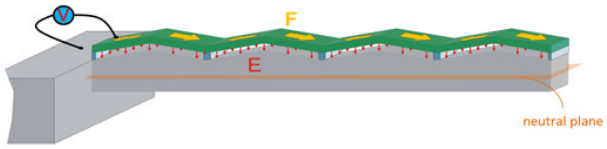


Fig. 1: Electrostatic bending actuator "Nano Electrostatic Drive" (NED): An electric potential *V* between two electrodes results in an electrostatic field *E* and a force *F*, which – due to the topology of the electrodes – results in a bending of the beam.

3. "Nano E-Drive"-Speaker for In-Ear-Applications

The NED-actuators are clamped on both sides and structured in a way that a S-shaped deformation occurs, see Fig. 2. Beams with opposite deflection curves are arranged in pairs, so that the volume between these actuators changes according to the signal voltage applied to these actuators.

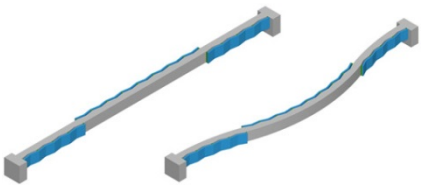


Fig. 2: Clamped-clamped NED-actuators with a S-shaped deflection curve in rest position (left) and deflected (right)

Multiple pairs of actuators are arranged on one chip, see Fig. 3. The chip is covered with a bottom and top wafer having slits to release the air. Separating the front and the backside of the chip, the acoustic short circuit is avoided and the chip emits audible sound. With this arrangement a variety of designs can be realized, not only squared but also circular footprints or any other desirable geometric form.

It has to be emphasized that with such an arrangement no additional diaphragm and no additional assembling steps are necessary. Furthermore the maximum displaced volume can be increased by "just" increasing the height of the beams, i.e. the thickness of the silicon on insulator (SOI)-wafer layer. The size of the chip and the footprint remains the same.

Due to the electrostatic approach and the absence of a diaphragm and any additional materials the NED audio transducer chip is fully CMOS compatible and the whole ecosystem of CMOS fabrication as used for integrated circuits can be utilized. Since no PZT is used the actuator is RoHS-compliant as well.

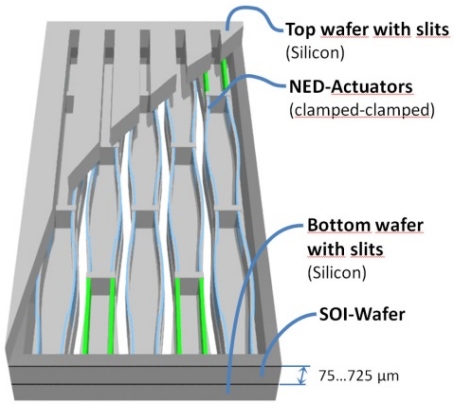


Fig. 3: NED audio transducer chip with multiple clamped-clamped actuators and a top and a bottom wafer with slits. The beams are shown in the deflected state.

3.1. First Test Structure

A first test structure of a NED audio transducer chip with four different beam lengths is depicted in Fig. 4. The microscopic image on the left shows four groups of beams, the longest beams in the lower half of the chip having a length of about 8 mm. The actuators have resonance frequencies between approx. 1 kHz and 25 kHz depending on the length of the beams. The beams are clamped on both sides. The length of the edge of the test structure is 10 mm.

The image on the right in Fig. 4 shows the chip with a top cover with biconcave-shaped slits above the center of two neighboring beams. The chip is glued on a printed circuit board and the electrical connection is made by wire bonding. With this structure electrical and acoustical functionalities have been verified for the first time.

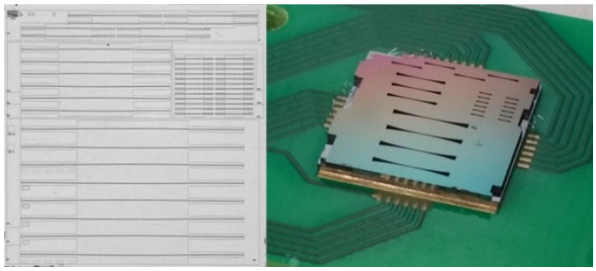


Fig. 4: First test structure of a NED audio transducer with a chip size of 10x10 mm². The microscopic image on the left shows eight 8 mm actuators (lower half), six respectively eight actuators of medium length (left upper half and top) and 24 short ultrasonic actuators (right upper half). On the right image the test structure is shown on a printed circuit board and with a top cover having biconcave-shaped slits.

3.2. Second Test Structure

A second test structure – optimized for audio reproduction – contains three sub-chips as depicted in Fig. 5. Each sub-chip having 14 pairs of beams with a height of 75 µm and a resonance frequency of about 9 kHz. Within the sub-chips different electrode designs have been tested. The bond wires to the printed circuit board are visible and on the chip itself the top cover with the slits can be seen.

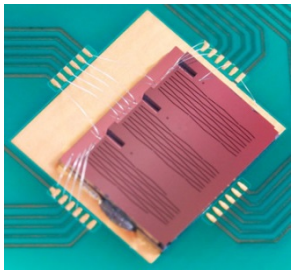


Fig. 5: Second test structure of a NED audio transducer with three sub chips, each having a size of approx. 6.2 x 2.4 mm².

The results of acoustic measurements in a 2cc-coupler correlated with the expected values for this specific design and reached almost 70 dB sound pressure level at resonance.

4. Design Options

Since no additional diaphragm is attached to the actuators, but the actuators themselves are used to displace the air, they can be arranged arbitrarily on the chip and many designs are conceivable.

A major advantage with respect to integration capabilities is a design where the vents for the air are along the edges of the chip. A possible configuration is shown in Fig. 6. The beams are configured in pairs as well and an excitation of the beams changes the volume between the individual beams in such a way that air flows in on one side and flows out on the opposite side of the chip and vice versa. With this design the top and bottom wafers are available for drive and signal processing electronics and additional MEMS sensors or actuators. Even a stack of multiple NED transducer chips is possible.

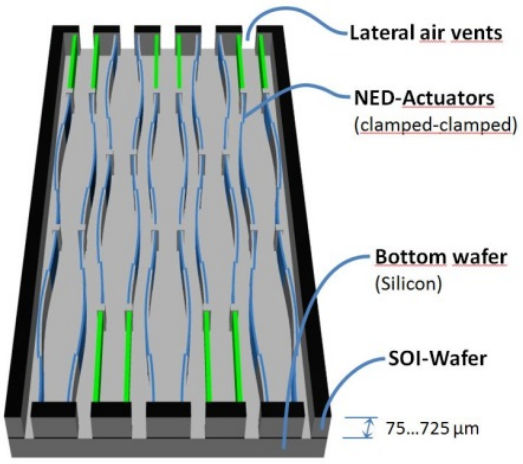


Fig. 6: Basic principle of a design with lateral air vents. Top wafer (not shown) and bottom wafer are available for electronic components or additional MEMS sensors and actuators.

Increasing the sound pressure level, which is one of the essential improvements necessary to cope with existing audio transducer technologies, can be accomplished in three ways: a) by increasing the number of beams, b) by increasing the maximum deflection of each beam and c) by increasing the height of the beams. A very attractive way is to increase the height of the beams, which is depicted in Fig. 7. In principle "just" the thickness of the SOI-wafer layer needs to be increased. The footprint of the chip and the size of the die will remain the same. The manufacturing costs, which are basically a function of the area of the chip, will change only marginal. However there are limitations regarding the process for deep silicon trench etching, thus the design of the beams would need to be adjusted.

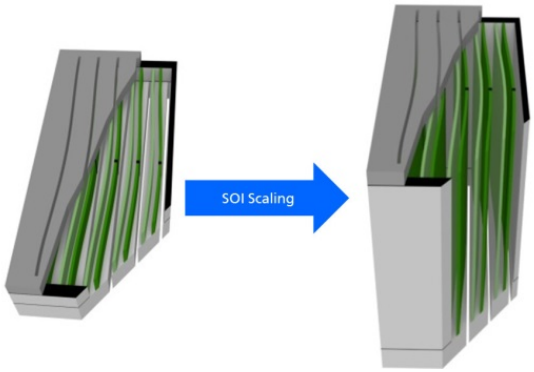


Fig. 7: Principle of increasing the height of the beams by increasing the thickness of the SOI-wafer.

5. Summary

Due to the manufacturing processes MEMS technology allows for light, small, energy efficient and low cost components. Due to the same manufacturing processes (CMOS) as for integrated circuits an easy integration of MEMS actuators and sensors with supplementary electronics is possible and a cost efficient mass production can be realized.

The presented "Nano Electrostatic Drive"-MEMS-Actuator is based on a bimorph-like bending actuator and allows for large deflections. Because of the small electrode gap low voltages are sufficient to generate high forces.

Two test structures of NED audio transducer chips were presented allowing audio reproduction in a small closed volume as it is the case for earphones and other in-ear audio devices such as hearables, in-ear-monitors and hearing aids. Due to the absence of a diaphragm any additional assembling steps are obsolete. The high degree of freedom regarding the design allows for miniaturization of in-ear transducer systems, and high price competitive transducers in mass markets.

Further development focuses on increasing sound pressure level and linearity by optimizing the actuator design.

Acknowledgement

The work shown here was funded in the project MEMSound by the German Federal Ministry of Education and Research (BMBF) in the funding measure "Validation of the technological and social innovation potential of scientific research - VIP +" under the grant number 03VP01800.

References

[1] Malcovati, P., Baschirotto, A.: The Evolution of Integrated Interfaces for MEMS Microphones. Micromachines 2018, 9, 323.

[2] Yole Développement: Acoustic MEMS and Audio Solutions, 2017 Report.

[3] Conrad, H. et al: A small-gap electrostatic micro-actuator for large deflections. Nature Communications 6, Art. No.: 10078 (2015)

Auswirkungen von Active-Noise-Cancelling Kopfhörern auf kognitive Leistungsfähigkeit und Mitarbeiterzufriedenheit im Bürokontext

B. J. Müller[1], A. Liebl[2]
[1] *Fraunhofer-Institut für Bauphysik, Stuttgart, Germany, Email: benjamin.mueller@ibp.fraunhofer.de*
[2] *Hochschule Döpfer, Regensburg, Germany, Email: a.liebl@hs-doepfer.de*

Zusammenfassung

Eine hohe Sprachverständlichkeit führt in offenen Büroumgebungen zu Problemen, da sie sich durch den Irrelevant Sound Effekt negativ auf die kognitive Leistungsfähigkeit auswirkt. Raumakustische Maßnahmen allein können unter Beibehaltung einer offenen Bürostruktur zu keinen ausreichenden Verbesserungen führen. In einer Probandenstudie mit einer Arbeitsgedächtnisaufgabe (Serial-Recall) wurde untersucht, ob sich Active-Noise-Cancelling (ANC)-Kopfhörer positiv auf die kognitive Leistungsfähigkeit und das subjektive Empfinden bei einer Großraumbüro-Geräuschkulisse auswirken. Es konnte dabei kein signifikanter Unterschied bezüglich der kognitiven Leistungsfähigkeit zwischen den Bedingungen mit aus- und eingeschalteten ANC bei aufgesetzten Kopfhörern sowie ohne Kopfhörer festgestellt werden. Allerdings hatte das ANC einen Einfluss auf das subjektive Empfinden. Die Hintergrundgeräusche wurden mit ANC beispielsweise als signifikant weniger lästig, die Konzentrationsfähigkeit als signifikant höher und der Abstand zur Störquelle als signifikant größer wahrgenommen als ohne.

1. Einführung

17 Millionen Deutsche sind im Büro beschäftigt [1]. Davon arbeiten nur 58 % in Einzel- oder Zweierbüros, die anderen 42 % in größeren Büroformen. Ginge es nach den Beschäftigten, sähen Bürolandschaften in Deutschland anders aus: 95 % aller Büroarbeiter würden sich für ein Ein- oder Zweierbüro entscheiden, wenn sie die Wahl hätten [2]. Trotzdem planen Unternehmen bei zukünftigen Immobilien aus wirtschaftlichen Gründen vor allem größere Bürolandschaften [1]. Die offenen Büroformen bringen viele Probleme im Bereich der Akustik, aber auch in anderen Bereichen mit sich. Durch den Einsatz raumakustischer Maßnahmen lassen sich nur bedingt Verbesserungen erzielen. Zwar wird versucht nicht nur die Nachhallzeit, sondern vor allem auch auch die Sprachverständlichkeit von störender oder irrelevanter Sprache im Büro zu reduzieren, jedoch kann die Raumakustik nur sehr begrenzt beeinflusst werden, wenn eine offene Bürostruktur gewahrt werden soll. Zur Verbesserung wären nämlich vor allem hohe, absorbierende Raumtrenn- und Schirmwände notwendig. Nicht nur subjektiv wird Lärm als großes Problem von den Mitarbeitern in offenen Büroumgebungen gesehen – Studien zeigen, dass er auch einen messbaren Einfluss auf den Körper hat. Das vegetative Nervensystem des Menschen wird ab einem Geräuschpegel von 60 dB(A) negativ beeinflusst, die psychische Verfassung und kognitive Leistungsfähigkeit sind schon ab 30 dB(A) betroffen. Besonders sprachlicher Störschall hat negative Auswirkungen, er führt schon ab einem Pegel von 35 dB(A) zu Leistungseinbußen des Arbeitsgedächtnisses [3]. Dieser Effekt wurde erstmals 1976 unter dem Namen Irrelevant Sound Effect (ISE) von Colle und Welsh untersucht [4]. Er beschreibt den Abfall der Kurzzeitgedächtnisleistung, wenn irrelevante Geräusche zu hören sind. Der ISE wirkt sich auf viele Bereiche des Gedächtnisses aus. Verschiedenste Studien seit 1976 zeigen negative Einflüsse auf das Merken von Zahlen [5], [6], das Vorlesen [7], das Fehler finden in einem Text [8], [9] sowie das Lösen von Rechenaufgaben [10] [11]. In mehreren Probandenstudien [12], [13], [10], [11] wurde der Effekt außerdem explizit bei Büro-Hintergrundgeräuschen nachgewiesen [4]. Um den ISE zu reduzieren, werden in Großraumbüros immer häufiger ANC-Kopfhörer eingesetzt. Aus diesem Grund wurde untersucht, inwieweit sich durch ANC-Kopfhörer die Störwirkung von Sprachschall in Büroräumen reduzieren lässt. Über einen Probandenversuch überprüft, ob diese zu einer Verbesserung der Arbeitsgedächtnisleistung beitragen können. Außerdem wurde im Zuge des Hörversuchs anhand von Fragebögen evaluiert, ob ANC-Kopfhörer einen Einfluss auf das subjektive Empfinden haben.

2. Methode und Versuchsplanung

2.1. Apparatur und Stimuli

Für die Durchführung des Hörversuchs im Labor wurden binaurale Aufnahmen aus einem besetzten Großraumbüro benötigt. Dafür wurden zunächst anhand des Oldenburger Satztests generierte Sätze am Fraunhofer IBP in einem Freifeldraum von einem männlichen Sprecher nachhallfrei aufgezeichnet. Als Großraumbüro wurde ein typisches Mehrpersonenbüro verwendet in welchem sich acht Büroarbeitsplätze befanden. Bei der Decke handelte es sich um eine Betondecke, der Boden war mit Teppich ausgestattet. Es befanden sich keine Schirmwände zwischen den Arbeitsplätzen. Die Nachhallzeit im Raum erfüllte die Anforderungen der Raumakustikklasse C nach VDI 2569:2016-02 - Entwurf.

Im Großraumbüro wurden drei typische Positionen (Abstand zum Zuhörer 2.79 m, 5.97 m, 11.57 m) für die Sprecher

ISBN 978-3-9812830-9-9
Der Autor hält die Urheberrechte für den gesamten Inhalt seines Manuskripts. Er überlässt dem VDT das Recht auf Veröffentlichung in gedruckter Form, Online oder als PDF-Download von https://tonmeistertagung.com

gewählt. Die Wiedergabe der Sprachsignale wurde mit einem Studiolautsprecher (Yamaha MSP3) realisiert, welcher eine ähnliche Richtcharakteristik wie ein menschlicher Sprecher aufweist und für jede Position unterschiedliche Satzfolgen mit einem Sprachpegel von 59 dB(A) in 1 m Entfernung nach ANSI/ASA S3.5-1997 (R2017) wiedergab. Die Aufnahme erfolgte mit dem Kunstkopf HEAD acoustics HMS III. Die Aufnahme wurde ohne ANC-Kopfhörer, mit ausgeschalteten ANC-Kopfhörern und mit angeschalteten ANC-Kopfhörern durchgeführt.

Der Hörversuch wurde im HiPIE (High Performance Indoor Environment)-Labor des Fraunhofer-Instituts für Bauphysik durchgeführt. Das Experimentalprogramm wurde mittels der Software PsyScope X Experiment Control System (Version B57) erstellt. Die Schalle wurden über das Versuchsprogramm abgespielt. Die Kopfhörerausgabe erfolgte über das HEAD acoustics PEQ V Frontend, welches die Entzerrung der Kunstkopfaufnahmen, eine kalibrierte Kopfhörerentzerrung für die verwendeten Sennheiser HD 600 Kopfhörer und die kalibrierte Wiedergabe der Schalle sicherstellte. Die Versuchspersonen (Vpn) wurden einzeln getestet. Zum Test der kognitiven Leistungsfähigkeit wurde der Serial-Recall Test verwendet. Die Aufgabe der Probanden war es, sich neun verschiedene nacheinander in zufälliger Reihenfolge präsentierte Zahlen von 1–9 zu merken und nach einer Pause von acht Sekunden wieder richtig wiederzugeben. Während der Pause wurde der Hintergrundschall weiter abgespielt, wodurch der Einfluss des Hintergrundgeräuschs auf die kognitive Leistungsfähigkeit gut abgebildet werden konnte. Die Eingabe erfolgte durch Anklicken der Zahlen auf dem Bildschirm. Nach der Übung mit sechs Durchgängen absolvierten die Probanden 12 Durchgänge pro Hintergrundschallbedingung. Insgesamt bearbeiteten die Probanden 54 Durchgänge. Am Ende einer jeder Bedingung wurden die Probanden aufgefordert, Fragen bezüglich der Schalle zu beantworten. Die Fragebögen wurden mit dem Web-basierten Umfragetool „LimeSurvey“ erstellt und im Browser ausgeführt. Die Hintergrundschallbedingungen wurden in randomisierter Reihenfolge dargeboten.

2.2. Versuchsdesign und Probandenstichprobe

Es wurde ein einfaktorielles Versuchsdesign mit vier Faktorstufen (Ruhebedingung: *Ruhe*, Kopfhörer – mit ANC: *ANCan*, Kopfhörer – ohne ANC: *ANCaus*, Kopfhörer nicht aufgesetzt: *Keine KH*) und Messwiederholung (within-subjects-design) verwendet. Allen Vpn wurden die selben Bedingungen präsentiert – es gab keine Kontrollgruppe. Als unabhängige Variable wurde der Hintergrundschall in den vier Faktorstufen *Ruhe*, *ANCan*, *ANCaus*, *Keine KH* variiert. Die abhängige Variable war die Fehlerrate (Prozentsatz der falsch eingegebenen Zahlen) bei der Serial-Recall Aufgabe. Im Hörversuch wurden die Daten von N = 21 Vpn erhoben. Die Vpn waren größtenteils Studierende verschiedener Fachrichtungen. Die Vpn waren zwischen 21 und 69 Jahre alt (M = 28.52 Jahre, MD = 25 Jahre). Es nahmen 11 weibliche und 10 männliche Vpn am Hörversuch teil. Die Teilnahme wurde vergütet.

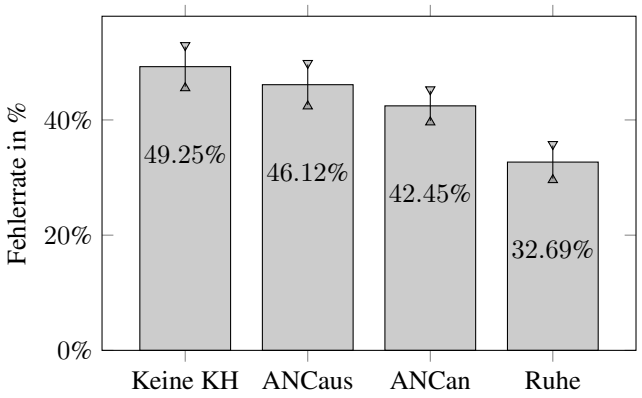


Abb. 1: Über alle VPN gemittelte Fehlerrate des Serial-Recalls. Die abgetragenen Fehlerbalken entsprechen der Standardabweichung.

3. Ergebnisse

Die Datenaufbereitung für den Serial-Recall-Test und die subjektiven Urteile erfolgte in Microsoft Excel 2016. Alle weiteren statistischen Analysen wurden danach mit der Software IBM SPSS Statistics (Version 19) durchgeführt. Als Signifikanzniveau wurde $\alpha = 0.05$ festgelegt. Es erfolgte eine Ausreißerkorrektur. Daten, die um mehr als die 2.5-fache Standardabweichung vom Mittelwert abwichen, wurden von der Auswertung ausgeschlossen.

3.1. Serial-Recall

Gemessen wurden die Fehlerraten (die falsch eingegebenen Zahlen) in den verschiedenen Bedingungen. Die einfaktorielle ANOVA (mit Messwiederholung) ergab einen signifikanten Haupteffekt des Faktors *Kopfhörer* auf die abhängige Variable *Fehlerrate (*$F(3, 60) = 8.64$, $p < .001$, $\eta^2 = .30$). Es zeigten sich beim Vergleich der Fehlerraten über t-Tests mit verbundener Stichprobe signifikante Unterschiede für die Vergleiche *Ruhe – Keine KH* ($t(20) = -4{,}01$, $p = .001$, $d = -0.88$), *Ruhe – ANCan* ($t(20) = -3.048$, $p = .006$, $d = -0.67$) und *Ruhe – ANCaus* ($t(20) = -2.858$, $p = .01$, $d = -0.62$). Die Vergleiche *Keine KH – ANCan* ($t(20) = 1.216$, $p = .238$, $d = 0.27$), *Keine KH – ANCaus* ($t(20) = 1.448$, $p = .163$, $d = 0.32$) sowie *ANCan – ANCaus* ($t(20) = -0.283$, $p = .780$, $d = -0.06$) ergaben keine signifikanten Ergebnisse. Die Mittelwerte sind in Abbildung 1 im Vergleich dargestellt. Durch die durchgeführten t-Tests konnte gezeigt werden, dass Kopfhörer – egal, ob mit an- oder ausgeschaltetem ANC – sich nicht signifikant auf die kognitive Leistungsfähigkeit auswirken.

3.2. Subjektive Urteile

3.2.1. Lästigkeit

Der Vergleich der Lästigkeit (siehe Abbildung 2) in den verschiedenen Hintergrundschallbedingungen über eine einfaktorielle ANOVA (mit Messwiederholung) ergab einen signifikanten Haupteffekt des Faktors *Hintergrundschall* auf die abhängige Variable *Lästigkeit* ($F(3, 60) = 132.24$, $p < .001$, $\eta^2 = .869$). Der Vergleich der wahrgenommenen Lästigkeit des Hintergrundschalls ohne Kopfhörer zu den Bedingungen mit Kopfhörer über paarweise t-Tests mit verbundener Stichprobe ergab signifikante Unterschiede zu *ANCan* ($t(20) = 4.176$, $p < .001$, $d = 0.91$) und zu *ANCaus* ($t(20) = 2.682$, $p =$

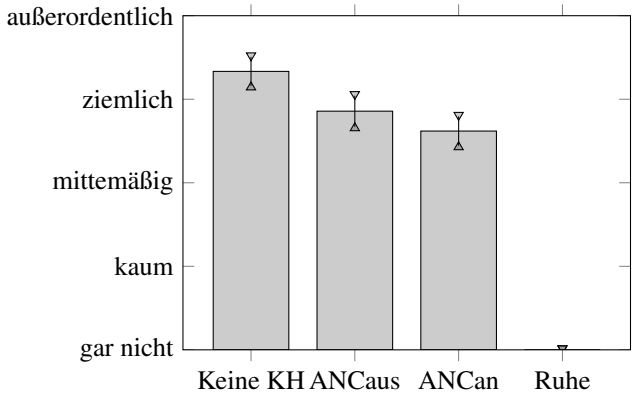


Abb. 2: Über alle Vpn gemittelter Wert der wahrgenommenen Sprecherlästigkeit. Die abgetragenen Fehlerbalken entsprechen der Standardabweichung.

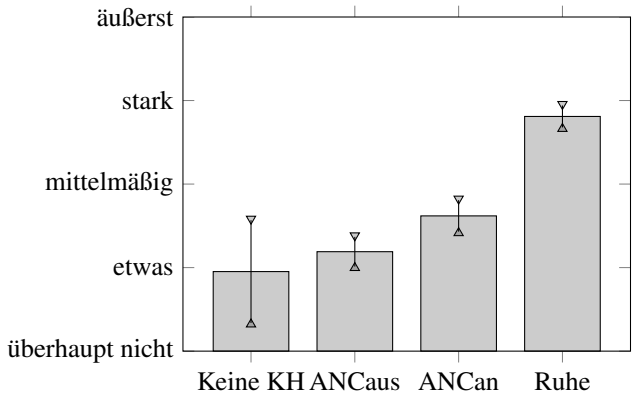


Abb. 3: Über alle Vpn gemittelter Wert des Konzentrationsvermögens. Die abgetragenen Fehlerbalken entsprechen der Standardabweichung.

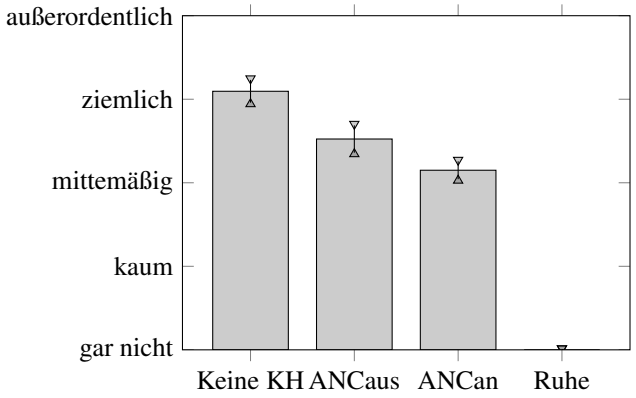


Abb. 4: Über alle Vpn gemittelter Wert der wahrgenommenen Sprecherlautheit. Die abgetragenen Fehlerbalken entsprechen der Standardabweichung.

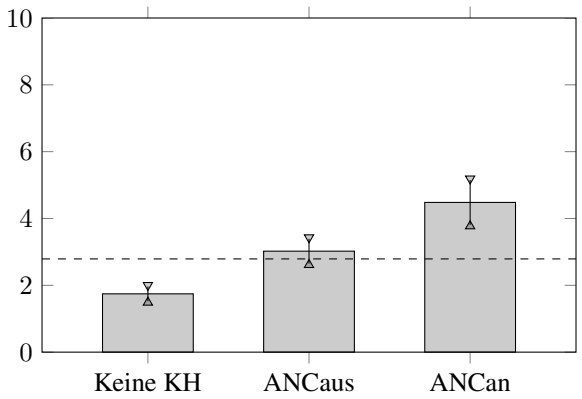


Abb. 5: Über alle Vpn gemittelter Wert des wahrgenommenen Abstands zum Sprecher. Die abgetragenen Fehlerbalken entsprechen der Standardabweichung. Die horizontale Linie zeigt den tatsächlichen Abstand des nächstgelegenden Sprechers.

.014, *d*= 0.59). Zwischen *ANCan – ANCaus* (*t*(20) = -1.096, *p* = .286, *d* = -0.24) konnten keine signifikanten Unterschiede festgestellt werden. Die Ruhebedingung wurde signifikant weniger lästig wahrgenommen.

3.2.2. Konzentrationsvermögen

Der Vergleich des Konzentrationsvermögens (siehe Abbildung 3) in den verschiedenen Hintergrundschallbedingungen über eine einfaktorielle ANOVA (mit Messwiederholung) ergab einen signifikanten Haupteffekt des Faktors *Hintergrundschall* auf die abhängige Variable *Konzentrationsvermögen* ($F(3, 60) = 37.977$, $p < .001$, $\eta^2 = .655$). Für die Vergleiche *Keine KH – ANCan* (*t*(20) = -3.005, *p* = .007, *d* = -0.66) sowie *ANCan – ANCaus* (*t*(20) = 2.121, *p* = .047, *d* = 0.46) zeigten sich signifikante Unterschiede im Konzentrationsvermögen.

3.2.3. Sprecherlautheit

Der Vergleich der Sprecherlautheit (siehe Abbildung 4) in den verschiedenen Hintergrundschallbedingungen über eine einfaktorielle ANOVA (mit Messwiederholung) ergab einen signifikanten Haupteffekt des Faktors *Hintergrundschall* auf die abhängige Variable *Lautheit* ($F(3, 57) = 163.369$, $p < .001$, $\eta^2 = .896$). Für die Vergleiche *Keine KH – ANCan* (*t*(19) = 7.285, $p < .001$, *d* = 1.63) sowie *Keine KH – ANCaus* (*t*(20) = 3.873, $p < .001$, *d* = 0.85) zeigten sich signifikante Unterschiede in der empfundenen Lautheit. Zwischen *ANCan – ANCaus* (*t*(19) = -1.926, *p* = .069, *d* = -0.43) konnten keine signifikanten Unterschiede festgestellt werden. Das Lautheitsempfinden in Ruhe war im Vergleich zu allen anderen Bedingungen signifikant geringer.

3.2.4. Abstand zum Sprecher

Der Vergleich des empfundenen Abstand des Sprechers (siehe Abbildung 5) in den verschiedenen Hintergrundschallbedingungen über eine einfaktorielle ANOVA (mit Messwiederholung) ergab einen signifikanten Haupteffekt des Faktors *Hintergrundschall* auf die abhängige Variable *Distanz* ($F(2, 38) = 11.047$, $p < .001$, $\eta^2 = .368$). In paarweisen t-Tests mit verbundener Stichprobe zeigten sich jeweils signifikante Unterschiede bezüglich der geschätzten Sprecherdistanz bei den Vergleichen der Bedingung *Keine KH* mit den Bedingungen *ANCan* (*t*(19) = -3.773, *p* = .001, *d* = -0.84) sowie *ANCaus* (*t*(19) = -4.325, $p < .001$, *d* = -0.97). Es zeigten sich ebenfalls signifikante Unterschiede bei den Vergleichen *ANCan – ANCaus* (*t*(19) = 2.179, *p* = .042, *d* = 0.49).

3.2.5. Langzeitstörung

Der Vergleich der Langzeitstörung (siehe Abbildung 6) in den verschiedenen Hintergrundschallbedingungen über eine einfaktorielle ANOVA (mit Messwiederholung) ergab einen signifikanten Haupteffekt des Faktors *Hintergrundschall* auf die abhängige Variable *Langzeitstörung* ($F(1.62, 30.76) = 65.573$, $p < .001$, $\eta^2 = .775$). In paarweisen t-Tests mit

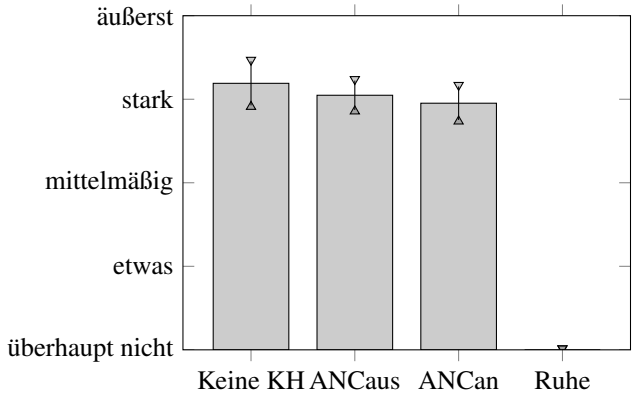


Abb. 6: Über alle Vpn gemittelter Wert der Langzeitstörung. Die abgetragenen Fehlerbalken entsprechen der Standardabweichung.

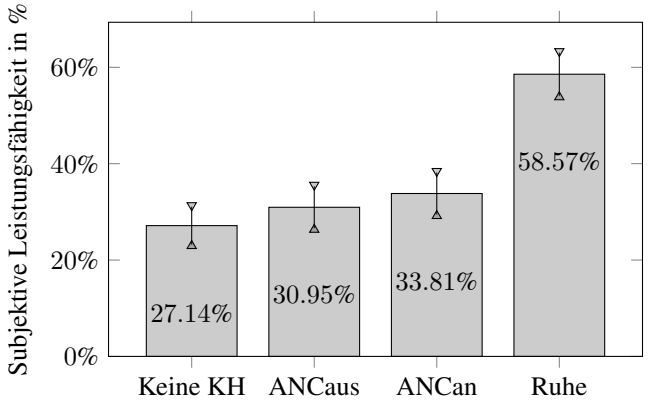


Abb. 7: Über alle Vpn gemittelter Wert der subjektiven Leistungsfähigkeit. Die abgetragenen Fehlerbalken entsprechen der Standardabweichung.

verbundener Stichprobe zeigten sich keine signifikanten Unterschiede bezüglich der Langzeitstörung bei den Vergleichen *Keine KH – ANCan* ($t(20) = 0.665$, $p = .514$, $d = 0.15$), *Keine KH – ANCaus* ($t(20) = 0.420$, $p = .679$, $d = 0.09$) und *ANCan - ANCaus* ($t(20) = -0.623$, $p = .54$, $d = -0.14$). Die Langzeitstörung in Ruhe war im Vergleich zu allen anderen Bedingungen signifikant geringer.

3.2.6. Subjektive Leistungsfähigkeit

Der Vergleich der subjektiv empfundenen Leistungsfähigkeit (siehe Abbildung 7) in den verschiedenen Hintergrundschallbedingungen über eine einfaktorielle ANOVA (mit Messwiederholung) ergab einen signifikanten Haupteffekt des Faktors *Hintergrundschall* auf die abhängige Variable *Leistungsfähigkeit* ($F(3, 60) = 30.520$, $p < .001$, $\eta^2 = .604$). In paarweisen t-Tests mit verbundener Stichprobe zeigten sich keine signifikanten Unterschiede bezüglich der empfundenen Leistungsfähigkeit bei den Vergleichen *Keine KH – ANCan* ($t(20) = -1.807$, $p = .086$, $d = -0.39$), *Keine KH – ANCaus* ($t(20) = -1.122$, $p = .275$, $d = -0.24$) und *ANCan – ANCaus* ($t(20) = 1.255$, $p = .224$, $d = 0.27$). Die subjektiv empfundene Leistungsfähigkeit in Ruhe war im Vergleich zu allen anderen Bedingungen signifikant besser.

4. Diskussion

Im Rahmen der Untersuchungen sollte überprüft werden, wie sich ANC-Kopfhörer auf die Arbeitsgedächtnisleistung unter Beeinträchtigung durch irrelevanten Sprachschall auswirken. Die Fehlerrate beim Serial-Recall in *Ruhe* war signifikant geringer, als unter anderen Bedingungen. Dies spricht für einen korrekten Versuchsaufbau- und ablauf. Im Gegensatz dazu fiel die Fehlerrate in den Bedingungen mit Kopfhörern nicht signifikant geringer aus, als bei der Bedingung *Keine KH*. Weder im Vergleich der Bedingung *Keine KH* mit *ANCaus*, noch mit *ANCan* fiel der Unterschied signifikant aus. Auch zwischen *ANCan* und *ANCaus* gab es keinen signifikanten Unterschied. Aus dem Ergebnis lässt sich schlussfolgern, dass ANC-Kopfhörer keinen Einfluss auf die Arbeitsgedächtnisleistung haben. Ein möglicher Grund dafür könnte sein, dass der Sprachpegel – solange er laut genug ist – für den Irrelevant Sound Effect nicht relevant ist, sondern ausschließlich die Sprachverständlichkeit [14]. Die Sprachverständlichkeit kann durch ANC-Kopfhörer sogar steigen, da Grundrauschen bis 600 Hz sehr gut reduziert wird, Sprache jedoch durch ihre Transienten weniger gut durch den aktiven Algorithmus ausgeblendet wird. Obwohl im Hörversuch keine Verbesserung der Arbeitsgedächtnisleistung festgestellt werden konnte, scheinen ANC-Kopfhörer das subjektive Empfinden der Probanden zu verbessern. Den störenden Sprachschall in der Bedingung *ANCaus* stuften die Probanden als signifikant weniger lästig ein als in *Keine KH*, mit aktiviertem ANC (*ANCan*) verringerte sich die Lästigkeit abermals signifikant. Auch das Konzentrationsvermögen in der Bedingung *ANCan* wurde als signifikant besser eingeschätzt, als in *ANCaus* oder *Keine KH*. In Bezug auf die subjektiv empfundene Privatheit können ANC-Kopfhörer ebenfalls von Nutzen sein. Die Probanden stuften die Sprecherlautheit mit Kopfhörern als geringer ein, was vor allem auf die Dämpfungswirkung der Kopfhörer zurückzuführen ist. Allerdings gab es hier keinen signifikanten Unterschied zwischen *ANCan* und *ANCaus*. Dies könnte darauf zurückzuführen sein, dass sich viele wichtige Sprachbestandteile im Frequenzbereich über dem durch ANC beeinflussten Bereich befinden. Der signifikante Unterschied zwischen dem wahrgenommenen Abstand des Sprechers (in der Bedingung *ANCan* wurde dieser als signifikant weiter entfernt wahrgenommen, als bei *ANCaus*) deutet trotzdem darauf hin, dass durch aktive Rauschunterdrückung eine Verbesserung zu verzeichnen ist. Die Ergebnisse der Befragung bezüglich Langzeitstörung und subjektiver Leistungsfähigkeit decken sich mit den Ergebnissen des Serial-Recalls. Hier konnten keine signifikanten Unterschiede zwischen *ANCan* und *ANCaus* sowie *Keine KH* festgestellt werden. Die Untersuchung zeigt, dass ANC nur in bestimmten Bereichen zusätzlich eine Verbesserung zur reinen passiven Schalldämmung in Bezug auf die subjektive Wahrnehmung bringt. Eine signifikante Verschlechterung durch ANC konnte jedoch in keinem Bereich festgestellt werden.

5. Schlussfolgerungen

Zusammenfassend kann festgehalten werden, dass ANC-Kopfhörer keinen signifikanten Einfluss auf die kognitive Leistungsfähigkeit bei der Arbeit in einem Großraumbüro

haben. Dies wird von den Probanden auch subjektiv so wahrgenommen, es wurden keine signifikanten Unterschiede bei der Befragung bezüglich subjektiver Leistungsfähigkeit gefunden. Für ANC-Kopfhörer kann es in Großraumbüros jedoch trotzdem sinnvolle Einsatzzwecke geben, da sie die Lästigkeit von Gesprächen anderer Mitarbeiter reduzieren und die wahrgenommene Privatheit steigern. Diese beiden Faktoren haben sich in vielen am Fraunhofer IBP durchgeführten Bürobefragungen als Hauptstörfaktoren erwiesen. In zusätzlichen Probandenstudien könnte eine umfangreichere Untersuchung der Auswirkungen von ANC-Kopfhörern auf die kognitive Leistungsfähigkeit durch andere Tests als den in dieser Arbeit verwendeten Serial-Recall durchgeführt werden.

6. Literatur

[1] S. Stadler, "Open Space Büros: Eine Studie über die Machbarkeit und Umsetzung von offenen Bürostrukturen", 2011.

[2] *bso-Studie 2015, Status quo der Büro-Arbeitsplätze in Deutschland*, 2015.

[3] *VDI 2569:2016-02 - Entwurf, Schallschutz und akustische Gestaltung im Büro*, Verein Deutscher Ingenieure, 2016.

[4] N. Venetjoki, A. Kaarlela-Tuomaala, E. Keskinen und V. Hongisto, "The effect of speech and speech intelligibility on task performance.", eng, *Ergonomics*, Bd. 49, Nr. 11, S. 1068–1091, 2006, ISSN: 0014-0139 (Print); 0014-0139 (Linking). DOI: 10.1080/00140130600679142.

[5] H. Colle, "Auditory encoding in visual short-term recall: effects of noise intensity and spatial location", *Journal of Verbal Learning and Verbal Behavior*, Bd. 19, Nr. 6, S. 722–735, 1980, ISSN: 0022-5371. DOI: 10.1016/S0022-5371(80)90403-X.

[6] P. Salamé und A. Baddeley, "Disruption of short-term memory by unattended speech: Implications for the structure of working memory", *Journal of Verbal Learning and Verbal Behavior*, Bd. 21, Nr. 2, S. 150–164, 1982, ISSN: 0022-5371. DOI: 10.1016/S0022-5371(82)90521-7.

[7] R. C. Martin, M. S. Wogalter und J. G. Forlano, "Reading comprehension in the presence of unattended speech and music", *Journal of Memory and Language*, Bd. 27, Nr. 4, S. 382–398, 1988, ISSN: 0749-596X. DOI: 10.1016/0749-596X(88)90063-0.

[8] N. D. Weinstein, "Noise and intellectual performance: A confirmation and extension", *The Journal of applied psychology*, Bd. 62, S. 104–111, 1977. DOI: 10.1037/0021-9010.62.1.104.

[9] D. M. Jones, C. Miles und J. Page, "Disruption of proofreading by irrelevant speech: Effects of attention, arousal or memory?", *Applied Cognitive Psychology*, Bd. 4, Nr. 2, S. 89–108, 1990. DOI: 10.1002/acp.2350040203.

[10] S. Banbury und D. C. Berry, "Habituation and Dishabituation to Speech and Office Noise", *Journal of Experimental Psychology Applied*, Bd. 3, 1997. DOI: 10.1037/1076-898X.3.3.181.

[11] S. Banbury und D. C. Berry, "Disruption of office-related tasks by speech and office noise", *British Journal of Psychology*, Bd. 89, S. 499–517, 1998. DOI: 10.1111/j.2044-8295.1998.tb02699.x.

[12] N. D. Weinstein, "Effect of noise on intellectual performance", *Journal of Applied Psychology*, Bd. 59, S. 548–554, 1974. DOI: 10.1037/h0037338.

[13] J. L. Loewen und P. Suedfeld, "Cognitive and Arousal Effects of Masking Office Noise", *Environment and Behavior*, Bd. 24, S. 381–395, 1992. DOI: 10.1177/0013916592243006.

[14] S. Banbury, W. J. Macken, S. Tremblay und D. M. Jones, "Auditory Distraction and Short-Term Memory: Phenomena and Practical Implications", *Human Factors*, Bd. 43, Nr. 1, S. 12–29, 2001. DOI: 10.1518/001872001775992462.

Evaluation der Wiedergabequalität von Soundbars mithilfe der MS-IPM Methodik

Theresa Liebl[1]

[1] *Institut für Rundfunktechnik GmbH, München, Germany, Email: liebl@irt.de*

Abstract

In 2016 wurde eine neue subjektive Versuchsmethodik zur Beurteilung von neuartigen Audio-Systemen, die ”Multiple Stimulus Ideal Profile Method” (MS-IPM) vorgestellt. Mit dieser Methodik können mehrere Wiedergabesysteme direkt miteinander verglichen werden, auch wenn keine explizite Referenz vorhanden ist. Evaluiert werden die Systeme in Bezug auf ihre subjektive Gesamtqualität und die Qualität verschiedener zur Fragestellung passender Attribute. Zusätzlich gibt jeder Proband für jedes Attribut einen Idealwert an, der nach eigener Ansicht einem hypothetischen idealen System entspricht.

Da das IRT an der Standardisierung solcher Methoden beteiligt ist, wurden die Besonderheiten der MS-IPM genauer untersucht. Dafür wurde ein Hörversuch entworfen, in dem die Wiedergabequalität von Soundbars beurteilt werden sollte.

Soundbars erfreuen sich seit einigen Jahren im Home-Entertainmentbereich steigender Beliebtheit. Diese kompakten Wiedergabegeräte, die mehrere Lautsprecher in einem Gehäuse vereinen, werden häufig damit beworben, dem Konsumenten echten Surround-Sound bieten zu können. Es sollte daher untersucht werden, ob Soundbars bei der Heimanwendung eine brauchbare Alternative zu einem Surround-System darstellen und die Audiowiedergabequalität von TV-Geräten verbessern können.

In mehreren Hörversuchen wurde mithilfe der MS-IPM die Wiedergabequalität von acht Soundbars, einem TV-Gerät und einem diskreten Mehrkanalsystem verglichen. Die dabei gewonnenen Erkenntnisse über die untersuchten Soundbars und die neue Versuchsmethodik werden im Folgenden beide dargestellt.

1. Einführung

Für spezielle Fragestellungen und neuartige Wiedergabesysteme stoßen etablierte subjektive Hörversuchsmethoden oft an ihre Grenzen, was die Weiter- und Neuentwicklung dieser Methoden sehr wichtig macht. In 2016 wurde eine neue Versuchsmethodik zur Beurteilung von Audio-Systemen, die ”Multiple Stimulus Ideal Profile Method” (MS-IPM) vorgestellt [1]. Für die Standardisierung neuer Methoden bedarf es ausführlicher Tests und gegebenenfalls Anpassungen der Entwürfe. Am IRT wurde daher eine Untersuchung zum Testen der MS-IPM durchgeführt. In dieser Untersuchung sollte die Wiedergabequalität von Soundbars im Vergleich zu herkömmlichen Wiedergabegeräten in einer Wohnzimmerumgebung untersucht werden.

Sowohl die Versuchsmethode als auch die Ergebnisse der Soundbaruntersuchung werden im Folgenden dargestellt. Eine detaillierte Darstellung der Versuchsergebnisse kann auch in [2] nachgelesen werden.

2. MS-IPM

Die ”Multiple Stimulus Ideal Profile Method” (MS-IPM) wurde entwickelt, um verschiedene Systeme miteinander zu vergleichen, auch wenn keine explizite Referenz gegeben ist. Die Methodik liefert Vergleichswerte zur Gesamtwiedergabequalität der untersuchten Systeme sowie zusätzliche Information über die Eigenschaften der Systeme durch die Verwendung verschiedener Attribute.

Ähnlich wie die MUSHRA Methode [3] nutzt MS-IPM die Darbietung mehrerer Stimuli gleichzeitig, um die zu untersuchenden Systeme miteinander zu vergleichen. Sowohl Gesamtwiedergabequalität als auch die Attribute werden auf einer 100-stufigen Skala bewertet. Relevante Attribute, um die Unterschiede zwischen den Systemen bestmöglich zu beschreiben werden vor dem Test aus etablierten Lexika ausgewählt. Zusätzlich dazu versucht die Methode zu ermitteln wie gut die untersuchten Systeme im Vergleich zu einem hypothetischen idealen System abschneiden. Dafür sollen die Probanden bei jedem Attribut einen Idealwert angeben der für dieses hypothetische ideale System steht, welches ihren Wünschen entspricht und ihrer Erfahrung nach im untersuchten Kontext das beste System wäre. Abhängig von den Eigenschaften der untersuchten Systeme und den Attributbewertungen kann der Idealwert von den Bewertungen der Systeme abweichen. Es kann nicht davon ausgegangen werden, dass dieser Idealwert dem am besten bewerteten System entspricht. Er kann sowohl über, als auch unter den Bewertungen dieses Systems liegen.

Die Kombination aus Bewertung der Gesamtwiedergabequalität und der Qualität der Systeme anhand verschiedener Attribute ermöglicht eine tiefer greifende Analyse und bessere Interpretation der Ergebnisse.

ISBN
978-3-9812830-9-9

Der Autor hält die Urheberrechte für den gesamten Inhalt seines Manuskripts. Er überlässt dem VDT das Recht auf Veröffentlichung in gedruckter Form, online oder als PDF-Download von https://tonmeistertagung.com

3. Versuchskonzept

Um Erfahrungen mit MS-IPM zu sammeln, wurde ein Hörversuch durchgeführt, bei dem die Wiedergabequalität von handelsüblichen Soundbars im Vergleich zu einem Fernsehgerät und einer 5.1-Lautsprecheraufstellung untersucht werden sollte. Ziel dieser Untersuchung war es sowohl die Wiedergabequalität dieser Systeme, als auch die Methodik an sich zu evaluieren.

Mit dem Hörversuch sollte evaluiert werden, ob Soundbars eine gute Alternative für eine 5.1-Lautsprecheraufstellung in einer Wohnzimmerumgebung darstellen und ob sie die Wiedergabequalität im Vergleich zu einem Fernsehgerät deutlich verbessern können. Die Unterschiede der Wiedergabesysteme wurden in zwei getrennten Hörversuchen für eine Stereo- und eine 5.1-Wiedergabe untersucht. Als Testmaterial wurde original TV-Material aus den Genres Sport, Dokumentation, TV-Show, Spielfilm und Musik verwendet.

Es wurden acht Soundbars aus unterschiedlichen Preisklassen und von verschiedenen Herstellern ausgewählt, welche auf dem Markt zum Zeitpunkt der Untersuchung zahlreich vertreten waren. Um die Wiedergabequalität dieser Systeme zu untersuchen, wurden sie mit einem Standard TV-Gerät mit integrierten Lautsprechern und einem Hochqualitativen 5.1-Setup verglichen.

MS-IPM versprach für diese Untersuchung die richtige Methode zu ein, da keine echte Referenz gegeben war. Es wäre reine Vermutung gewesen das 5.1-System als bestes System anzunehmen und somit als Referenz zu verwenden. Vor den Untersuchungen war nicht auszuschließen, dass eine Soundbar unter den gegebenen Bedingungen besser bewertet werden würde. Ein Standard A/B-Vergleich, wie er in ITU-R BS.1284-1 [4] beschrieben ist wäre auch eine passende Methodik gewesen die Systeme miteinander zu vergleichen, diese Methode wäre allerdings deutlich zeitaufwändiger gewesen als die MS-IPM. Außerdem versprach die Kombination aus Bewertung der Gesamtwiedergabequalität und verschiedener Attribute eine deutlich detailliertere Beschreibung der Systeme und ihrer Unterschiede.

4. Testvorbereitung

4.1. Testmaterial

Die Unterschiede zwischen den Systemen sollten für den Anwendungsfall des normalen Fernsehzuschauers zu Hause untersucht werden. Daher wurden typische Fernsehinhalte in Stereo und 5.1 als Testmaterial verwendet, um realistische Bedingungen für die Untersuchung zu schaffen. Ein kurzer Ausschnitt in Stereo und 5.1 wurde aus den Genres Sport, Dokumentation, TV-Show, Spielfilm und Musik ausgesucht. Bei Musik wurde dabei noch zwischen Klassik und Pop unterschieden. Insgesamt wurden somit 12 Ausschnitte mit einer Länge von 10 bis 20 Sekunden ausgewählt.

4.2. Versuchsaufbau

Für den Hörversuch wurde ein Raum mit Absorbern akustisch optimiert, um einer typischen Wohnzimmerakustik zu entsprechen. Das 5.1-Lautsprechersystem wurde nach den Anforderungen der ITU-R BS.775-3 [5] aufgestellt und die Abhörposition der Probanden während des Hörversuchs lag im Sweet Spot dieser Aufstellung. Das Fernsehgerät wurde auf einem Regal platziert welches speziell für die Untersuchung angefertigt wurde (siehe Abb. 1). In diesem Regal konnten neben dem TV-Gerät noch der Center-Lautsprecher des 5.1-Systems und vier Soundbars gleichzeitig untergebracht werden. Die Soundbars wurden über und unter dem Center angeordnet und die Position der einzelnen Geräte wurde für jeden Probanden variiert. Dadurch konnten alle acht Soundbas untersucht werden und jede Soundbar wurde mindestens einmal an jeder Position im Regal getestet. Dadurch sollte ein möglicher Einfluss der Soundbar-Position im Regal auf die Ergebnisse des Versuchs minimiert werden. Bei der Positionierung der Soundbars und der Absorber im Versuchsraum wurde darauf geachtet die individuellen Anforderungen der einzelnen Geräte zu berücksichtigen. Es wurden also notwendige Reflexionspunkte an den Wänden freigehalten und die Entfernungsangaben der Geräte zum Abhörort eingehalten. Während des Versuchs wurde das Regal mit den Wiedergabegeräten mit einem akustisch transparenten Stoff verdeckt, um einem möglichen optischen Einfluss bei der Bewertung der Systeme auszuschließen.

Abb. 1: Hörversuchsraum

4.3. Versuchssoftware

Die Versuchsmethodik wurde in einer webbasierten Versuchsumgebung des IRT implementiert. Darüber konnten die Probanden die Testsequenzen starten, loopen und ihre Bewertungen abgeben. Über eine MIDI-gesteuerte optische Matrix wurde ein nahtloses Umschalten zwischen den Wiedergabegeräten ermöglicht.

Die grafische Benutzeroberfläche der Versuchssoftware ist in Abb. 2 und Abb. 3 zu sehen.

4.4. Attributauswahl

Die nötige Auswahl relevanter Attribute für die MS-IPM-Untersuchung wurde nach den entsprechenden Empfehlungen [1] durchgeführt. Vier Experten trafen sich zu einem Probehören im Versuchsraum. Sie hatten dabei Zeit sich mit den Wiedergabesystemen und dem ausgewählten Testmaterial vertraut zu machen. Aus etablierten Lexika [6], welche eine Auswahl an möglichen Attributen boten, wurden für die Untersuchung relevante Attribute ausgewählt. Dabei wurde darauf geachtet Attribute zu wählen, welche die auffälligsten Unterschiede und die Eigenschaften der Wiedergabesysteme am besten beschreiben. Anschließend wurde die Auswahl diskutiert und fünf zu evaluierende Attribute für den Hörversuch ausgewählt (siehe Tab. 1). Die Attribute wurden in Deutsch übersetzt und die genauen Beschreibungen sowohl in die Versuchssoftware integriert als auch in ausgedruckter Form für die Probanden zur Verfügung gestellt.

Attribut	Beschreibung	Skala
Umhüllung *(für 5.1)*	Bist du vom wiedergegebenen Klang umgeben und gibt er dir ein Gefühl von Räumlichkeit?	komplett umhüllt / nicht umhüllt
Breite *(für Stereo)*	Die Breite des Klangbildes (der wahrgenommene Winkel). Die Breite eines raumbezogenen Nachhalls soll nicht in die Bewertung miteinfließen.	breit / schmal
Natürlichkeit	Geräusche, die mit einer hohen Klangtreue wiedergegeben werden. Akustische Instrumente, Stimmen und Geräusche klingen wie in der Realität. Das Geräusch entspricht der Erwartung des originalen Geräusches ohne Klangfärbung, räumliche Färbung oder Verzerrung. „nichts hinzugefügt – nichts fehlt.“ Die Klangquelle ist klar im Raum einzuordnen. Der Klang kommt nahe an die Erfahrung der Wahrnehmung des originalen Klangs heran.	natürlich / unnatürlich
Detailliertheit	Ein hochaufgelöster Klang, reich an Details. Instrumente, Stimmen, usw. können einfach auseinandergehalten werden. Die Musik hat viele Details. Details, die nicht gemessen werden können. Details, die der Musik „Seele“ geben. Es kann sich um kleine hörbare Feinheiten handeln: Das Atmen eines Sängers, Finger wandern über Gitarrensaiten, die Klappen einer Klarinette, das Geräusch von Klavierhämmern bei Saitenanschlag.	detailliert / nicht detailliert
Dosiger Klang (Canny)	Das Hörbeispiel klingt als würde es in einer Dose oder einer Röhre abgespielt werden. Der Klang ist charakterisiert durch hervorstechende und schmalbandige Resonanzen im Mittenbereich.	dosig / nicht dosig
Bassstärke	Der relative Pegel des Bassbereichs, d.h. der tiefen Frequenzen. Beispiele: männliche Stimmen, Bassgitarre, Bassdrum, Timpani und Tuba. Sollte nicht mit Basstiefe verwechselt werden, welche die Ausdehnung der tiefen Bassfrequenzen bezeichnet.	stark / schwach

Tab. 1: Attributauswahl für dem Soundbar-Hörversuch

5. Versuchsdurchführung

Der Abhörpegel wurde für alle Probanden und Versuchsteile auf einen festen Wert von 64 dB(A) eingestellt. Teilgenommen haben an beiden Versuchsteilen je 24 Probanden mit Hörversuchserfahrung. Beide Teile wurden in zwei Sessions mit ca. 45 Minuten Länge aufgeteilt, um Ermüdungseffekte zu vermeiden. Die Teilnehmer konnten den Versuch jederzeit unterbrechen und zu einem späteren Zeitpunkt fortsetzen.

Der Versuch wurde in folgender Reihenfolge durchgeführt:

- Einführung der Probanden
- Training zur Gesamtwiedergabequalität
- Bewertung der Gesamtwiedergabequalität
- Training zur Bewertung der Attribute und des Idealwerts
- Bewertung der Attribute und des Idealwerts

Im ersten Schritt fand eine Einführung der Probanden mit einer schriftlichen Anleitung und mündlichen Hinweisen statt. Dabei wurde der grundsätzliche Testablauf erklärt und die Aufgabe, sowie die Methodik genau beschrieben. Im zweiten Schritt hatten die Probanden Zeit sich alle Testbeispiele und Systeme genau anzuhören, um mit dem Test-set und der Bedienung der Software für die Bewertung der Gesamtwiedergabequalität vertraut zu werden (siehe Abb. 2). Im Anschluss daran konnten noch offene Fragen geklärt werden.

Die Bewertung der Gesamtwiedergabequalität im dritten Schritt wurde für alle Systeme gleichzeitig in einem Multi-Stimulus-Vergleich vorgenommen. Jedes Trial bestand aus einem Testbeispiel anhand dessen alle Systeme bewertet wurden. Die Reihenfolge der Beispiele und Systeme war für jeden Probanden zufällig.

Nach der Bewertung der Gesamtwiedergabequalität hatten die Probanden noch einmal Zeit sich mit der Bewertung der einzelnen Attribute und der Wahl des Idealwerts vertraut zu machen. Der Regler für die Idealwert-Bewertung wurde in der Testsoftware in Gelb hervorgehoben, daneben wurde eine genaue Beschreibung des zu bewertenden Attributes angezeigt (siehe Abb. 3).

Nach dem Training folgte die Bewertung der ausgewählten Attribute für jedes Testbeispiel und alle Systeme. Für jedes Attribut war die Reihenfolge der Beispiele und Systeme zufällig. Die Probanden bewerteten alle sechs Beispiele für ein Attribut bevor mit das nächsten abgefragt wurde.

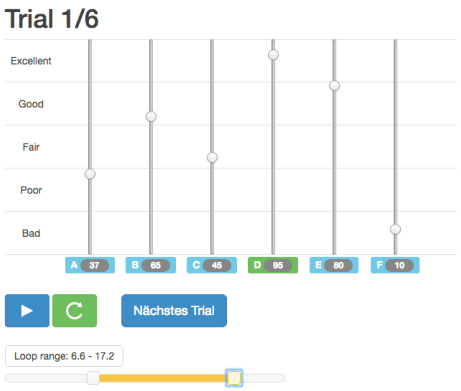


Abb. 2: Benutzeroberfläche der Gesamtwiedergabequalitätsbewertung

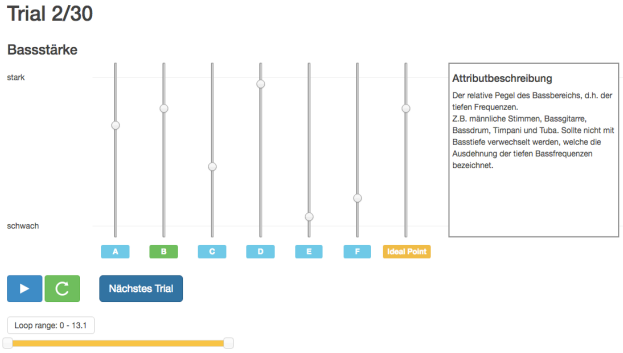


Abb. 3: Benutzeroberfläche der Attributbewertung

6. Ergebnisse

6.1. Gesamtwiedergabequalität

Abb. 4 und Abb. 5 zeigen die Mittelwerte mit 95% Konfidenzintervall der Bewertungen der Gesamtwiedergabequalität für die 5.1- und Stereo-Wiedergabe, gemittelt über alle Beispiele und Probanden.

Es ist deutlich zu erkennen, dass sowohl für 5.1- als auch für Stereo-Wiedergabe das TV-Gerät deutlich am schlechtesten bewertet wurde. Das 5.1-System erreicht die besten Bewertungen, allerdings nicht mit einem signifikanten Unterschied zu einer der Soundbars (SB2), welche von den acht untersuchten Soundbars am besten bewertet wurde. Vor allem bei 5.1-Wiedergabe konnte sich diese deutlich vom Rest der Soundbars abheben. Die Bewertungen der restlichen sieben Soundbars bewegen sich hauptsächlich im mittleren Bereich der Bewertungsskala.

6.2. Attributbewertung

Für einen detaillierteren Vergleich der Systeme wird die Auswertung der Attributbewertungen und die Wahl des Idealwertes herangezogen. Die Bewertungen des Idealwerts wurden für jedes Attribut über alle Samples und Probanden gemittelt. Dadurch entsteht ein ideales Profil welches ein ideales System darstellt, das von den Probanden erstellt wurde.

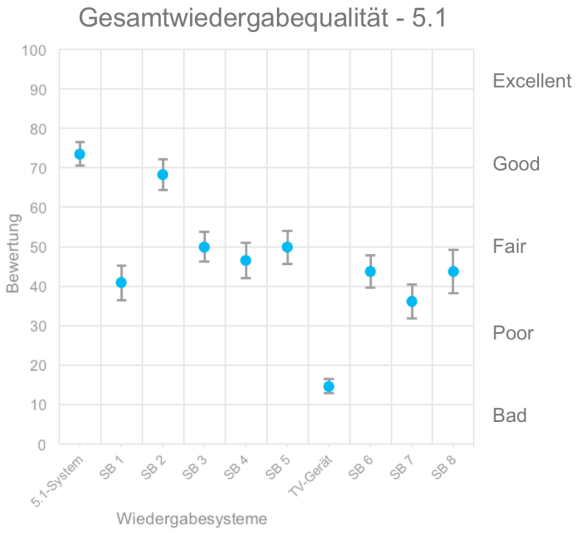


Abb. 4: Bewertung der Gesamtwiedergabequalität bei 5.1-Wiedergabe, gemittelt über alle 24 Probanden und alle Beispiele

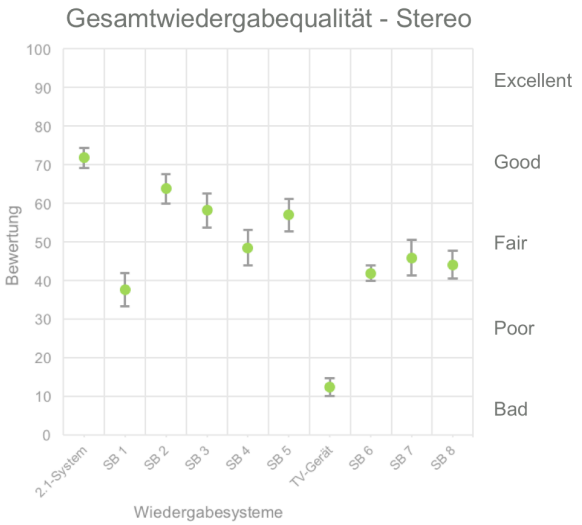


Abb. 5: Bewertung der Gesamtwiedergabequalität bei Stereo-Wiedergabe, gemittelt über alle 24 Probanden und alle Beispiele

Eine Zusammenstellung der Ergebnisse der Attributbewertung für 5.1-Wiedergabe ist in Abb. 6 zu sehen. In diesem Spinnennetz-Diagramm sind die Ergebnisse für die fünf Attribute und die zehn Wiedergabegeräte, gemittelt über alle Beispiele und Probanden, sowie das ideale Profil gemeinsam dargestellt.

Diese Zusammenstellung zeigt die Qualitäten und Unterschiede der Systeme auf einen Blick. Dem Fernsehgerät fehlt es demnach nicht nur an Detailliertheit und Natürlichkeit, sondern auch an umhüllenden Eigenschaften. Es scheint zudem einen ausgeprägten dosigen Klang und nahezu keine Bassstärke zu besitzen. Das 5.1-System und die SB 2 kommen dem idealen Profil (in Gelb dargestellt) am nächsten. Die beiden können sich bei fast allen Attributen vom Rest der Soundbars im Test abheben. Nur bei der Bassstärke können mehrere Systeme das Ideal erreichen.

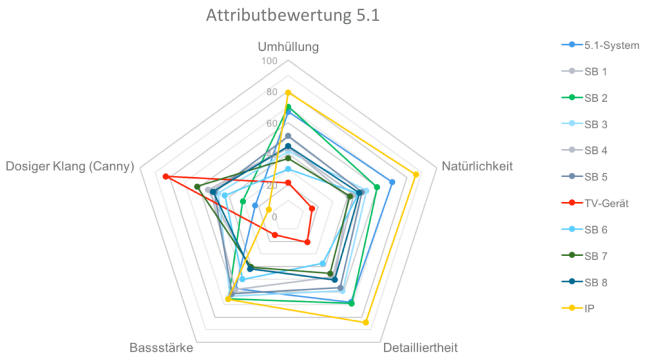


Abb. 6: Ergebnis der Attributbewertung und Wahl des Idealwertes bei 5.1-Wiedergabe, gemittelt über alle 24 Probanden und alle Beispiele

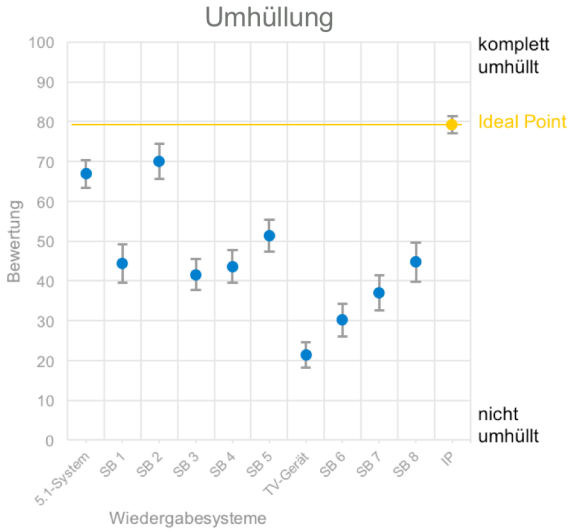


Abb. 7: Ergebnis der Attributbewertung und Wahl des Idealwertes für das Attribut Umhüllung bei 5.1-Wiedergabe, gemittelt über alle 24 Probanden und alle Beispiele

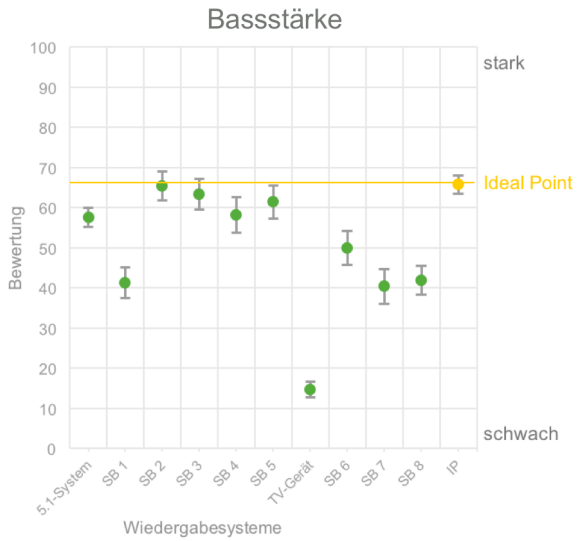


Abb. 8: Ergebnis der Attributbewertung und Wahl des Idealwertes für das Attribut Bassstärke bei 5.1-Wiedergabe, gemittelt über alle 24 Probanden und alle Beispiele

Abb. 7 und Abb. 8 zeigen detailliertere Ergebnisse der Attributbewertung beispielhaft für die beiden Attribute Umhüllung und Bassstärke.

In Abb. 7 sind die Attributbewertungen und der Idealwert für Umhüllung dargestellt. Es ist deutlich zu sehen, dass keines der untersuchten Systeme das Ideal erreicht, welches durch die gelbe Linie gekennzeichnet ist. Das 5.1-System und die SB 2 kommen der Linie sehr nahe, aber die restlichen Systeme sind weit von diesem, von den Probenden gewünschten, idealen System entfernt. In Abb. 8 sind der Idealwert und die Bewertungen für das Attribut Bassstärke zu sehen. Der gemittelte Idealwert liegt hier bei einem Wert von 66 auf der Bewertungsskala, was deutlich tiefer ist als bei Umhüllung, wo er bei 79 liegt. Dies zeigt, dass ein ideales System nicht zwangsläufig an hohen Bewertungen festzumachen ist. Auch Bewertungen im mittleren Bereich können optimal sein, wenn für die Probanden das ideale System in diesem Bereich liegt. Vier der untersuchen Soundbars erreichen für das Attribut Bassstärke den Idealwert. Überraschend an den Ergebnissen ist, dass das 5.1-System den Idealwert nicht erreicht.

Genauere Auswertungen aller Attribute und eine detailliertere Beschreibung können in [2] nachgelesen werden.

7. Fazit

Die Untersuchungsergebnisse zeigen, dass Soundbars in einer Wohnzimmerumgebung eindeutig die Audiowiedergabequalität eines herkömmlichen TV-Gerätes verbessern können. Sowohl für Stereo, als auch für 5.1-Wiedergabe liefern die Soundbars eine deutlich bessere Qualität. Es konnte auch gezeigt werden, dass eine Soundbar in Wohnzimmer durchaus auch eine Alternative zu einem disktreten Mehrkanalsystem darstellen kann.

Es konnte bestätigt werden, dass sich die MS-IPM für diese Art von Untersuchung sehr gut eignet. Die Attributbewertungen mit Wahl des Idealwertes ermöglichten einen besseren Einblick in die Qualitäten der Systeme und die Erwartungen der Probanden an diese Systeme. Die Bewertung ohne Referenz und auch die Wahl des Idealwertes waren für die Probanden eine Herausforderung und erforderten etwas Übung. Die Unterschiede zwischen den Systemen und zu den Erwartungen der Probanden konnten dadurch aber besserer eingeordnet werden.

Die Auswahl relevanter Attribute für die jeweils gegebene Untersuchung ist ein wichtiges Element in der Vorbereitung, um am Ende aussagekräftige Ergebnisse zu erhalten. Dafür sollte genug Zeit eingeplant und mit Bedacht ausgewählt werden. Außerdem sind das Training und eine gute Einführung der Probanden entscheidend, vor allem wenn die Probanden mit dieser Versuchsmethodik noch nicht vertraut sind.

Literatur

[1] Zacharov, N., Pike, C., Melchior, F., and Worch, T., “Next Generation Audio System Assessment using the Multiple Stimulus Ideal Profile Method”. Proceedings of QoMEX 2016, Lisbon, Portugal, 2016

[2] Wakan, S., “Untersuchungen zur Wiedergabequalität von Soundbars mithilfe der Evaluationsmethodik MS-IPM“. Bachelor Thesis, Hochschule der Medien Stuttgart, 2017

[3] Recommendation ITU-R BS.1534-3, “Method for the subjective assessment of intermediate quality levels of coding systems”. ITU, Geneva, Switzerland, 2015

[4] Recommendation ITU-R BS.1284-1, “General methods for the subjective assessment of sound quality”. ITU, Geneva, Switzerland, 2003

[5] Recommendation ITU-R BS.775-3, “Multichannel stereophonic sound system with and without accompanying picture”. ITU, Geneva, Switzerland, 2012

[6] Pedersen, T. H., and Zacharov, N., “The Development of a Sound Wheel for Reproduced Sound.” In 138th Convention of the Audio Engineering Society, 2015

Verfahren zur Verbesserung der Bass-Wiedergabe mit optimiertem Zeitverhalten

O. Kreuer[1], D. Leckschat[1], C. Epe[1]
[1] *Hochschule Düsseldorf, Germany, Email:*
oliver.kreuer@study.hs-duesseldorf.de
dieter.leckschat@hs-duesseldorf.de

Abstract

Vorgestellt wird ein neues Verfahren zur verbesserten Wiedergabe tiefer Frequenzen bei Lautsprechersystemen. Geschlossene Systeme zeichnen sich gegenüber Bassreflexsystemen durch ein besseres Impulsverhalten und eine präzisere Wiedergabe aus. Demgegenüber steht eine tiefer liegende Grenzfrequenz des Reflexsystems bei gleichbleibendem Gehäusevolumen durch die Verwendung eines speziell abgestimmten Resonators. Das neue Verfahren vereint die Vorzüge beider Systeme. Es ermöglicht eine tiefreichende Basswiedergabe bei verhältnismäßig geringem Gehäuse-Volumen und bietet zudem ein optimiertes Zeitverhalten.

1. Einleitung

Bei Lautsprechersystemen strebt man seit Anbeginn des Lautsprecherbaus an, aus kleinen Volumen möglichst viele tieffrequente Anteile herauszuholen. Geschlossene Gehäuse waren für diesen Zweck bisher ungeeignet, da sie ein sehr großes Volumen benötigen, um einen entsprechenden Tiefgang zu erreichen. Bassreflexsysteme hingegen haben sich in der Vergangenheit stark durchgesetzt, da sie durch die Verwendung eines Resonators weniger Volumen benötigen und mehr Tiefgang erzeugen. Allerdings weisen Bassreflexgehäuse gegenüber einem geschlossenen Gehäuse aufgrund des Laufzeitunterschiedes zwischen Resonator und Chassis ein schlechteres Impulsverhalten auf. Das neue System basiert, der besseren Präzision wegen, auf einem kleinen, geschlossenen Lautsprechergehäuse, dessen Frequenzgang durch Signalverarbeitung zu tiefen Frequenzen hin verbessert werden soll. Unter Verwendung der bereits bekannten „Linkwitz-Transformation“, in Kombination mit einer dynamischen Anpassung der Grenzfrequenz, können besonders tiefreichende Bässe bei kleinen Gehäusevolumen erzeugt werden. Die Grenzen des Verfahrens legen der mögliche Membranhub des Chassis und die zur Verfügung stehende Verstärkerleistung fest.

2. Linkwitz-Transformation

2.1. Prinzip

Die Linkwitz-Transformation [1] beschreibt ein Filter, welches auf beliebige geschlossene Gehäuse angewendet werden kann. Der Grundgedanke ist, ein bestehendes Gehäuse mit einer bestimmten (hohen) Einbauresonanzfrequenz und -güte auf eine neue (tiefere) Abstimmung zu „transformieren“ und dabei die Resonanzfrequenz und die Einbaugüte beliebig wählen zu können. Dabei ist die Umsetzbarkeit der Transformation stark abhängig von den Thiele-Small-Parametern und der Belastbarkeit des verwendeten Lautsprecherchassis.

2.2. Biquad

Betrachtet man den Lautsprecher als Hochpass 2. Ordnung, so besitzt dieser ein Nullstellenpaar im Reellen bei $f = 0\ Hz$ und ein konjugiert komplexes Polstellenpaar bei der Resonanzfrequenz f_c mit der Güte Q_{tc}. Die folgende Formel 1 stellt die Übertragungsfunktion des Lautsprechers als Hochpass 2. Ordnung dar:

$$H(s) = \frac{s^2}{s^2 + \frac{2\pi f_c}{Q_{tc}} \cdot s + (2\pi f_c)^2} \qquad (1)$$

Bei der hier beschriebenen Linkwitz-Transformation handelt es sich um ein Biquad-Filter, also ein Filter 2. Ordnung, welches entweder analog (z.B. mit einem Operationsverstärker) oder digital mit einem IIR-Filter umgesetzt werden kann.
Um die ursprüngliche Abstimmung des Gehäuses auf eine Neuabstimmung zu transformieren, muss man einerseits die Polstellen der Originalabstimmung bei der Resonanzfrequenz kompensieren. Dies wird erreicht, indem man die Nullstellen des neuen Filters auf die Position der Polstellen des Lautsprechers legt. Die Polstellen des Filters ordnet man in der komplexen s-Ebene derart an, dass sich eine gewünschte neue Resonanzfrequenz und wählbare, optimierte Resonanzgüte ergibt. Mit der neuen Resonanzfrequenz f_t und der ursprünglichen Resonanzfrequenz f_c lässt sich eine Verstärkung A_{DC} berechnen:

$$A_{DC} = 40 \cdot \log_{10} \frac{f_c}{f_t}\ [dB]$$

Formel 2 beschreibt das Produkt der Übertragungsfunktionen des Linkwitz-Entzerrers und des Lautsprechers. Sie zeigt die Kompensation der originalen Polstellen des Lautsprechers durch die neuen Nullstellen in Form von Kürzen des Nennerpolynoms des Lautsprechers mit dem Zählerpolynoms der Transformation:

ISBN
978-3-9812830-9-9
Der Autor hält die Urheberrechte für den gesamten Inhalt seines Manuskripts. Er überlässt dem VDT das Recht auf Veröffentlichung in gedruckter Form, Online oder als PDF-Download von https://tonmeistertagung.com

$$H(s) = \frac{s^2 + \frac{2\pi f_c}{Q_{tc}} \cdot s + (2\pi f_c)^2}{s^2 + \frac{2\pi f_t}{Q_t} \cdot s + (2\pi f_t)^2} \cdot \frac{s^2}{s^2 + \frac{2\pi f_c}{Q_{tc}} \cdot s + (2\pi f_c)^2} \quad (2)$$

Man erhält die Übertragungsfunktion (Formel 3) eines neuen Hochpasses 2. Ordnung mit der zuvor gewählten Resonanzfrequenz und Güte:

$$H(s) = \frac{s^2}{s^2 + \frac{2\pi f_t}{Q_t} \cdot s + (2\pi f_t)^2} \quad (3)$$

Mit der Linkwitz-Transformation wird also die Neuabstimmung eines zuvor auf eine viel höhere bzw. andere Resonanzfrequenz abgestimmten geschlossenen Lautsprechers ermöglicht. Das Filter nutzt, durch die Verstärkung unterhalb der originalen Resonanzfrequenz, den Membranhub zur erweiterten Basswiedergabe aus. Hierdurch erreicht das geschlossene Gehäuse mehr Tiefgang, ohne auf eine präzise Wiedergabe verzichten zu müssen. In Abbildung 1 ist die Equalizer-Kurve eines Beispiels zu sehen, bei dem ein geschlossener Lautsprecher mit einer ursprünglichen Resonanzfrequenz von 93 Hz und einer Einbaugüte von 1,42 (blau) auf eine Frequenz von 30 Hz mit einer Güte von 0,7 (rot) transformiert wurde. Die gelbe Kurve zeigt dabei die Filterkurve, die im tieffrequenten Bereich um bis zu 20 dB verstärkt.

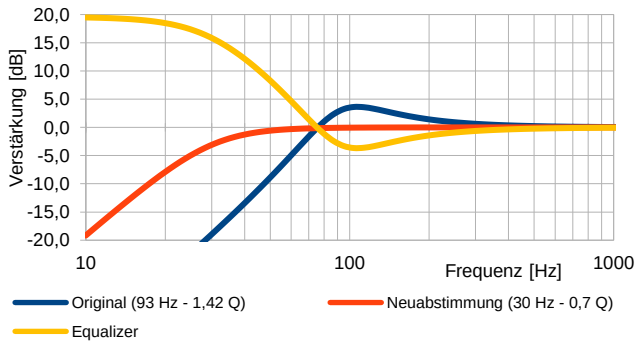


Abb. 1: Equalizer-Kurven der Linkwitz Transformation für eine Neuabstimmung eines geschlossenen Gehäuses von 93 Hz auf 30 Hz. [2]

2.3. Problematik

Das Konzept der Linkwitz-Transformation funktioniert bis zu einem gewissen Punkt sehr gut. Allerdings kann man bereits an den hierfür nötigen Verstärkungsfaktoren erkennen, dass man bei einer sehr tiefen Abstimmung an Grenzen stößt.
Bewirkt die Transformation unter Verwendung eines 100 Watt-Verstärkers beispielsweise eine Anhebung von 20 dB (100-fache Leistung), so wird dem Verstärker bereits bei 1 Watt im nicht angehobenen Bereich, eine Leistung von 100 Watt im tieffrequenten Bereich abverlangt. Das würde bedeuten, dass der Verstärker bereits bei 1 Watt voll ausgelastet wäre und jeder Pegel darüber hinaus zur Übersteuerung führen würde.
Andersherum würde bei ausreichender Verstärkerleistung irgendwann die Auslenkung des Lautsprechers so groß sein, dass Verzerrungen auftreten oder der Lautsprecher an seine mechanischen bzw. thermischen Grenzen stößt.
Es findet also keinerlei Schutz des Lautsprechers oder des Verstärkers statt. Das Filter arbeitet statisch und wird bei hohen Pegeln also problematisch.

3. Dynamische Linkwitz-Transformation

3.1. Idee

Die Idee ist nun, das Filter in Abhängigkeit der problematischen Parameter, also der maximalen Verstärkerleistung, der maximalen Auslenkung des Lautsprechers oder möglichen Verzerrungen, zu verändern. Bevor sich einer dieser Parameter in den gefährlichen Bereich bewegt, würde also die zuvor beliebig gewählte Abstimmfrequenz des Systems nach oben verschoben werden, was eine geringere Verstärkung, allerdings auch weniger Tiefgang mit sich zieht.

3.2. Steuerung

3.2.1. Steuergröße

Zuerst muss eine Steuergröße festgelegt werden, anhand derer die „Betriebsgrenzen“ für die Verschiebung der Grenzfrequenz des Filters festgelegt werden können und die später im Betrieb das System steuern kann. So kann beispielsweise das Audiosignal am Eingang des Systems als Referenz verwendet werden. Nachteil dieses Prinzips ist, dass alle Parameter des Verstärkers und Lautsprechers bekannt sein müssen. So muss man Auskunft über Verstärkungsfaktor, maximale Ausgangsleistung, maximale Belastbarkeit, maximaler Hub und Verzerrungen haben. Sind alle Faktoren bekannt, kann das Eingangssignal als Steuergröße dienen.
Auch Sensorik am Lautsprecher zur Messung der aktuellen Auslenkung kann eine mögliche Referenz sein, wird aber hier nicht behandelt.

3.2.2. Betriebsgrenzen

Mit der gewählten Steuergröße können nun die zwei benötigten Grenzen festgelegt werden:

Untergrenze: Die Auslenkung, bis zu der das System komplett statisch auf der tiefsten gewünschten Abstimmfrequenz betrieben werden kann, ohne dass Verstärker oder Lautsprecher Schaden nehmen oder Verzerrungen auftreten.

Obergrenze: Die Auslenkung, ab der das System auf die originale Abstimmfrequenz des Gehäuses abgestimmt sein sollte, damit keine Beschädigungen entstehen. Im Optimalfall kann das System selbst bei der Obergrenze noch tiefer abgestimmt sein.

In der unten angeführten Abbildung 2 kann man die Equalizer-Kurvenschar einer dynamischen Linkwitz-Transformation betrachten. Die Verstärkung der Linkwitz-Transformation im Bassbereich nimmt mit steigendem Pegel ab einer bestimmten Grenze ab und nähert sich einer definierten maximalen Spannung an. Beim Erreichen der Obergrenze findet keine Verstärkung des Bassbereichs mehr

statt, sondern nur noch die Kompensation der ursprünglichen Polstellen des Lautsprechergehäuses.

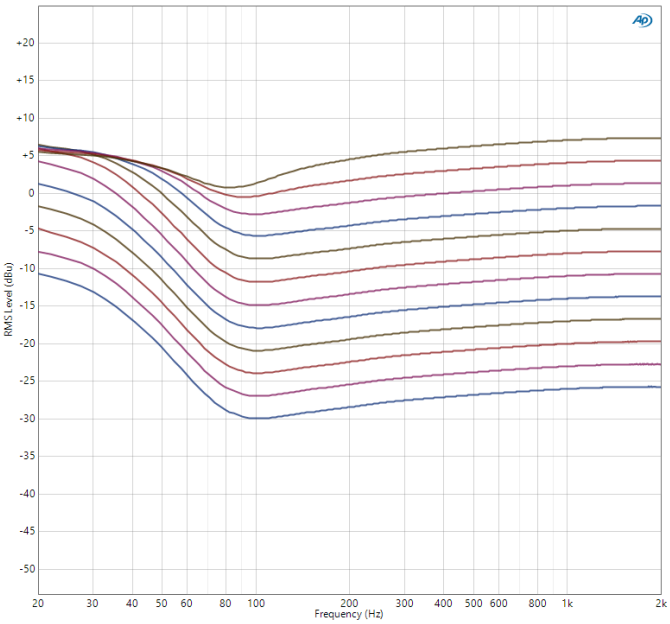


Abb. 2: Equalizer-Kurvenschar der pegelabhängigen (dynamischen) Linkwitz-Transformation.

Abbildung 3 zeigt dagegen die verschiedenen Amplituden-frequenzgänge des Prototyp-Lautsprechers bei steigendem Pegel. Dabei ist zu erkennen, dass der Lautsprecher bis zum Erreichen einer bestimmten Grenze (hier etwa 80 dB) tiefe Frequenzen bis 30 Hz wiedergeben kann. Ab dieser Grenze setzt die Regelung ein und verschiebt die Grenzfrequenz der Linkwitz-Transformation mit steigendem Pegel nach oben, bis die Obergrenze und somit die originale Resonanzfrequenz des Lautsprechers erreicht ist.

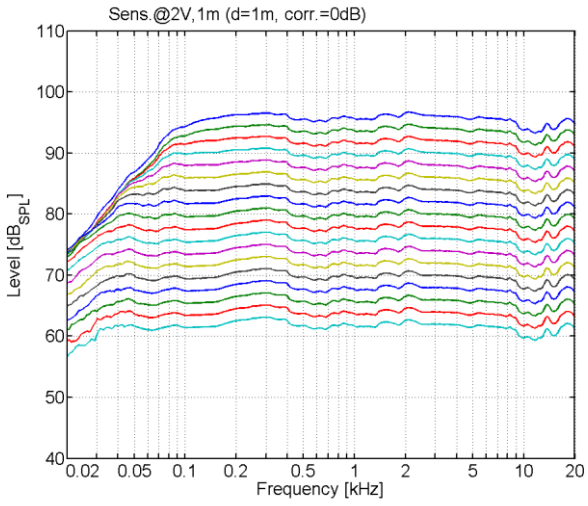


Abb. 3: Steigende Frequenzamplitudengänge eines abgestimmten Systems mit einer Ober- und Untergrenze.

3.2.3. Regelung und Interpolation

Wie bereits beschrieben wird die Abstimmfrequenz des Systems in Abhängigkeit von der gewählten Steuergröße verschoben. Betrachtet man das Filter in der digitalen Ebene als IIR-Filter [3], so bedeutet dies eine Veränderung der Polstelle des Filters, also des Nennerpolynoms in folgender Filterfunktion:

$$H(z) = \frac{b_0 + b_1 \cdot z^{-1} + b_2 \cdot z^{-2}}{1 + a_1 \cdot z^{-1} + a_2 \cdot z^{-2}} \quad (4)$$

Will man die Polstelle verschieben, so müssen lediglich die Koeffizienten a_1 und a_2 simultan verändert werden. Es werden zwei Koeffizienten-Paare errechnet. Eines für die oberste Abstimmfrequenz und eines für die untere Abstimmfrequenz.

Befindet sich die Steuergröße nun unterhalb der Untergrenze, treten die Koeffizienten für die tiefe Abstimmfrequenz in Kraft.

Befindet sich die Steuergröße allerdings zwischen Unter- und Obergrenze, so werden die Koeffizienten-Paare untereinander linear interpoliert und mit ansteigendem Pegel der Lautsprecher und/oder der Verstärker geschützt.

4. Prototyp

Zur Vorführung wurde ein Prototyp (Abbildung 4) angefertigt, der die Stärken des Systems genauer darstellen und hörbar machen sollte. Es handelt sich hierbei um einen 3-Wege Lautsprecher, bei dem Hoch- und Mitteltontreiber passiv voneinander getrennt sind. Tieftöner und Mittel-/Hochtonkombination sind aktiv getrennt und besitzen jeweils einen eigenen Verstärkerkanal. Die Entzerrung der Lautsprecher sowie die Umsetzung des Filters wird durch einen ADAU1467 Sigma DSP und externe AD/DA-Wandler realisiert. Das Volumen für den Tieftontreiber beträgt etwa 7 Liter, was eine Resonanzfrequenz f_c von 85 Hz und eine Güte Q_{tc} von etwa 1,29 mit sich zieht (siehe Abbildung 5). Mit Hilfe des Systems konnte der Lautsprecher auf 30 Hz mit einer Güte von 0,7 abgestimmt werden und nähert sich bei hohen Lautstärken immer weiter der Originalabstimmung von 85 Hz mit einer Güte von ebenfalls 0,7 an.

Abb. 4: Prototyp für die Vorführung des Systems auf der Tonmeistertagung 2018.

4.1. Messergebnisse

Die in Abbildung 1 bereits behandelten Kurven beschreiben die Linkwitz-Transformation in der Theorie, Abbildung 5 zeigt die gemessenen Amplitudenfrequenzgänge des Prototyps im Vergleich.

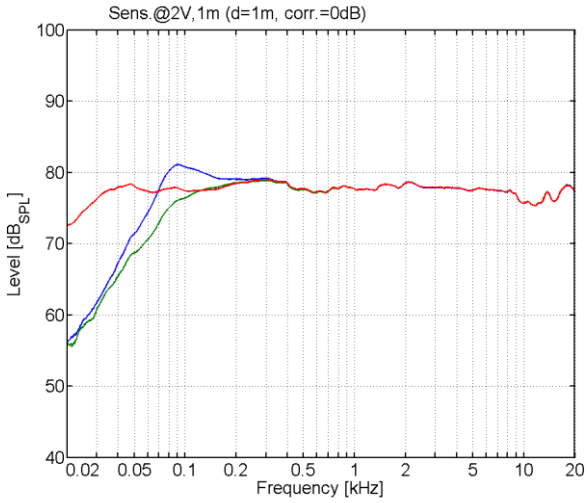


Abb. 5: Messergebnisse des Prototyp-Amplitudenfrequenzgangs ohne Transformation mit hoher Güte (blau), mit der höchsten Abstimmung bei 85 Hz (grün) und der tiefsten Abstimmung bei 30 Hz (rot).

Kleine geschlossene Gehäuse mit hoher Güte, so wie es bei dem Prototypen der Fall ist, besitzen nicht nur im Frequenzbereich eine Überhöhung an der Resonanzfrequenz (Abbildung 5), sondern weisen auch ein starkes Einschwingen im Zeitbereich auf, welches in Abbildung 6 veranschaulicht wird. Im Vergleich zum Einschwingverhalten des Lautsprechers bei der Originalabstimmung (blau) ist zu erkennen, dass die Impulsantwort des Systems, welches auf 30 Hz mit der Güte 0,7 abgestimmt ist (rot), deutlich schwächer ausschwingt und sich somit impulstreuer verhält.

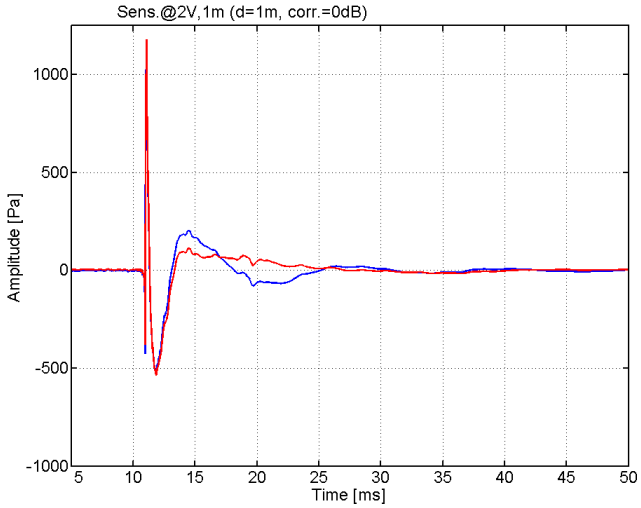


Abb. 6: Impulsantworten des Lautsprechers ohne Transformation (85 Hz) mit hoher Güte von 1,3 (blau) und der tiefsten Abstimmung bei 30 Hz und der Güte 0,7 (rot).

4.2. Subjektive Bewertung

Mit der vorgestellten prototypischen Implementierung konnten bislang nur informelle Hörversuche durchgeführt werden. Der Lautsprecher mit vorgeschalteter Linkwitz-Transformation ist, gemessen an seinem vergleichsweise kleinen geschlossenen Gehäuse, durch den weitreichenden Tiefbass sehr eindrucksvoll anzuhören. Die Basswiedergabe ist, je nach Raumakustik des Abhörraums, sehr präzise und weist selbst bei tieffrequenten, impulsstarken Anteilen ein beeindruckendes Zeitverhalten auf.

5. Fazit

Die Methode der dynamischen Linkwitz-Transformation ermöglicht eine deutlich erweiterte Basswiedergabe in kompakten, geschlossenen Lautsprechergehäusen. Mit ihrer Hilfe kann ein verbessertes Bassverhalten, sowie eine Verbesserung der Impulsantwort des Lautsprechers erzielt werden. Die Bassanhebung der Transformation ist lediglich durch die thermische und mechanische Belastbarkeit des Lautsprechers sowie die maximale Verstärkerleistung begrenzt. Eine pegelabhängige Regelung verhindert das Überschreiten dieser Grenzen bei hohen Pegeln, indem sie die Grenzfrequenz des Filters nach oben verschiebt. Das Verfahren lässt sich überall dort einsetzen, wo geringes Volumen, aber gleichzeitig eine präzise, tiefreichende Wiedergabe gefordert ist. So kann es beispielsweise Verwendung in HiFi-Lautsprechern, Studio-Monitoren, Soundbars, Beschallungsanlagen oder Leistungsverstärkern finden.

6. Literatur

[1] Webseite von Siegfried Linkwitz
URL: http://www.linkwitzlab.com

[2] Linkwitz Transform Calculator von miniDSP
URL: http://www.minidsp.com

[3] Wikipedia Artikel zu Digitalen Biquad Filtern
URL: http://en.wikipedia.org/wiki/Digital_biquad_filter

Personal WFS – Wiedergabe binauraler Schallsignale durch ein Individuum-bezogenes Wellenfeldsynthese-System

Kerstin Fuhrmann[1], Prof. Dr.-Ing. Dieter Leckschat[2], Dipl.-Ing. Christian Epe M.Sc.[3]
[1] *Hochschule Düsseldorf, Germany, Email: kerstin_fuhrmann@gmx.net*
[2] *Hochschule Düsseldorf, Germany, Email: dieter.leckschat@hs-duesseldorf.de*
[3] *Hochschule Düsseldorf, Germany, Email: christian.epe@hs-duesseldorf.de*

Zusammenfassung

Um eine binaurale Wiedergabe über Lautsprecher zu ermöglichen, wird konventionell das Verfahren der Crosstalk-Cancellation verwendet. Bei dieser Methode wird der Hörer allerdings in seiner Bewegungsfreiheit beeinträchtigt, da bereits kleine Kopfbewegungen den Höreindruck erheblich verschlechtern, sofern kein Trackingsystem verwendet wird. Durch die verwendeten Lautsprecher ist zudem oftmals das Sichtfeld eingeschränkt. Das Projekt „Personal WFS – Wiedergabe binauraler Schallsignale durch ein Individuum-bezogenes Wellenfeldsynthese-System“ zeigt einen Lösungsansatz, um die Wiedergabe binauraler Signale über Lautsprecher mit Hilfe der Wellenfeldsynthese (WFS) zu verbessern. Anstatt eines Sweetspots wird mit Hilfe der WFS und einer geeigneten Lautsprecheranordnung eine „Sweetarea“ geschaffen, in welcher der Hörer mehr Bewegungsfreiheit besitzt und zudem sein Sichtfeld kaum beeinflusst wird. Der kompakte Aufbau hält die sogenannte Personal WFS außerdem mobil und kostengünstig.

1. Einführung

Die Personal WFS ist ein Lösungsansatz zur Wiedergabe binauraler Schallsignale über ein Individuum-bezogenes Wellenfeldsynthese-System. Das System ist zudem durch seinen modularen Aufbau mobil gehalten.

Zunächst werden verschiedene Verfahren zur binauralen Wiedergabe thematisiert. Daraufhin folgt die Konzeption der Personal WFS und ein kleiner Exkurs in die Theorie der Wellenfeldsynthese. Im Anschluss wird der Aufbau der Personal WFS konkret erläutert. Abschließend werden die Evaluation und Messergebnisse der Personal WFS vorgestellt.

2. Binaurale Wiedergabe

2.1. Wiedergabe über Kopfhörer

Das bekannteste Wiedergabeverfahren für binaurale Schallsignale ist das über Kopfhörer. Dabei setzt der Hörer einen Kopfhörer auf oder ein In-Ear-System ins Ohr ein. Jedes Ohr erhält dadurch ein unabhängiges Schallsignal, das durch eine Aufnahme oder mittels eines Binauralrenderers und HRTFs (Head Related Tranfer Function) entstanden ist. Torso, Kopf und Raum haben somit keinen Einfluss auf das Hörereignis. Allerdings findet auf diese Weise keine Körperschallübertragung statt. Dies führt dazu, dass der Hörer sich nicht so sehr in die virtuelle Szenerie eingebunden fühlt wie in einer realen Situation [1]. Des Weiteren kann ein Gegenstand am Kopf die Bewegungsfreiheit des Hörers einschränken.

2.2. Crosstalk-Cancellation (ohne Tracking)

Crosstalk-Cancellation ist ein weiteres Verfahren zur Wiedergabe binauraler Schallsignale. Zwecks Vergleichbarkeit wird hier die Variante ohne Tracking vorgestellt. Der grundlegende Gedanke der Crosstalk-Cancellation ist die Wiedergabe binauraler Tracks über eine meist stereophone Lautsprecheranordnung. Würden Binauraltracks ohne Crosstalk-Cancellation, somit ohne jegliche Signalverarbeitung über solch einen Aufbau abgespielt werden, würde das Schallsignal des linken Wiedergabekanals sowohl von dem linken Ohr als auch mit einer Laufzeitverzögerung von dem rechten Ohr wahrgenommen werden (**Abb. 2**). Gleiches gilt umgekehrt für das rechte Schallsignal. Es entsteht ein Übersprechen, Crosstalk, auf das jeweilige contralaterale Ohr. So können die binauralen Informationen nicht richtig wahrgenommen werden und es entsteht ein schmales, frontales virtuelles Schallereignis. Die Erzeugung der korrekten räumlichen Illusion des Binauraltracks ist nicht möglich [1].

Abb. 1: Aufbau der Personal WFS

Bei der Verwendung von Crosstalk-Cancellation wird durch Signalverarbeitung das Übersprechen auf das contralaterale Ohr kompensiert. Dennoch ist die korrekte Wiedergabe auf einen kleinen räumlichen Bereich, einen sogenann-

ISBN
978-3-9812830-9-9
Der Autor hält die Urheberrechte für den gesamten Inhalt seines Manuskripts. Er überlässt dem VDT das Recht auf Veröffentlichung in gedruckter Form, online oder als PDF-Download von https://tonmeistertagung.com

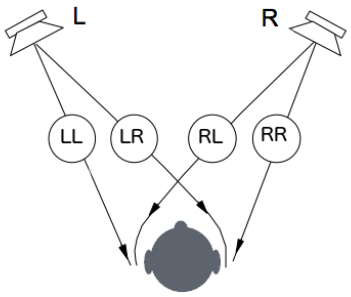


Abb. 2: Veranschaulichung der Laufzeitdifferenzen, die zwischen dem linken und rechten Ohr entstehen. LL und RR entsprechen den Signalwegen, die für das linke beziehungsweise rechte Ohr bestimmt sind. LR und RL illustrieren den Signalweg zum contralateralen Ohr.

ten Sweetspot, begrenzt. An allen Positionen außerhalb des Sweetspots funktioniert die Crosstalk-Cancellation auf Grund falscher Abstimmung der Laufzeitunterschiede nicht. Somit ist der Hörer an eine Hörposition fest gebunden und hat keine Bewegungsfreiheit, ohne große Qualitätsverluste der virtuellen Szene hinzunehmen. Schon leichte Kopfbewegungen zerstören die geschaffene Illusion. Durch die Crosstalk-Cancellation ist im Gegensatz zur Wiedergabe über Kopfhörer eine Körperschallübertragung möglich, aber das Sichtfeld kann durch die Lautsprecher stark eingeschränkt sein.

2.3. Binaural Sky

Der Binaural Sky ist 2005 von dem IRT (Institut für Rundfunktechnik) entwickelt und in den Folgejahren optimiert worden. Der Binaural Sky besteht aus einem kreisförmig angeordneten Lautsprecherarray, das über dem Kopf des Hörers platziert ist. Um eine 3D-Audiowiedergabe zu ermöglichen, werden virtuelle Quellen mittels einer Wellenfeldsynthese erzeugt. Diese fokussierten Quellen (siehe Abschnitt 4. Theorie WFS) sind direkt vor den Ohren des Hörers positioniert [2]. Dieser Aufbau bietet zwar freie Sicht, ist allerdings ortsgebunden zu errichten, da es einer festen Installation über dem Kopf des Hörers bedarf.

3. Konzeption Personal WFS

„Personal“ bedeutet in diesem Zusammenhang Individuumbezogen, also für einen Hörer ausgelegt.

Die Idee der Personal WFS ist die Schaffung eines Systems, das die Wiedergabe binauraler Schallsignale durch virtuelle Schallquellen, die durch eine Wellenfeldsynthese an den Ohren des Hörers erzeugt werden, realisiert. Die generierten virtuellen Quellen sind wie im Falle des Binaural Skys fokussiert (siehe Abschnitt 2.3 Binaural Sky). Hierdurch werden, ähnlich wie bei der Wiedergabe über Kopfhörer, die Einflüsse des Torsos und des Kopfes gemindert. Des Weiteren soll das System mit wenig Aufwand an verschiedenen Orten auf- und abgebaut werden können. Dies wird durch den modularen Aufbau aus zwei Lautsprecherzeilen, einem Laptop und einem 19” Rack mit benötigter Hardware realisiert.

Die beiden Lautsprecherzeilen sind vertikal links und rechts neben dem Hörer in einem Abstand von 120 cm zueinander angeordnet (**Abb. 1**). Der Abstand zwischen einem Lautsprecherarray und dem Ohr des Hörers beträgt ca. 50 cm. Die Entfernung variiert mit der Kopfgröße.

Typischerweise ist eine Wellenfeldsynthese horizontal aufgebaut. Bei der hier verwendeten vertikalen Umsetzung werden allerdings die Eigenschaften eines Linienstrahlers ausgenutzt. So wird anstatt eines Sweetspots, wie dieser bei dem Crosstalk-Cancellation-Verfahren vorliegt, eine Sweetarea auf der horizontalen Ebene geschaffen. Das heißt, es liegt ein größerer Bereich vor, in dem eine korrekte Wiedergabe des Binauraltracks funktioniert. In diesem Bereich nimmt der Hörer bei leichten Vor- und Rückwärtsbewegungen sowie moderaten Drehungen des Kopfes keine erheblichen Qualitätsverluste der auditiven Illusion wahr. Genauere Erläuterungen sind im Abschnitt 6.2 Horizontale Winkelabhängigkeit zu finden.

4. Theorie WFS

Die Reproduktion realer Schallfelder durch Überlagerung von vielen, nah beieinander positionierten Lautsprecherschallsignalen zu einem realen, räumlich ausgedehnten Schallfeld wird als Wellenfeldsynthese bezeichnet [3]. Es wird zwischen zwei verschiedenen Arten virtueller Quellen unterschieden: Distanzquellen und fokussierte Quellen. Distanzquellen sind virtuelle Quellen, die hinter der Lautsprecherzeile platziert sind. Diese Quellenart lässt sich im Gegensatz zu den fokussierten Quellen vollständig synthetisieren. Fokussierte Quellen werden vor der Lautsprecherzeile positioniert und eine korrekte Synthese ist nur zwischen virtueller Quelle und Hörposition möglich [4].

5. Aufbau der Personal WFS

Der portable Aufbau des Personal WFS-Systems gliedert sich in drei Kategorien: Software, Hardwarekomponenten und Lautsprecherzeilen. Abbildung 3 illustriert den gesamten Signalfluss.

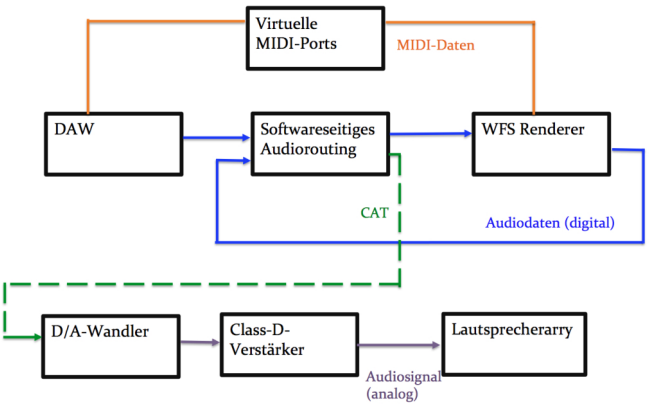


Abb. 3: Signalfluss des Aufbaus der Personal WFS
Der obere Teil wird komplett softwareseitig auf einem Laptop umgesetzt. Die Schnittstelle zu den Hardwarekomponenten bildet eine Netzwerkübertragung.

5.1. Software

Das rechnerinterne Audiorouting und die Verwendung von virtuellen MIDI-Ports halten das System kompakt. Auf diese Weise wird der Hardwareballast minimiert und nur noch ein Laptop benötigt. Das ist zuvor noch nicht der Fall gewesen.

Der in mehreren Abschlussarbeiten der Hochschule Düsseldorf (ehemalige Fachhochschule Düsseldorf) eigens entwickelte und weiterentwickelte WFS-Renderer bildet die Basis für die Personal WFS. Der Renderer generiert aus Quellsignalen und Positionsdaten der virtuellen Quelle für jeden Lautsprecher ein eigenes Signal. Um den Renderer mit den Quellsignalen, also den Audiospuren, zu speisen, wird eine DAW (Digital Audio Workstation) genutzt. In dieser wird ebenfalls die Entzerrung des Lautsprecherfrequenzgangs vorgenommen. Die Position der virtuellen Quelle im virtuellen Raum wird über MIDI-Steuerdaten (Music Instrument Digital Interface) zwischen dem WFS-Renderer und der DAW ausgetauscht. Die daraus berechneten Lautsprechersignale werden anschließend softwareseitig an den Dante Controller geroutet.

5.2. Hardware

Als Schnittstelle zwischen Soft- und Hardware wird eine Netzwerkübertragung via Dante verwendet. Daraufhin werden die analogen Audiosignale von dem D/A-Wandler an die 16 Class-D-Verstärker-Chip-Module übertragen. Diese Stereoverstärker liefern an einer 8-Ohm-Last 25 W pro Kanal. Schließlich erfolgt die Übertragung an die 32 Lautsprecher.

5.3. Lautsprecherzeilen

Die zwei Lautsprecherzeilen bestehen aus je 16 Breitbandchassis, wovon jedes ein eigenes ein Liter großes Gehäuse besitzt (**Abb. 4**). Die Treiber sind in einem Abstand der Membranmitten von 92,5 mm eingebaut. Damit das System transportabel bleibt, sind sehr leichte Treiber mit einem Eigengewicht von 110 g pro Stück verbaut.

Abb. 4: Lautsprecherarray in der Bauphase
Jeder Treiber wird in ein eigenes 1-L-Volumen eingebaut.

Abbildung 5 zeigt den Frequenzgang der Mittelung aller Treiber in einem 1-L-Volumen ohne Gehäusebedämpfung (rot) im Vergleich zu dem Frequenzgang des bedämpften Gehäuses inklusive der Entzerrung durch ein DAW-Plug-In (blau). Die Bedämpfung des Gehäuses und die Entzerrung ermöglichen eine recht lineare Übertragung des Schallsignals.

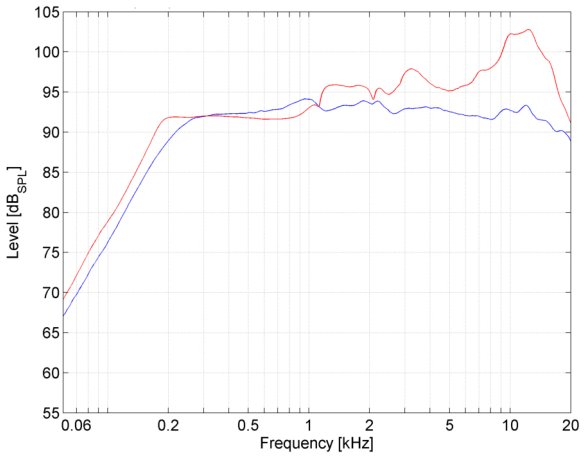


Abb. 5: Vergleich der Frequenzgänge
Rot: Mittelung aller Treiber in einem 1-L-Gehäuse ohne Bedämpfung
Blau: bei aktiver Personal WFS an der virtuellen Quelle (Bedämpfte Gehäuse inkl. Entzerrung)

6. Evaluation und Messergebnisse

6.1. Funktionsfähigkeit der WFS

Die Funktionsfähigkeit der Wellenfeldsynthese lässt sich anhand folgender Überlegung bestätigen:
Wird keine Wellenfeldsynthese verwendet, entsteht zwischen dem mittleren Lautsprecher und dem äußersten Lautsprecher des Arrays eine Laufzeitdifferenz von fast 1 ms. Abbildung 6 verdeutlich dies.

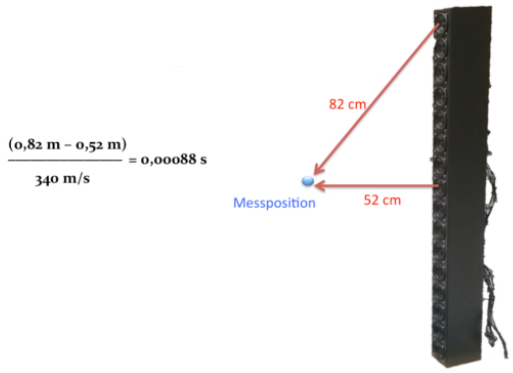


Abb. 6: Veranschaulichung der Laufzeitdifferenz, die zwischen dem mittleren und äußersten Lautsprecher derselben Zeile entstehen, wenn keine WFS verwendet wird.

Der Vergleich der Messungen (**Abb. 7**) der Impulsantwort des Lautsprechers auf der 0°-Achse (horizontal und vertikal) an der Position der fokussierten Quelle und des davon am weitesten entfernten Lautsprechers derselben Zeile zeigt, dass bei aktiver Wellenfeldsynthese nur eine marginale Laufzeitdifferenz von 0,04 ms vorliegt. Diese ist auf eine Ungenauigkeit der sehr schwierig exakt an der fokussierten Quellen auszurichtenden Mikrofonposition zurückzuführen.

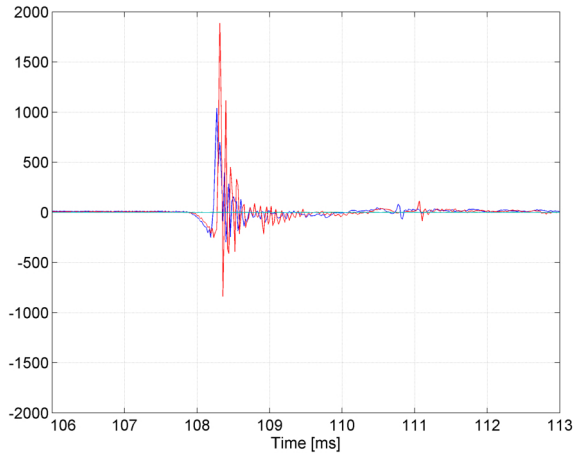


Abb. 7: Impulsantwort des Lautsprechers auf der 0°-Achse an der virtuellen Quelle und des davon am weitesten entfernten Lautsprechers derselben Zeile bei aktiver Wellenfeldsynthese

6.2. Horizontale Winkelabhängigkeit

Bewegt der Hörer seinen Kopf vor und zurück oder leicht nach links beziehungsweise rechts, ist in dem Fall des Sweetspots keine korrekte Wahrnehmung des binauralen Schallereignisses mehr möglich. Liegt hingegen eine Sweetarea vor, wird der Höreindruck nicht gemindert. Das Messergebnis (**Abb. 8**) der horizontalen Winkelabhängigkeit des Frequenzgangs zeigt, dass zwischen der 0°-Position, an der virtuellen Quelle, und 10° horizontal keine gravierenden spektralen Unterschiede entstehen. Es ist lediglich eine leichte Bündelung der hohen Frequenzen bei 10° zu verzeichnen. Dies ist der Eigenschaft der Linienquelle geschuldet. Erst ab 30° ist ein Einbruch von ca. 5 dB bei der Übertragung der Frequenzen ab 5 kHz festzustellen. Somit ist messtechnisch gezeigt, dass eine horizontale Sweetarea vorliegt.

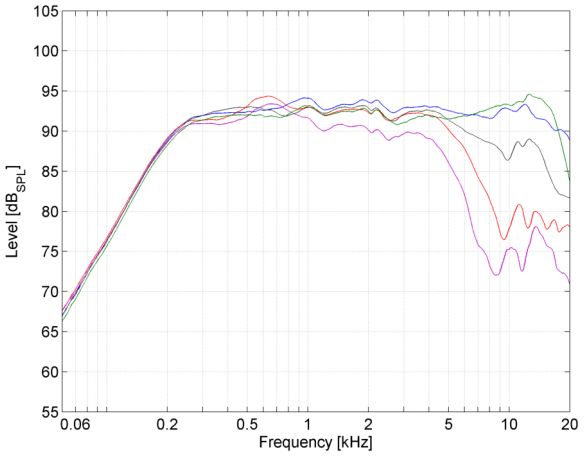


Abb. 8: Horizontale Winkelabhängigkeit bei aktiver Personal WFS Blau: 0° (an virtueller Quelle) Grün: 10° Schwarz: 30° Rot: 45° Lila: 60°

6.3. Subjektive Evaluation

Zum Zeitpunkt dieser Veröffentlichung liegt noch keine fundierte Evaluation mit Probanden vor. Das System ist auf der Tonmeistertagung 2018 in einem Showroom präsentiert worden. Nachfolgend ist das Feedback der Tagungsteilnehmer, die die Personal WFS getestet haben, zusammengefasst.

Meist ist von einer korrekten Entfernungs- und Richtungsortung des Schallereignisses berichtet worden. Zudem seien leichte Kopfbewegungen möglich und die Testhörer haben sich gut in die virtuelle Szenerie integriert gefühlt. Lediglich die frontale Ortung hat noch Mängel aufgewiesen, welche allerdings sehr wahrscheinlich einem HRTF-Datensatz geschuldet sind, der nicht mit dem Kopf der meisten Hörer übereinstimmt. Die frontale Ortung desselben Demotracks war über Kopfhörer abgehört ebenfalls schlecht.

7. Zusammenfassung

Zusammengefasst lässt sich sagen, dass bei der Wiedergabe binauraler Schallsignale über die Personal WFS eine Sweetarea anstatt eines Sweetspots entsteht. Hierdurch sind leichte Kopfbewegungen nach vorne und hinten möglich. Der Hörer wird weder in seiner Bewegungsfreiheit durch Gegenstände am Kopf noch in seinem vorderen Sichtfeld eingeschränkt. Der komplette Aufbau ist leicht zu errichten und dadurch nicht ortsgebunden. Durch diese Vorteile ist eine Anwendung in der Augmented Reality denkbar.

8. Literatur

[1] Nicol,R.: Binaural Technology. Audio Engineering Society Inc., 2010

[2] IRT: Neue Anwendung der Wellenfeldsynthese: „The Binaural Sky“, 2006
URL: http://hauptmikrofon.de/theile/2005-2_BinauralSky_IRT-Kurzartikel-2005.pdf

[3] Slavik, K.M., Weinzierl, S.: Handbuch der Audiotechnik. Springer-Verlag, Berlin Heidelberg, 2008

[4] Boone, M.M., Verheijen, E.N.G.: Sound Reproduction Applications with Wave Field Synthesis. Audio Engineering Society, 1998

Sprachverständlichkeit um jeden Preis? Originalton bei Spiel- und Dokumentarfilm

F. Andriessens[1]

[1] *Dipl. Tonmeister f. audiovisuelle Medien Felix Andriessens (AES / bvft / VDT), Germany, eMail: felix@filmtonmeister.com*

Abstract

Sprachverständlichkeit bei Film- und Fernsehproduktionen ist ein Parameter, mit dem neben Sprecher*innen und Darsteller*innen unmittelbar auch alle Filmton-Gewerke zu tun haben – vom Set bis zur Mischung, von O-Tonmeister*innen bis zu Komponist*innen. Er betrifft technische Parameter genauso wie gestalterische. Auch die Planungen des Produktions- und aller Bild-Departments haben weitreichenden Einfluss darauf. Das ist nicht immer allen Beteiligten bewusst. Dieser Text zeigt die Stellen auf, an denen besonders auf die Erhaltung von Sprachverständlichkeit geachtet werden muss, und zeigt unter Abwägung von Kosten und Nutzen Wege dorthin auf.

1. Das Ungleichgewicht in der Wahrnehmung von Bild und Ton

Sprachverständlichkeit bei Film- und Fernseh-Produktionen stellt vor allem deswegen eine besondere Herausforderung dar, weil solche Produktionen in ihren Abläufen naturgemäß sehr bildfixiert sind. Es gibt also ein großes Ungleichgewicht in der bewussten Aufmerksamkeit auf Bild und auf Ton, der sich durch viele Bereiche zieht.

Dabei macht jedes dieser beiden Elemente, die ja nur zusammen den „added value“ eines Filmes [1], also mehr als die Summe seiner Einzelteile ergeben, nicht nur gemäß George Lucas „50% der Kino-Erfahrung“ aus (Zitat George Lucas: "Sound is 50 percent of the movie going experience, and I've always believed audiences are moved and excited by what they hear in my movies at least as much as by what they see."). Ton hat dieser Theorie zur Folge mindestens das gleiche Wirkungsgewicht wie Bild.

Und da wir nun unsere Augen schließen können, unsere Ohren jedoch zumindest in physischer Hinsicht in gleicher Form nicht, unterliegt die durchschnittliche menschliche Alltagswahrnehmung dem Eindruck, dass Ton immer „da ist“. Für das Medium Film aber gibt es diese Naturgegebenheit nicht. Ton „zu haben“ bedeutet hier immer Arbeit, zusätzlich zur Erschaffung des bewegten Bildes. Eine Arbeit, die vermutlich auch wegen dieser vermeintlichen Naturgegebenheit von Ton häufig selbst von professionellen Film-Herstellern unterschätzt wird.

Dies, obwohl Tongestaltung vom reinen Personalaufwand her für einen durchschnittlichen Spielfilm weit kostengünstiger zu erreichen ist als Bildgestaltung, wie im Folgenden schon anhand der typischen Arbeitskraft-Verteilung zwischen Ton und Bild bei einer Spielfilm-Produktion demonstriert werden wird. Wenden wir uns dazu den Geschehnissen am Drehort und ihren Auswirkungen auf die Sprachverständlichkeit zu.

2. Sprachverständlichkeit bei der Originaltonaufnahme

2.1. Am Drehort: Spielfilm

Beim Spielfilm wird für das Bild mit Hilfe einer mehrköpfigen Kostüm-Abteilung, einer mehrköpfigen Ausstattungs-/ Requisiten-Abteilung und einer mehrköpfigen Szenenbild-Abteilung (die für den Bau der Kulisse zuständig ist), zunächst eine Bild-Szenographie entworfen.

Dazu kommen dann noch eine mehrköpfige Beleuchter-Crew, die Abteilung Kamera-Bühne (die sich um Dollys, Kräne und dergleichen kümmert), und dann die eigentliche, meist zwei- bis dreiköpfige Kamera-Crew.

Diesem Personal- und Vorbereitungsaufwand auf Bild-Seite - Schauspieler*innen finden in dieser Betrachtung übrigens keine Erwähnung, weil sie beiden Seiten, Bild wie Ton, gleichermaßen angehören - stehen auf Ton-Seite nun in der Regel lediglich ein/e Tonmeister*in und ein/e Ton-Assistent*in gegenüber.

Abb. 1: Typische Spielfilm-Drehsituation

Selbst die Tatsache, dass in der Ton-Postproduktion auch bei kleineren Spielfilm-Projekten meist mindestens fünf bis sechs Filmtonschaffende (Dialog-Editor, Sounddesigner, Geräuschemacher*in (engl. Foley Artist), Geräusch-Tonmeister*in (engl. Foley-Mixer), Mischtonmeister, ADR-Tonmeister [2], ganz zu schweigen von Komponist*innen)

ISBN
978-3-9812830-9-9

Der Autor hält die Urheberrechte für den gesamten Inhalt seines Manuskripts. Er überlässt dem VDT das Recht auf Veröffentlichung in gedruckter Form, online oder als PDF-Download von https://tonmeistertagung.com

am Werk sind, kann an einer Überzahl in der Bild-„Mannschaft“ nichts ändern.

Da es bei der Herstellung eines Filmes nun aber nicht um ein physisches Kräftemessen zwischen Bild und Ton geht, sondern um das gemeinsame Erzählen einer Bewegt-Bild-Geschichte auf eine Art, die die Zuschauer*innen vor allem emotional erfasst und bindet, liegt das Argument nahe, dass dieser Vergleich von schierer Arbeitskraft nicht zielführend ist. Zumal bestimmte Abläufe im Bildbereich nun einmal schon in rein physischer Hinsicht einen bestimmten Aufwand an Arbeitskraft zu ihrer Erledigung erfordern.

Dieses Argument ist kaum zu widerlegen. Und dennoch ist der beschriebene grundlegende Unterschied von Personalstärke schon am Drehort Sinnbild und Verstärker einer bei der Produktionsplanung und am Set vorherrschenden Sichtweise, die den Ton als filmisches Ausdrucksmittel permanent in die zweite Reihe stellt. Um so mehr, wenn man bedenkt, dass die Dreharbeiten zu dem Zeitpunkt meist schon längst gelaufen sind, zu dem die Ton-Postproduktions-Mitarbeiter*innen engagiert werden.

Die auf dieser Sichtweise aufbauende Haltung hat aus dramaturgisch-künstlerischer und letzten Endes auch wirtschaftlicher Sicht große Nachteile, führt sie doch dazu, dass das Potential der tonseitigen „50% Wirkung“ des Filmes, und damit auch das Vermarktungspotential des Endproduktes beschnitten wird.

Dabei ist es häufig effektiver, angesichts des beschriebenen Personalaufwandes aber in jedem Fall wirtschaftlich effizienter, dramaturgische Elemente über den Ton zu erzählen als über das Bild. Oder, um es anhand zweier Beispiele zu formulieren:

Ein unscharfes Bild mit dramatisch-wackeliger Kameraführung ist schnell und kostengünstig hergestellt. Lässt man einen Sounddesigner hier die Geräuschkulisse einer Schlacht oder eines gewaltigen, jagenden Monsters erstellen, hat man gute Chancen, einen starken emotionalen Effekt zu einem sehr günstigen Preis zu erreichen - günstiger zumindest als die Beschäftigung hunderter Komparsen für die Schlacht, oder einer Heerschar von VFX-Künstlern für die Gestaltung des Monsters.

2.2. Auswirkungen auf die Sprachverständlichkeit

Welche Rolle spielt das alles für das Thema Sprachverständlichkeit?

Die beschriebene Grundsituation hat zur Folge, dass die Drehsituation schon in der Vorbereitungsphase in erster Linie nach dem optischen Wert beurteilt wird. Ob eine mittelalterliche Location neben einer Zugstrecke oder einer Autobahn, ein Kostüm mit klimpernden Reißverschlüssen oder ein permanent quietschendes Requisit akustisch brauchbar oder gar störend sind, ist nur in zweiter Linie von Interesse.

Eine Drehort-Vorabbesichtigung durch das Original-Ton-Team muss bisweilen erkämpft werden, falls der/die Tonmeister*in überhaupt rechtzeitig dafür gebucht wird, und so ist es keine Seltenheit, dass es dazu nicht kommt, oder die Bezahlung dieser Vorbereitungszeit in Frage gestellt wird.

In der eigentlichen Drehsituation werden zur Einsparung von Drehzeit und/oder zur Erweiterung der editorischen Möglichkeiten viele Szenen direkt in naher und totaler Einstellungsgröße gleichzeitig gedreht, „durchgemastert“, wie man sagt, was die Möglichkeiten zur dynamischen optimalen Mikrofonpositionierung (sprich: zum erfolgreichen „Ton-Angeln“) erheblich einschränkt, oder gar vollständig zunichte macht.

Action-Sets sind darüber hinaus legendär für ihre geringe Ausbeute an verwendbarem O-Ton - neben den genre-unabhängig häufig schon von sich aus akustisch ungünstigen Drehorten rauscht hier ein Lüfter für den dramatischen Wind-Effekt im Haar der Darsteller*innen, dort surrt die Drone für eine möglichst effektreiche Einstellung aus der Vogelperspektive, dazu gibt es noch reichlich Kommandos während der laufenden Einstellung von Regie, Tier-Trainer*innen und Stunt-Koordinator*innen.

So darf man sich also die durchschnittliche Dreh-Situation für einen Spielfilm vorstellen: raumakustisch von sich aus schon häufig suboptimal, reich an Nebengeräuschen, mit stark eingeschränkten Möglichkeiten zur optimalen Mikrofonplatzierung, bei geringen und sinkenden Vorbereitungszeiten.

Für die Sprachverständlichkeit von unter solchen Umständen entstandenen Aufnahmen ist nichts Gutes zu erwarten. Wie sehen die Bedingungen bei Dokumentarfilmdrehs aus?

2.3. Am Drehort: Dokumentarfilm

Im Vergleich zum Spielfilm könnte die personelle Situation bei einem Dokumentarfilm-Dreh unterschiedlicher kaum sein: während beim Spielfilm ungefähr 90% der Belegschaft am Drehort ausschließlich mit der Bildgestaltung beschäftigt sind, werden die drei beim Dokumentarfilm vorkommenden Abteilungen Regie, Kamera und Ton in der Regel von jeweils einer Person ausgefüllt. Es scheint also ein Gleichgewicht gegeben.

Abb. 2: Typische Teamzusammensetzung beim Dokumentarfilm-Dreh

Darüber hinaus ist bei einem Dokumentarfilm-Dreh für alle Beteiligten klar, dass die gefilmten und aufgenommenen Situationen in der Regel nicht wiederholt werden können, damit also einmalig und unwiederbringlich sind, und alle Departments unbedingt Rücksicht aufeinander nehmen müssen. Nicht wenige Dokumentarfilm-Kameraleute äußern sogar von sich aus die Ansicht, dass der Ton im Dokumentarfilm wichtiger ist als das Bild, um die gewünschte Geschichte zu erzählen.

Die Rahmenbedingungen unterscheiden sich also schon in dieser Hinsicht stark.

Dafür ist die eigentliche Drehsituation beim Dokumentarfilm noch stärker von Unwägbarkeiten geprägt als beim Spielfilm: niemand weiß so ganz genau, welches tonliche Element sich in einer Situation im Nachhinein als das Wichtigste herausstellen wird, welche Entwicklung im Hintergrund auf einmal das Potential für die Erklärung eines wichtigen Umschwungs in der Handlung bekommt, welcher Ton off-Kamera eine Stimmung oder Situation so klar darstellt, dass er eigentlich unabdingbar ist.

Und die Drehorte sind häufiger noch als im Spielfilm akustisch ungünstig, sei es durch Hintergrundlärm, sei es durch die akustischen Eigenschaften der Räume dort. Das bedeutet, dass es für jede/n Dokumentarfilm-Original-Tonmeister*in einmal die Situation geben wird, einen Augenblick zu spät mit dem Mikrofon zu sein, vielleicht eine Sekunde zu spät auf „Record“ gedrückt zu haben, ohnmächtig anhören zu müssen, wie vielleicht doch noch ein Auto durch den Text der in Tränen aufgelösten Hauptfigur fährt, etc.

Es gilt also mehr noch als beim Spielfilm-Dreh, dass man sich beim Dokumentarfilm-Dreh nicht zu sehr von der Idee einer sauberen, „reinen“ Tonaufnahme mit optimaler Sprachverständlichkeit abhängig machen darf.

Die entscheidende Frage ist: ist eine optimale Sprachverständlichkeit überhaupt notwendig?

2.4. Verständlichkeit vs. emotionaler Gehalt

Ein Bonmot der Filmtongestaltung lautet: „Die Wahrheit klingt schlecht und ist Mono“, und es beschreibt eine Hörkonvention sehr treffend. Während sich der Ausspruch eher auf einen tongestalterischen Aspekt bezieht, als Beispiel sei die geheime (und ihrer dramaturgischen Rolle als „Wahrheit“ entsprechend verrauschte und verzerrte) Bandaufnahme in Francis Ford Coppolas „The Conversation“ von 1974 genannt, entstammt diese Hörkonvention eben genau der Erfahrung, dass die Aufnahme eines „wahren“ Momentes selten technisch einwandfrei ist. Diese Konvention lässt sich also gestalterisch nutzen. Umgekehrt bedeutet sie aber auch eine gewisse Freiheit bei der Original-Ton-Aufnahme. Denn es bringt eben nichts, eine Geschichte technisch einwandfrei zu erzählen, die emotional keinen Gehalt hat. Das gilt sowohl für den Dokumentarfilm, als auch für den Spielfilm.

Liefert ein/e Schauspieler*in eine hoch emotional wirkende, aber vernuschelte oder tontechnisch ungünstige Performance, ist deswegen selbst bei der Spielfilm-Produktion keine „Passt-immer“-Lösung zur Hand, obwohl dort - zumindest in technischer Hinsicht - prinzipiell jeder Ton austauschbar ist, weil die Darsteller*innen sich in der Regel für Synchron-Aufnahmen nach dem Dreh bereithalten müssen. Häufig muss man dann entschieden werden zwischen einem emotionalen, aber akustisch schlecht verständlichen Ton, und einer Synchron-Aufnahme mit hoher Sprachverständlichkeit und geringerem emotionalen Inhalt.

2.5. Schlecht verständlicher Originalton: gemeinsame Ursachen für Spiel- und Dokumentarfilm

Die Tonaufzeichnung bei Dokumentar- und Spielfilm unterliegt unterschiedlich häufig optisch bedingten Einschränkungen. Die rein tontechnische Problemstellung bleibt aber immer die gleiche: es geht darum, in Absprache mit dem Bildschnitt und unter Berücksichtigung der Drehsituation einen Workflow und ein Aufnahme-Setup zu entwickeln, das hinreichend flexibel ist, aber gleichzeitig so „schlank“ wie möglich, und das über den gesamten Dreh beibehalten werden kann. Mit Workflow ist dabei das Folgende gemeint: sowohl Originalton-technische Details wie die Wahl der Kanalbelegung als auch die Art der Verkoppelung von Bild und Ton sowie der Weg nicht nur vom Ton-Recorder in den Bildschnitt, sondern auch von dort in die Ton-Postproduktion, unter Beibehaltung von sinnvoll gewählten Metadaten über diesen gesamten Weg.

Dies alleine ist bei jedem Projekt und bei jeder neuen Kombination von Kamera-Art, Drehbedingungen, Ton-Recorder-Typ, Bildschnitt-System und Ton-Schnitt-System aufs neue eine Herausforderung, deren Bewältigung durch Tests bereits vor jedem Dreh erfolgen sollte. Denn scheitert ein Workflow, kann das erhebliche Auswirkungen auf die Sprachverständlichkeit haben. Es nützt nichts, acht Spuren kristallklaren O-Ton aufzuzeichnen, wenn dieses Material hinterher nicht mehr unter vertretbarem Aufwand synchron mit dem Bild verkoppelt werden kann, und stattdessen das Kamera-Mikrofon verwendet werden muss.

Dieser Schritt im Ablauf ist also nicht weniger wichtig als alle folgenden, wie das Mikrofon so nah wie möglich an die Protagonisten zu bekommen, zumindest aber innerhalb des Hallradius, diesen Abstand über eine Szene möglichst wenig zu verändern, um die Schnittmöglichkeiten möglichst wenig einzuschränken, und die Sprache der Protagonisten über einen verlässlichen Übertragungsweg innerhalb der Systemdynamik des Recorders aufzuzeichnen.

Geangelter, drahtgebundener Ton ist dabei nach wie vor das Mittel der Wahl, sowohl bei Spiel- als auch Dokumentarfilm, lässt er doch im Gegensatz zu per Funk übertragenem Anstecker-Ton dynamische Reaktionen auf das Verhalten der Protagonisten zu, sowohl was die Positionierung des Mikrofons angeht, als auch, was die Mikrofonverstärker-Einstellungen betrifft; von der Robustheit des Übertragungsweges per Kabel ganz zu schweigen.

Wie bereits angedeutet, lassen manche Drehsituationen die Positionierung einer oder mehrerer drahtgebundener Angeln gar nicht erst zu, häufig aus optischen Gründen, beim Dokumentarfilm mit mehreren Protagonisten schon aus logistischen Gründen, also beispielsweise wegen nicht im Voraus zu beantwortenden Frage, wer wann und wie weit entfernt von der tonmeisterlichen Position spricht.

Hier müssen dann per Funk übertragene Ansteck-Mikrofone verwendet werden, deren Vorverstärkung am Sender während der eigentlichen Drehsituation nicht mehr geändert werden kann; die anhaltende Ausbreitung von drahtloser Datenübertragung zu Lasten der für reine Sprachübertragung verfügbaren Frequenzbereiche im Rahmen der Digitalen Dividende I und II hat seit 2010 für viele O-Tonmeister*innen schmerzlich spürbar gemacht, wie anfällig für Störungen dieser Übertragungsweg überdies ist. Von einstreuenden Funk-Schärfen der Kameraabteilung und ähnlichem Ungemach ganz zu schweigen.

An typischen Begrenzungsfaktoren für Sprach-verständlichkeit im Film- und Fernsehbereich fallen also schon am Drehort nach den bisherigen Betrachtungen für Spiel- wie Dokumentarfilm an:

- gescheiterter Workflow,
- emotionale, aber in der Sprachverständlichkeit reduzierte Performance von Protagonisten
- Umgebungslärm aus der weiteren Umgebung (meist Motorenlärm),
- Lärm durch das Drehteam und/oder deren Maschinen (Sprache, Schritte, Surren, Pfeiffen, Brummen, etc.),
- Lärm durch Requisiten und/oder Kostüme,
- ungünstige Akustik (Resonanzen, lange Nachhallzeiten),
- „offiger“ Angel-Mikrofon-Ton durch Mehrkamera-Auflösungen,
- verraschelte Ansteckmikrofone,
- gestörte Funkübertragungen (Rauschen, Aussetzer).

Was bedeutet das nun für die Sprachverständlichkeit in der dem Dreh folgenden Stufe in der Filmproduktion, die Postproduktion?

3. Sprachverständlichkeit in der Postproduktion

3.1. Dialogschnitt – warum nicht immer gleich ADR?!

Einige der beschriebenen Störfaktoren für die Sprachverständlichkeit kann die/der Bild-Editor*in und/oder der/die Dialog-Editor*in in der Postproduktion ausschalten: durch

- Restauration,
- Austausch durch Alternativtakes/-Versatzstücke oder – im Falle einer Spielfilm-Produktion -
- Nachsynchronisation (ADR, engl. für automated dialogue replacement, automatisiertes Dialog-Ersetzen).

Dies immer natürlich mit dem Risiko des Verlustes von emotionalem Gehalt gegenüber dem Original-Take, gelegentlich (wenn auch selten) natürlich auch mit der Chance, eine optisch gute, aber von der sprachlich-emotionalen Ebene her ungünstige Einstellung nachträglich aufzuwerten.

Dass ADR beim Dokumentarfilm in der Regel nicht möglich ist, erklärt sich von selbst. Warum aber werden Originalton-Stellen mit schlechter Sprachverständlichkeit in einem Spielfilm nicht immer und ohne Zögern durch Nachsynchronisation ersetzt, wenn eine Rettung durch die Restauration oder das Einsetzen von Alternativtakes nicht möglich ist?

Um das zu verstehen, lohnt es sich, die ADR-Situation genauer anzuschauen: Darsteller*innen, die sich zuvor möglicherweise wochenlang auf mehrere Drehtage vorbereitet hatten, in dieser Zeit immer mehr in ihrer Film-Rollen aufgegangen sind, und während des eigentlichen Drehs jede ihrer Einstellung als Adrenalin-treibenden Moment unter hohem Druck und vor einem ganzen Film-Team als Zuschauer erleben, haben zum üblichen Zeitpunkt von Synchron-Aufnahmen Monate später bereits eine längere Zeit ohne weitere Beschäftigung mit ihrer damaligen Rolle hinter sich. Sie befinden sich vielleicht schon im nächsten Projekt, und werden nun eines Tages quasi „aus heiterem Himmel“ angerufen und gebeten, sich an einem bestimmten Tag zu einer bestimmten Zeit in einem bestimmten Tonstudio einzufinden. Bestenfalls kennen sie diesen Ort bereits aus vergangenen Projekten; schlimmstenfalls ist ihnen die gesamte Situation unbekannt, in einem akustisch stark entkoppelten, nachhallarmen Raum, alleine vor einem Mikrofon, zu einem ihnen zu dem Zeitpunkt noch unbekannten Bild ihrer selbst, synchron den gleichen (oder einen optimierten) emotionalen Ausdruck bei erhöhter Sprachverständlichkeit abzurufen und vorzuführen.

Dazu werden sie in dieser Situation häufig von Regisseur*innen angeleitet, die diesen Moment schon seit Wochen mit Nervosität, manche mit Angst, wenige mit Vorfreude, auf sich und „ihren Film“ zukommen sehen, an dem sie ja nun schon möglicherweise seit Jahren arbeiten, wenn man die Drehbuch-Phase mitrechnet. Diese Regisseur*innen haben sich in der Bildschnitt-Phase häufig sehr stark an den emotionalen Ausdruck einer sprachlich zu ersetzenden, am Set vermutlich hart erarbeiteten Szene gewöhnt, leider auch an die schlechte Tonqualität, und gehen deswegen nun um so widerwilliger, teilweise mit Unverständnis und nur auf Drängen der/des Dialog-Editoren/in in die Synchron-Aufnahmesituation.

Diesem Gemisch aus Nervosität und tendenziell destruktiver Stimmung auf allen Seiten werden die wenigsten Darsteller*innen innerhalb eines so kurzen Zeitraums wie dem der ADR-Aufnahmesituation eine Performance entgegensetzen können, die der zu ersetzenden Aufnahme in emotionalem Gehalt gleichwertig oder überlegen ist.

Es ist also nicht verwunderlich, dass nicht reihenweise Spielfilme bei dem geringsten Verdacht auf verminderte Sprachverständichkeit nachsynchronisiert werden, sondern ADR als Mittel schon aus diesen Gründen gerne gemieden wird. Und darüber hinaus werden die zusätzlich anfallenden Kosten für Studio-Miete, ADR-Tonmeister*innen- und Editor*innen-Gagen und Reisekosten der Darsteller*innen von den Produktionsleiter*innen nur zu gerne eingespart, wenn sie denn im Budget überhaupt jemals eingeplant waren.

Überall dort nun, wo

- der emotionale Inhalt über die technische Sprachverständlichkeit gestellt wird, wo
- Dialog-Editor*innen keine brauchbaren Alternativ-Takes zur Verfügung haben und dann auch noch sprachverständlich Kritisches nicht durch ADR ersetzt werden darf, oder wo
- aus zeitlichen Gründen keine Möglichkeit oder einfach auch
- kein Budget für Restauration, Austausch oder Nachsynchronisation vorhanden ist,

kommt es bereits zu diesem Zeitpunkt bei der Erstellung der Filmtonspur zur Beeinträchtigung der Sprach-verständlichkeit.

Weitere Beeinträchtigungen zu verhindern, ist eine der Aufgaben aller weiteren Filmton-Gestalter*innen.

3.2. Sounddesign und Musik

Sprache von den Darsteller*innen am Drehort oder aus dem Synchron-Studio ist zwar ein sehr wichtiger Teil der Filmtonspur, aber natürlich bei weitem nicht der einzige. Sie wird durch weitere gestalterische und dramaturgische Elemente ergänzt:

Zum einen wären da weitere Sprache in Form von zusätzlichen Menge-/Masse-Sprach-Aufnahmen (für Off-Screen-Hintergründe und für während des Drehs ursprünglich textlose Statistenrollen, deren Sprache in der finalen Tonmischung meist unverständlich bleiben wird, aber zumindest als Klangfarbe erwünscht ist) und Voice-Overs, die nach dem Dreh in Tonstudios aufgenommen werden.

Darüber hinaus gibt es neben der Sprache die (sich nicht nur zeitlich, sondern auch teilweise inhaltlich überschneidende) Ebenen Foleys, Effekte, und Atmos, die zusammen das Sounddesign bilden, sozusagen die akustische Szenographie, sowie die Musik, die sich ihrerseits inhaltlich zum Teil mit dem Sounddesign überschneidet.

3.3. Vor der Mischung und in der Mischung

All diese Elemente müssen nun zeitlich, spektral und spatial so angeordnet werden, dass sie sich in kritischen Punkten nicht verdecken. Und Sprache ist immer kritisch, wie auch eine alte Mischtonmeister-Regel sinngemäß besagt: „Wichtig bei der Filmtonmischung ist zuerst der Dialog, dann der Dialog, dann die Musik, dann der Dialog, und dann (eventuell) das Sounddesign.“

Oder, um es mit dem französischen Film-Theoretiker Michel Chion zu sagen: „Sound in film is voco- and verbocentric, above all, because human beings in their habitual behaviour are as well. When in any given sound environment you hear voices, those voices capture and focus your attention before any other sound (wind blowing, music, traffic). Only afterward, if you know very well who is speaking and what they‘re talking about, might you turn your attention from the voices to the rest of the sounds you hear. So if these voices speak in an accessible language, you will first seek the meaning of the words, moving on to interpret the other sounds only when your interest in meaning has been satisfied.“ [3]. Kurz: Dialog führt.

Komponist*innen, Sounddesigner*innen und Mischtonmeister*in müssen also so gestalten, dass die Sprache nicht von musikalischen oder tongestalterischen Effekten überlagert wird – zumindest an Stellen, an denen es auf Sprachverständlichkeit ankommt.

Vergleicht man Filme aus den 1960er Jahren mit Filmen aus den letzten zehn bis zwanzig Jahren, merkt man: die Filmtonspur ist - bedingt auch durch die technischen Möglichkeiten wie verbessertem Signal-Rausch-Abstand und Möglichkeiten der spatialen Platzierung - in den letzten Jahrzehnten inhaltlich immer dichter geworden.

Für diese Art der verdichteten Gestaltung ist es hilfreich, ein Wiedergabeformat zur Verfügung zu haben, das jenseits von Mono, jenseits von Stereo, jenseits von Surround arbeitet. Je mehr Lautsprecher zur Verfügung stehen, um Schallquellen zu positionieren, desto leichter ist klangliche Transparenz zu erreichen. Diese Arbeit der spatialen Platzierung wird schon länger nicht mehr erst durch Mischtonmeister*innen erledigt, sondern ist - auch angesichts schrumpfender Ton-Postproduktions-Budgets, bei denen die Miete für das Misch-Atelier einer der größten Kostenfaktoren ist – schon seit längerer Zeit Teil der Arbeit von Sounddesigner*innen und Komponist*innen geworden, die am Ende im Mischatelier häufig nur noch finalisiert wird.

Hat der/die Dialog-Editor*in nun also gute Arbeit geleistet, und aus den Roh-Aufnahmen vom Set und dem vom Bild-Editorial geschnittenen Ton zusammen mit eventuellen weiteren Sprachaufnahmen eine verständliche Dialog-Spur gebaut und in die Mischung gebracht, ist es die Aufgabe aller anderen Filmtonschaffenden, diese Verständlichkeit nicht aufs Neue zu zerstören.

Zusammengefasst wird es für die Sprachverständlichkeit in der Ton-Postproduktion in den folgenden Situationen schwierig:

bei

- zeitlicher Überschneidung von Musik und/oder Tongestaltung mit Dialog,
- spektraler Überschneidung von Musik und/oder Tongestaltung mit Dialog,

- spatialer Überschneidung von Musik und/oder Tongestaltung mit Dialog.

Vor allem gleichzeitige Überschneidungen in mehr als einem dieser Parameter sind im Sinne einer optimalen Sprachverständlichkeit nach Möglichkeit zu vermeiden.

3.4. Demokrit vs. Grillparzer

„Ständige Arbeit wird leichter durch Gewöhnung“ [4] ist als Zitat des griechischen Philosophen Demokrit überliefert. Gewöhnung nun gerade ist aber ein großer Feind der Herstellung und/oder Erhaltung von Sprachverständlichkeit – hier gilt eher Grillparzers „Ach die Gewohnheit ist // Ein lästiges Ding, selbst an Verhaßtes fesselt sie!“ [5].

Denn genau wie Zuschauer*innen sehen Dialog-Editor*innen einen Film nur einmal zum ersten Mal. Aber während sich für die Zuschauer*innen bei diesem ersten Mal jede Filmszene unbewusst in „sprachlich verständlich“ oder „sprachlich nicht verständlich“ aufteilt, sich also entscheidet, ob alle entscheidenden Sprach-Informationen verstanden werden können, sehen und hören Dialog-Editor*innen die Sprache jeder Einstellung des Filmes während der Schnitt-Arbeit gefühlt unzählige Male. Was beim ersten Sehen vielleicht nicht zu verstehen war, erklärt sich beim zweiten Mal, möglicherweise schon durch den jetzt bekannten Verhandlungsablauf.

Ein Gewöhnungseffekt lässt sich also nicht vermeiden, behindert aber die Fähigkeit zur Einschätzung von Sprachverständlichkeit sehr stark. Die aufkommenden Möglichkeiten zur technischen Messung von Sprachverständlichkeit in der Postrpoduktion können hier eine wichtige Hilfe darstellen.

Die Problematik der Gewöhnung betrifft auch und gerade auch Mischtonmeister*innen. Am Ende tragen sie die Verantwortung dafür, die Tonspur so zu gewichten, dass sowohl nicht-sprachliche Elemente in der zu erwartenden Abhörumgebung der Zuschauer*innen zur Geltung kommen, als auch, dass eine verständliche Sprachspur zu hören ist.

Auch hier können neue referenzsignalfreie Messmethoden für Sprachverständlichkeit auf Basis von neuronalen Netzen [6] eine große Hilfe darstellen. Am wichtigsten aber ist die aurale Prüfung des Ergebnisses unter möglichst gleichen Abhörbedingungen wie bei der Mehrzahl der Zuschauer*innen.

4. Nach der Postproduktion

Und das ist ein ganz entscheidender Aspekt: die bei den Zuschauer*innen herrschenden Abhörbedingungen liegen nicht mehr in der Hand der Filmton-Gestalter*innen, und lassen sich in ihrer Vielseitigkeit nur begrenzt simulieren. Was im Kinobereich trotz eigentlich existierender Normen / Zertifizierungen schon schwierig genug ist, ist für den Fernsehbereich fast unmöglich: wenn am Ende ein Flachbild-Fernseher seinen „Klang“ in ein akustisch nicht behandeltes Wohnzimmer mit einem Hallradius von 1m gibt, und die Zuschauer*innen 2m entfernt auf dem Sofa sitzen, ist das der Sprachverständlichkeit nicht zuträglich. Das um so mehr, wenn die Hörfähigkeit der Zuschauer*innen im Höhenbereich altersbedingt bereits herabgesetzt ist.

Die beliebte Spruch „Fix it in the mix“, „Reparier es in der Mischung“, erweist sich hier einmal mehr als sinnbefreit.

5. Zusammenfassung

Sprachverständlichkeit bei Spiel- und Dokumentarfilm ist das Resultat von Teamarbeit, einer Teamarbeit, bei der die tonmeisterliche Arbeit nur ein Teilaspekt ist. Sie betrifft Kamera wie Ausstattung, Requisite wie Maske, Licht wie Kamerabühne, also alle am Film beteiligten Departments.

Entscheidend für ein optimales Ergebnis ist bei dieser Teamarbeit frühe und umfassende Kommunikation der Tonabteilung durch alle beteiligten Departments. Dafür ist aber auch eine frühe und möglichst umfassende Einbindung mindestens des Set-Ton-Departments, möglichst jedoch des gesamten Ton-Departments inkl. der Postproduktions-Abteilung, notwendig.

Trotz der erfreulichen Entwicklung von Algorithmen zur Messung von Sprachverständlichkeit gilt nach wie vor: diese Algorithmen haben ihre Grenzen. Sie können helfen, bei Gewöhnung und unter Zeitdruck leichter zu entscheiden, ob der Dialog-Anteil einer Mischung zeitlich, spektral und spatial ausreichend getrennt ist von anderen Ton-Elementen, und damit ausreichend verständlich. Was sie nicht messen können, sind die emotionale Aussagekraft von Tongestaltung. Und das ist die Kraft, die Tongestaltung im Film unmittelbarer noch als Bildgestaltung hat: emotionale Einbindung, die Immersion des Zuschauers.

Vom deutschen Film-Mischtonmeister Milan Bor wird der Ausspruch kolportiert, dass der amerikanische Zuschauer den Film verstehen wolle, der deutsche Zuschauer eher den Dialog.

Abb. 3: Schlechte Abhörbedingungen in einem Behelfskino der Deutschen Luftwaffe während eines Auslandseinsatzes.

Das ist eine natürlich überspitzte Beschreibung einer Situation, in der der Schwanz mit dem Hund wedelt, in der die „Kju-Ssih“, wie man neudeutsch sagt, also die QC,

Quality Control, zu deutsch Qualitätskontrolle, das in der Tongestaltung angelegte Lagerfeuer wegen „Knacksern im Soundtrack“ zurückgehen lässt, also nur auf das Signal, aber nicht auf den Inhalt schaut.

Entscheidend ist für jede Tongestaltung immer der Kontext, in dem sie geschieht: wenn das Ausdrücken bestimmter Emotionen im gegebenen Zusammenhang wichtiger ist als die Sprachverständlichkeit, ist auch eine textlich unverständliche Dialogstelle verwendbar und hinnehmbar. Diese inhaltlich-dramaturgische Entscheidungen können Algorithmen nicht - zumindest noch nicht – fällen. Sie müssen deswegen von Originaltonmeister*innen, Dialogeditor*innen und letzten Endes von den Mischtonmeister*innen getroffen werden.

6. Literatur

[1] Chion, M.: Audio-Vision. Columbia University Press, New York, 2004, S. 5

[2] Berufsbilder im Filmton auf der Homepage der Berufsvereinigung Filmton e.V., URL: https://www.bvft.de/info/berufsbilder/

[3] Chion, M.: Audio-Vision. Columbia University Press, New York, 2004, S. 6

[4] Demokrit von Abdera: Fragment 241.

[5] Grillparzer, F.: Sappho. Wallishausser, Wien, 1819, S.80

[6] R. Huber, J. Ooster, B. T. Meyer: Single-Ended Speech Quality Prediction Based on Automatic Speech Recognition. Journal Audio Engineering Society, Jg. 66, Nr. 10, 2018, S.759–769

Verständlichkeit von Fernsehton im Tonkanalformat 3.0 – Einfluss verschiedener Abhörpegel auf das eingestellte Sprache-Hintergrund-Verhältnis

Sebastian Goossens[1], Theresa Liebl[1]

[1] *Institut für Rundfunktechnik GmbH, München, Germany*

Abstract

Um die Sprachverständlichkeit im Fernsehen zu verbessern, wird auch die Einführung eines Tonkanalformats 3.0 diskutiert. Bei einer ordentlichen 3.0 Mischung sollte der Center-Kanal ausschließlich Sprache enthalten. Nach ersten Einschätzungen könnte dadurch eine verbesserte Sprachverständlichkeit (bzw. eine verminderte Höranstrengung) gewährleistet werden.

Um Auswirkungen dieses Formats auf die Sprachverständlichkeit bei den Zuschauern zu evaluieren, wurden am IRT Vergleichstests durchgeführt. Dabei zeigte sich, dass bei einer 3.0-Audiomischung eine manuelle Steuerung des Center-Kanal-Pegels genutzt wird, um die Sprachverständlichkeit von TV-Inhalten zu verbessern. Ebenso hat die Abhörlautstärke einen Einfluss auf den gewählten Center-Kanal-Pegel.

Es wurden drei unterschiedliche Abhörpegel (50, 60 und 70 dB(A)) angeboten. Auf einer Soundbar und einem 3.0-Wiedergabesystem sollte der Center-Pegel einer 3.0-Mischung so eingestellt werden, dass eine optimale Sprachverständlichkeit erreicht wird. Die Ergebnisse zeigen deutlich, dass bei einem geringeren Abhörpegel der Pegel des Center-Kanals im Vergleich zur Originalmischung deutlich angehoben wurde.

Dies zeigt die Wichtigkeit eines separaten Sprachkanals in TV-Produktionen. Eine praktische Umsetzung ist aktuell im Rundfunk allerdings schwierig, da ein reiner Sprachkanal meist nicht vorliegt und auch nicht übertragen werden kann. Zudem erlauben die meisten Endgeräte keine einfache Pegelanhebung des Centerkanals. Gelöst werden könnte dieses Problem langfristig durch eine objektbasierte Audioproduktion und Übertragung.

1. Einführung

Die Sprachverständlichkeit des Fernseh-Programms ist für die öffentlich-rechtlichen Rundfunkanstalten ein wichtiges Thema. Seit einigen Jahren werden die Zuschauerbeschwerden dazu immer mehr, was für die Rundfunkanstalten einen dringenden Handlungsbedarf bedeutet. Es werden verschiedenste Ansätze verfolgt, um das Problem zu lösen.

Bei diesen Bemühungen die Sprachverständlichkeit im Fernsehen zu verbessern wird auch die Einführung eines Tonkanalformats 3.0 diskutiert. Bei einer 3.0-Mischung sollte, laut ARD/ZDF Sprachverständlichkeits-Empfehlung [3], der Center-Kanal möglichst nur Sprache enthalten. Nach ersten Einschätzungen könnte allein dadurch eine verbesserte Verständlichkeit von TV-Programmen auf dem richtigen Wiedergabesystem gewährleistet werden. Eine 3.0-Produktion wäre weniger aufwendig als beispielsweise eine 5.1-Produktion, bei der dieser positive Effekt auch gegeben wäre. Auch für die Zuschauer wäre eine 3.0-Wiedergabe weniger aufwendig, da beispielsweise eine einfache Wiedergabe über Soundbars möglich wäre.

Um zu evaluieren, welche Auswirkungen auf die Sprachverständlichkeit eine 3.0-Audiomischung bei den Zuschauern zu Hause wirklich hat und ob mit 3.0 eine Verbesserung der Verständlichkeit von TV-Programmen bzw. eine Verminderung der Höranstrengung erzielt werden kann, wurde am IRT eine Versuchsreihe mit mehreren Hörversuchen durchgeführt.

2. Erste Versuchsreihe zum 3.0-Format

Auf der DAGA 2017 wurde eine Untersuchung des IRT präsentiert, die sich mit der Sprachverständlichkeit des Tonkanalformats 3.0 im Fernsehen beschäftigt hat [1].

2.1. Vergleich 3.0 zu Stereo

Bei dieser Untersuchung wurden verschiedene Varianten von 3.0-Mischverfahren - zwei Upmix-Verfahren und ein 3.0-Handmix - mit einer Stereomischung verglichen.

Die Ergebnisse zeigten, dass keine eindeutige Tendenz zu einer der untersuchten 3.0-Varianten oder der Stereo-Mischung vorhanden ist. Allerdings ist hierbei zu beachten, dass in der Untersuchung rein nach der Höranstrengung und der Sprachverständlichkeit, nicht nach dem Gesamteindruck oder dem Klangbild gefragt wurde. Diese Aspekte könnten durchaus ein anderes Ergebnis bewirken. Es ist außerdem zu berücksichtigen, dass sich die Probanden bei der Untersuchung im Sweet Spot der 3.0-Lautsprecheraufstellung befanden und sich somit auch bei der Stereo-Wiedergabe der Center-Kanal in der Bildmitte befand. Bei einer Abhörposition außerhalb des Sweet Spots ist davon auszugehen, dass der Zuhörer bzgl. der Sprachverständlichkeit der Mischungen eher von dem festen Center-Kanal der 3.0-Mischungen profitieren würde.

ISBN
978-3-9812830-9-9

Der Autor hält die Urheberrechte für den gesamten Inhalt seines Manuskripts. Er überlässt dem VDT das Recht auf Veröffentlichung in gedruckter Form, Online oder als PDF-Download von https://tonmeistertagung.com

2.2. Vergleich der Wiedergabesysteme

Bei dieser Untersuchung wurde die Verständlichkeit einer 3.0-Mischung auf verschiedenen Wiedergabesystemen - drei Soundbars, zwei 3.0-Lautsprecheraufstellungen und einem Flachbildfernseher - untersucht.

Bezüglich der Sprachverständlichkeit bzw. der Höranstrengung gab es keine eindeutige Präferenz zu einem der untersuchten Wiedergabesysteme.

Bei diesem Ergebnis ist allerdings zu beachten, dass viele Probanden Probleme mit der reinen Bewertung der Verständlichkeit hatten. Die verschiedenen Wiedergabesysteme wiesen im Vergleich deutliche Unterschiede in der Klangqualität auf, was einen großen Einfluss auf den Gesamteindruck hatte und die Bewertung erschwerte.

2.3. Centerpegel

Bei dem dritten Versuch wurde überprüft, ob eine manuelle Steuerung der Lautstärke des Center-Kanals von den Probanden genutzt wird, um eine Verbesserung der Verständlichkeit bzw. eine Verminderung der Höranstrengung zu erzielen. Es wurde davon ausgegangen, dass der Center-Kanal einer 3.0-Mischung größtenteils Sprache enthält und dass durch eine Anhebung des Pegels eine verbesserte Verständlichkeit bewirkt werden kann.

Im Ergebnis wurde für alle Testsequenzen eine deutliche Anhebung des Center-Kanal-Pegels vorgenommen. Im Schnitt über alle Testsequenzen lag diese Anhebung bei 2,6 dB.

Bei der Auswertung dieser Untersuchung gab es Anzeichen dafür, dass die Abhörlautstärke einen Einfluss auf den gewählten Pegel des Center-Kanals haben könnte. Auch eine in 2014 durchgeführte Untersuchung zur Verständlichkeit von „Voice-Over-Voice“-Passagen im Fernsehen [2] zeigte einen deutlich erkennbaren Einfluss unterschiedlicher Abhörpegel auf das gewünschte Mischungsverhältnis zwischen zwei Sprachspuren in Übersetzungspassagen.

Daher wurde entschieden, den Versuch zur Anpassung der Lautstärke des Center-Kanals einer 3.0-Mischung noch einmal mit verschiedenen fest eingestellten Abhörpegeln zu wiederholen, um die sich abzeichnende Tendenz zu bestätigen.

3. Versuch zum Einfluss der Abhörpegel

Aufbauend auf den bereits durchgeführten Untersuchungen sollte ein weiterer Hörversuch den Einfluss der Abhörlautstärke bei der manuellen Einstellung einer gut verständlichen Mischung im Tonkanalformat 3.0 in einer Wohnzimmerumgebung deutlich machen. Dazu sollte bei drei unterschiedlichen Abhörpegeln die gewünschte Lautstärke des Center-Kanals so eingestellt werden, dass ohne Höranstrengung eine gute Verständlichkeit der Mischung gewährleistet ist.

3.1. Welche Wiedergabepegel sind relevant?

Hierfür musste geklärt werden, in welchem Bereich die Abhörlautstärke beim Versuch variiert werden sollte. Im Jahr 2014 waren in 21 europäischen Studios bzw. Regieräumen die bei der Produktion eingestellten Abhörpegel gemessen worden. Die Ergebnisse lagen zwischen 64 und 77 dB(A). Im Mittel wurde bei der Produktion ein Pegel von 73 dB(A) eingestellt.

Des Weiteren wurden im Fernsehvorführraum des IRT (ITV – Showroom) mit 10 Personen Pegelmessungen durchgeführt. [2] Der Altersdurchschnitt der Teilnehmer lag bei dieser Untersuchung bei 42,7 Jahren, wobei eine Bandbreite von 23 bis 58 Jahren abgedeckt wurde.

Den Teilnehmern wurden jeweils 3 verschiedene Programme dargeboten und sie sollten die Lautstärke des Fernsehers so einstellen, wie es ihnen angenehm erschien. Bei dieser eingestellten Lautstärke wurde dann mit einem Audio- und Akustik-Analysator (NTI Audio XL2) an der Position des Hörers eine Pegelmessung vorgenommen. Die daraus resultierenden Werte für jedes Programm und jeden Probanden wurden anschließend gemittelt, woraus sich die mittlere Abhörlautstärke von 51 dB(A) ergab.

Aufgrund dieser Erkenntnisse wurden für den neuen Versuch 50, 60 und 70 dB(A) als feste Wiedergabepegel gewählt, da 50 dB(A) die ungefähre Abhörlautstärke für TV-Inhalte in einer Wohnzimmerumgebung ist und 70 dB(A) etwa die durchschnittliche Abhörlautstärke beim Mischen im Produktionsstudio.

3.2. Versuchsrahmen

Die Anpassung der Mischungsverhältnisse wurde auf zwei verschiedenen Wiedergabegeräten durchgeführt, welche in der ersten Versuchsreihe [1] bereits verwendet wurden. Ein 3.0-Heimkinosystem von B+W und eine Soundbar von Yamaha. Die gewählte Soundbar war die einzige aus einer Reihe von Testgeräten, welche die manuelle Ansteuerung des Center-Kanals ermöglichte. Bei der 3.0-Anlage konnte die Ansteuerung über den angeschlossenen A/V-Receiver erfolgen.

Abb. 1: Hörversuchsumgebung; Aufstellung der Wiedergabegeräte

Durchgeführt wurde der Hörversuch in einem Raum, der akustisch einem guten Wohnzimmer angepasst wurde. Das 3.0-Lautsprechersystem wurden in einem Stereodreieck

aufgestellt, die Soundbar wurden in einem Regal unter einem Fernseher platziert, welcher für die optische Darbietung der Testsequenzen genutzt wurde. Die Anordnung der Wiedergabegeräte ist in Abb. 1 zu sehen.

Die Abhörposition während der Hörversuche befand sich im Sweetspot der 3.0-Aufstellung. Für die Hörversuche wurden die Wiedergabesysteme mit einem akustisch transparenten Stoff verdeckt, um die Versuchsteilnehmer optisch nicht zu beeinflussen.

Gesteuert wurde der Hörversuch durch eine browser-basierte Testsoftware des IRT. Über einen Laptop konnten die Probanden den Befehl zur Wiedergabe der Testsequenzen ausführen und die Lautstärke des Center-Kanals verändern (siehe Abb. 2).

Als Versuchsmaterial wurde real ausgestrahltes TV-Material aus verschiedenen Genres verwendet. Bei der Auswahl wurde speziell auf eine schlechte Sprachverständlichkeit des Ausgangsmaterials und ein breites Spektrum an Inhalten geachtet. Es wurden fünf Testsequenzen zwischen 15 und 30 Sekunden Länge aus den Genres Sport, Dokumentation, Spielfilm und TV-Show ausgewählt.

3.3. Versuchsdurchführung

Jede Testsequenz wurde in den drei gewählten Abhörpegeln 50, 60 und 70 dB(A) sowohl auf der B+W-Anlange, als auch auf der Yamaha Soundbar wiedergegeben. So ergaben sich insgesamt 30 zu bewertende Sequenzen. Die Aufgabe der Probanden war es mit Hilfe eines stufenlosen Reglers (siehe Abb. 2) für jede Testsequenz das gewünschte Mischungsverhältnis zwischen dem Center und dem linken und rechten Kanal einzustellen. Es war dabei auf eine klar verständliche Sprache und eine geringe Höranstrengung zu achten. Die Startposition des Reglers war bei jeder Sequenz auf einer zufälligen Höhe. Die Laustärke des Center-Kanals konnte in einem Bereich zwischen -6 und +6 dB im Vergleich zur Originalmischung verändert werden.

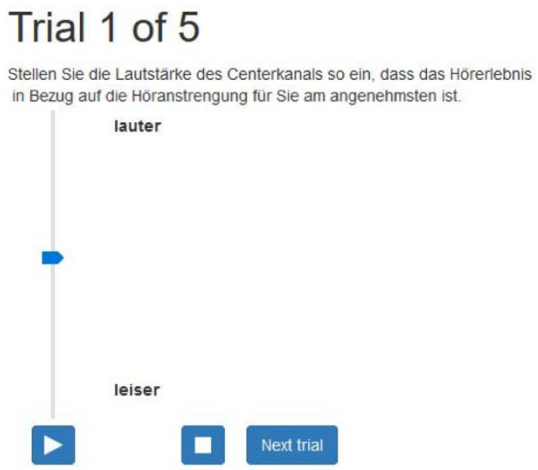


Abb. 2: Oberfläche der Hörversuchssoftware

Vor dem eigentlichen Versuch wurde mit jedem Probanden ein kurzes Training durchgeführt, um vorab einen Einblick in das Testverfahren zu geben.

An dem Hörversuch haben 20 Personen zwischen 23 und 60 Jahren mit normalem Hörvermögen teilgenommen.

3.4. Ergebnisse

In Abbildung 3 und 4 sind die Ergebnisse des Hörversuchs in Form von Boxplot-Diagrammen dargestellt. Die Grafiken zeigen die eingestellten Center-Kanal-Pegel für jede Testsequenz auf der Yamaha Soundbar (Abb. 3) und der 3.0-Anlage von B+W (Abb. 4), gemittelt über alle Teilnehmer für die drei Abhörpegel (50 dB(A) in blau, 60 dB(A) in rot und 70 dB(A) in grün). Die rot-gestrichelte Linie in der Mitte, am 0-Punkt auf der y-Achse, gibt die Original-Mischungsverhältnisse an.

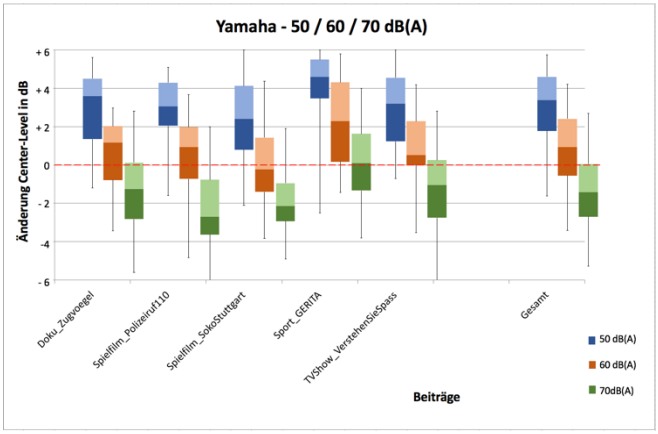


Abb. 3: Ergebnisse des Hörversuchs als Boxplot-Diagramm; Eingestellter Center-Kanal-Pegel auf der Yamaha Soundbar für die drei untersuchten Abhörpegel 50, 60 und 70 dB(A) und alle Testsequenzen

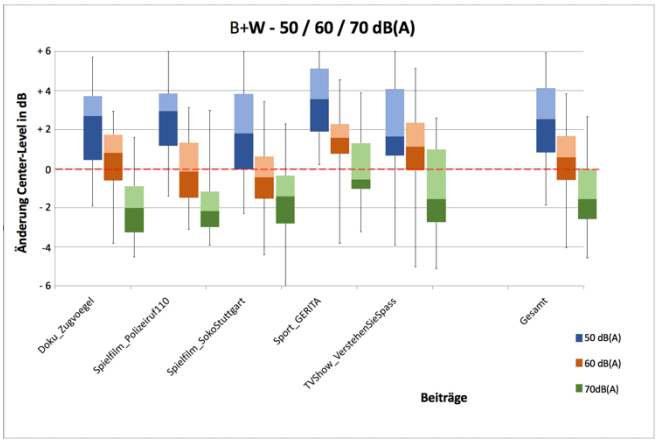


Abb. 4: Ergebnisse des Hörversuchs als Boxplot-Diagramm; Eingestellter Center-Kanal-Pegel auf der 3.0-Anlage von B+W für die drei untersuchten Abhörpegel 50, 60 und 70 dB(A) und alle Testsequenzen

Tabelle 1 zeigt die Mittelwerte der eingestellten Mischungsverhältnisse für die drei Abhörpegel auf beiden

Wiedergabesystemen, gemittelt über alle Versuchsteilnehmer und Testsequenzen. An den Werten in der Tabelle ist zu sehen, dass auf beiden Wiedergabegeräten ähnliche Ergebnisse erzielt werden konnten und dass sich eine eindeutige Tendenz abzeichnet. Bei einem geringeren Abhörpegel wurde der Pegel des Center-Kanals im Vergleich zu Originalmischung deutlich angehoben.

Wiedergabesystem	Abhörpegel	Mittelwert
Yamaha Soundbar	50 dB(A)	+3,0 dB
Yamaha Soundbar	60 dB(A)	+0,8 dB
Yamaha Soundbar	70 dB(A)	-1,3 dB
B+W 3.0-Anlage	50 dB(A)	+2,4 dB
B+W 3.0-Anlage	60 dB(A)	+0,9 dB
B+W 3.0-Anlage	70 dB(A)	-1,3 dB

Tabelle 1: Gesamtmediane für die drei untersuchten Abhörpegel auf den beiden Wiedergabesystemen, gemittelt über alle Versuchsteilnehmer und Testsequenzen

Auch die beiden Boxplot-Diagramme zeigen diese Tendenz deutlich. Bei 50 dB(A) Schalldruckpegel am Abhörort wurde für alle Testsequenzen eine deutliche Anhebung des Center-Kanal-Pegels vorgenommen. Bei 60 dB(A) Abhörpegel verschieben sich die Werte bei allen Sequenzen etwas nach unten, der Center-Pegel bewegt sich hier näher am Null-Punkt, das heißt nahe der Originalmischung. Bei 70 dB(A) liegen die eingestellten Pegel-Werte des Center-Kanals sogar vermehrt unterhalb der Null-Linie. Dies ist bei beiden Wiedergabesystemen in gleicher Weise der Fall.

Zu erkennen ist auch, dass bei der Testsequenz aus dem Bereich Sport der Pegel des Center-Kanals auf beiden Wiedergabesystemen und bei allen Abhörpegeln etwas höher gewählt wurde als bei den anderen Testsequenzen. Dies konnte bereits in der ersten Untersuchung [1] mit variablem Abhörpegel festgestellt werden.

Die Ergebnisse des Hörversuchs zeigen deutlich, dass bei einer geringeren Wiedergabelautstärke der Pegel des Center-Kanals bei allen Inhalten deutlich angehoben wurde, um die Höranstrengung zu vermindern und die Sprachverständlichkeit zu verbessern. Eine bei höherem Abhörpegel erstellte Tonmischung kann im Studio also gut verständlich sein, bei einer leiseren Wiedergabe im Wohnzimmer wäre allerdings ein lauterer Center- oder Sprach-Kanal-Pegel nötig, um eine gute Sprachverständlichkeit zu gewährleisten. Um auch zu Hause gut verständliche Tonmischungen sicherzustellen, wäre es daher sinnvoll, Mischungen schon im Tonstudio bei einem niedrigen Pegel gegenzuhören und die Verständlichkeit der Sprache zu überprüfen.

Bei der Interpretation der Ergebnisse ist allerdings auch zu beachten, dass bei der Auswahl der Testsequenzen für die Untersuchung speziell auf eine schlecht verständliche Sprache geachtet wurde. Würde man optimal verständliche Tonsignale verwenden, könnten die Ergebnisse dieses Hörversuchs anders ausfallen. Für schlecht verständliches Material und eventuell auch für Zuschauer mit einer Hörminderung würde die Möglichkeit der Pegel-Anhebung des Center-Kanals jedoch einen deutlichen Vorteil bieten und das Fernsehprogramm für bestimmte Zuschauergruppen attraktiver machen.

4. Umsetzung der Ergebnisse

Neben dem reinen Mischungsverhältnis haben allerdings auch noch weitere Faktoren einen Einfluss auf die Verständlichkeit von TV-Inhalten [3]. Neben oftmals unsauber aufgenommener Sprache und zu lauten Hintergrundgeräuschen nehmen auch die Abhörbedingungen der Zuschauer einen erheblichen Einfluss. Unterschiedlichste Tonwiedergabegeräte kommen für Rundfunkinhalte zum Einsatz, die räumlichen Bedingungen in den Wohnzimmern beeinflussen die Wiedergabequalität und somit die Verständlichkeit und auch das individuelle Hörvermögen der Zuschauer ist ein ausschlaggebender Faktor.

Um allen unterschiedlichen Anforderungen bei den Zuschauern zu Hause möglichst gut gerecht zu werden und eine optimale Sprachverständlichkeit für alle gewährleisten zu können, wären individuelle Lösungen nötig. Eine Möglichkeit die Verständlichkeit von TV-Produktionen bei den Konsumenten verbessern zu können, wäre die Übertragung eines separaten Sprachkanals, welcher wirklich ausschließlich Sprache enthält.

Eine Umsetzung dieser Lösung ist aktuell allerdings schwierig, da ein separater Sprachkanal von Produktionen meist nicht vorhanden ist und auch nicht einfach gesendet werden kann. Zudem ist die manuelle Steuerung des Center-Kanals einer 3.0- oder auch 5.1-Mischung bisher nur auf wenigen Endgeräten möglich.

5. Objektbasierte Audioproduktion

Eine Lösung für diese Problemstellungen wäre eine Umsetzung mit objektbasierter Audioproduktion. Aktuell finden Fernsehproduktionen rein kanalbasiert statt. Dabei werden mehrere Audiospuren in einer DAW gemischt und in einer festen Kanalbelegung, welche für ein spezifisches Lautsprechersetup ausgelegt ist (z.B. Stereo oder 5.1), gespeichert (siehe Abb. 5). Jeder Audiokanal der fertigen Mischung muss von einem Lautsprecher wiedergegeben werden. Diese festgelegte Mischung wird zum Endnutzer übertragen, welcher am Ende so gut wie keine Möglichkeit mehr hat die Mischung an seine Bedürfnisse anzupassen.

Eine objektbasierte Audioproduktion hingegen bietet die Möglichkeit erstellte Mischungen individuell anzupassen.

Hierbei werden die Audiospuren in einer DAW oder einem Editor wie gewohnt gemischt. Gespeichert und übertragen wird die fertige Mischung dann allerdings nicht in einer festen Kanalbelegung, sondern in Form von Einzelspuren, sog. Objekten, mit verknüpften Metadaten als objektbasierte Audioszene. Die Metadaten enthalten die nötigen Informationen für die erstellte Mischung, wie Pegel und

Position der einzelnen Objekte. Ein Renderer übernimmt am Ende dann die Zusammensetzung der gespeicherten Szene in die gewünschte Mischung, angepasst auf das jeweils angeschlossene Wiedergabesystem. Dadurch kann gewährleistet werden, dass die empfangene und gehörte Mischung immer dem jeweiligen Format des Endgerätes entspricht, was unerwünschte Veränderungen in der erstellten Mischung durch Up- oder Downmix-Prozesse verhindert.

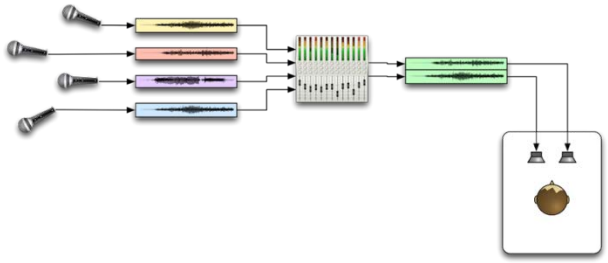


Abb. 5: Schematische Darstellung einer kanalbasierten Audioproduktion

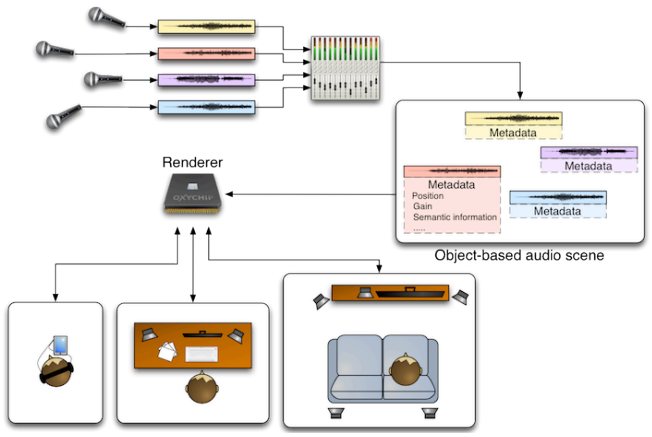


Abb. 6: Schematische Darstellung einer objektbasierten Audioproduktion

Für den hier angedachten Anwendungsfall könnte der Ersteller der Mischung in den Metadaten eines Objektes also festhalten, dass es sich um eine reine Sprachspur handelt, wo das Sprach-Objekt in der Szene platziert sein soll, wie hoch der Pegel dieser Sprachspur in der Mischung sein soll und, dass dem Endnutzer die Möglichkeit gegeben wird diesen Pegel in einem bestimmten Bereich zu verändern. Die Sprachspur wird also nicht fest mit den restlichen Anteilen der Mischung vermischt, sondern als eigenes Objekt in der Gesamtszene übertragen. Informationen über die Möglichkeit der Ansteuerung dieses Objektes können dann von der Produktion bis zum Zuschauer zu Hause mitgegeben werden und der Pegel je nach angeschlossenem Endgerät individuell angepasst werden. Aktuell kann eine solche Anpassung auf Smartphones bereits über eine App erfolgen, bei einer Wiedergabe am PC über einen im Browser integrierten Player. Bei der Wiedergabe über ein Fernsehgerät ist es allerdings vom jeweiligen Gerät abhängig, ob eine Ansteuerung eines einzelnen Objektes möglich ist. Standard-Fernsehgeräte haben dies noch nicht integriert, einige Hersteller treiben die Entwicklung dahingehend aber bereits voran.

Vorteil dieser Art der Produktion und Übertragung wäre also, dass Sprache als eigenes Objekt übertragen wird, Informationen zur Ansteuerung mitgegeben werden können und der Pegel zu Hause individuell angepasst werden kann. Was eine schnelle Umsetzung dieser Technik jedoch erschwert ist, dass eine Anpassung in der Aufnahme und Produktion stattfinden müsste und dass die Unterstützung einer Pegel-Steuerung von Einzelobjekten von den Endgeräten abhängig ist.

6. Fazit

Der zusätzlich durchgeführte Hörversuch mit verschiedenen Abhörlautstärken hat eindeutig gezeigt, dass diese einen deutlichen Einfluss auf das gewünschte Mischungsverhältnis zwischen Sprache und Hintergrund hat, wenn eine gute Verständlichkeit von Fernsehinhalten bzw. eine geringe Höranstrengung gewährleistet werden soll. Das „richtige“ Mischungsverhältnis ist offenbar von der Wiedergabe-lautstärke abhängig. Ein zusätzliches Abhören von Mischungen bei geringeren Pegeln im Studio und eventuelles Anpassen der Mischung könnte somit schon eine Verbesserung in der Verständlichkeit für die Zuschauer bewirken. Um jedoch für alle Zuschauer zu Hause eine geringe Höranstrengung gewährleisten zu können, wäre es nötig einen separaten sauberen Sprachkanal zu übertragen, mit der Möglichkeit diesen individuell im Pegel zu verändern. Die Umsetzung dieser Lösung wäre am effektivsten mit einer objektbasierten Audioproduktion und Übertragung möglich.

7. Literatur

[1] Liebl, T., Heinz, S., Hartmann, C., „Kann durch das Tonkanalformat 3.0 die Sprachverständlichkeit im Fernsehen verbessert werden?“. DAGA, Kiel, 2017

[2] Liebl, T., Goossens, S., Krump, G., „Verbesserung der Sprachverständlichkeit, speziell bei Voice-Over-Voice-Passagen“, Tonmeistertagung, Köln, 2014

[3] Sprachverständlichkeit im Fernsehen - Empfehlung für Programm und Technik
URL: `https://www.irt.de/de/publikationen/technische-richtlinien.html`

Automatische Überwachung der Sprachverständlichkeit im Rundfunkmaterial

Rainer Huber, Hannah Baumgartner, Niko Moritz, Stefan Goetze
Fraunhofer-Institut für Digitale Medientechnologie (IDMT), Institutsteil Hör-, Sprach- und Audiotechnologie, Oldenburg, Germany, Email: Rainer.Huber@idmt.fraunhofer.de

Abstract

Dieser Beitrag stellt einen referenzsignalfreien Ansatz zur Vorhersage von Höranstrengung von einkanaligen Mischungen von Sprache und Atmosphäre vor - Höranstrengung ist eine der Sprachverständlichkeit verwandte Größe. Der verwendete "single-ended" Ansatz basiert auf der Auswertung von Posteriogrammen, die durch ein System zur automatischen Spracherkennung erzeugt werden. Hintergrundgeräusche führen auch bei der automatischen Spracherkennung zu einer erhöhten Erkennungsunsicherheit - ähnlich wie bei menschlichem Verstehen. Diese Unsicherheit ist mit einem geeigneten Maß quantifizierbar und kann als Grundlage zur Bewertung der Höranstrengung dienen. In zwei Studien wurden Fernsehsignale von 20 Probanden hinsichtlich ihrer subjektiven Höranstrengung bewertet und mit den Ergebnissen der Posteriorgramm-Auswertung verglichen. Die Korrelationen zwischen menschlichen Bewertungen und Modellbewertungen für die beiden Datensätze liegen bei $r = 0.86$ bzw. $r = 0.90$. Die Methode ermöglicht somit eine automatische, objektive Überwachung der Höranstrengung von komplexen Sprachsignalmischungen.

1. Einführung

Die Sprachverständlichkeit von Rundfunkmaterial ist von vielen Faktoren abhängig: von der Aufnahme, der Bearbeitung, der Mischung, aber auch von individuellen Gehöreigenschaften, der heimischen Technik und den akustischen Gegebenheiten am Ort des Empfängers. Die Bewertung ist höchst subjektiv und die Problematik komplex. All diese Aspekte können in der Produktion sicherlich nicht Beachtung finden, allerdings ist eine grundsätzlich verständliche Mischung eine wichtige Voraussetzung bei der Qualitätssicherung von Rundfunkmaterial.

1.1. Verständlichkeit vs. Höranstrengung

Spachverständlichkeit wird gemeinhin als der Anteil verstandener Wörter in Prozent definiert. Doch auch wenn ein Zuschauer noch annähernd jedes Wort versteht, kann die für das Verstehen aufgewandte Höranstrengung als lästig empfunden werden. Es ist davon auszugehen, dass der typische Zuhörer sich in keiner Weise für das Verstehen von Text innerhalb einer Sendung anstrengen möchte, und noch mehr ist davon auszugehen, dass es im Fernsehen generell keine Passagen geben sollte, in denen tatsächlich ganze Textanteile nicht verstanden werden können. Insofern eignet sich die Größe „Sprachverständlichkeit“ schlecht zur Bewertung von Rundfunkmaterial. Eine der Sprachverständlichkeit verwandte Größe, die „Höranstrengung“ kann dagegen den für Broadcastanwendungen relevanten Bereich recht gut abbilden: Höranstrengung beschreibt eine typischerweise subjektiv kategorial beurteilte Empfindungsgröße. Beispiele für Bewertungsskalen zur Messung von Höranstrengung sind die von der International Telecommunication Union (ITU) in der Empfehlung ITU-T P.800 [1] beschriebene fünfstufige kategoriale Skala oder die am Hörzentrum Oldenburg von Schulte und Kollegen entwickelte 13-stufige Skale mit 7 verbalen Kategorien [2] (s. **Abb. 2**). Sprachverständlichkeit und Höranstrengung korrelieren miteinander (negativ), doch bietet die Höranstrengung gegenüber der Sprachverständlichkeit den Vorteil, noch zwischen Konditionen differenzieren zu

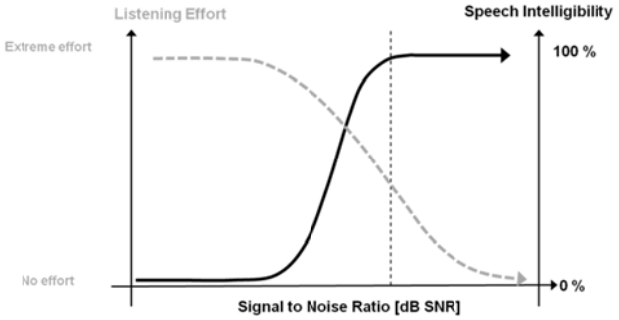


Abb. 1: Schematischer Verlauf von Höranstrengung („Listening Effort“, grau gestrichelt) und Sprachverständlichkeit („Speech Intelligibility“, schwarz durchgezogen) als Funktion des SNR. (Aus [3])

können, bei denen die Sprachverständlichkeit bereits (nahezu) 100% beträgt, so z.B. bei für Film- und Fernsehanwendungen realistischen, d.h. kleinen, positiven

ISBN
978-3-9812830-9-9
Der Autor hält die Urheberrechte für den gesamten Inhalt seines Manuskripts. Er überlässt dem VDT das Recht auf Veröffentlichung in gedruckter Form, online oder als PDF-Download von https://tonmeistertagung.com

Signal-zu-Rausch-Verhältnissen (SNR). **Abb. 1** (entnommen aus [3]) verdeutlicht dies anhand schematischer Verläufe von Höranstrengung und Sprachverständlichkeit als Funktionen des SNR.

1.2. Modellierung

Wie schon bei technischen („instrumentellen") Sprach- und Audioqualitäts-Messmethoden, so sind auch für die Messung von Höranstrengung prinzipiell sowohl referenzbasierte als auch referenzfreie Methoden zur Modellierung denkbar. Referenzbasierte (auch als „intrusiv" oder „double-ended" bekannte) Ansätze vergleichen Ein- und Aussgangssignal des zu bewertenden Systems miteinander, um dadurch ein Maß z.B. für die Klangqualität eines Audioübertragungssystems zu erhalten. Im Gegensatz dazu verwenden referenzfreie (auch als „nicht-intrusive" oder „single-ended" bekannte) Ansätze nur das Ausgangssignal des zu bewertenden Systems. Bezogen auf den Mischvorgang als zu bewertendes „System" lassen sich die Methoden folgendermaßen erläutern:

- ⇨ Referenzbasiert: Das reine Sprachsignal ist sowohl Eingangssignal als auch Referenzsignal des Systems, das gemischte Signal entspricht dem Ausgangssignal des Systems.
- ⇨ Referenzfrei: Nur das Ausgangssignal des zu bewertenden Systems - in unserem Beispiel das gemischte Signal - steht zur Verfügung.

In der Praxis stehen die Ursprungssignale der Mischung bzw. das reine Sprachsignal nicht immer zur Verfügung, daher sind referenzfreie Methoden universeller einsetzbar als referenzbasierte Methoden.

2. Methode

Dieser Beitrag beschäftigt sich mit der Anwendung eines referenzfreien Verfahrens zur Bestimmung der Höranstrengung von Rundfunkmischungen.

Grundsätzlicher Ansatz dieser Methode ist es, den Grad der Unsicherheit eines Systems zur automatischen Spracherkennung (ASR) bei der Erkennung von gestörter Sprache zu quantifizieren [4, 5, 6]. Das ASR-System verwendet ein tiefes neuronales Netz, um sogen. Phonem-posterior-Wahrscheinlichkeiten (oder kurz „Posteriorgramme") zu berechnen. Der menschlichen Wahrnehmung ähnlich, vermindern Störungen der Sprache wie Verzerrungen oder Hintergrundgeräusche die Erkennungssicherheit des ASR-Systems, was sich in „verschmierten" Posteriorgrammen widerspiegelt. Der Grad dieser Posteriorgrammverschmierung wird durch ein geeignetes mathematisches Maß, dem sogen. *M*-Measure von Hermansky und Co-Autoren [7], quantifiziert. (Einzelheiten zum ASR-System und dem *M*-Measure folgen in den Abschnitten 2.3 und 2.4; Illustrationen von Posteriorgrammen sind in [4] enthalten.) Es konnte gezeigt werden, dass das *M*-Measure sehr gut mit subjektiven Höranstrengungs-Messdaten verschiedener Datensätze korreliert [5, 6].

2.1. Höranstrengungsdaten – Datensatz I

2.1.1. Stimuli

Es standen getrennte Tonspuren (Sprache/Atmosphäre) von verschiedenen deutschen TV-Produktionen zur Verfügung. Aus diesem Material wurden 234 Audio-Clips mit einer mittleren Länge von ca. 10s herausgeschnitten und mit stark variierenden (z.T. auch unrealistischen) Mischungsverhältnissen gemischt, um einen möglichst großen Bereich der erwarteten Höranstrengung abzudecken.

2.1.2. Probanden und Bewertungsprozedur

20 normalhörende Probanden im Alter von 20-30 Jahren (Median = 25 Jahre), 10 davon männlich, 10 weiblich, nahmen an der Studie teil. Sie bewerteten die empfundene Höranstrengung der 234 Audio-Clips auf der in **Abb. 2** gezeigten Skala auf einem Touch-Screen. Die Frage, die an die Probanden gestellt wurde, lautete: „Wie anstrengend ist es für Sie, die Sprache zu verstehen?" Die rein akustische Darbietung erfolgte über Kopfhörer (Sennheiser HD 650) in einer Hörkabine. Die gewählten Höranstrengungskategorien wurden auf einen Zahlenwert aus dem Intervall [1, 13] abgebildet. (1 = „mühelos" … 13 = „extrem anstrengend") (vgl. **Abb. 2**)

Abb.2: Bewertungsskala für Höranstrengung nach [2]

2.2. Höranstrengungsdaten – Datensatz II

2.2.1. Stimuli

Es standen wiederum getrennte Tonspuren (Sprache/ Atmosphäre) von weiteren verschiedenen TV-Produktionen zur Verfügung. Aus diesem Material wurden 210 Audio-Clips mit einer mittleren Länge von ca. 10s herausgeschnitten und mit stark variierenden Mischungs-

verhältnissen gemischt. Im Gegensatz zum Datensatz I enthielten die Atmosphären-Tonspuren jedoch keine Sprache.

2.2.2. Probanden und Bewertungsprozedur

20 normalhörende Probanden im Alter von 20-29 Jahren (Median = 23 Jahre), 9 davon männlich, 11 weiblich, nahmen an der Studie teil. Die Bewertungsprozedur war dieselbe wie beim Datensatz I (s.o.).

2.3. Posteriorgramm-Generierung

Zur Generierung der Posteriorgramme wurde dasselbe ASR-System wie in [6] verwendet, das daher hier nur kurz beschrieben werden soll (für Details siehe [6]):

Als akustische Merkmale dienen die 10ms-Kurzzeit-Energien einer 40-kanaligen Mel-Filterbank. Es werden jeweils -15…+15 10ms-Zeitblöcke (= 310 ms) zusammengefasst und einem tiefen *Time-Delay Neural Network* (TDNN) als Merkmalsvektor übergeben. Dieses tiefe TDNN verfügt über sieben verdeckte Schichten mit jeweils 700 *rectified linear units*. Die Ausgabeschicht besteht aus 6448 Neuronen, d.h. eins je Triphon (ein Triphon ist eine Sequenz aus drei Phonemen). Trainiert wurde das Netz mit einer eigenen Sprachdatenbasis, bestehend aus 1000 Stunden ungestörter Sprache. Diese Datenbasis wurde durch Mischung mit verschiedenen Störgeräuschen bei unterschiedlichen SNR auf einen Umfang von ca. 8000 Stunden erweitert.

2.4. Posteriorgramm-Maß

Von der Ausgabe des tiefen TDNN, d.h. dem Posteriorgramm, wird die „mean temporal distance“ oder kurz „*M*-Measure“ nach Hermansky et al. [7] berechnet. Das *M*-Measure berechnet den durchschnittlichen mathematischen Abstand zwischen zwei Vektoren von Phonem-Posteriors $p_{t-\Delta t}$ und p_t (d.h. zwei Spalten des Posteriorgramms) mit einem zeitlichen Abstand Δt:

$$M(\Delta t) = \frac{1}{T-\Delta t}\sum_{t=\Delta t}^{T} D(p_{t-\Delta t}, p_t), \qquad (1)$$

T ist dabei die zeitliche Länge des analysierten Posteriorgramms (welche gleich der Länge der analysierten Audiodatei ist, d.h. ca. 10s in dieser Studie). D ist die symmetrische Kullback-Leibler-Divergenz zwischen zwei Vektoren x und y mit den Komponenten $x(i)$ und $y(i)$:

$$D(x,y) = \sum_{i=1}^{N} x(i)\log\left(\frac{x(i)}{y(i)}\right) + \sum_{i=1}^{N} y(i)\log\left(\frac{y(i)}{x(i)}\right) \qquad (2)$$

In dieser Studie ist N gleich der Dimensionalität der TDNN-Ausgabeschicht (6448) und M wurde berechnet für Δt =350 bis 800 ms (in 50ms-Schritten) und anschließend gemittelt. Dies ergibt den finalen Höranstrengungs-Prädiktor $\overline{M}$.

3. Ergebnisse

3.1. Datensatz I

Subjektiv gemessene, über die Probanden gemittelte Höranstrengung-Bewertungen (LE-MOS - „Listening Effort-Mean Opionion Scores“) werden den entsprechenden Werten des Höranstrengungs-Prädiktors $\overline{M}$ in **Abb. 3** in Form eines Scatter-Plots gegenüber gestellt. Werden die $\overline{M}$-Werte entsprechend der rot gestrichelt eingezeichneten Regressionsfunktion (Polynom 3. Grades) auf die subjektive LE-MOS-Skala transformiert, so ergeben sich ein Korrelationskoeffizient nach Pearson von r = 0,9, ein Rangkorrelationskoeffizient nach Spearman von rs = 0,85 sowie eine Standardabweichung von einer perfekten Vorhersage von SD = 1,66 LE-MOS-Skaleneinheiten. Es zeigen sich einige Datenpunkte rechts-oberhalb der Ausgleichskurve. Bei näherer Inspektion dieser Daten stellte sich heraus, dass es sich dabei überwiegend um Audioclips handelt, die Sprache in der Atmosphäre enthalten. (Rote Kreise in **Abb. 3**.) Die Höranstrengung solcher Signale wird von der instrumentellen Methode also überwiegend unterschätzt. Das liegt daran, dass die Methode nicht zwischen Nutz- und Störsprache unterscheiden kann. Werden solche Audioclips aus dem Datensatz entfernt, so ergibt sich ein Korrelationswert von r = 0,95.

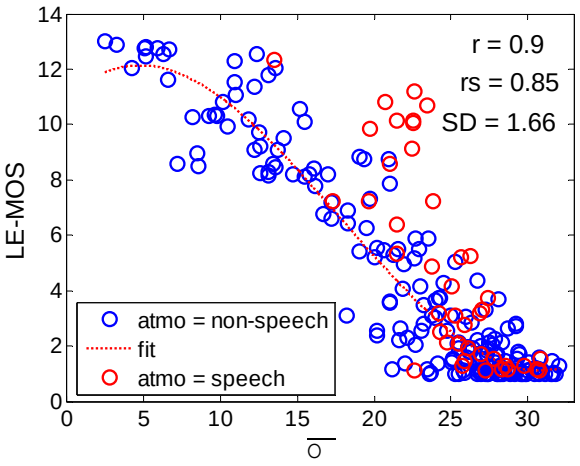


Abb. 3: Ergebnis der Höranstrengungsvorhersage für den Datensatz I. Über die Probanden gemittelte subjektive Bewertungen („Listening Effort Mean Opion Scores“ – LE-MOS) auf der Ordinate sind über den entsprechenden Werten des instrumentellen Höranstrengungs-Prädiktors $\overline{M}$ auf der Abszisse aufgetragen. Blaue Kreise: Bewertungen von Audioclips ohne Sprache in der Atmosphäre. Rote Kreise Bewertungen von Audioclips mit Sprache in der Atmosphäre (zumeist „voice over voice“). Rote, gepunktete Kurve: Nicht-lineare Regressionsfunktion (Polynom 3. Grades). r: Pearson-Korrelationskoeffizient. rs: Spearman-Rangkorrelations-koeffizient. SD: Standardabweichung von einer perfekten Vorhersage in LE-MOS-Skaleneinheiten.

3.2. Datensatz II

Die Ergebnisse für den Datensatz II sind in **Abb. 4** dargestellt. Obwohl die Audioclips keine Sprache in den

Atmosphären enthalten, ergibt sich hier eine geringere Korrelation zwischen subjektiv gemessenen und berechneten Höranstrengungsdaten von $r = 0{,}86$.

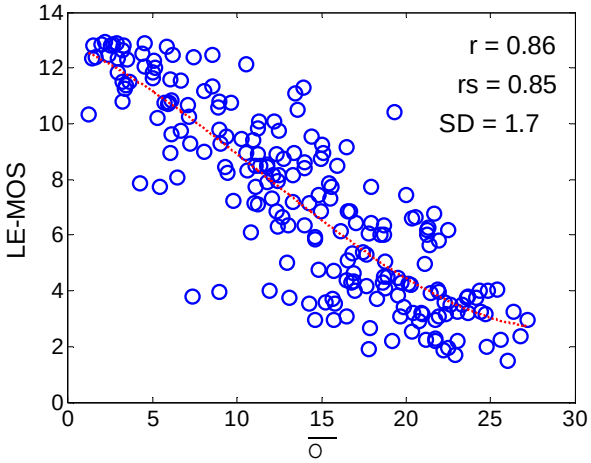


Abb. 4: Ergebnis der Höranstrengungsvorhersage für den Datensatz II. Sonst wie **Abb. 3**

3.3. Kombination der Datensätze I+II

Um zu testen, wie konsistent die Zusammenhänge zwischen subjektiven LE-MOS-Daten und instrumentellen Vorhersagewerten $\bar{M}$ über die Datensätze hinweg sind und um letztlich eine einheitliche Abbildungsfunktion f: $\bar{M}$→LE-MOS zu erhalten, wurden die beiden Datensätze kombiniert. Die Beziehung zwischen LE-MOS-Daten und entsprechenden $\bar{M}$-Werten für die Vereinigungsdatenmenge ist in **Abb. 5** dargestellt. Wie in der Abb. zu sehen ist, ergänzen sich die Datenpunkte beider Datensätze (blau: Datensatz 1 (ohne Clips mit Sprache in der Atmosphäre), rot: Datensatz 2) sehr gut zu einer homogenen Gesamtdatenmenge. Die Ausgleichsfunktion zwischen subjektiven und berechneten Datenwerten lässt sich sehr gut durch eine einfache lineare Funktion darstellen. Die Korrelation beträgt $r = 0{,}89$. Die Gleichung für die Ausgleichsfunktion lautet:

$$\text{LE-MOS} = -0{,}4*\bar{M} + 13 \qquad (3)$$

Wenn man zusätzlich den Wertebereich des gem. Gleichung (3) transformierten $\bar{M}$-Wertes auf den Bereich der LE-MOS-Skala [1, 13] begrenzt, erhält man schließlich den in **Abb.** 6 gezeigten Zusammenhang zwischen subjektiv gemessenen und durch das Modell vorhergesagten Höranstrengungswerten.

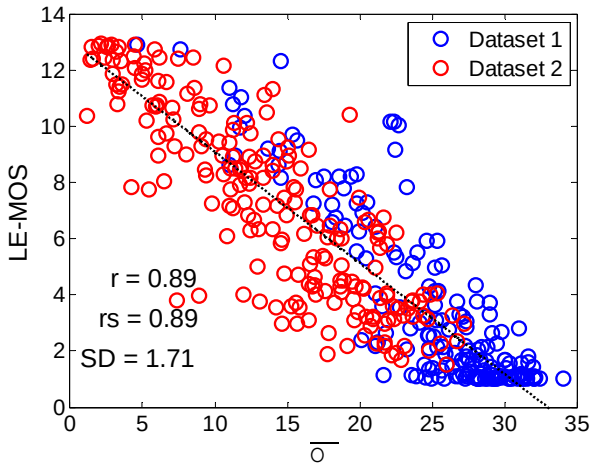


Abb. 5: Ergebnis der Höranstrengungsvorhersage für die kombinierten Datensätze I+II. Blau: Datensatz I; rot: Datensatz II. Schwarz, gepunktet: Lineare Regressionsfunktion. Sonst wie **Abb. 3**

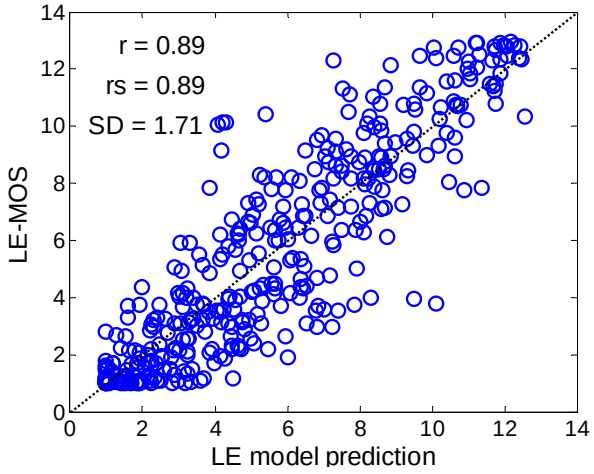


Abb. 6: Ergebnis der Höranstrengungsvorhersage für die kombinierten Datensätze I+II durch den finalen Modell-Prädiktor. Schwarz, gepunktet: Identitätsfunktion.

4. Diskussion

Die Anwendung der referenzfreien instrumentellen Methode zur Vorhersage von Höranstrengung nach [6] scheint auch für den Einsatz bei Broadcast-Anwendungen sehr vielversprechend zu sein, insbesondere, da die bisherigen Ergebnisse der Höranstrengungsvorhersage ohne jegliches Training oder Optimierung der Methode auf die vorliegenden Daten erzielt wurden. Weitere Verbesserungen der Ergebnisse sind denkbar, wenn der Spracherkenner auch mit typischen Hintergrund-„Stör"-Signalen, wie sie im Rundfunkmaterial auftreten, trainiert würde.

Eine weitere Verbesserung der Ergebnisse könnte möglicherweise mit Hilfe einer automatischen Sprachaktivitätserkennung erzielt werden. Bislang wurden auch (z.T. längere (> 1s)) Sprachpausen bei der Berechnung des Höranstrengungs-Prädiktors mit einbezogen, obwohl die

tatsächliche Höranstrengung während Sprachpausen nicht definiert ist. Mithilfe der automatischen Sprachaktivitätserkennung könnten Sprachpausen als solche erkannt und aus dem Signal entfernt werden. Voraussetzung wäre eine ausreichende Robustheit der Sprachaktivitätserkennung gegenüber Hintergrundgeräuschen.

Eine Einschränkung der Methode betrifft die Bewertung von Sprache, die von konkurrierender Sprache im Hintergrund (voice over voice) überlagert wird. Hier unterschätzt die Methode die empfundene Höranstrengung. Dies ist erklärbar, da die Methode nicht zwischen Nutz- und Störsprache unterscheiden kann. Die entsprechenden Posteriorgramme von zwei Sprechern weisen nur geringradige Verschmierungen auf und die Phoneme (bzw. Triphone) werden vom ASR-System weiterhin überwiegend erkannt. Erst bei einer Überlagerung von vielen Sprechern („babble noise“) ist mit einer erhöhten Erkennungsunsicherheit des ASR-Systems zu rechnen.

Der bei einzelnen (verständlichen) Störsprechern zusätzlich auftretende Effekt des „informational masking“, der neben der energetischen Maskierung zu einer weiteren Erhöhung der Höranstrengung beiträgt, kann durch den hier verwendeten instrumentellen Ansatz nicht abgebildet werden.

Eine mögliche Abmilderung dieser Einschränkung des Verfahrens könnte durch eine unabhängige Detektion von Sprache im Hintergrund vermittelt werden. In einem solchen Fall könnte ein Korrekturwert (Malus) zur Höranstrengungsschätzung hinzuaddiert werden. Folgende Klassifizierung der Signale ist hierfür nötig: (1) nur Nutzsprache, (2) Sprache mit nicht-Sprache im Hintergrund, (3) Nutzsprache mit Störsprache im Hintergrund.

Eine weitere Einschränkung der Methode trifft die Tatsache, dass das referenzfreie Höranstrengungsmodell im Moment noch als „Mono“-Modell agiert und so den Einfluss einer räumlichen Verteilung der Quellen außer Acht lässt. Insbesondere mit Blick auf Surround und 3D-Produktionen ist das ein Mangel, den es in Form einer Binauralisierungsstufe noch zu beheben gilt.

Die Korrelation, die zwischen den subjektiven Höranstrengungsbewertungen und den Modellvorhersagen des zweiten Datensatzes erzielt wurde, ist mit $r = 0{,}86$ deutlich geringer als die des ersten Datensatzes (ohne Clips mit Sprache in der Atmosphäre) mit $r = 0{,}95$. Die Ursache für diesen Unterschied ist unklar.

5. Schlussfolgerungen

Die vorgestellte referenzfreie, instrumentelle Methode zur Schätzung der empfundenen Höranstrengung hat sich insgesamt als geeignet zur Anwendung auf Rundfunk-Audiomaterial herausgestellt. Bei der Bewertung von Audioclips mit Sprache im Hintergrund muss berücksichtigt werden, dass die Methode die tatsächliche Höranstrengung tendenziell unterschätzt. Die Methode bietet Verbesserungspotenzial, z.B. durch Trainieren des Spracherkenners mit Sprache mit rundfunktypischen Hintergrundgeräuschen, durch Anwendung eines automischen Sprachaktivitätserkenners und der Erweiterung durch eine Binauralisierungsstufe, welche der Räumlichkeit verschiedener Produktionsformate gerecht wird.

6. Danksagung

Diese Studie wurde durchgeführt im Rahmen des vom BMBF geförderten Projekts SITA; FKZ: 01/S17017B.

7. Literatur

[1] ITU-T: Recommendation P.800. Methods for Subjective Determination of Transmission Quality. Geneva, 1996

[2] Schulte, M., Meis, M., Wagener, K.: Listening Effort and Speech Intelligibility. In Proceedings 8th EFAS Congress / 10th Congress of the German Society of Audiology, 2007.

[3] Klink, K., Schulte, M., Meis, M.: Measuring listening effort in the field of Audiology – a literature review of methods, part 1. Z Audiol 51 (2) (2012) 60–67

[4] Huber, R., Spille, C., Meyer, B.T.: Single-Ended Prediction of Listening Effort Based on Automatic Speech Recognition. In Proceedings Interspeech 2017, 1168-1172

[5] Huber, R., Krüger, M., Meyer, B. T.: Single-ended prediction of listening effort using deep neural networks. Hear. Res. 359 (2018), 40-49

[6] Huber, R., Pusch, A., Moritz, N., Rennies, J., Schepker H., Meyer, B.T.: Objective Assessment of a Speech Enhancement Scheme with an Automatic Speech Recognition-Based System. In Proceedings ITG Conference on Speech Communication (2018), 86-90

[7] Hermansky, H., Variani, E., Peddinti, V.: Mean temporal distance: Predicting ASR error from temporal properties of speech signal. In Proceedings ICASSP 2013, 38th IEEE Int. Conf. Acoust. Speech Signal Process. doi: 10.1109/ICASSP.2013.6639105

Verband Deutscher Tonmeister e.V.

AES Reviewed Paper at Tonmeistertagung 2018

Presented * by VDT in cooperation with the
Central European Region of the Audio Engineering Society (AES).

Investigations on the Impact of Distance Cues in Virtual Acoustic Environments

Christoph Pörschmann[1], Johannes M. Arend[1,2], Philipp Stade[1,2]
[1] *TH Köln, Institute of Communications Engineering, Cologne, Germany*
[2] *TU Berlin, Audio Communication Group, Berlin, Germany*

Abstract

In this paper, we analyze different auditory distance cues in dynamic binaural synthesis. We compare the contributions of sound intensity, direct-to-reverberant ratio (DRR), and near-field cues. For the auralization, we use the BinRIR method, which allows to generate binaural room impulse responses (BRIRs) for dynamic binaural synthesis based on one measured omnidirectional room impulse response (RIR). With BinRIR, applying a simple geometric model, the listener position can be freely adjusted and the distance cues can be adapted separately. Furthermore, near-field head-related impulse responses (HRIRs) can be applied for direct sound and early reflections if the listener is very close to the virtual sound source. In a listening experiment, we presented stimuli at different distances in four synthesized rooms. In one condition, the stimuli contained natural distance-dependent intensity cues, and in another condition, the stimuli were normalized in loudness. The results showed that even for loudness-normalized stimuli, an adequate distance perception can be obtained by adapting the DRR. The influence of near-field HRIRs, which were also tested in the experiment, is weak.

1. Introduction

Sound intensity, direct-to-reverberant ratio (DRR), as well as the spectrum influence auditory distance perception of sound sources [1] [2] [3]. Furthermore, it is contrarily discussed in literature if and to what extent binaural cues are relevant for the perceived sound source distance in the near field [4] [5]. Generally speaking, perceived source distance increases with decreasing level. In anechoic conditions, the relationship between level and distance from a sound source to the receiver is characterized by the 6 dB law for each doubling of the source distance. In reverberant conditions, this decrease is reduced due to reflections and reverberation. However, as intensity also depends on the source signal, it has to be regarded as a relative distance cue. Thus, the presented signal needs to be compared to a reference in order to judge distance.

In reverberant conditions, the ratio of energy reaching a listener on the direct path to the energy reaching the listener via the reflecting surfaces (DRR) is inversely related to the distance of the sound source. As the DRR is independent of the source signal, it provides absolute distance information [2]. Furthermore, changes of the DRR are typically related to a modified initial time delay gap (ITDG), which describes the temporal difference between the direct sound and the first strong reflection.

Distance perception is generally most accurate when both DRR and level cues are available. However, when analyzing the cues isolated, intensity cues provide more accurate information than DRR only [2]. In highly reverberant environments, though, both cues can provide equally accurate information for distance discrimination. On the contrary,

* Please note that the AES Reviewed Papers at Tonmestertagung can be published by both, AES and VDT, in print, online and as PDF download.

ISBN 978-3-9812830-9-9

studies analyzing just-noticeable-differences (JNDs) showed that the JNDs for the DRR are comparably large. According to [2] they are in a range of 2–3 dB for 0 and +10 dB DRR and about 6–8 dB at -10 and +20 dB DRR. This suggests that the principal role of the DRR is to provide a rough absolute distance information, rather than to support fine distance discriminations which can be detected from small changes in overall intensity [3]. According to the the pressure-discrimination hypothesis, the JNDs in source distance are determined by the ability to discriminate changes in sound pressure. According to [6] for broadband noise, the smallest detectable change in level is approximately 0.4 dB.

The spectrum of the signal serves as another distance cue. For larger distances (sound source distances more than 15 m) dissipation causes a low-pass filtering characteristic which increases towards higher distances and by this influences perceived distance. Thus sounds with decreased high-frequency components relative to low-frequency components are perceived to be further away [2]. As this cue is strongly influenced by the spectrum of the source signal, it can be regarded as a relative distance cue. In the near field (sound source distance less than 1 m) the spectrum also changes with distance because diffraction and head-shadowing effects lead to a low-pass filtering characteristic which increases towards lower distances [7] [8]. Additionally, for close sound sources diffraction and head shadowing induce a distance-related change in interaural time differences (ITDs) and interaural level differences (ILDs) [7] [8]. While the influence of distance on the ITDs is relatively low, ILDs change substantially over distance.

Distance estimation is highly relevant for headphone-based virtual acoustic environments (VAEs) which are applied in various areas like audio engineering, telecommunications, or architectural acoustics to create a natural room impression. Furthermore, the use of VAEs is well suited for applying listening experiments and presenting stimuli from different distances and rooms in a natural environment. For example, in the context of distance estimation Zahorik [9] analyzed contributions of different cues to distance estimations based on one measured room. Brungart [10] assessed distance estimation of near-field sound sources in a VAE.

Varying sound source distances can be considered relatively easy in simulation-based systems (e.g. [11]). However, depending on the application, the auralization often relies on binaural room impulse responses (BRIRs), which are either measured with artificial heads for different head orientations or rely on an arbitrary BRIR synthesis. If the BRIRs are acquired with an artificial head, measurement times for the acquisition of a circular (or even a spherical) set of BRIRs are often quite high and increase even more if data at different listening positions and distances needs to be captured. Thus, the use of parametric models allowing to control DRR, sound intensity, and near-field cues separately in order to consider distance modifications on measured data can be beneficial. In a previous publication, we presented such an approach called BinRIR [12] [13]. This algorithm allows to generate BRIRs for dynamic auralization based on one omnidirectional room impulse response (RIR). The approach uses a parametric model for the BRIR synthesis and spatializes an omnidirectional RIR by extracting direct sound, early reflections, and diffuse reverberant parts from the RIR. The synthesis allows for a free shift of the listener position in the room. Additionally, near-field head-related impulse responses (HRIRs) can be applied if the sound source or one of the reflections are very close to the listener. Thus, this approach can be used to auralize arbitrarily chosen listener positions in the virtual room. This allows to create naturally sounding stimuli from measured data and to separately control the different cues which affect distance perception.

Applying BinRIR, we investigated the impact of intensity, DRR, and near-field cues on distance estimation in a listening experiment. In the form of a multi-stimulus comparision test, we tested distance perception for four rooms and different distances between source and listener. Additionally we varied the presented distance cues. The results of the experiment show that intensity and DRR have a significant impact on distance estimation. Even for loudness-normalized stimuli, listeners were able to judge distance appropriately. Near-field cues, however, did not significantly support distance estimation.

The paper is organized as follows: Section 2 describes the BinRIR algorithm and explains how the listener position shifts and the near-field HRIRs are applied. This algorithm provides the basis for the psychoacoustic study. Section 3 presents the listening experiment in which we examined distance estimation for four different rooms. In section 4, we analyze in which way the different cues contribute to distance estimation. Finally, section 5 concludes the paper.

2. BinRIR Method

For the listening experiment presented in this paper, we synthesized stimuli with the BinRIR method [12] [13]. BinRIR generates BRIRs based on a measured omnidirectional RIR and allows to separately modify different distance cues. The basic structure of BinRIR is shown in Fig. 1. The method only applies to frequency components above 200 Hz. For lower frequencies, the interaural coherence of a typical BRIR is nearly one and the omnidirectional RIR can be maintained.

BinRIR requires only one measured omnidirectional RIR to synthesize an arbitrary BRIR dataset. To obtain a BRIR, we use predictable information from geometrical acoustics as well a perception-motivated simplified description of the diffuse sound field. For this, the RIR is split into two different parts. Onset detection is used to identify the direct sound in the ommnidirectional RIR. This section starting with the onset is windowed (5 ms followed by 5 ms raised cosine offset ramp). The following time section is assigned to the early reflections and the transition towards the diffuse reverberation. In order to determine sections with strong early reflections in the omnidirectional RIR, the energy is calculated in a sliding window of 8 ms length and time sections which contain high energy are marked. Peaks which are 6 dB above the RMS of the sliding window are determined and assigned to geometric reflections. Each

* Please note that the AES Reviewed Papers at Tonmestertagung can be published by both, AES and VDT, in print, online and as PDF download.

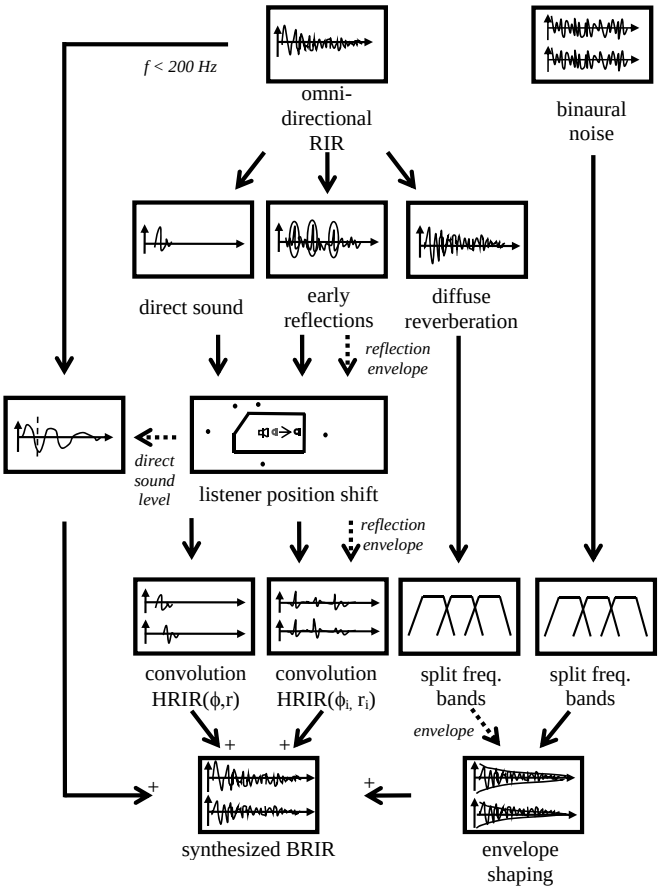


Fig. 1: Block diagram of the BinRIR algorithm for synthesizing a BRIR based on a single omnidirectional RIR.

section comprises both the amplitude and the spectral shape of the reflection. Thus, all specific characteristics of the reflections, e.g. edge diffraction and the acoustical properties of the surfaces, are adequately considered. In contrast, the incidence directions of the synthesized reflections are chosen by the algorithm and base on a spatial reflection pattern adapted from a shoebox room with non-symmetric positioned source and receiver. Each windowed section of the RIR is convolved with the appropriate HRIR. For this, a measured set of HRIRs of a Neumann KU100 artificial head [14] is used. To obtain interim directions between the given directions of the HRIR set, an interpolation by means of spherical harmonics transform is performed [15]. By this a binauralized version of the early geometric reflections is obtained.

All sections of the RIR which were not detected as geometric reflections are assigned as diffuse. Several studies have shown that after the so-called perceptual mixing time, the measured (binaural) room impulse response can be simplified [16][17]. In BinRIR the diffuse reverberation part is synthesized by shaping binaural noise according to the envelope of the diffuse part of the omnidirectional RIR. The diffuse part is considered reaching the listener temporarily and spatially equally distributed and is not influenced by shifts of the listener position. Following recent studies (e.g. [18] [19])

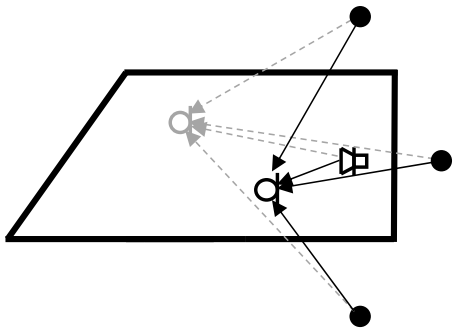


Figure 1: Basic principle of the listener position shifts (LPS) applying mirror images: The amplitude as well as the temporal structure of the direct sound and the early reflections are adapted accordingly. The receiver is moved from an initial position (grey) to a modified position (black). By this, the paths of direct sound and reflections are changed.

Fig. 2: Basic principle of the listener position shifts. Mirror images are applied demonstrating the modification of the amplitude and the temporal structure of direct sound and early reflections. The receiver is moved from an initial position (grey) to a modified position (black). By this, the paths of direct sound and reflections are changed.

BinRIR overlaps sections with early reflections and diffuse parts instead of using a fixed mixing time. Cross-fading is applied in order to merge sections with geometric reflections and sections with diffuse reverberation.

The possibility to consider listener position shifts in BinRIR based on a simple geometric model is of specific relevance for this study. For this, the distance between the listener and each of the mirror images is determined based on the delay between the corresponding peak of the reflection and the peak of the direct sound. In a next step, a shifted position of the listener is considered and amplitudes (based on the $1/r$ law), distances, and directions of incidence are recalculated for each or the reflections (Fig. 2). Thus by shifting the listener position the amplitudes and time delays of the direct sound and the reflections are affected. As a consequence the ITDG and the DRR are modified as well by the listener position shifts.

Furthermore, HRIRs measured at different distances can be considered in BinRIR. Thus, in addition to the used far-field HRIRs [14], near-field HRIRs, as well measured with a Neumann KU100 at distances between 0.25 m and 1.50 m are used [8]. Regarding the distance of the direct sound and each of the reflections, the HRIR is chosen from the set that fits the distance of the reflections best. However, for distances above 2 m, far-field HRIRs are chosen.

The BRIR synthesis is repeated for constant shifts in the azimuth angle (e.g. 1°) for the direct and the reflected sound. Thus, a circular set of BRIRs is created, which can be used for dynamic binaural synthesis. Since binaural cues are absent in the measured omnidirectional RIR, the BinRIR algorithm incorporates several inaccuracies and errors. For example, the directions of incidence of the synthesized early reflections are not in line with the original ones. Hence, differences in perceptual spatial properties (e.g. envelopment) between the original and the synthesized room may occur. In [12] [13] we already performed a detailed technical analysis as well as a perceptual comparison to a measured reference and examined the performance of the algorithm. Thus these issues are not further discussed and analyzed in the present paper.

* Please note that the AES Reviewed Papers at Tonmestertagung can be published by both, AES and VDT, in print, online and as PDF download.

3. Listening Experiment

We performed a multi-stimulus comparison test. In two different sessions the listeners estimated the auditory distances of normalized and non-normalized stimuli. By this, we examined the impact of source intensity, DRR, and near-field cues to distance estimation. Generally, intensity cues and DRR both significantly contribute to distance estimation [2] [3]. However, the influence of DRR and intensity strongly depend on the room and thus, it was an essential part of the study to compare the contributions for acoustically different rooms. Finally, the listening experiment served to test if the synthesis of the BinRIR algorithm allows to adapt the perceived distance appropriately. Thus, we included stimuli synthesized with a measured BRIR in the test, for which we already found a good perceptual correlation to the synthesis with BinRIR [20].

3.1. Participants

In total, 34 adults (5 female, 29 male) aged between 18 and 46 years (M = 24.6 years, SD = 5.07) took part in the experiment. Most of them were students at TH Köln. All listeners already participated in previous listening experiments and thus were familiar with dynamic binaural synthesis. None of the subjects reported hearing problems.

3.2. Setup

The experiments took place in the anechoic chamber at TH Köln, which ensured a low background noise level of less than 20 dB(A). Furthermore, this room is completely different to any of the investigated rooms and potential influences of audiovisual room convergence were thus avoided. The experiment was set up, controlled, and executed running the software Scale [21]. The listeners had to enter their ratings on a touch-screen computer (iPad). The stimuli were presented with dynamic binaural synthesis using the SoundScape Renderer [22] and a Polhemus Fastrak head tracking system to consider full circle horizontal head movements with a resolution of 1°. The stimuli were presented via AKG K-601 headphones. No headphone equalization was applied to the stimuli.

3.3. Materials

Rooms We used measurement data from four different rooms [23]. Table 1 gives an overview about the rooms and their properties. The Control Room 7 (CR7) is the main control room for radio drama production at WDR Broadcast Studios in Cologne. We measured the RIRs in the sweet spot, right in front of the mixing console. The Large Broadcast Studio (LBS) and the Small Broadcast Studio (SBS) are as well located at the WDR in Cologne and are used for various recordings of concerts and performances. TGC is a training room of a local dance club in Cologne. The loudspeaker types varied between the four rooms. In CR7, we used the installed Bowers & Wilkens 803D loudspeakers. In LBS and SBS, the source was a full PA stack involving an AD Systems Stium Mid/High unit combined with three AD Systems Flex 15b subwoofers. In TGC, one AD Systems Flex 15 speaker was applied. Please refer to [23] for further information on the geometry of the rooms and the positions of the loudspeakers and the listener.

The measurements comprised both, a circular reference set of BRIRs and an omnidirectional impulse response, measured at the pivot position of the artificial head. In this study, the binaural measurements, which were performed in steps of 1° on the horizontal plane with a Neumann KU100 artificial head, were only used to compare the results of the BinRIR synthesis to a binaurally measured reference. A Microtech Gefell M296S microphone was used for the omnidirectional measurement in TGC; for the other omnidirectional measurements, an Earthworks M30 microphone was applied.

Distances BRIR sets describing different distances between sound source and receiver were synthesized with the BinRIR algorithm. The listener position was shifted as described in section 2 according to so-called distance factors (DFs) on a line between sound source and listener. The DFs correspond to the quotient between the synthesized stimulus distance and the distance of the measured RIR. Thus, DF = 1 is equivalent to the measuring distance between source and receiver $\text{Distance}_{\text{SrcRec}}$ of the measurement (see Table 1). For each room, we tested the following DFs: 0.125, 0.25, 0.5, 1, and 2. Additionally, the measured BRIR – in the following named original – was included in the experiment. Moreover, for close distances in which near-field cues might influence distance estimation, a variant assessing the appropriate near-field HRIR for the direct sound and another variant using only far-field HRIRs were synthesized. Depending on $\text{Distance}_{\text{SrcRec}}$, a varying number of near-field stimuli was used for the different rooms. For LBS, we presented only one stimulus using near-field HRIRs, for SBS and TGC two, and for CR7 three near-field stimuli.

Direct-to-reverberant ratio (DRR) The DF strongly influences the DRR. We calculated the DRR as the ratio of the energy reaching the listener within the first 7.5 ms (corresponding to a distance of 2.55 m) after the direct sound

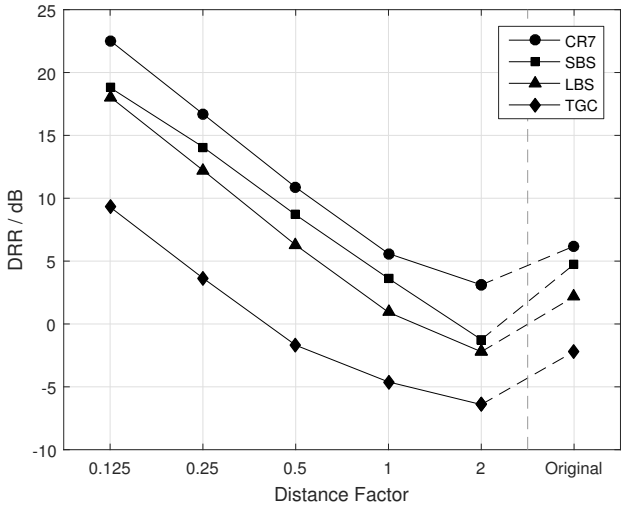


Fig. 3: Direct-to-reverberant ratios (DRRs) for the four different rooms and the varying distance factors (DFs). The highest DRRs can be observed for the rooms with the shortest RT_{60}. In CR7, all synthesized positions are inside the critical distance. On the contrary, for TGC, only the DF = 0.125 and DF = 0.25 are inside the critical distance.

* Please note that the AES Reviewed Papers at Tonmestertagung can be published by both, AES and VDT, in print, online and as PDF download.

Room	Volume	Area	RT_{60}	$Distance_{SrcRec}$
Control Room 7 (CR7)	168 m^3	60 m^2	< 0.25 s	2.7 m
Large Broadcast Studio (LBS)	6100 m^3	579 m^2	1.8 s	13.0 m
Small Broadcast Studio (SBS)	1247 m^3	220 m^2	0.9 s	7.0 m
TGC Training Room (TGC)	191 m^3	67 m^3	2.3 s	6.8 m

Tab. 1: Main properties of the measured and synthesized rooms. The RT_{60} corresponds to the mean reverberation time in the frequency range between 500 and 1000 Hz.

and the energy of the reverberant parts. As only the time differences between the direct sound and the floor reflection is smaller than this value, all the geometric reflections from walls and ceiling are in this context considered as a part of the reverberation. The results for the different rooms are illustrated in Fig. 3. It can be observed that the DRR shows maximal values of up to 22.5 dB for CR7 and generally increases towards small distances. Thus, it can be assumed that in the CR7, especially for nearby sound sources, the influence of the room on distance perception is small. On the contrary, in TGC the DRR is in a range from -6.4 dB to 8.7 dB for all conditions, which probably allows the listener to exploit DRR for distance estimation in all conditions. The influence of near-field cues on the DRR are not shown in Fig. 3. However, they did not exceed 1.1 dB.

Stimuli A looped drum and a guitar sequence were used as test signals. These signals have already been used in previous listening experiments [e.g. 24]. In the listening experiment we presented two conditions of each stimulus. In one condition, the stimuli were loudness normalized according to ITU-R BS. 1770 [25], the other one comprised loudness differences with increasing sound pressure levels towards lower distances as calculated by the geometric model. The playback-level was calibrated to a Leq of 60 dB(A) SPL for the normalized stimuli and for the stimuli with a DF of 1. A maximal Leq of 75.2 dB(A) SPL was obtained for a DF of 0.125 in CR7. The Leq(A) for non-normalized stimuli depending on room and DF is given in Fig. 4. The levels vary most for CR7, especially towards lower distances. In total, level differences of more than 15 dB can be observed here. In TGC, which was the most reverberant room in our study, the level differences between all distances were less than 4 dB. The test signal had only small influence on the SPL; differences between drums and guitar were mostly below 1 dB. Only for low distances in CR7, some differences are about 2 dB. The use of near-field HRIRs did as well only marginally influence the SPL. Maximal deviations of less than 1 dB were observed here.

3.4. Procedure

We performed a multi-stimulus comparison test. In two different sessions, the participants estimated the distances of the normalized and the non-normalized stimuli on a seven-point category scale with the categories 'very close', 'close', 'rather close', 'medium', 'rather distant', 'distant', 'very distant'. The subjects were allowed to rate interim values between the given categories. As in earlier experiments [20] the equidistance between the categories was underlined by the visual presentation. One half of the subjects started with the conditions with the normalized stimuli, the other half with the non-normalized stimuli. In each session, every participant had to rate the stimuli for the four different rooms and the two different test signals. The order of conditions was randomized within each session. For each condition we presented the stimuli for the different DFs, the near-field stimuli, and the binaurally measured reference in one multi-stimulus comparison. The scale was displayed on the tablet computer (iPad) and results were given by setting a slider to the appropriate position. Several test sliders were shown at the same time and the subjects were able to switch between the corresponding stimuli of the DFs as often as required. No additional information on the type, shape or size of the rooms or on the reference was given to the subjects. The experiment started with a short introduction including several test trials in order to make the subjects familiar with the test procedure and the stimuli. It included stimuli with the largest and smallest simulated distance and was used for anchoring as well.

4. Results

We investigated the impact of sound intensity, DRR, and near-field cues on distance estimation. Generally, the subjects stated that distance estimation of the stimuli was easy to perform. However, some participants reported that a few of the close stimuli were perceived being unnatural regarding their tone color or their naturalness.

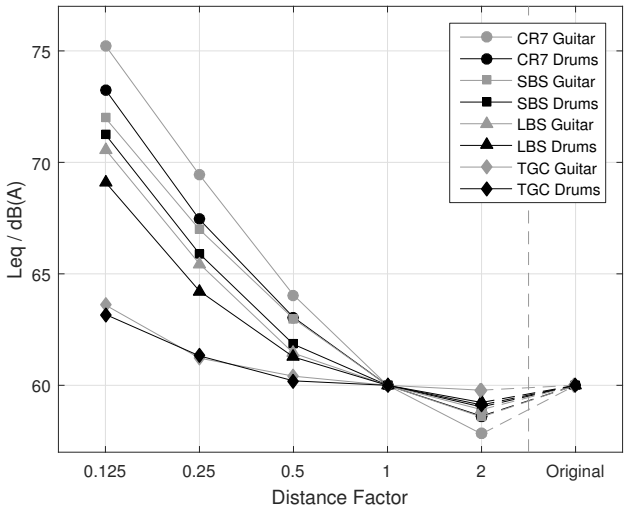


Fig. 4: Levels (Leq) in dB(A) of the non-normalized stimuli depending on the distance factor (DF) for the four rooms. The Leqs vary most for the CR7, especially towards smaller distances, and are minimal for the TGC room at higher distances. Differences between the guitar (grey) and drums (black) source signal do not exceed 2 dB.

* Please note that the AES Reviewed Papers at Tonmestertagung can be published by both, AES and VDT, in print, online and as PDF download.

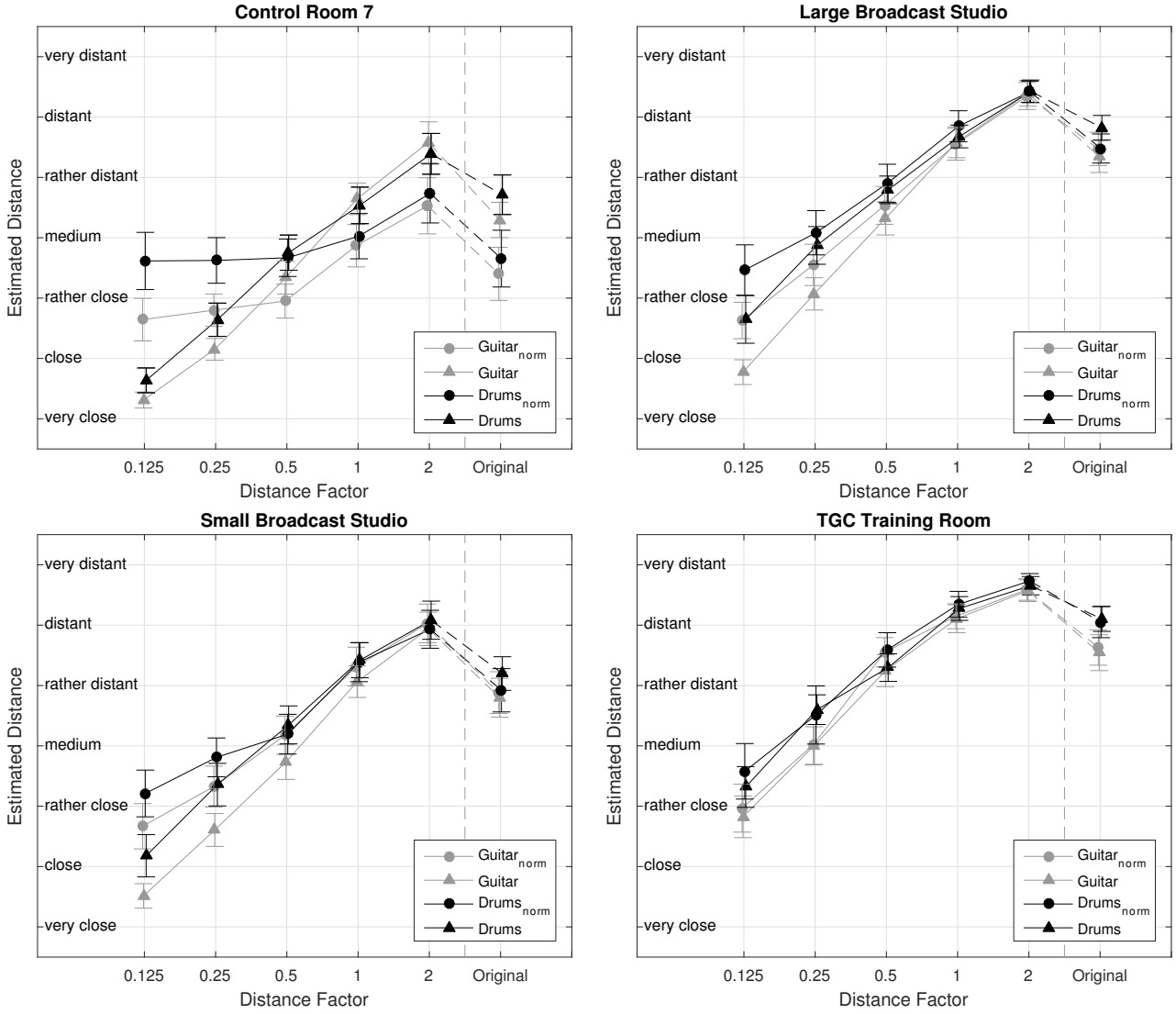


Fig. 5: Estimated distance (mean values and 95 % confidence intervals) for Control Room 7, Large Broadcast Studio, Small Broadcast Studio, TGC Training Room Cologne. On the x-axis, the different distance factors (DF) and the original measurement are plotted.

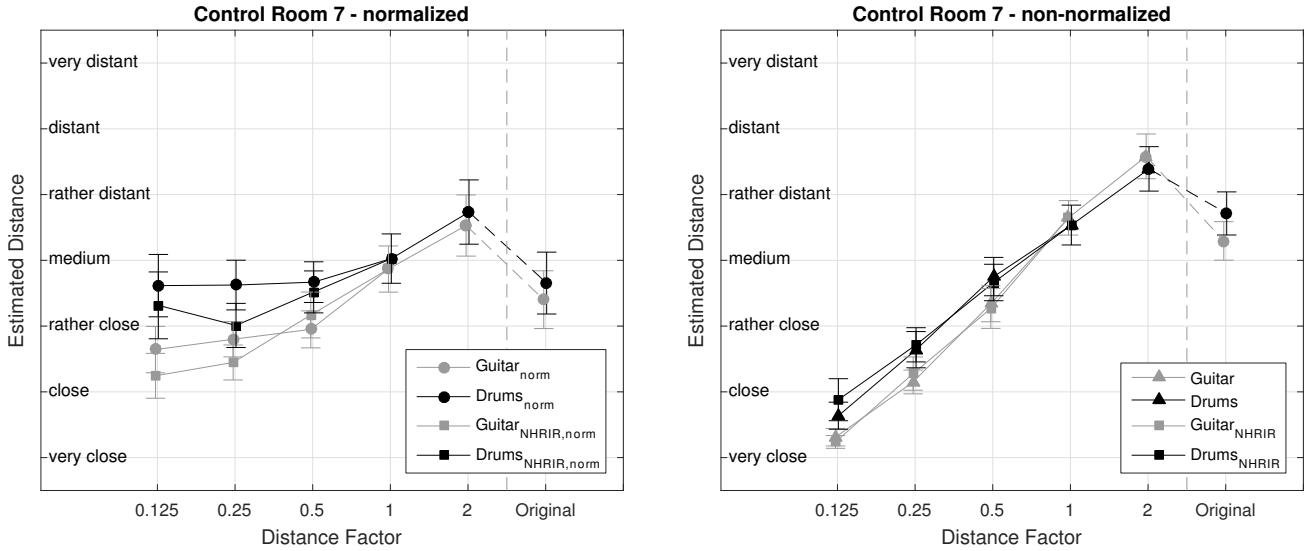


Fig. 6: Impact of near-field HRIRs on estimated distance (mean values and 95 % confidence intervals) for Control Room 7. On the x-axis, the different distance factors (DF) and the original measurement are shown. The left plot shows the loudness-normalized condition, the right one the non-normalized conditions with natural intensity differences.

* Please note that the AES Reviewed Papers at Tonmestertagung can be published by both, AES and VDT, in print, online and as PDF download.

First, we performed a Greenhouse-Geiser corrected [26] repeated measures ANOVA with the factors DF, room, test signal, and loudness normalization. A varying number of stimuli assessing the near-field HRIRs was presented for the different rooms. Thus, in order to get a balanced test design, these stimuli were not part of the ANOVA. The analysis yielded significant main effects for all factors. Main factors with highest effect size were DF ($F(4, 132) = 448$, $p < .001$, $\eta_p^2 = .93$) and room ($F(3, 99) = 137$, $p < .001$, $\eta_p^2 = .81$). The effect sizes of test signal ($F(1, 33) = 36.1$, $p < .001$, $\eta_p^2 = .52$) and normalization ($F(1, 33) = 25.5$, $p = .003$, $\eta_p^2 = .44$) were lower. Surprisingly, the main effect of normalization had the lowest effect size. Furthermore, various significant interaction effects were observed. Strongest was the interaction between DF and normalization ($F(4, 132) = 59.0$, $p < .001$, $\eta_p^2 = .64$). Thus in a next step, we further analyzed the data by performing a nested repeated measures ANOVA for each room. Here, we found no significant effect of normalization for TGC ($p = .13$). For SBS ($p < .001$) and LBS (p=.008), there was a significant main effect of normalization which, however disappeared after a subsequent removal of the closest distances from the analysis. We performed the ANOVA for LBS without the DF of 0.125 and for SBS without the DFs of 0.125 and 0.25. In this case, no significant main effect of normalization remained for LBS ($p = .16$) and SBS ($p = .32$). For CR7 even when removing close stimuli (DFs of 0.125 and 0.25) from the analysis a significant effect of normalization ($p < .001$) remained. It can be concluded that for SBS, LBS and TGC normalizing the stimuli only had significant influence on distance estimation for close stimuli which have high DRRs (in this study above 14 dB, see Fig. 3). For these stimuli, large level differences due to the normalization occur. In our study for all these stimuli the level difference exceeded 6 dB(A) (see Fig. 4).

In a next step, we calculated the mean values and the 95 % confidence intervals. The results for the four tested rooms are shown in Fig. 5. Generally, for all rooms and test signals, a good differentiation in estimated distance between the presented distances was achieved. Regarding the loudness-normalized stimuli, distance estimation varied only slightly in comparison to the non-normalized ones. This coincides with the findings of Kolarik [2] who found that both DRR and intensity can serve as robust distance cues.

For further statistical analysis we investigated the influence of the near-field cues and compared the distance estimations for stimuli with near-field HRIRs for close distances to the ones which only far-field HRIRs. In Fig. 6 the respective mean values of estimated distance with 95 % confidence intervals are shown. We performed t-tests with Hochberg-correction which showed only for two of the conditions in CR7 a significant influence (drums, DF = 0.25, loudness normalization, $p <$.001; guitar, DF = 0.125, loudness normalization, $p =$.007). In this room, the simulated distances for which near-field HRIRs were considered range from 0.34 m at a DF of 0.125 to 1.35 m at a DF of 0.5. Generally DRR is higher for CR7 than for the other rooms, and thus reverberation probably completely masks the influence of near-field cues in the other rooms. For three out of four tested rooms and for all of the non-normalized stimuli, no significant influence of the near-field HRIRs was noted. It can be stated that if at all, near-field cues only marginally influence distance estimation in reverberant environments.

Finally, we performed t-tests with Hochberg-correction [27] in order to compare the stimuli auralized with the measured BRIRs (originals) to the stimuli generated using synthetic BRIRs with DF of 1. The analysis showed a significant difference for 5 out of 16 conditions at the 0.05 level (CR7: guitar, non-normalized; SBS: guitar, normalized; TGC room: guitar, normalized; guitar, non-normalized; drums, non-normalized). However, even though slight differences in distance estimation for the original BRIR exist, we can assume that the BinRIR algorithm allows to estimate distance comparably to a binaurally measured reference. Similar effects have already been observed in an earlier study [20].

5. Conclusion

We applied the BinRIR algorithm to investigate the contribution of source intensity, DRR, and near-field cues to distance estimation. Based on measured omnidirectional RIRs, we synthesized stimuli for four different rooms at varying source distances. The listening experiment showed that for all investigated rooms, the desired control of the perceived distance could be achieved.

As shown in previous studies, intensity has a strong influence on the results. This manifests for example in the differences that were observed between the normalized and the non-normalized stimuli. Furthermore, the DRR contributes strongly to the estimated distance as well. Even for loudness-normalized stimuli, subjects were able to estimate distance appropriately. In TGC, no significant contribution of the normalization as a main factor could be shown, and for SBS and LBS, normalization only had influence on the results for close distances. Finally, the study delivers information on the relevance of near-field cues for distance estimation. The use of near-field HRIRs created additional variance in the results but no significant trend of their influence could be observed.

The results of this study are relevant for applications in which VAEs are used to present stimuli at different distances, for example in radio dramas or games in which a plausible impression of the environment is desired. Often, a distance control independent of the sound intensity is desired here. Both controlling DRR and intensity allows to change the perceived distance. In future studies it might be interesting to investigate the influence of moving listeners and sound sources on distance estimation. In augmented reality environments specific aspects arise: future research on distance perception might be related to influences of mismatches between real and synthesized objects.

6. Acknowledgment

The research presented in this paper has been carried out in the Research Project MoNRa which was funded by the Federal Ministry of Education and Research in Germany. Support Code: 03FH005I3-MoNRa. We thank the participants of the listening experiments for their patience and commitment.

* Please note that the AES Reviewed Papers at Tonmestertagung can be published by both, AES and VDT, in print, online and as PDF download.

7. References

[1] Blauert, J., *Spatial Hearing - Revised Edition: The Psychoacoustics of Human Sound Source Localisation*, MIT Press, Cambridge, MA, 1997.

[2] Kolarik, A. J., Moore, B. C. J., Zahorik, P., Cirstea, S., and Pardhan, S., "Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss," *Attention, Perception, & Psychophysics*, 78, pp. 373–395, 2016.

[3] Zahorik, P., Brungart, D. S., and Bronkhorst, A. W., "Auditory distance perception in humans: A summary of past and present research," *Acta Acustica united with Acustica*, 91(3), pp. 409–420, 2005.

[4] Brungart, D. S., Durlach, N. I., and Rabinowitz, W. M., "Auditory localization of nearby sources. II. Localization of a broadband source." *Journal of the Acoustical Society of America*, 106, pp. 1956–1968, 1999.

[5] Shinn-Cunningham, B., "Distance cues for virtual auditory space," *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia*, (December), pp. 227–230, 2000.

[6] Miller, G. A., "Sensitivity to Changes in the Intensity of White Noise and Its Relation to Masking and Loudness," *The Journal of the Acoustical Society of America*, 19(4), pp. 609–619, 1947.

[7] Brungart, D. S. and Rabinowitz, W. M., "Auditory localization of nearby sources. Head-related transfer functions," *Journal of the Acoustical Society of America*, 106(May), pp. 1465–1479, 1999.

[8] Arend, J. M., Neidhardt, A., and Pörschmann, C., "Measurement and Perceptual Evaluation of a Spherical Near-Field HRTF Set," in *Proc. of the 29th Tonmeistertagung - VDT Int. Conv.*, 2016.

[9] Zahorik, P., "Assessing auditory distance perception using virtual acoustics." *The Journal of the Acoustical Society of America*, 111(4), pp. 1832–1846, 2002.

[10] Brungart, D. S. and Simpson, B. D., "Auditory localization of nearby sources in a virtual audio display," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 107–110, 2001.

[11] Schröder, D. and Vorländer, M., "RAVEN: A Real-Time Framework for the Auralization of Interactive Virtual Environments," *Forum Acusticum*, pp. 1541–1546, 2011.

[12] Pörschmann, C. and Wiefling, S., "Perceptual Aspects of Dynamic Binaural Synthesis based on Measured Omnidirectional Room Impulse Responses," in *International Conference on Spatial Audio*, 2015.

[13] Pörschmann, C., Stade, P., and Arend, J. M., "Binauralization of Omnidirectional Room Impulse Responses - Algorithm and Technical Evaluation," in *Proceedings of the DAFx 2017*, pp. 345–352, 2017.

[14] Bernschütz, B., "A Spherical Far Field HRIR / HRTF Compilation of the Neumann KU 100," in *Proc. of the 39th DAGA*, pp. 592–595, 2013.

[15] Bernschütz, B., *Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording*, Dissertation, TU Berlin, 2016.

[16] Pörschmann, C. and Zebisch, A., "Psychoakustische Untersuchungen zu synthetischem diffusen Nachhall Psychoacoustic Investigations on synthetically created diffuse Reverberation," in *Proc. of the 27th Tonmeistertagung - VDT Int. Conv.*, pp. 539–550, 2012.

[17] Lindau, A., Kosanke, L., and Weinzierl, S., "Perceptual evaluation of physical predictors of the mixing time in binaural room impulse responses," *Journal of the Audio Engineering Society*, 60(11), pp. 887–898, 2012.

[18] Stade, P. and Arend, J. M., "Perceptual Evaluation of Synthetic Late Binaural Reverberation Based on a Parametric Model," in *AES Conference on Headphone Technology*, pp. 1–8, 2016.

[19] Coleman, P., Franck, A., Menzies, D., and Jackson, P. J. B., "Object-based reverberation encoding from first-order Ambisonic RIRs," in *Proceedings of 142nd AES Convention*, pp. 1–10, Berlin, Germany, 2017.

[20] Pörschmann, C. and Stade, P., "Auralizing Listener Position Shifts of Measured Room Impulse Responses," *Proceedings of the DAGA 2016*, pp. 1308–1311, 2016.

[21] Vazquez Giner, A., "Scale - Conducting Psychoacoustic Experiments with Dynamic Binaural Synthesis," *Proceedings of the DAGA2015*, pp. 1128–1130, 2015.

[22] Geier, M., Ahrens, J., and Spors, S., "The soundscape renderer: A unified spatial audio reproduction framework for arbitrary rendering methods," in *Proceedings of 124th Audio Engineering Society Convention 2008*, pp. 179–184, 2008.

[23] Stade, P., Bernschütz, B., and Rühl, M., "A Spatial Audio Impulse Response Compilation Captured at the WDR Broadcast Studios," in *Proc. of the 27th Tonmeistertagung - VDT Int. Conv.*, pp. 551–567, 2012.

[24] Bernschütz, B., Vázquez Giner, A., Pörschmann, C., and Arend, J. M., "Binaural reproduction of plane waves with reduced modal order," *Acta Acustica united with Acustica*, 100(5), pp. 972–983, 2014.

[25] ITU-R BS.1770-4, "Algorithms to measure audio programme loudness and true-peak audio level," 2015.

[26] Greenhouse, S. W. and Geisser, S., "On methods in the analysis of profile data," *Psychometrika*, 24(2), pp. 95–112, 1959.

[27] Hochberg, Y., "A sharper bonferroni procedure for multiple tests of significance," *Biometrika*, 75(4), pp. 800–802, 1988.

* Please note that the AES Reviewed Papers at Tonmestertagung can be published by both, AES and VDT, in print, online and as PDF download.

AES Reviewed Paper at Tonmeistertagung 2018

Presented* by VDT in cooperation with the

Central European Region of the Audio Engineering Society (AES).

Spatial vs. Static sound - the differences it makes for 360° videos, highlighted by heatmaps

S. Rothe[1], M. Rieger[2]

[1] *Ludwig-Maximilians-Universität München, Germany, Email: sylvia.rothe@ifi.lmu.de*

[2] *VRTONUNG, Munich, Germany, Email: info@vrtonung.de*

Abstract

Cinematic virtual reality has been increasing in popularity over the last years. While watching 360° movies with head mounted displays, viewers can freely choose the direction of view, and thus the visible section of the movie. So, the viewer has a spatial visual experience, however, for the most 360° movies currently the sound is not spatial. For validating the influence of spatial sound, a user study was conducted, where the viewing behaviour for VR videos with spatial and non-spatial sound was compared. We used heatmaps of viewing directions and applied statistical analysis methods to spatiotemporal data. Using approaches of spatial statistics – analysis of space-time cubes and Getis Ord Gi* statistic – we found out that the behaviour for both conditions was often similar. However, in the spatial case, objects with sound attract the viewer's attention for a longer time.

1. Introduction

360° movies are attracting widespread interest and have many possible applications, e.g. telling stories about exciting locations in the world or ancient places of interest in history. Especially, museums and other educational institutions can take advantage of this as well as documentaries or travel experiences who bring the user to places, they can't go to.

In **cinematic virtual reality (CVR)** the viewer watches 360° movie using a **head mounted display (HMD)** or other VR devices. Thus, the viewer is inside the scene, and can freely choose the direction of view. Accordingly, the viewer determines the visible section of the movie – the **field of view (FoV).** Therefore, it is not always possible to show the viewer what is important for the story. Several conventional methods of filmmaking for guiding the viewer's line - such as close ups or zooms - are not practicable in CVR.

In our work we examine where the participants are looking at without any search tasks. We investigate if they follow several sound cues. In contrast to traditional movies, it is unclear if the source of a sound is in the FoV of the viewer, it depends on the viewing direction. In our experiments we compare the viewing behaviour for movies with spatial sound (Ambisonics, .tbe-file) and non-spatial (stereo .wav-file) sound. We are interested whether sound can attract the viewer's attention and if there is a difference watching a movie with spatial and non-spatial sound.

Non-spatial or static means that the sound always remains the same, even if the viewer moves the head and the field of view is changing. This sort of sound is also called head-locked stereo and here represents a downmix also named as static binaural stereo. So, the soundtrack for every user of this group always remained identical and is not depending on the head movement of the viewer.

Spatial means that the sound, just like the image, changes through head-movements of the viewer. The direction of the sound is identical with the source of the sound in the movie. Synonyms would be 3D audio, immersive sound, dynamic binaural stereo. The soundtrack was perceived as something different for each user. While the so-called 360° sound field was the same for each user, the sound for the current direction

*: Please note that the AES Reviewed Papers at Tonmeistertagung can be published by both, AES and VDT, in print, online and as PDF download.

ISBN 978-3-9812830-9-9

was rendered in real time depending on the head-movement of the viewer.

To explore this, we logged the head movements of the participants and visualized them in several ways. Similar to geodata, our logged data have two space coordinates on the surface of a sphere (from -90° to 90°, from -180° to 180°, fig. 1) and a time coordinate (timecode of the movie). The head direction is a vector (λ, φ) where λ represents the yaw and φ the pitch.

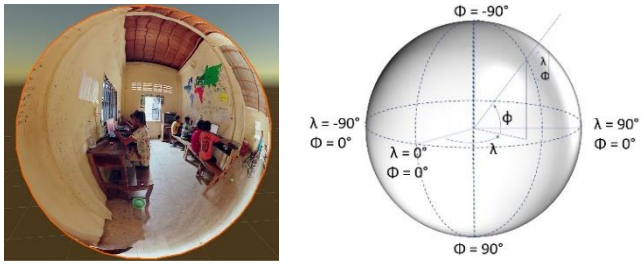


Fig. 1. The sphere of a 360° movie has the same coordinate system used for the earth.

For studying our collected data, we developed an analysing tool for movies which generates a heatmap for every timecode. Basically, there are several options for displaying the heatmap, e.g. via HMD or using flat displays. The perspective of the re-searcher, to explore the data on a sphere via HMD seemed unsuitable to us, since just a small part of the data is visible.

For presenting a full 360° image on a flat screen, the movie is represented by an **equirectangular projection** of the 2-dimensional sphere to a flat plane:

$$P: \begin{pmatrix}\lambda\\ \varphi\end{pmatrix} \rightarrow \begin{pmatrix}x\\ y\end{pmatrix}, \quad \lambda, x \in [-\pi, +\pi], \quad \varphi, y \in \left[-\frac{\pi}{2}, +\frac{\pi}{2}\right]$$

Our tool uses this approach and transforms the logged data in the same way. The projected data are visualized on the movie using **heatmaps**.

Using equirectangular projection for the movie and the data has the advantage of seeing the whole picture at once. However, the further an area is away from the horizontal 0°-line, the more the area is **distorted** (fig. 2).

Although we took the deformation into account when developing our heatmap tool, we assume that for our experiments there was no big influence of the distortion on the results, because head movements far from the horizontal line were very rare in our dataset. As can be seen in fig. 9, there were only few data outside the 20°-area. This was caused by the composition of the movie - all points of interest were near the horizontal line (0°) - and by the fact that we used head tracking and not eye tracking. Comfortable head movements range in a 30°-area only [1], whereas eye movements deviate further from the 0°-line. For this reason, we neglected the distortion in the inferential statistical analysis part of this paper.

Fig. 2. A manual generated heatmap for studying the distortion in the heatmap tool.

Heatmaps are suitable for finding clusters in the data, but they are not sufficient for inferential statistics to make statements about confidence. Therefore, **statistical methods** are needed for determining the significance of the clusters – for identifying hotspots.

The term **hotspot** is used in the literature in different ways. In our work, a hotspot is a cluster with high values identified by statistical methods using confidence levels. In our dataset, the value is the number of views in a space-time segment.

Our collected data are spatiotemporal data – data which have a space and a time component. This type of data is often used in geographical researches, so we applied some methods of spatial statistics which are used in geography: space-time cube analysis and hotspot analysis using Getis-Ord Gi* statistic.

2. Related Work

Much research in recent years has focused on presence in virtual reality environments – which can be adapted to CVR. Spatial sound leads to a higher level of presence [2,3] and increases the sense of place [4]. In our research, we focused on finding out where the viewer is looking at and if spatial sound changes the viewing behaviour.

In 1996 Pausch et al.[5] examined how the attention of the viewer can be drawn to a desired spot. The orientation of the head was logged, and conventional histograms were used for illustrating the data. The results show that most people hardly turned their head. The authors visualized this by histograms showing the rotation angle. Similarly, Sheikh et al. [6] connected several cues (motion, gestural and audio cues) to the main character of a scene. The head orientation was recorded and the percentage of people who had seen the target over time was evaluated. In their experiments, the cues with an audio component were proved to be more helpful than just visual cues, even if the sound was not fully spatialized. The results were displayed by diagrams showing the time for seeing the target. In histograms, tables and diagrams specific values were presented. With our approach, we want to illustrate the viewing behaviour in heatmaps and analyse the data using spatial statistical methods.

Spatial audio can be used for solving tasks in virtual and augmented reality. The use of spatial sound improves the results in search and navigation tasks [7,8]. Van der Burg et al. [9] showed that audio cues (pop) synchronized to a salient visual cue (pip) reduces the search time, even if the audio cue

does not have any location information. Emil R. Høeg et al. [10] enhanced this experiment to virtual reality with sound cues from the same direction as the visual cue. They demonstrated that binaural cues lead to shorter search times, even though the visual cue was not always visible at the moment the audio cue was presented. In the experiments, the participants were given a search task in an abstract VR environment. In our study in comparison to [9,10], we move closer to a real cinematographic setting by using a realistic scene instead of abstract symbols and by not giving a concrete task to the participants but letting them choose freely what to do next.

To investigate the viewing direction for every timecode in the movie, we use space-time cubes (STC) and the Getis-Ord Gi*statistic. STCs were introduced by Hägerstrand in 1970 [11] (Hägerstraand 1970) and can be used for analysing geographical data [12–14]. The Getis-Ord Gi* statistic was established by Getis and Ord [15,16] for analysing spatial data. Songchitruksa et al. [17] used the Getis-Ord spatial statistic for identifying hotspots.

3. Analysing Methods

3.1. Heatmaps

In our experiments, the centre of the field of view (FoV) was recorded, and a heatmap was generated for every time interval in the movie. The heatmap represents how frequently viewers looked at a certain area segment within a specific time frame. For this we mapped the interior of a circle to each viewpoint and used a gradient function to compute the colour and the opacity of the overlapping areas.

Our heatmap tool is implemented in HTML5 using the HTML5 video element. This element has an additional HTML-tag: the track tag which was originally designed for text tracks defining subtitles. We used this track for generating the heatmap (fig. 3). A track tag requires a vtt-file (video text track). The recording script logged the head direction 60 times per second, which can be adjusted. The data were written in vtt-files.

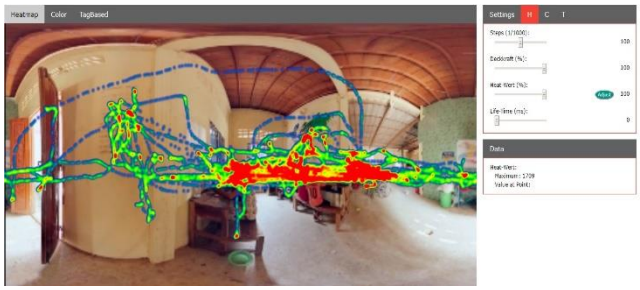


Fig. 3. View of the movie in the heatmap mode.

Investigating the heatmap, several parameters can be adjusted, e.g. the time interval, the cell radius, the scaling and the opacity (heatmap mode, fig.3).

Inspecting heatmaps is a first approach for finding conspicuities and clusters in the data. However, heatmaps do not provide any information about the significance of the results. For this it needs examinations using statistical methods.

3.2. Space-time Cubes

A method of investigating spatiotemporal data is to analyse space-time cubes (STC). In these cubes two coordinates represent the space and the third one represents the time. With this technique the data can be visualized and explored in a comfortable way (fig.4).

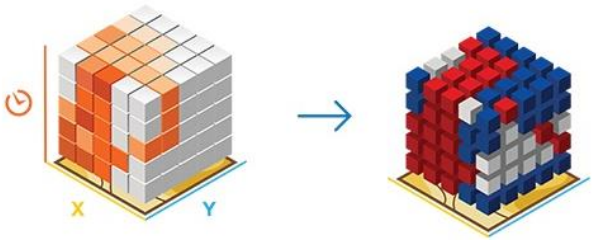


Fig. 4. A space-time cube has one time and two space coordinates, left: the density of a value is shown for every space-time section, right: the result of a hotspot analysis (significant hotspots are red) [18]

For calculating the STC and applying spatial statistical methods, the GIS software ArcGIS Pro was used. We converted the vtt-file into an Excel file which was imported into ArcGIS to employ the implemented STC method for analysing the data.

In a first step, we inspected space-time cubes showing the counts of incidents – in our case "how often" users looked at a certain area segment within a specific time frame, as shown in fig. 5.

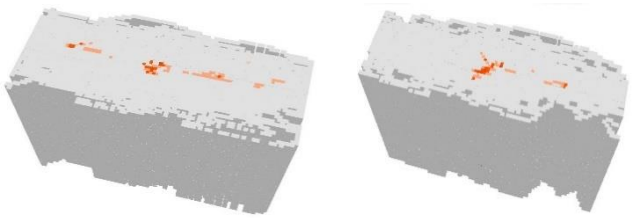


Fig. 5. The space-time cubes for both groups (left: movie with spatial sound, right: movie with non-spatial sound), there is one time coordinate and two space coordinates. For investigating the data, the slices of the cube are useable.

In this way we could see for every time interval (we used one second) where the most views are. However, the correlation between the neighbouring points (in space and time) were not considered in this step. For this we will make use of the hotspot analysis.

A problem of STC for spherical data is the distortion caused by the projection from the sphere to a plane. The further an area is away from the horizontal 0°-line the more relevant is the distortion. However, in our dataset this distortion is negligible, because nearly all data are close to the 0°-horizontal line (fig. 9). Therefore, we used the projected data also in the next step – the statistical analysis - for finding differences in the viewing behaviour watching a movie with spatial and non-spatial sound.

3.3. Getis-Ord Gi* Statistic

For finding statistically significant hotspots, methods of spatial statistics can be applied. These methods take into account the neighbouring relations between the space-time segments in the STC.

The collected data are point incident data, these are points connected to an event – in our case the viewer looked at this point. We were interested in significant clusters. To find such clusters, we used the Getis-Ord Gi* statistic [15]. This statistical method requires values for the investigated points. In order to use it, the incident data were aggregated and incident counts established. The incident counts - the number of views within a segment - are the attribute values which are analysed by the method.

The Getis-Ord Gi* statistic is given as:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \overline{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2\right]}{n-1}}}$$

where x_j is the attribute value for point *j*, $w_{i,j}$ is the spatial weight between point *i* and *j*, *n* is equal to the total number of points and:

$$\overline{X} = \frac{\sum_{j=1}^{n} x_j}{n}$$

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \left(\overline{X}\right)^2}$$

Using the spatial statistic tools of the ArcGIS Pro Software again, we generated STCs displaying hotspots (fig. 6).

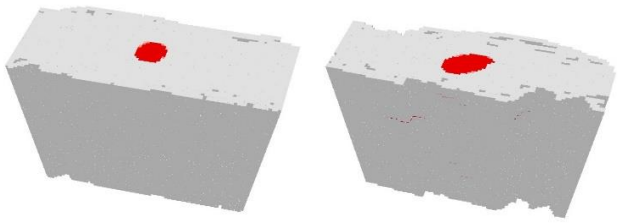


Fig. 6. The space-time cubes with significant hotspots for both groups, (left: movie with spatial sound, right: movie with non-spatial sound), there is one-time coordinate and two space coordinates.

For every point in the STC the p value can be displayed by double-clicking on the point. The p value represents the probability that the observed pattern was created randomly. A small p value means that the pattern is most likely caused by a cluster. Segments with p values smaller than 0.01, which means 99% confidence, are displayed in red.

Using the 3D mode of the GIS tool for a more detailed investigation of the data was very time consuming with our dataset. The 3-dimensional STC was suitable for an overview. However, it was difficult to navigate in the 3D cube due to the size of our dataset. We used an Acer Predator G3-710 (Intel i7-6700, 3.4GHz, 16GB RAM). Navigating in smaller datasets was not a problem. Switching to the 2D mode and modifying the time slices by a slider we could analyse the clustered areas in a comfortable way (fig. 7).

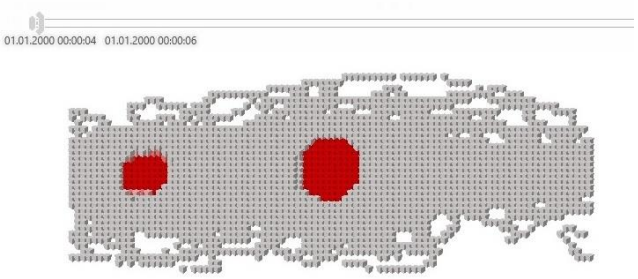


Fig. 7. Using a slider in the 2D-mode we could navigate in the STC changing the time slice. The red areas are significant hotspots.

With this approach we compared the two datasets. We changed the time slices step by step and compared the hotspots generated by using spatial sound with the hotspots generated by using non-spatial sound.

4. User Study

4.1. Material

For our study we utilized a 9:26 min movie ("Crossing Border") made by young filmmakers. This movie was produced with spatial sound and contains many cues for guiding the viewer, for example speaking persons, movements and sounds. The participants sat on a swivel chair while watching the movie via head mounted display (Samsung Gear VR with Samsung Galaxy S6) and headphones. After watching the movie, a short, unstructured interview followed to find hints if the viewer realized something regarding sound.

4.2. Participants

Our study is based on a between-subject test design. The movie was shown to two groups. One group, called group spatial here (20 participants aged between 23 and 65, mean=31.5, 10 women and 10 men), watched the movie with spatial sound, the other group, called group non-spatial here (20 participants aged between 21 and 67, mean=34.7, 8 women and 12 men), with non-spatial sound. Each group contained 6 participants watching CVR movies for the first time, and 3 experienced participants. The other participants watched CVR movies occasionally. There was no special task for the participants.

Before even starting with the session, the test persons were not told what the study is about to avoid any psychological influences with expectations of 3d audio compared to stereo audio. If we told the viewers about the comparisons we wanted to analyse, they would immediately listen more for the soundtrack than a normal consumer. Therefore, group spatial ranged from sound engineers who afterwards weren't sure if the sound was spatial, to laypersons who immediately realized that the sound field was rotating with head-movement.

4.3. Results

Analysing the tracked head movements in total, it is noticeable that in both groups the movements are very limited. We inspected the yaw and pitch angles of head movements separately. The yaw angles (turning the head to the side) were similar for both groups. Most of the time the participants looked straight forward and seldomly turned around (fig. 8). These results resemble those of [5].

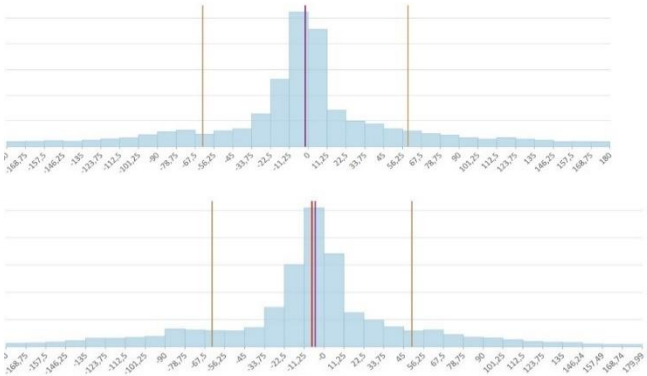


Fig. 8. The participants looked barely left and right in both groups, above: spatial sound, below: non-spatial sound (lines: red - mean, blue - median, yellow - standard deviation), the x axis shows the angle of head movements and the y axis the number of views (the views are recorded 60 per second).

Besides, participants barely moved their heads downwards or upwards (fig.9). Mostly the participants were keeping their head straight ahead, barely more than 15° downwards or 5° upwards.

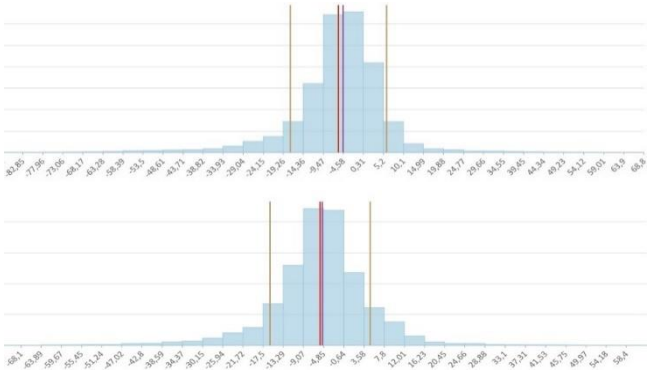


Fig. 9. The participants looked barely up and down in both groups, above: spatial sound, below: non-spatial sound (lines: red - mean, blue - median, yellow - standard deviation), the x axis shows the angle of head movement and the y axis the number of views (the views are recorded 60 per second).

One reason for that is the type of our collected data – we tracked the head movements. Near the horizontal line it is easy to move the head for changing the viewing direction. However, for looking at the ground or sky it is more comfortable to move the head just slightly and expand the viewing direction by eye movements. The sector in which a person moves the head effortlessly is around 30° in both directions [1].

Another reason for the small covered section is the setup of the movie - nearly all points of interest are nearby the horizontal line.

At first sight, the heatmaps for both groups look very similar. When persons were speaking or moving, there were areas of high data density for both groups (fig. 10).

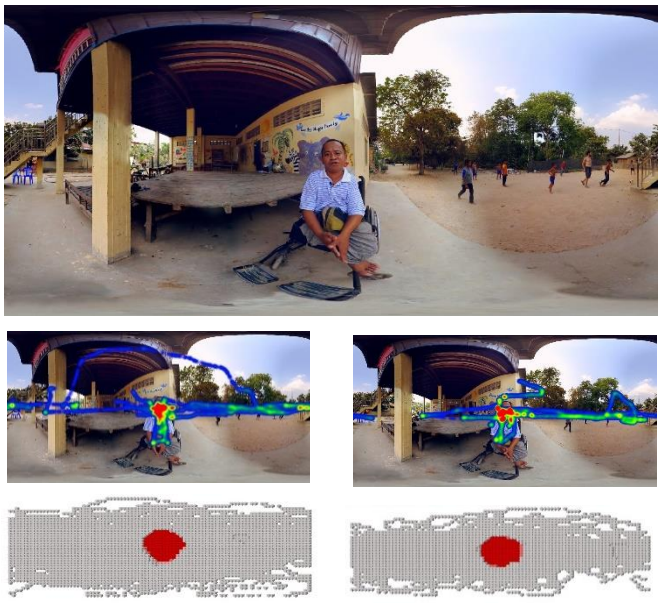


Fig. 10. Speaking or moving people attract the viewer's attention of both groups very similar (left: spatial sound, right: non-spatial sound), top: the scene of the movie, middle: heatmaps, bottom: significant hotspots.

Using the Getis-Ord Gi* statistic for identifying hotspots in the STC we found more hotspots in the data of the group which watched the movie with non-spatial sound (99 hotspots in the group spatial and 149 hotspots in the group non-spatial). Afterwards the STCs were compared, it is possible to navigate through the time slices for a closer inspection. In this way, more difference between the two groups could be found. If in a scene two or more cues occurred simultaneously (for example: speaking in one place and movements in another), the hotspots were split more often in the group that watched the movie with non-spatial sound. For example, figure 11 shows that, in addition to the hotspots around the mopeds, a hotspot around the sign has formed.

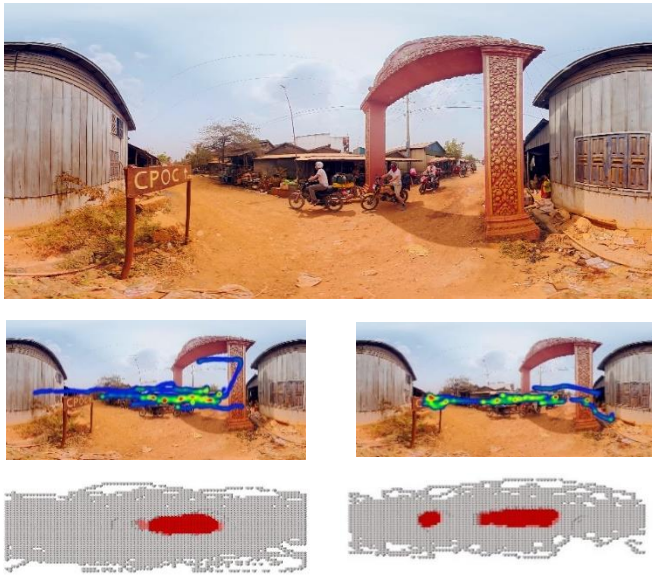


Fig. 11. Slices of the STC (left: movie with spatial sound, right: movie with non-spatial sound), top: the scene of the movie, middle: heatmaps, bottom: significant hotspots.

When people are talking to the camera, the viewer in CVR has the impression that the person is talking to him. In nearly all cases the viewers with non-spatial sound were looking to the speaker approximately 8s before starting to look around. In the case of spatial sound, it took 1-3s longer before watching for other details. In figure 12 - 15 an example is shown. The girl is talking to the camera/viewer about her life. The scene starts at 2:34 (fig.12). At 2.38 (spatial) and 2:39 (non-spatial) nearly all participants are looking at the girl (fig.13). After the girl started talking, the hotspots have not changed for 12s in both groups (fig.14). However, the participants of the group non-spatial started earlier to look around as figure 15 shows.

Fig. 12. Scene image and significant hotspot for time code 2:34, start of speaking (left: spatial, right: non-spatial sound).

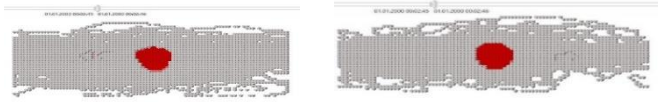


Fig. 13. Hotspot for time code 2:38 (left: spatial) and 2:39 (right: non-spatial), the associated image is nearly the same as in fig. 15.

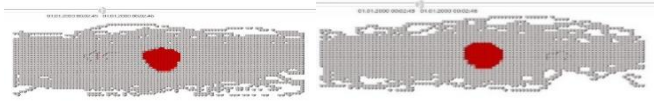


Fig. 14. Hotspot for time code 2:46, 12 s after the girl started to speak (left: spatial, right: non-spatial), the associated image is nearly the same as in fig. 15.

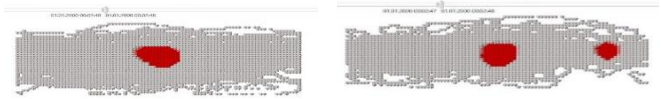


Fig. 15. Hotspot for time code 2:48, 14s after the girl started to speak (left: spatial, right: non-spatial), the associated image is nearly the same as in fig. 15.

We could find very similar results for other timecodes in the movie, where people spoke to the viewer.

5. Discussion and Limitations

5.1. Results highlighted by the Heatmap

Comparing the viewing behaviour watching CVR videos with spatial and non-spatial sound, often the behaviour was similar for both groups. Summarizing we found the following results:

- Speaking or active people draw the attention of the viewer with spatial and non-spatial sound.
- With spatial sound, the viewers observing a person stay a few seconds longer
- Details without sounds draw more attention in the case of non-spatial sound.
- If the viewers saw a scene with a setting that has already been shown earlier through the experience, such as an interview with the same protagonist in the same environment, both groups started looking away from the interviewee faster than at the first time.

We used heatmaps and spatial statistical methods for inspecting the data. Heatmaps have the advantage of being connected directly to the playing movie. So, it is very comfortable to inspect the data visually. However, the discovered results must be verified for significance using statistical methods. For finding statistically significant results we applied the Getis-Ord Gi* statistic analysing STCs. On this way we could determine significant hotspots in our dataset and compare the viewing behaviour for spatial and non-spatial sound.

The applied methods projected the movie and the data to a flat screen. This causes a distortion in areas far away from the 0°-horizontal line. In our heatmap tool this was considered, so the calculation was also correct for areas away from the horizontal line.

However, the problem of distortion was theoretically present in our spatial statistical analysis - using the STC. Since in our dataset all points are close to the horizontal 0°-line we neglected the distortion. As the STC has proved of value for inspecting tracking data, it is worth to expand this method by considering the distortion.

5.2. Results highlighted by the Survey

Since group spatial was watching talking protagonists longer, we gathered interesting opinions from the subjects why this is. One stated that she “felt rude looking away”. Like in real-life, people would look at a person who is talking the oneself. If they turned away, they wouldn’t see the fellow human being anymore, but would still hear the person talking from the side. There seems to be a relation in VR, meaning that spatial sound gives the viewers a bigger feel of presence from the virtual person that non-spatial sound is not capable of, due to the lack of localisation during head-tracking. So spatial sound seems to unconsciously remind the viewer of another person still being nearby.

Another finding can’t be highlighted by the heatmap. Some viewers of group non-spatial were confused by the voice-over since it doesn’t give an audible cue if the talking human is visible inside of the scene or not and thus a voice-over. In both

cases, the dialogue was perceived mono as in head localisation and not spatial. But since group spatial could hear the interviewee spatialized with head-tracking, they immediately knew that the person was visible in the scene and a voice-over in cut-scenes if it was mono, so head-locked. This presents an intuitive way of making the viewer understand, what kind of elements are visible in the current environment and the distinction of film language.

5.3. Limitations

In our study, we tracked the head movements – not the eyes. Through that, we could follow the viewing direction in general, however not in detail. This was sufficient for a first approach and the experiments led to evaluable data. Tracking the eyes instead the head would give more detailed information but also needs more efforts for considering the distortion.

The content used in this case-study had its main action focused mainly at 0° azimuth, so there was little motivation for the viewer, to rotate more than 180° around the z-axis. But the invisible zone around the viewer is important for triggering head-movement with spatial sound and therefore might show more differences when comparing its soundtrack to a stereo-downmix.

Since the stereo-downmix represented a binaural print of the 0° position azimuth and 0° elevation, it already contained spatial information and does not represent a classic stereo mix which is also available for loudspeaker playback.

6. Conclusion

In this research we investigated where the viewer is looking in cinematic virtual reality. We compared the behaviour watching movies with spatial and non-spatial sound and found that sound draws the attention of the viewer even if it is non-spatial. These results can be used for integrating cues in a movie for guiding the attention of the viewer to things which are important for the story.

Our approach of using spatial statistical methods has proved of value and we could investigate if clusters of viewing directions are notable. This method is usable for all spatiotemporal datasets, especially head and eye tracking data. Further investigations are necessary to explore the viewers' behaviour in cinematic virtual reality.

7. References

[1] Dreyfuss, H.; others The measure of man: human factors in design; Whitney Library of Design New York, 1967;

[2] Poeschl, S.; Wall, K.; Doering, N. Integration of spatial sound in immersive virtual environments an experimental study on effects of spatial sound on presence. In 2013 IEEE Virtual Reality (VR); IEEE, 2013; pp. 129–130.

[3] Larsson, P.; Västfjäll, D.; Olsson, P.; Kleiner, M. When what you see is what you hear: Auditory-visual integration and presence in virtual environments. Proc. 10th Annu. Int. Work. Presence, Oct. 25-27, 2007, Barcelona, Spain 2007, 11–18.

[4] Dreyfuss, H.; others The measure of man: human factors in design; Whitney Library of Design New York, 1967;

[5] Pausch, R.; Snoddy, J.; Taylor, R.; Watson, S.; Haseltine, E. Disney's Aladdin. In Proceedings of the 23rd annual conference on Computer graphics and interactive techniques - SIGGRAPH '96; ACM Press: New York, New York, USA, 1996; pp. 193–203.

[6] Brown, A.; Sheikh, A.; Evans, M.; Watson, Z. Directing attention in 360-degree video. In IBC 2016 Conference; Institution of Engineering and Technology, 2016; p. 29 (9 .)-29 (9 .).

[7] Rumiński, D. An experimental study of spatial sound usefulness in searching and navigating through AR environments. Virtual Real. 2015, 19, 223–233, doi:10.1007/s10055-015-0274-4.

[8] Gunther, R.; Kazman, R.; MacGregor, C. Using 3D sound as a navigational aid in virtual environments. Behav. Inf. Technol. 2004, 23, 435–446, doi:10.1080/01449290410001723364.

[9] Van der Burg, E.; Olivers, C. N. L.; Bronkhorst, A. W.; Theeuwes, J. Pip and pop: Nonspatial auditory signals improve spatial visual search. J. Exp. Psychol. Hum. Percept. Perform. 2008, 34, 1053–1065, doi:10.1037/0096-1523.34.5.1053.

[10] Hoeg, E. R.; Gerry, L. J.; Thomsen, L.; Nilsson, N. C.; Serafin, S. Binaural sound reduces reaction time in a virtual reality search task. In 2017 IEEE 3rd VR Workshop on Sonic Interactions for Virtual Environments (SIVE); IEEE, 2017; pp. 1–4.

[11] Hägerstraand, T. WHAT ABOUT PEOPLE IN REGIONAL SCIENCE? Pap. Reg. Sci. 2005, 24, 7–24, doi:10.1111/j.1435-5597.1970.tb01464.x.

[12] Kraak, M. J. The space - time cube revisited from a geovisualization perspective 2003, 1988–1996.

[13] Kristensson, P. O.; Dahlback, N.; Anundi, D.; Bjornstad, M.; Gillberg, H.; Haraldsson, J.; Martensson, I.; Nordvall, M.; Stahl, J. An Evaluation of Space Time Cube Representation of Spatiotemporal Patterns. IEEE Trans. Vis. Comput. Graph. 2009, 15, 696–702, doi:10.1109/TVCG.2008.194.

[14] Diansheng Guo; Jin Chen; MacEachren, A. M.; Ke Liao A Visualization System for Space-Time and Multivariate Patterns (VIS-STAMP). IEEE Trans. Vis. Comput. Graph. 2006, 12, 1461–1474, doi:10.1109/TVCG.2006.84.

[15] Getis, A.; Ord, J. K. The Analysis of Spatial Association by Use of Distance Statistics. Geogr. Anal. 1992, 24, 189–206, doi:10.1111/j.1538-4632.1992.tb00261.x.

[16] Ord, J. K.; Getis, A. Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. Geogr. Anal. 1995, 27, 286–306, doi:10.1111/j.1538-4632.1995.tb00912.x.

[17] Songchitruksa, P.; Zeng, X. Getis–Ord Spatial Statistics to Identify Hot Spots by Using Incident Management Data. Transp. Res. Rec. J. Transp. Res. Board 2010, 2165, 42–51, doi:10.3141/2165-05.

[18] Visualize Space Time Cube in 3D—Help | ArcGIS Desktop Available online: http://desktop.arcgis.com/en/arcmap/latest/tools/space-time-pattern-mining-toolbox/visualizecube3d.htm (accessed on Jul 5, 2018).

[19] Dreyfuss, H.; others The measure of man: human factors in design; Whitney Library of Design New York, 1967;

Automatisches Audiotracking in der Praxis

Dirk Ziervogel[1], Alexander Nickol[2]
[1] *Projektleitung & Fachplanung, Email: dirk.ziervogel@neumannmueller.com*
[2] *Projektleitung & Fachplanung, Email: alexander.nickol@neumannmueller.com*

Abstract

Nach einer Einführung in die Themenbereiche Ortung und Tracking geht es voranging um deren Implementierung in die Tontechnik auf dem aktuellen Stand der Technik. Problemstellungen und praxistaugliche Lösungsansätze werden unter anderem im Hinblick auf optimale Lautsprecherpositionierung und deren Anzahl, Adaption der Laufzeiten mittels automatisierter Mappingverfahren und notwendige Anpassungen, Arbeitsweisen mit statischen und bewegten Quellobjekten sowie Trackingsysteme deren Workflow und Schnittstellen besprochen. Im Anschluss ist ein Erfahrungsaustausch ausdrücklich erwünscht.

1. Ortung vs. Tracking

Ortung oder auch akustische Lokalisation ist ein passiver Vorgang der Richtungs- und Entfernungslokalisation zur räumlichen Erfassung akustischer Objekte. Dabei sind Laufzeit- und Pegeldifferenzen am Ohr, individuell unterschiedliche Resonanzen an der Ohrmuschel, die erste Wellenfront und Reflexionsmuster sowie Klangfarben und Erfahrung die zentralen Aspekte.

Tracking beschreibt die aktive gleichzeitige Verfolgung bewegter Objekte. Die Systeme liefern Beobachtungsdaten zu Ort (xyz), Lage und Ausrichtung, Geschwindigkeit und Beschleunigung. Dabei sind je nach Einsatz unterschiedliche Abfrageraten von einem konkreten Zeitpunkt bis zu zyklischen Intervallen möglich. Durch spezifische Algorithmen wird auch eine Vorrauschau möglich.

1.1. Ortung in der Beschallung

Die Betrachtung der akustischen Lokalisation in der Beschallung ist nötig um einen korrekten Bezug zwischen akustischem und visuellem Eindruck herzustellen. Zielsetzung dabei ist es für den Zuhörer ein möglichst natürliches Hörerlebnis zu schaffen. Dafür wird versucht einen Ortungsbezug zur tatsächlichen Quelle und nicht zum Lautsprecher herzustellen. Bei bspw. sehr breiten Bühnen oder örtlich komplexeren Darbietungen ist die korrekte Ortung ein wichtiger Aspekt für die Orientierung des Zuschauers. Darüber hinaus soll Ortung für einen möglichst großen Teil des Publikums funktionieren und somit allen Zuhörern gleichberechtigt eine korrekte akustische Lokalisation der Quellen ermöglichen. Durch einen genauen Abgleich vom akustischen und visuellen Eindruck erleichtert man dem Zuhörer deutlich den Zugang zu transportierten Inhalten. Denn je niedriger die Diskrepanz zwischen akustischer Lokalisation und visuellem Eindruck ist, desto geringer ist die notwendige Anpassungsleistung des Zuschauers. Der Konzentrationsaufwand und die Ermüdung sinken entsprechend und das Bühnengeschehen ist für den Zuschauer deutlich einfacher nachzuvollziehen. Bei komplexen akustischen Ereignissen verbessert sich die Verständlichkeit und Transparenz durch eine differenzierte Abbildung der tatsächlichen Verteilung der Quellen auf horizontaler, vertikaler und zeitlicher Ebene.

1.2. Adaption der Laufzeiten

Das klassische akustische Einmessverfahren stellt eine relativ aufwändige Variante zur Ermittlung der benötigten Lautzeitdifferenzen dar. Denn jede Lautsprecherposition muss für jeden benötigten Ortungsbezug einzeln erneut eingemessen werden. Entsprechend hoch kann der zeitliche Aufwand bei großen Systemen mit mehreren Ortungsbezügen werden. Die technische Umsetzung der Signalverteilung und Verzögerung erfolgt dann mittels einer statischen Delay-Matrix. Die Ergebnisse sind allerdings häufig kompromissbehaftet, da die Wahl eines geeigneten Messpunktes großen Einfluss hat und der gewünschte Ortungsbezug meist nicht für alle Positionen gleichermaßen gut funktioniert. Zudem ist der Aufwand bei größeren Systemen mit vielen Lautsprechern und mehreren Ortungsbezügen entsprechend hoch.

Eine Alternative dazu stellen automatisierte Verfahren dar. Neue Delay-Matrix-Systeme wie bspw. SoundHub, DS100 oder L-ISA sind in der Lage Informationen zu Lautsprecherpositionierung, Ausrichtung und Abstrahlverhalten sowie objekt- oder zonenbasierten Ortungsbezügen zu verarbeiten. Dabei werden unterschiedliche Delay-Matrix-Verteilungen entsprechend der internen Algorithmen je Ortungsobjekt oder Zone bezogen auf die Lautsprecherpositionen automatisch ermittelt. Somit lassen sich sehr schnell komplexe Setups auch mit unterschiedlichen Ortungsbezügen erstellen wie bspw. in Abb. 1 und 2. Allerdings müssen die Daten mit großer Sorgfalt eingepflegt und gegebenenfalls auf die örtlichen Gegebenheiten angepasst werden um funktionierende Ergebnisse zu erhalten. Anpassungen oder Korrekturen einzelner Lautsprecherpositionen können bei Bedarf manuell vorgenommen werden.

ISBN
978-3-9812830-9-9
Der Autor hält die Urheberrechte für den gesamten Inhalt seines Manuskripts. Er überlässt dem VDT das Recht auf Veröffentlichung in gedruckter Form, online oder als PDF-Download von https://tonmeistertagung.com

Abb.1: DS100 - Big Band Setup mit Ortungsbezügen zu allen einzelnen Instrumenten

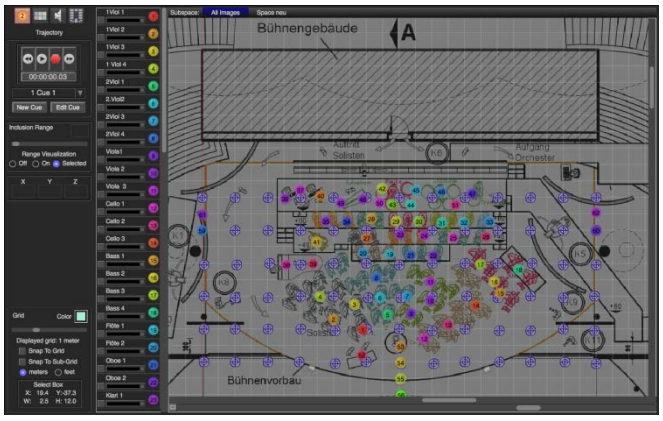


Abb.2: SoundHub - Orchester mit Ortungsbezügen zu allen einzelnen Instrumenten und Instrumentengruppen

1.3. Bewegte Objekte

Mit Hilfe der automatisierten Verfahren werden dynamische Ortungsbezüge mit sich bewegenden Objekten möglich, da die Martixverteilung automatisiert angepasst werden kann. Hier lässt sich in objekt- und zonenbasierte Systeme unterscheiden. Bei objektbasierenden Systemen findet eine kontinuierliche Anpassung des „Mappings“ auf die Lautsprecher für das sich bewegende Objekt statt. Dagegen ändert sich der Ortungsbezug bei zonenbasierten Systemen durch Überblendung zwischen den jeweils festen Martixverteilungen entsprechend der Ortungszone. Durch moderne Algorithmen sind auch schnelle Bewegungen ohne wahrnehmbare Artefakte möglich geworden.

Bewegungsabläufe lassen sich entweder durch zuvor programmierte Automationsspuren in DAW´s oder bspw. Q-Lab erzeugen oder durch manuelle Positionierung der Ortungsobjekte in der Matrix-Software. Die Anbindung eines Trackingsystems ermöglicht es die tatsächlichen Bewegungen der Objekte in Echtzeit zu verarbeiten.

2. Tracking

Als Tracking bezeichnet man im Allgemeinen eine Liste von Punkten im Koordinatensystem, mit denen eine exakte räumliche Positionierung von Objekten möglich wird

Seinen Ursprung fand das Tracking in der Industrie um Logistikprozesse zu überwachen und zu automatisieren. Heute dienen Trackingsysteme als Werkzeug um z.B. sportliche Abläufe zu analysieren, Handelsprozesse zu optimieren und auch im Bereich der Ton- Licht –und Videotechnik Prozesse zu automatisieren. Hierfür gibt es verschiedenste Systeme mit unterschiedlichen Lokalisierungsverfahren und anwendungsspezifischen Differenzen in Bezug auf Präzision. Für die Anwendungszwecke in der Veranstaltungsbranche sind hauptsächlich RTLS (real time locating system) Systeme mit folgenden Verfahrensweisen von Interesse:

AOA –Angle of Arrival beschreibt die Positionsbestimmung der ankommenden Signale über den Einfallswinkel zu den jeweiligen Ankern. Die Positionsbestimmung erfolgt über Triangulation.

TOA – Time of Arrival beschreibt die Positionsbestimmung über die Signallaufzeit zwischen zwei Ankern.

TDOA – Time Difference of Arrival beschreibt die Positionsbestimmung aus der Differenz der Empfangszeitpunkte der gesendeten Signale zu den Ankern.

Es gibt Systeme die mehrere Verfahrensweisen miteinander kombinieren.

2.1. Einrichtung eines Trackingsystems

Ein Trackingsystem besteht zumeist aus drei Komponenten. Einem Server, einer definierten Zahl an Ankern (Empfänger) und Sensoren (Sender). Beim Aufbau liegt das Hauptaugenmerk auf der Ankerpositionierung und der Frage ob 2D oder 3D getrackt werden soll? Bei einem 2D Tracking ist es ausreichend, wenn alle Anker in einer Höhenebene positioniert werden. Wohingegen es beim 3D Tracking darauf ankommt Anker in den verschiedensten Ebenen zu installieren um eine gute Performance zu erhalten. Bei beiden Verfahren spielen Umgebungsvariablen wie bspw. reflektierende Oberflächen, Abschirmungen durch Metalle oder direkte Sichtverbindungen eine große Rolle (Line of sight).

Nach entsprechender Ankerpositionierung muss nun das komplette System eingemessen werden. Auch hier gibt es Unterschiede in Bezug auf manuelle oder automatisierte Einmessverfahren. Es hat sich jedoch herausgestellt, dass akkurates manuelles Einmessen, bspw. mit Hilfe eines Tachymeters, die optimale und exakteste Methode darstellt. Alle gemessenen Werte werden als XYZ Koordinaten angegeben, wobei penibel darauf zu achten ist, dass positive und negative Werten nicht verwechselt werden.

Um die Funktionsfähigkeit zu verifizieren, muss der zu ortende Bereich mit einem Sensor abgeschritten werden. Anhand einer Wegeaufzeichnung lässt sich nachvollziehen wie gut das System tatsächlich arbeitet. Treten DropOuts aufgrund von Reflexionen auf oder werden Signale blockiert, müssen Ankerpositionen angepasst werden. Beispielhaft sind entsprechende Visualisierungen und Statusinformationen in Abb. 3 und 4.

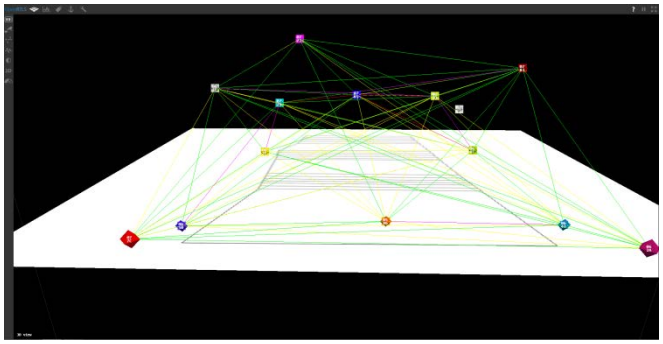


Abb.3: OpenRTLS - 3D View der Ankerpositionen

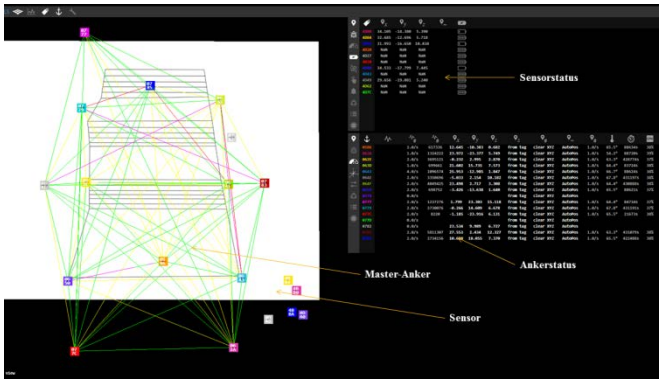


Abb.4: OpenRTLS - 2D View mit Sensor- und Ankerstatus

3.2 Anbindung und Anpassung

Für die Anbindung eines Trackingsystems an eine Delay-Matrix ist es wichtig zu wissen, welche Datenformate verarbeitet werden können. Standardformate wie OSC, Midi, ArtNet finden hier häufig Anwendung. Gelegentlich sind Anpassungen der Systeme aneinander im Hinblick auf das Koordinatensystem nötig.

Mittels externer Software besteht außerdem die Möglichkeit Rohdaten direkt zu verarbeiten. Durch Filterung ist sowohl eine Glättung als auch Vorausschau möglich.

3. Lautsprecherpositionierung

Für eine gute Lautsprecherpositionierung sollte man sich zunächst Gedanken darüber machen was abgebildet werden soll und welche Ebenen dafür nötig sind. Grundsätzlich lässt sich zwischen symmetrischen und asymmetrischen Anordnungen unterscheiden. Auch klassische Systemanordnungen wie in Abb. 5 mit Left, Center, Right, Near-, In-, OutFill und Delay´s profitieren in hohem Maße von objektorientiertem Time- und Level-Mapping je Lautsprecherposition. Vor allem bei komplexem Programm wie bspw. klassischem Orchester lässt sich dadurch eine differenzierte horizontale Verteilung und zeitliche Tiefenstaffelung umsetzten und somit eine deutlich transparentere Abbildung schaffen. Aufgrund der relativ geringen Anzahl an Lautsprecherpositionen funktioniert der differenzierte Ortungsbezug nicht überall gleichermaßen gut. Um ein homogeneres Bild für einen größeren Publikumsbereich zu erzielen, empfiehlt es sich mit mehr Lautsprecherpositionen zu arbeiten wie in Abb. 6. Große Systeme können durch mehrere Kleine ersetzt werden wobei auf regelmäßige Abstände und ein breites Abstrahlverhalten zu achten ist. Ziel dabei ist es, dass sich Zuhörer möglichst im Abdeckungsbereich mehrerer Lautsprecher befinden. Zum einen wird dadurch eine nochmals differenziertere Abbildung aber vor allem ein homogenerer Ortungsbezug für das Publikum ermöglicht und direkte Ortung einzelner Lautsprecherpositionen verringert. In den seitlichen Randbereichen ergibt sich eine natürliche Abnahme der komplexen Abbildung einhergehend mit dem visuellen Eindruck, da man sich nur noch im Abdeckungsbereich weniger Lautsprecher befindet.

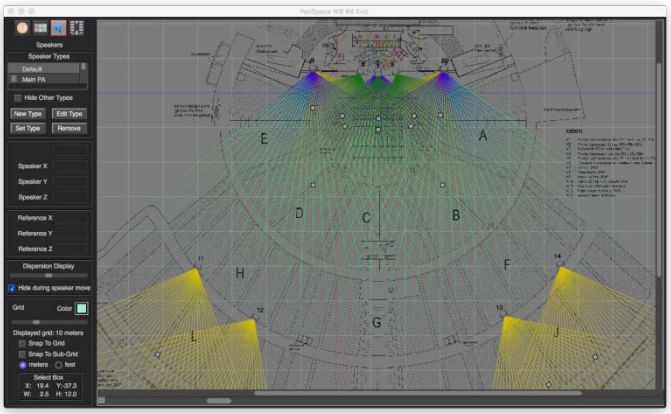


Abb.5: SoundHub - klassische Lautsprecheranordnung

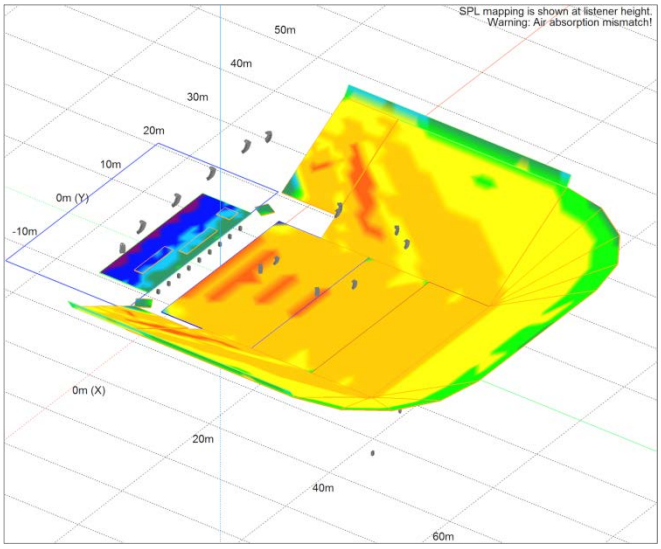


Abb.6: ArrayCalc - symmetrische Lautsprecheranordnung mit mehr Positionen

Bei asymmetrischer Lautsprecheranordnung wird meist eine Hauptbeschallung durch Einzellautsprecher an ortungsrelevanten Positionen ergänzt, siehe Abb. 7. Die Positionierung richtet sich ausschließlich nach der erwünschten Ortung, wobei versucht wird einen primäreren Bezug herzustellen. Dadurch werden 3-dimensionale Staffelungen und autarke Einzelpositionen möglich. Breitabstrahlende Lautsprecher an möglichst vielen Positionen sind auch hier vorteilhaft um einen homogenen Ortungsbezug zu ermöglichen aber auch die nötige Flexibilität zu erhalten. Allerdings wird meist eine

Integration der Lautsprecher in das Bühnenbild notwendig wie in Abb. 8 zu sehen ist.

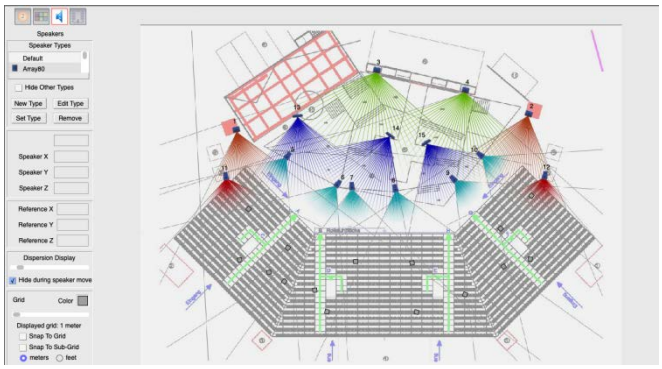


Abb.7: SoundHub – asymmetrische Lautsprecheranordung

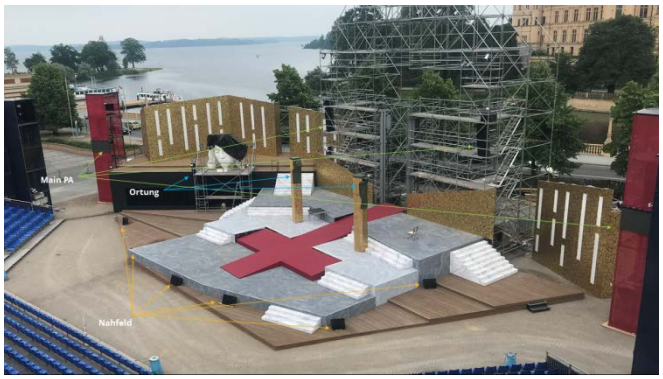


Abb.8: asymmetrische Lautsprecheranordung ins Bühnenbild integriert

4. Objekt- und zonenbasierte Ansätze

Um sich ein klares Bild der unterschiedlichen Ansätze zu erschließen, sollte man sich zunächst die Begriffe des Quell- und Ortungsobjektes sowie der Ortungszone etwas genauer anschauen.

Das Quellobjekt bezeichnet das reelle akustische Objekt. Es kann durch ein oder mehrere Mikrofone abgenommen werden, wobei diese dann als Mix zusammengefasst und mit einem entsprechenden internen Timealignment versehen werden können. So können nicht nur einzelne Instrumente sondern auch ganze Instrumentengruppen zu einem Quellobjekt werden.

Das Ortungsobjekt (Abb. 9) repräsentiert ein zugeordnetes Quellobjekt bezogen auf eine konkrete Position. Entsprechend kann es einen einzigen Kanal oder sogar einen ganzen Mix repräsentieren. Die Ortung bezieht sich aber immer auf eine einzelne Position, die sowohl statisch als auch dynamisch sein kann. Bei Bewegungen des Ortungsobjekts wird nun kontinuierlich die Matrixverteilung angepasst. Der tatsächliche Ortungsbezug eines Quellobjekts entspricht hier der Position des zugeordneten Ortungsobjekts.

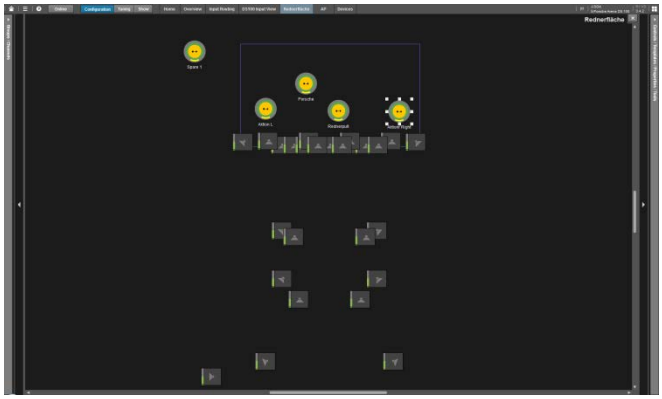


Abb.9.: R1– Ortungsobjekte

Die Ortungszone (Abb. 10) beschreibt einen definierten Bereich für Quellobjekte mit einem statischen Ortungsbezug. Die Zugehörigkeit der Quellobjekte zur Zone ist dynamisch, wobei eine doppelte Zuordnung nicht möglich ist. Über die Größe und Anzahl der Zonen lässt sich die „Auflösung“ der Ortung definieren. Eine Annäherung zum objektbasierten Ansatz wird durch viele kleine Zonen möglich. In zonenbasierten Systemen folgt der Ortungsbezug der aktuell zugeordneten Ortungszone. Bei Bewegungen wird hier zwischen den statischen Matrixverteilungen entsprechend der aktuellen Zone überblendet, wobei sich die Überblendzeiten frei definieren lassen.

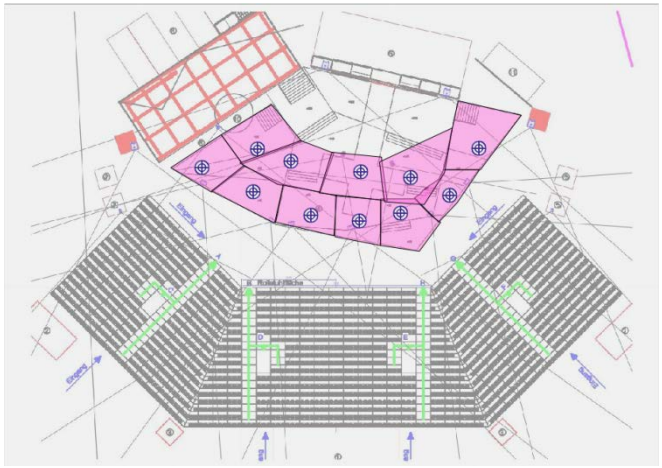


Abb.10: SoundHub - Ortungszonen

Somit lässt sich sagen, dass objektbasierte Systeme Bewegungen je nach Schnelligkeit des Algorithmus direkt abbilden. Bei zonenbasierten Systemen fällt die Abbildung von Bewegungen in Abhängigkeit von Größe, Anzahl der Zonen und der Überblendzeiten unterschiedlich fein aus. Das wiederum kann bei schlechten Trackingergebnissen von Vorteil sein, da Zonenbasierte Systeme mit großen Zonen etwas träger und stabiler auf ungewollte Bewegungen reagieren können.

5. Erfahrungen aus der Praxis

Detaillierte Vorbereitungen in Bezug auf Lautsprecheranordnung, Matrixstruktur und Trackingperformance ist die Grundlage für ein gut

funktionierendes System. Ein automatisches Audiotrackingsystem kann erst dann beeindrucken, wenn von Anfang an jedes Detail beachtet wird. Viele Daten die verarbeitet werden müssen finden ihren Ursprung in einem gemeinsamen Koordinatensystem, daher sind genaue Messinstrumente zum Einmessen und gegebenenfalls zur Korrektur der Planungsgrundlage zwingend notwendig!

Immissionsschutz bei Veranstaltungen

Benjamin Bernschütz[1], Jörn Latz[2]
[1] *Technische Hochschule Mittelhessen, Fachbereich MuK, Wiesenstr. 14, 35390 Gießen*
E-Mail: benjamin.bernschuetz@muk.thm.de
[2] *Kramer Schalltechnik GmbH, Otto-von-Guericke-Str. 8, 53757 Sankt-Augustin*
E-Mail: j.latz@kramer-schalltechnik.de

Zusammenfassung

Das Thema Immissionsschutz bei Veranstaltungen gewinnt zunehmend an Relevanz. Eine immer dichtere Besiedelung bei gleichzeitig steigendem Angebot an Veranstaltungen und Konzerten führt regelmäßig zu Lärmkonflikten. Eine Veranstaltung muss auf der einen Seite zur praxisgerechten Realisierung immer mit einem gewissen Mindestversorgungspegel beschallt werden. Auf der anderen Seite sind die umliegenden Anwohner vor schädlichen Umwelteinwirkungen zu bewahren und die entsprechenden Immissionsschutzanforderungen zu berücksichtigen. Im Umgang mit den Regelwerken und mit den organisatorischen Zusammenhängen sind Technikdienstleister und Toningenieure oftmals unsicher und nicht hinreichend vertraut. Schalltechnischen Gutachtern und Behörden sind andererseits häufig die Bedürfnisse der Produktion und die Möglichkeiten moderner Beschallungstechnik nicht vollumfänglich bekannt. Hier wird in der Praxis regelmäßig großes Potential verschenkt. Eine enge Kooperation zwischen Behörden, schalltechnischen Gutachterbüros und Technikdienstleistern kann zu einer erheblichen Verbesserung der Situation aller Beteiligten führen. Der vorliegende Artikel liefert eine Übersicht über die Grundlagen des Immissionsrechts, die Messtechnik und die Beurteilung der schalltechnischen Situation. Weiterhin werden Minderungsmaßnahmen zur Reduktion der Immissionsbelastung und der Einsatz vernetzter Messtechnik zur Echtzeitüberwachung der Immissionssituation bei Events diskutiert. Der Artikel richtet sich vorwiegend an Technikdienstleister und Toningenieure bzw. auch an Veranstalter und Agenturen, um sie für die Thematik zu sensibilisieren, ihnen einen Überblick zu geben und den Einstieg in diese etwas unübersichtliche Materie zu erleichtern.

1. Einführung

Bedingt durch eine immer dichtere Besiedelung in den Ballungsräumen und eine stetig steigende Anzahl an Events kommt es immer häufiger zu Lärmkonflikten. Bei Open Air Veranstaltungen ist die schalltechnische Situation besonders kritisch, wobei auch zunehmend Indoor Events problematischer werden und sich beispielsweise die Lärmbeschwerden bezüglich Schallimmissionen im tieffrequenten Bereich deutlich häufen.

2. Abgrenzung von Personenkreisen

Die erste Unsicherheit besteht oft darin, die verschiedenen Personenkreise abzugrenzen und die dafür geltenden Richtlinien zuzuordnen. Grundsätzlich können bei einer Veranstaltung drei verschiedene Personenkreise abgegrenzt werden:

2.1. Beschäftigte auf der Veranstaltung

Von der Thekenkraft bis zum Toningenieur fallen die Beschäftigten auf einer Veranstaltung unter die konventionellen Regularien des Arbeitsschutzes. Das anzuwendende Regelwerk ist die LärmVibrationsArbSchV [1]. Hierin sind verschiedene Auslösewerte und Schutzmaßnahmen festgeschrieben. Die Beschäftigten dürfen nur einer gewissen Lärmdosis pro Tag ausgesetzt werden und bestimmte Spitzenpegel dürfen zu keinem Zeitpunkt überschritten werden. In diesem Kontext ist die Arbeit von Toningenieuren im Live-Bereich, die über lange Zeiträume hohen Schalldruckpegeln ausgesetzt sind, ohne dabei durchgehend Gehörschutz tragen zu können, als äußerst problematisch anzusehen. Jedoch ist dies ein eigenständiges Thema und losgelöst vom Immissionsschutz zu betrachten.

2.2. Besucher des Events

Eine weitere direkt beteiligte Personengruppe sind die Besucher der Veranstaltung. Der Schutz der Besucher fällt in den Bereich der Verkehrssicherungspflicht des Veranstalters. Hier liefert die DIN 15905-5 [2] entsprechende Vorgehensweisen und Richtwerte, die einer Gehörschädigung des Publikums vorbeugen sollen. Die DIN 15905-5 [2] würde im Streitfall vom Gericht als Stand der Technik herangezogen und hat somit als DIN-Norm impliziten Gesetzescharakter. Hier treten sehr oft Missverständnisse bei Toningenieuren und Technikdienstleistern auf: Die DIN 15905-5 [2] bezieht sich ausschließlich auf den Schutz des anwesenden Publikums und hat mit Immissionsschutz – bei dem es immer um die umliegende Nachbarschaft geht – überhaupt nichts zu tun. Auch die Messungen im Rahmen der DIN 15905-5 [2] haben nichts mit Immissionsschutz zu tun.

2.3. Umliegende Anwohner

Die dritte Personengruppe sind die umliegenden Anwohner bzw. die Nachbarschaft. Dies betrifft Personen, die indirekt beteiligt sind und im Einwirkbereich der Veranstaltung leben oder arbeiten. Genau diese Personengruppe steht im Wirkungsbereich des Immissionsschutzes. Hier sind die Regularien erheblich komplexer und die Messverfahren anspruchsvoller als bei den ersten beiden Personengruppen.

ISBN
978-3-9812830-9-9
Der Autor hält die Urheberrechte für den gesamten Inhalt seines Manuskripts. Er überlässt dem VDT das Recht auf Veröffentlichung in gedruckter Form, Online oder als PDF-Download von https://tonmeistertagung.com

3. Interessenkonflikt und Regulierung

Der Veranstalter möchte auf der einen Seite einen möglichst hohen Innenpegel auf seinem Gelände oder in seiner Venue erzielen. Zumindest ist er daran interessiert, einen Mindestversorgungspegel für das Publikum sicherzustellen, der eine praxisgerechte Durchführung der Veranstaltung ermöglicht.

Auf der anderen Seite stellt eine Veranstaltung für die umliegenden Anwohner eine Lärmquelle dar. Es muss gewährleistet sein, dass die Anwohner vor schädlichen Umwelteinwirkungen geschützt sind; so ist es im Bundesimmissionsschutzgesetz BImSchG [3] festgesetzt.

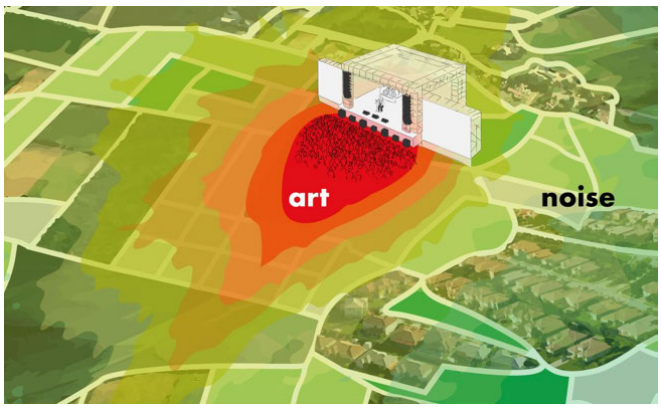


Abb. 1: Eine treffende Illustration der Situation. Auf dem Veranstaltungsgelände wird der emittierte Schall als Kunst, beim Anwohner hingegen als Lärm wahrgenommen. [Quelle: d&b audiotechnik]

Hier stellt die Behörde mit Rückgriff auf verschiedene Gesetze, Verwaltungsvorschriften und Normen die regulierende Instanz dar. Der Veranstalter hat sein Event bei der Behörde zu beantragen und erhält im Idealfall eine Genehmigung zur Durchführung der Veranstaltung. In der Genehmigung können neben einer Vielzahl möglicher Auflagen auch den Immissionsschutz betreffende Auflagen erteilt werden, da die Behörde den Anwohner vor schädlichen Umwelteinwirkungen zu bewahren hat.

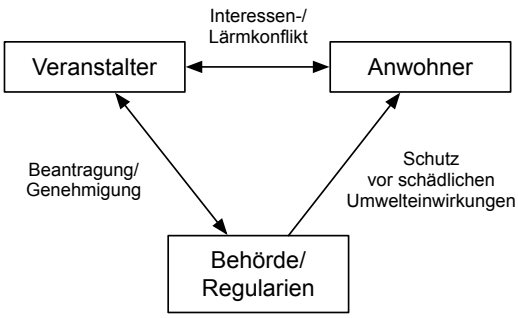


Abb. 2: Interessen- oder Lärmkonflikt zwischen Veranstalter und Anwohner. Behörde und Regularien als regulierende Instanz.

Veranstalter sehen die Behörden oft kritisch und verstehen Sie als Verhinderer und Gegner. Dies ist allerdings eine einseitige und auch falsche Perspektive. In der Regel haben die Behörden keine grundlegenden Einwände gegen die Durchführung von Veranstaltungen und möchten Veranstaltungen im Sinne der Kulturförderung oft sogar explizit ermöglichen und unterstützen. Hat der Veranstalter eine Genehmigung von der Behörde erhalten, in der bestimmte Spielzeiten und Immissionsrichtwerte festgesetzt sind, so steht die Einhaltung der Auflagen zwar einerseits auf der Liste der Pflichten des Veranstalters – andererseits erwachsen dem Veranstalter aus der Genehmigung und den Auflagen auch verschiedene Rechte. Er darf nämlich die angegebenen Spielzeiten und Immissionsrichtwerte vollständig ausschöpfen und die Anwohner haben dieses hinzunehmen. Ein bei Anwohnern weit verbreiteter Irrglaube besteht darin, dass eine Veranstaltung unzulässig sein muss, sobald sie wahrnehmbar ist. Es besteht jedoch seitens des Anwohners kein Anspruch darauf „nichts“ zu hören. Die Behörde wägt im Prozess der Genehmigung einer Veranstaltung jeweils genau ab, welche Immissionsbelastungen für den Anwohner zumutbar und hinzunehmen sind.

Bei umbauten Veranstaltungsstätten wie Hallen, Stadien, Sälen, Clubs oder Arenen kann eine Pauschalgenehmigung im Rahmen der Baugenehmigung erteilt werden. Aber auch hier sind ggf. bei einzelnen besonders geräusch- oder zeitintensiven Veranstaltungen Ausnahmegenehmigungen erforderlich bzw. möglich – je nach Sichtweise.

4. Schalltechnische Sachverständige

4.1. Position und Aufgaben

Der Prozess von der Beantragung bis zur Durchführung eines Events wird oftmals von einem schalltechnischen Sachverständigen begleitet. Der Sachverständige hat eine beratende und aufklärende Funktion gegenüber dem Veranstalter, übernimmt meist die fachliche Kommunikation mit den entsprechenden Behörden und führt notwendige Abstimmungen durch. Schalltechnische Sachverständige kennen die Regelwerke, haben technische Sach- und Fachkenntnisse und sprechen die entsprechende Fachsprache mit allen Beteiligten.

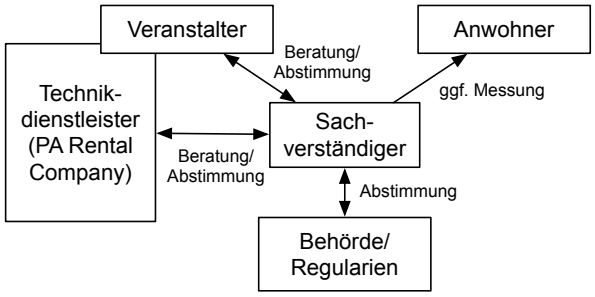


Abb. 3: Einordnung der Position des schalltechnischen Sachverständigen und des Technikdienstleisters in das Dreieck aus Abbildung 2.

Die Anforderung der Begleitung der Veranstaltung durch einen Sachverständigen erfolgt meist implizit durch die Behörde, indem ggf. Immissionsprognosen und/oder Messungen gefordert werden. Der Sachverständige oder eine benannte Messstelle, vgl. Abschnitt 4.2, wird vom Veranstalter beauftragt und bezahlt. Dennoch ist der Sachverständige stets ein unabhängiger Gutachter, der die Situation neutral beschreibt und beurteilt. Er berät den Veranstalter konstruktiv, welche Schritte zu gehen sind und wie die schalltechnische Situation z. B. durch geeignete Minderungsmaßnahmen optimiert werden kann. Er kann ggf. durch entsprechende Argumentation und durch die

Darlegung von notwendigen Fakten die Abwägungsprozesse der Behörden unterstützen, wenn es beispielsweise um die Beantragung einer Ausnahmegenehmigung geht. Der Gutachter wird aber ebenso neutrale Fakten darbieten, die ggf. darlegen könnten, dass eine Veranstaltung nicht durchführbar bzw. möglicherweise nicht genehmigungsfähig ist. Aus diesem Grunde sollte ein Veranstalter möglichst frühzeitig in Kontakt mit dem Gutachter treten und die Situation darlegen. Oftmals kann in Vorgesprächen und groben Abschätzungen bereits im Vorfeld geklärt werden, ob eine Veranstaltung überhaupt in die Nähe der Durchführbarkeit rückt oder ob die Situation verhältnismäßig aussichtslos ist. Die Entscheidung über eine Genehmigungsfähigkeit trifft aber stets die Ortsbehörde. Der Sachverständige liefert immer nur den Abwägungsprozess unterstützende Fakten.

Ebenso wird ein Sachverständiger bei laufender Veranstaltung in der Regel nicht selbstständig aktiv in das PA-System oder Mischpult eingreifen, um den Pegel abzusenken oder ganz abzuschalten. Bei laufender Veranstaltung berät der Sachverständige den Veranstalter und die Techniker und dokumentiert Messergebnisse und besondere Vorkommnisse. Nach der Veranstaltung erstellt der Sachverständige einen Bericht in gutachterlicher Form, den er dem Veranstalter und üblicherweise auch parallel der Behörde zukommen lässt.

Modernere Sachverständigenbüros, die viele Projekte im Eventbereich betreuen, gehen in der Beratungsfunktion zu Minderungsmaßnahmen mittlerweile deutlich über die klassische Schallschutzwand hinaus. Sie treten im Vorfeld der Veranstaltung in Austausch mit dem Technikdienstleister und erarbeiten gemeinsam mit dem Technikdienstleister auf die Veranstaltung und Immissionssituation angepasste Beschallungskonzepte, die nicht nur den Innenbereich der Veranstaltung berücksichtigen, sondern gleichzeitig auch die Immissionssituation im Umfeld optimieren, vgl. Abschnitt 12. Solche Büros beschäftigen mitunter auch Personal, das originär aus der Veranstaltungs- bzw. Beschallungstechnik stammt und sich mit PA-Systemdesign und den allgemeinen Herausforderungen der Technikdienstleister sehr gut auskennt, sodass hier eine ganz konkrete und lösungsorientierte Beratung und Kooperation auf Augenhöhe stattfinden kann.

4.2. Benannte Messstellen und Sachverständige

Zunehmend werden für die Messungen und für Prognosen Sachverständige einer benannten Messstelle nach § 29 b BImSchG [3] gefordert. Hier ist durch strenge Akkreditierungsprozesse der Messstellen sichergestellt, dass die Sachverständigen unabhängig sind und über entsprechende Sach- und Fachkenntnis verfügen. Notwendige Sach- und Fachkenntnisse umfassen dabei sowohl die rechtlichen und regulatorischen Rahmenbedingungen sowie auch den Umgang mit der erforderlichen Mess- und Simulationstechnik. Sachverständige einer benannten Messstelle – zumindest aber der oder die Messstellenleiter – haben immer ein einschlägiges Ingenieurstudium absolviert. Weiterhin existieren z. B. auch öffentlich bestellte und vereidigte Sachverständige, die jedoch häufig ohnehin einer benannten Messstelle angehörig sind.

Sofern seitens der Behörde nicht explizit eine benannte Messstelle oder ein anderweitig anerkannter Sachverständiger angefordert ist, könnte grundsätzlich jeder „Sachverständige“ (kein geschützter Begriff) die Messungen oder Prognosen durchführen. Sofern besondere Anforderungen an den Sachverständigen bestehen, teilt die Behörde das mit. Es gibt durchaus auch kompetente Sachverständige außerhalb der benannten Messstellen. Veranstaltern sei an dieser Stelle aber dringend davon abgeraten allzu preiswerte oder wesentlich zu optimistische Experten für diese Aufgaben zu berufen. Die Behörden lehnen einzelne Gutachten bzw. Gutachter durchaus ab, sofern die Prognosen oder Messungen nicht lege artis erscheinen. Dies kann zu erheblichen organisatorischen Problemen und am Ende zu deutlich höheren Gesamtkosten führen. Bei benannten Messstellen gibt es diese Probleme in aller Regel nicht.

Nicht selten lassen sich auch in schwierigen Situationen am Ende Möglichkeiten finden, die Veranstaltung auf seriösem Wege und mit immissionsrechtlichem Fundament durchführbar zu machen. Seriöse Sachverständige mit entsprechender Erfahrung können hier wichtige Beiträge für eine regelkonforme Lösung liefern. Bei benannten Messstellen kann man beispielsweise relativ sicher davon ausgehen, dass die entsprechende Erfahrung vorhanden ist. Selbstverständlich können aber auch Gutachter außerhalb der benannten Messstellen eine vergleichbar gute Beratung liefern.

Insbesondere in schalltechnisch kritischen Situationen kann es mitunter sinnvoll sein, im Großumfeld regional ansässige Sachverständige oder Messstellen zu beauftragen, da diese regelmäßig mit den jeweiligen Ortsbehörden zu tun haben und meist schon ein entsprechendes Vertrauensverhältnis besteht. Zudem kennen regionale Sachverständige und Messstellen die meisten Locations und die ggf. damit zusammenhängenden Problemlagen bereits. Sie führen bei neuen Locations auch eher eine Ortsbesichtigungen durch, sofern die Location nicht unbedingt hunderte von Kilometern entfernt liegt.

Trotz aller Bemühungen gibt es aber auch Situationen und Konstellationen, die nicht funktionieren können. Auch das kommt durchaus vor und dann kann die Veranstaltung in dieser Form oder in dieser Location nicht durchgeführt werden.

5. Position der Technikdienstleister

Bei den vom Veranstalter beauftragten Technikdienstleistern steht der Immissionsschutz nur sehr selten im Fokus der Bemühungen. Der Immissionsschutz wird oft vollständig ignoriert und nicht als Teil des eigenen Aufgabengebiets verstanden. Beschaller sind häufig ausschließlich darauf fokussiert, den Innenbereich der Veranstaltung gut zu versorgen, indem sie einen hohen Maximalpegel, eine gute Coverage und einen gleichmäßigen Frequenzgang auf der Zuhörerfläche anstreben. Das Umfeld wird dabei oftmals überhaupt nicht betrachtet und es werden nicht selten versehentlich grobe und häufig auch vermeidbare Fehler gemacht, die in schalltechnisch kritischen Situationen die praxisgerechte Durchführung der Veranstaltung gefährden oder zu Beschwerdelagen führen können.

Die Beschallungsanlage ist in der Regel der maßgebliche Emittent bei Veranstaltungen. Moderne Beschallungstechnik bietet – entsprechende Kompetenzen der Betreiber vorausgesetzt – vielfältige Möglichkeiten, um Richtwirkung zu erzielen, den Nutzschall auf den Publikumsbereich zu konzentrieren und Störschall im Umfeld zu reduzieren. Dabei profitieren alle Beteiligten. Der Beschaller kann einen höheren Innenpegel erzielen, die Anwohner haben ihre Ruhe, die Behörden können die Genehmigung der Veranstaltung besser rechtfertigen und es kommt ggf. im Idealfall gar nicht erst zu einer Beschwerdelage. Der Veranstalter hat somit insgesamt weniger Probleme und kann alle Seiten zufrieden stellen.

Aus o. g. Gründen sollte ein professioneller und kundenorientierter Technikdienstleister durchaus die Berücksichtigung der Immissionssituation als Teil seines originären Aufgabengebiets verstehen. Dafür sind genauere Kenntnisse der Regularien des Immissionsschutzes erforderlich. In der Regel wird die Nützlichkeit oder die Notwendigkeit dieser Kenntnisse erheblich unterschätzt. Spezifische und für den Technikdienstleister zielgruppengerecht aufbereitete Informationen zum Thema Immissionsschutz zu bekommen ist mitunter auch etwas schwierig.

Selbstverständlich kann hier ein schalltechnischer Gutachter entsprechend vermitteln und in Kooperation mit dem Beschaller Konzepte entwickeln. Ein gewisses Basiswissen und Grundverständnis für die Zusammenhänge sollte ein professioneller Technikdienstleiter jedoch in jedem Falle mitbringen.

6. Behörden und Regelwerke

6.1. Zuständige Behörden

Zuständig für die Regulierung von Immissionsschutzangelegenheiten und die Genehmigung von Veranstaltungen im Bezug auf die Immissionssituation sind jeweils die Ortsbehörden. In der Regel ist das örtliche Umweltamt der Stadt (oder des Kreises), in der die Veranstaltung stattfindet, für die Entscheidungsfindung und Genehmigung zuständig. Die Ordnungsämter sind dabei meist das ausführende Organ. Sie übernehmen die Kommunikationsfunktion bzw. Korrespondenzen und sind ggf. auch während der Veranstaltung vor Ort, um die Einhaltung von Auflagen zu kontrollieren. In der Regel ist das Ordnungsamt nicht dauerhaft während der Veranstaltung vor Ort. Es finden meist eher stichprobenartige Kontrollen statt oder es werden Kontrollen durch eine akute Beschwerdelage ausgelöst. Bei Beschwerdelagen kann auch lokal die Polizei zur Klärung der Situation eingesetzt sein. Dies variiert je nach Gemeinde. In der Regel entscheiden die Umweltämter im Vorfeld der Veranstaltung in der Genehmigungsphase über die Genehmigungsfähigkeit und mögliche Auflagen. Auch nach der Veranstaltung wird dort über mögliche Konsequenzen bei Nichteinhaltung von Auflagen entschieden. Während laufender Veranstaltungen, die häufig abends stattfinden, haben die Ordnungsbehörden jedoch in der Regel die Möglichkeit und die Befugnis, die Veranstaltung bei Nichteinhaltung von in der Genehmigung festgesetzten Auflagen abzubrechen oder anderweitige akute Konsequenzen wie z. B. ein Herunterpegeln der Beschallungsanlage zu veranlassen. Die genauen Zuständigkeiten und Befugnisse variieren je nach Gemeinde.

6.2. Regelwerke

Es existieren verschiedene Gesetze, Verwaltungsvorschriften und DIN-Normen, die im Zusammenhang mit dem Thema Immissionsschutz relevant sind. Im Zuge der Globalisierung soll der Immissionsschutz zukünftig europaweit betrachtet werden. Momentan ist die Gesetzgebung hier jedoch noch nicht einheitlich und in Deutschland gilt somit nach wie vor die deutsche Gesetzgebung.

An oberster Stelle steht somit immer noch das deutsche Bundes-Immissionsschutzgesetz BImSchG [3]. Darin sind bewusst keine spezifischen Angaben wie beispielsweise Immissionsrichtwerte zu finden. Im BImSchG [3] ist lediglich der allgemeine Schutz vor schädlichen Umwelteinwirkungen bzw. der Anspruch auf gesunde Wohnverhältnisse festlegt. Zur Durchführung des Gesetzes wird auf die allgemeinen Verwaltungsvorschriften verwiesen.

Die wichtigste Verwaltungsvorschrift in diesem Kontext ist die Technische Anleitung zum Schutz gegen Lärm oder kurz TA Lärm [4]. Die TA Lärm [4] hat bundesweite Rechtskraft und legt sehr genaue Details, wie z. B. Immissionsrichtwerte, Messverfahren, Ausnahmen etc., fest. Weiterhin wird aus der TA Lärm [4] heraus direkt auf die Anwendung einzelner DIN-Normen wie beispielsweise DIN 45680 [5] zur Messung und Beurteilung tieffrequenter Immissionen verwiesen, die dadurch automatisch Gesetzescharakter erhalten.

Ferner existieren zusätzlich auf Länderebene die Landes-Immissionsschutzgesetze wie z. B. das Landes-Immissionsschutzgesetz NRW [7]. Darin werden beispielsweise der Schutz der Nachtruhe oder die Benutzung von Tongeräten reguliert. Diese individuellen Gesetze der Länder sind zwar landesspezifisch definiert, unterscheiden sich aber nicht allzu wesentlich voneinander.

Veranstaltungen fallen grundsätzlich in die Kategorie Freizeitlärm. Lediglich Sportveranstaltung können ggf. davon abweichend in die Kategorie Sportlärm fallen und nach dem entsprechenden Regelwerk [8] beurteilt werden. Als Verwaltungsvorschrift wäre bei Veranstaltungen grundsätzlich die TA Lärm [4] heranzuziehen. Die Verwaltungsvorschrift TA Lärm [4] entstammt originär dem Bereich Gewerbe- und Industrielärm und schließt die Beurteilung von Freizeitgeräuschen explizit aus. Somit wurden in den Bundesländern spezifische Regelwerke zum Freizeitlärm erlassen, die grundlegend auf der TA Lärm [4] basieren, jedoch an einigen Stellen spezifische Anpassungen enthalten.

Die Arbeitsgemeinschaft Immissionsschutz des Bundes und der Länder (LAI) hat eine allgemeine Freizeitlärmrichtlinie unter dem Namen Freizeitlärmrichtlinie LAI [9] erarbeitet, die von einigen Bundesländern, wie beispielsweise Hessen, direkt und unverändert übernommen wurde. Andere Bundesländer haben ausgehend von dieser Basis eigene Modifikationen vorgenommen. Ein Beispiel dafür ist der Runderlass Freizeitlärm des Bundeslandes NRW [10].

Es sind somit in jedem Bundesland die jeweils länderspezifischen Regelwerke zu beachten, was bei Tourneen

eine Herausforderung darstellen kann. Bei internationalen Tourneen wird es entsprechend nochmals deutlich komplexer. Hier ist es ratsam jeweils lokale Sachverständige zur Beratung hinzuzuziehen.

Die entsprechenden Regelwerke sind grundsätzlich sehr kurz und kompakt abgefasst. Aber der Teufel steckt im Detail und in den möglichen Auslegungen. Es existieren umfangreiche Bücher mit Auslegungen und Gerichtsurteilen zu den Regelwerken. Die Materie ist juristisch deutlich komplexer als sie für den Techniker zunächst zu sein scheint.

6.3. Trennung der Lärmarten

Im Immissionsschutz ist grundsätzlich zwischen verschiedenen Lärmarten wie z. B. Verkehrslärm, Schienenlärm, Industrielärm, Freizeitlärm, Sportlärm u .s .w. zu unterscheiden.

Etwas kurios erscheint dabei die Tatsache, dass alle Lärmarten formal streng getrennt behandelt werden. Läge also ein Immissionsort direkt neben einer großen Autobahn und der Verkehr würde die Immissionen der Veranstaltungen bei weitem maskieren, so spielt dies zur Beurteilung keine Rolle, da es sich hier um unterschiedliche Lärmarten handelt. Die Immissionsrichtwerte aus dem Freizeitlärmanteil sind formal völlig unabhängig von der Autobahn einzuhalten. Bei einer Messung führt dies in der Praxis jedoch häufig zu problematischen Situationen, da der Anteil der Veranstaltung nicht eindeutig von der Autobahn getrennt oder ggf. gar nicht erfasst werden kann. Die eindeutige Trennung der Lärmarten funktioniert leider nur bei Juristen. Techniker haben damit in der Praxis etwas größere Schwierigkeiten. Die Trennung der Lärmarten ist zwar vom Gesetzgeber vorgegeben und gewollt, führt aber folglich in der Praxis oft zu Problemen. Hier müssen dann nach Rücksprache mit der Behörde geeignete Ersatzimmissionsorte zur Durchführung einer aussagekräftigen Messung gefunden werden.

7. Immissionsprognose

Der erste Schritt im Genehmigungsprozess nach der Beantragung einer Veranstaltung besteht häufig in der Durchführung einer Immissionsprognose. Ist eine bestimmte Location mit äquivalenten Eventformaten bei den Behörden bekannt oder handelt es sich um eine wiederholte Durchführung einer Veranstaltung, so kennen die Behörden die Situation. Hier ist meistens ein Brief des Sachverständigen ausreichend, der eine vergleichbare Situation zu bekannten Konstellationen bestätigt. Wird eine Location jedoch ganz neu für die Nutzung durch Events erschlossen oder handelt es sich um neue Veranstaltungsformate in einer bekannten Location (z. B. ein Rock-Konzert oder ein EDM Event statt der üblichen Kammermusikabende), so wird die Behörde in der Regel eine Immissionsprognose anfordern. Die Prognose wird vom schalltechnischen Sachverständigen erstellt.

Imissionsprognosen werden rechnergestützt nach der DIN ISO 9613-2 [11] durchgeführt. Dabei werden Schallausbreitungsberechnungen unter Berücksichtigung von Abstand, Abschirmung und Reflexion (z. B. durch Gebäude), Topographie, Bodeneffekten und Dissipation (Luftdämpfung) durchgeführt. Es existieren grundsätzlich präzisere Verfahren zur Schallausbreitungsprognose, jedoch ist das Verfahren nach DIN ISO 9613-2 [11] in den einschlägigen Regelwerken vorgeschrieben und somit im Kontext behördlicher Genehmigungsverfahren nicht verhandelbar.

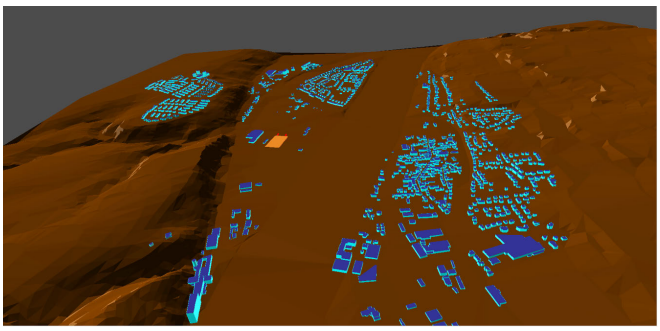


Abb. 4: Topographisches Geländemodell mit Gebäuden in der Prognosesoftware MAPANDGIS [12]. Die Publikumsfläche der Veranstaltung ist orange gekennzeichnet.

Am Anfang der Prognose steht die Modellierung der Topographie des Geländes und die Digitalisierung von Hindernissen (z. B. Gebäude, größere Wände, ggf. Vegetation). Dabei entsteht ein 3D-Geländemodell, in dem auch die Bodenbeschaffenheiten über entsprechende Absorptionsfaktoren berücksichtigt werden. Die Digitalisierung des Modells ist in der Regel mit erheblichem Zeitaufwand verbunden und somit ein zu berücksichtigender Kostenpunkt, den der Veranstalter zu tragen hat, bevor überhaupt klar ist, dass die Veranstaltung stattfinden kann.

Geländetopographien und teils auch Gebäudedaten werden von einigen Bundesländern als Geodatensätze zur Verfügung gestellt. Derartige Datensätze vereinfachen die Modellierung erheblich, wobei die Daten immer gut kontrolliert werden sollten. In jedem Falle sollte vor der Modellierung stets eine umfangreiche Ortsbesichtigung im Bereich der Veranstaltung sowie auch im Umfeld bei den Anwohnern stattfinden. Geodatensätze, Luftbilder und Karten sind oftmals nicht ganz aktuell. So können innerhalb von wenigen Wochen z. B. ganze Hallenkomplexe aufgestellt oder abgerissen werden, die maßgeblichen Einfluss auf die Schallausbreitung und Abschirmung haben. Hier kommt es regelmäßig zu größeren Überraschungen bei der Ortsbesichtigung.

In das Geländemodell wird die Beschallungsanlage der Veranstaltung als Emittent aufgenommen. Wesentlich für eine zutreffende Prognose ist, dass die Emissionsansätze möglichst genau auf die geplante Veranstaltung zutreffen. Dies bezieht sich auf die Positionierung, Schallleistung und Richtwirkung der Beschallungsanlage sowie auch auf das Programmaterial. Das Programmaterial wird hierbei in Form von typischen Ausgangsspektren berücksichtigt.

Meist werden Beschallungsanlagen als einfache Punkt- oder Linienquellen gemäß DIN ISO 9613-2 [11] mit grober sektorweiser Richtwirkung in Oktavfrequenzbändern modelliert. Schallquellen werden in typischer Prognosesoftware lediglich energetisch und nicht komplex summiert, da im Lärmschutzbereich üblicherweise keine kohärenten Quellen vorkommen. Diese Art der simplifizierten Modellierung wird den physikalischen Eigenschaften

größerer Beschallungsanlagen allerdings nur wenig gerecht. Quellen in einem verbundenen Beschallungssystem strahlen kohärent ab und können vielfach durch gezieltes Processing in ihrer Richtwirkung beeinflusst werden, was bei den o. g. simplifizierten Ansätzen ggf. zu einer deutlich erhöhten Prognoseunsicherheit führen kann. Eine gute Übersicht zu diesem Thema liefern Belcher und Shabalina [13].

Beispielsweise gehen der Audiosystemhersteller d&b audiotechnik und SoundPlan, ein Anbieter für Schallprognosesoftware, einen neuen gemeinsamen Weg mit ihren Softwarelösungen NoizeCalc [14] bzw. SoundPLAN noise [15]. Die Beschallungsanlage inklusive aller Komponenten und Eigenschaften wird hier direkt aus dem Tool ArrayCalc [16] für das PA-Systemdesign exportiert und als Quelle mit komplexer Addition (Amplitude und Phase) in die Prognose integriert. Mit diesem Vorgehen können die tatsächlichen physikalischen Eigenschaften einer Beschallungsanlage erheblich besser abgebildet werden als mit den üblichen simplifizierten Punkt- und Linienquellen der klassischen Prognose. Die Unsicherheiten und sogar der Aufwand für die Modellierung sind deutlich geringer, da das Systemdesign in der Regel ohnehin durchgeführt wird und als Datensatz vorliegt. Einzig mit einer Lösung dieser Gestalt können relativ zuverlässige Relationen zwischen Innenpegel (z. B. am F.o.H.) und dem zu erwartenden Pegel an den Immissionsorten hergestellt werden, wo klassische Immissionsprognosetools bisher vollständig versagen.

Mit ausreichend Erfahrung und guten Emissionsansätzen lassen sich aber auch mit regulärer Immissionsprognosesoftware zutreffende Prognosen erstellen. Grundlegende Emissionsansätze bezüglich der Schallleistung (je nach Größe der Fläche/Besucherzahl), Richtwirkung und Ausgangsspekten (je nach Programmaterial) sind in der VDI 3770 [17] zu finden. In Kombination mit Erfahrungswerten vergleichbarer Veranstaltungen und geschickter Modellierung der Beschallungsanlage gelingen in der Regel zutreffende Prognosen für das Umfeld.

Neben der Beschallungsanlage müssen ggf. je nach Anforderung der Behörde zusätzlich noch der An- und Abreiseverkehr der Besucher, Parkplatzverkehr, Besucherströme oder Auf- und Abbauarbeiten in die Prognose einbezogen werden. Hier wird die Modellierung der Emissionssituation mitunter relativ aufwendig.

Aus der Prognose leiten sich häufig Empfehlungen für den maßgeblichen Immissionsort (vgl. Abschnitt 8) und ggf. weitere Immissionsorte ab, da aus der Berechnung die Zonen mit höchster Immissionsbelastung hervorgehen. Für die betreffenden Immissionsorte kann anhand der Berechnungen überprüft werden, ob die Immissionsrichtwerte bei der geplanten Veranstaltung eingehalten werden oder ob die Veranstaltung bestenfalls mit einer Ausnahmegenehmigung durchführbar wäre, vgl. Abschnitt 10. In letzterem Falle wird von der Behörde auf Basis der Prognose abgewogen, welche Immissionsrichtwerte für die Veranstaltung festgesetzt werden müssen oder ob die Veranstaltung überhaupt genehmigungsfähig ist.

Anhand der Immissionsprognose können seitens der Behörden auch Vorgaben für den Innenpegel der Veranstaltung (z. B. am F.o.H.) festgelegt werden. Über die Prognose werden dafür die Immissionspegel in Relation zum Innenpegel gesetzt. Hier ist jedoch besonders große Vorsicht geboten, da dieses Verhältnis mit großen Unsicherheiten behaftet ist, sofern die Beschallungsanlage nicht tatsächlich als komplexwertige kohärente Quelle mit ihren vollständigen physikalischen Eigenschaften modelliert wird. Im näheren Umfeld der Quellen (auf dem Veranstaltungsgelände) sind die Ergebnisse der Simulation ansonsten nicht realistisch und es kommt zu unrealistischen Pegelwerten.

Alle Immissionsorte werden bei der Prognose üblicherweise bei Mitwindbedingungen betrachtet, um so immer mit den ungünstigsten Bedingungen zu rechnen. Generell sind Prognosen immer als pessimale Abschätzung (worst-case) auszulegen, um in jedem Falle auf der sicheren Seite zu liegen.

Neben Immissionsorten und den anzustrebenden Immissionsrichtwerten können im Rahmen der Prognose auch geeignete klassische Schallminderungsmaßnahmen wie z. B. mobile Schallschutzwände geplant werden.

Um einen Überblick über die Gesamtsituation zu erhalten und die Zonen größter Immissionsbelastung zu lokalisieren kann optional eine Lärmkarte zur Visualisierung der Schallpegelverteilung im Umfeld bzw. Großumfeld der Veranstaltung erstellt werden, vgl. Abbildung 5. Bei einer Lärmkarte werden meist äquivalente A-bewertete Dauerschallpegel in Form von Farben visualisiert. Zu berücksichtigen sind in diesem Fall immer noch mögliche Zuschläge, die bis zu 12 dB ausmachen können, vgl. Abschnitt 9.3. In der Regel müssen dabei gleich mehrere Lärmkarten für verschiedene Geschosshöhen und/oder Bespielungsvarianten erstellt werden.

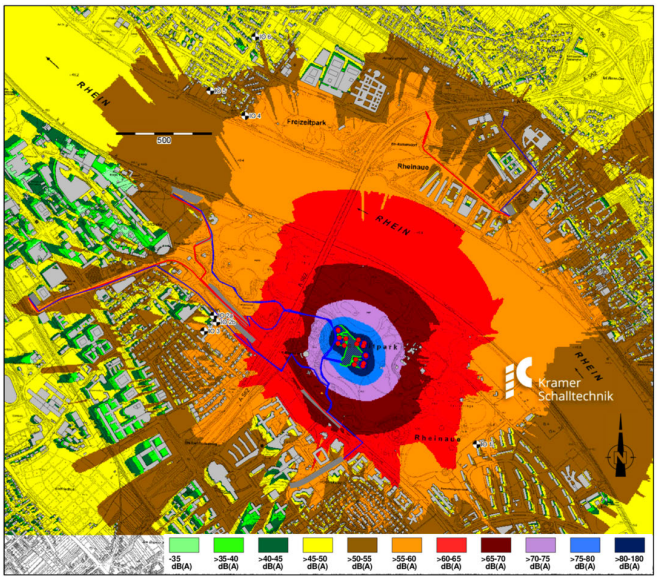


Abb. 5: Beipiel für eine Lärmkarte zur übersichtlichen Darstellung der Immissionsbelastung im Umfeld einer Open Air Veranstaltung. Erstellt wurde die Karte in der Software MAPANDGIS [12]. In der Prognose wurden neben den Beschallungsquellen (rote Punkte) auch Zufahrtswege und Publikumsströme berücksichtigt (Linien im Umfeld).

Eine Prognose sollte immer von Beginn an in enger Abstimmung mit der Behörde und den Technikdienstleistern durchgeführt werden. So kann zielführendes Arbeiten sichergestellt und es können unnötige Iterationen vermieden

werden. Die Prognose wird üblicherweise in gutachterlicher Form mit textlicher Erläuterung der Situation der Behörde zugeleitet und sie wird in der Regel über einen sogenannten „Grünstempel“ zum Bestandteil der amtlichen Genehmigungsunterlagen der Veranstaltung erklärt.

8. Maßgeblicher Immissionsort

In Genehmigungen mit Auflagen zum Immissionsschutz wird mindestens ein Punkt im Umfeld der Veranstaltung ganz explizit mit Straße, Hausnummer und meist auch Geschosshöhe benannt, an dem bestimmte Richtwerte einzuhalten sind. Dieser Punkt wird als maßgeblicher Immissionsort bezeichnet. Er liegt in der Regel dort, wo die ungünstigste Kombination aus Abstand, Abschirmung, Geschosshöhe und Schutzanspruch aufeinandertreffen. Diesen Punkt kann man z. B. über die Immissionsprognose mithilfe von Lärmkarten lokalisierten, vgl. Abbildung 5.

Weiterhin geht man davon aus, dass die Immissionsrichtwerte auch an allen anderen Punkten eingehalten werden, sofern sie am maßgeblichen Immissionsort eingehalten werden. Dieser weist ja bereits die ungünstigste Kombination aller Einflussfaktoren auf.

Der maßgebliche Immissionsort muss somit nicht unbedingt die räumlich nächstgelegene schutzbedürftige Nutzung sein. Liegt beispielsweise in 200 m Entfernung zum Festivalgelände eine Betriebsleiterwohnung im Gewerbegebiet und ist auch noch durch eine große Halle gegen das Veranstaltungsgelände abgeschirmt, so ist diese Situation ggf. deutlich unkritischer als eine schutzbedürftige Nutzung in einem reinen Wohngebiet in 500 m Entfernung ohne Abschirmung. Der maßgebliche Immissionsort würde in diesem Beispiel also im weiter entfernten Wohngebiet liegen. Ferner kommt es auf die Lage zur Bühne an. Schutzbedürftige Nutzungen in Hauptabstrahlrichtung der Bühne werden in der Regel stärker belastet als Nutzungen hinter der Bühne oder seitlich der Bühne.

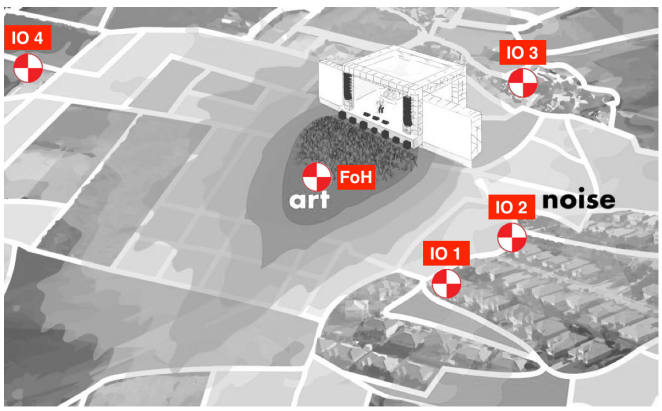


Abb. 6: Lage der Immissionsorte (IO) um das Festivalgelände und zusätzlicher Punkt am F.o.H. [Eigenständig modifizierte Darstellung. Quelle des Hintergrundbildes: d&b audiotechnik]

Nicht direkt jedes Haus oder gar jeder Raum eines Hauses gilt als schutzbedürftige Nutzung, was die Sache zusätzlich verkompliziert. Schlaf- und Wohnräume oder sonstige Räume zum dauernden Aufenthalt gelten als schutzbedürftige Nutzungen. Abstellräume oder Flure sind hingegen beispielsweise keine schutzbedürftigen Nutzungen. Küchen sind nur dann schutzbedürftige Nutzungen, sofern es sich um eine Wohnküche handelt, u. s. w.

Es ist nicht immer leicht, die tatsächliche Schutzbedürftigkeit einer Nutzung festzustellen. Dieser Faktor kann aber im Grenzfall genau den entscheidenden Unterschied ausmachen. Auch hier wird wieder ersichtlich, dass eine genaue Ortsbesichtigung wesentlicher Bestandteil der Arbeit eines Sachverständigen ist. Es hat sich schon mancher vermeintlich höchst kritischer Immissionsort als verlassener Kuhstall oder als längst verlassenes oder gar bereits abgerissenes Gebäude herausgestellt. Teils sind aber auch neue Gebäude errichtet, die noch in keiner Karte verzeichnet und in keinem Luftbild zu sehen sind. In Zeiten von Modulbauweise entstehen Gebäude innerhalb weniger Tage und werden auch umgehend bezogen. Somit entsteht in kürzester Zeit ein neuer Immissionsort.

Aus der Diskussion um den maßgeblichen Immissionsort wird auch eine weitere große Herausforderung ersichtlich: Bereits die Überschreitung der Richtwerte bei einem einzigen Anwohner kann dazu führen, dass die Veranstaltung formal nicht genehmigungsfähig ist. So kann auch ein einziger berechtigter Lärmkonflikt mit einem engagiertem Beschwerdeführer ein ganzes Festival kippen.

Oftmals werden in Genehmigungen gleich mehrere Immissionsorte um das Veranstaltungsgelände herum benannt. Damit wird z. B. möglichen Prognoseunsicherheiten und unerwarteten Wetterlagen entgegengewirkt. Ggf. sollen dabei auch beratungsresistente Dauerbeschwerdeführer durch die Erteilung besonderer Aufmerksamkeit zufriedengestellt werden.

Mitunter kommt es auch vor, dass Pegel für den F.o.H. oder einen anderen Punkt auf dem Veranstalungsgelände festgesetzt sind. Diese Vorgaben stehen nicht in Zusammenhang mit der DIN 15905-5 [2] und dem Schutz des anwesenden Publikums. Die Vorgaben stammen aus der Immissionsprognose, in der eine rein rechnerische Relation zwischen Immissionspegeln und dem Pegel am F.o.H. hergestellt wurde. Hier sei nochmals der Hinweis gegeben, dass diese Relation mit den klassischen Verfahren der Quellansätze und Berechnung nach DIN 9613-2 [11] mit großer Unsicherheit behaftet ist und es aus der reinen Berechnung kaum zu sinnvollen Ergebnissen kommt. In manchen Fällen können somit befremdliche Auflagen für die Pegel am F.o.H. oder an anderen Referenzpunkten innerhalb des Veranstaltungsgeländes zustande kommen. Sofern die auferlegten Richtwerte an einem Referenzpunkt auf dem Veranstaltungsgelände unrealistisch für eine Durchführung der Veranstaltung sind, sollte dringend frühzeitig mit den Behörden Rücksprache gehalten werden.

Bei gewöhnlichen und eher gemäßigten Veranstaltungen ist bereits nur durch die Anwesenheit des Publikums ein durchgehender Innenpegel in der Größenordnung von 80 dB(A) zu erwarten, sodass Beschallungspegel unterhalb von 90 dB(A) als Mindestversorgungspegel auf der Publikumsfläche grundsätzlich kaum sinnvoll realisierbar sind. Bei Rock- und Popkonzerten und vielen anderen Veranstaltungsformaten liegen die Pegel für eine veranstaltungstypische Durchführbarkeit oft deutlich darüber.

9. Richtwerte, Beurteilungszeiten und Beurteilungspegel

9.1. Richtwerte

Die Richtwerte an den Immissionsorten ergeben sich aus verschiedenen Faktoren. Wesentlich ist der Schutzanspruch, der sich aus der Gebietsausweisung des Flurstücks ergibt, auf dem sich das betreffende Gebäude befindet. Die Gebietsausweisung wird von der Behörde ermittelt oder festgelegt und ergibt sich beispielsweise aus einem Flächennutzungsplan (FNP) oder einem Bebauungsplan (B-Plan). Der FNP lässt hierbei gewisse Abwägungs- und Interpretationsspielräume. Ein rechtskräftiger B-Plan ist wesentlich unflexibler – hier sind die Gebietsausweisungen eindeutig festgeschrieben. Es kommt mitunter vor, dass für Teile der Umgebung B-Pläne und für andere Bereiche nur FNP oder überhaupt keine Pläne existieren. In Abbildung 7 ist eine Übersicht über die Immissionsrichtwerte im Außenbereich am Beispiel der Freizeitlärmrichtlinie LAI aufgeführt.

	Tagsüber Werktag außerhalb Ruhezeit	Tagsüber Werktag innerhalb Ruhezeit, Sonn- und Feiertage	Nachts (lauteste Nachtstunde)
Industriegebiete (IN)	**70 dB(A)**	**70 dB(A)**	**70 dB(A)**
Gewerbegebiete (GE)	**65 dB(A)**	**60 dB(A)**	**50 dB(A)**
Kern- Dorf- Mischgebiete (MI)	**60 dB(A)**	**55 dB(A)**	**45 dB(A)**
Allgemeine Wohngebiete (WA)	**55 dB(A)**	**50 dB(A)**	**40 dB(A)**
Reine Wohngebiete (WR)	**50 dB(A)**	**45 dB(A)**	**35 dB(A)**
Kurgebiete (z.B. Krankenhäuser)	**45 dB(A)**	**45 dB(A)**	**35 dB(A)**

Abb. 7: Exemplarische Immissionsrichtwerte im Außenbereich gemäß Freizeitlärmrichtlinie LAI [9].

Aus den Richtwerten in Abbildung 7 wird ersichtlich, dass es verschiedene Gebietsabstufungen gibt. Dabei ist nicht relevant in welchem Gebiet die Veranstaltung selbst stattfindet, sondern in welchem Gebiet der Immissionsort liegt. Die o. g. Immissionsrichtwerte gelten im Außenbereich und zwar formal in einem Abstand von $d = 0,5\,\mathrm{m}$ zum geöffneten Fenster. Die Immissionsrichtwerte beziehen sich auf den Beurteilungspegel L_r, dessen Basis der A-bewertete äquivalente Dauerschallpegel L_Aeq bildet und der noch verschiedene Zuschläge enthalten kann, vgl. Abschnitt 9.3. Es handelt sich beim Immissionsrichtwert also grundlegend um einen auf einen Mittelwert bezogenen Pegel. Zusätzlich sollen einzelne Geräuschspitzen die Immissionsrichtwerte tags um nicht mehr als 30 dB(A) sowie nachts um nicht mehr als 20 dB(A) überschreiten. Dieser Zusatz ist als Spitzenpegelkriterium bekannt.

9.2. Beurteilungszeiten

Neben den Immissionsrichtwerten ist die Beurteilungszeit t_B ein weiterer wichtiger Einflussfaktor. Folgende Beurteilungszeiten werden z. B. in der Freizeitlärmrichtline LAI [9] angesetzt:

Werktage:

- Tags, außerhalb der Ruhezeit (08.00 Uhr bis 20.00 Uhr): $t_\mathrm{B} = 12$ Stunden
- Tags, innerhalb der Ruhezeiten (06.00 Uhr bis 08.00 Uhr und 20.00 Uhr bis 22.00 Uhr): $t_\mathrm{B} = 2$ Stunden
- Nachts (22.00 Uhr bis 06.00 Uhr): $t_\mathrm{B} = 1$ Stunde, lauteste Nachtstunde

Sonn- und Feiertage:

- Tags, außerhalb der Ruhezeit (09.00 Uhr bis 13.00 Uhr und 15.00 Uhr bis 20.00 Uhr): $t_\mathrm{B} = 9$ Stunden
- Tags, innerhalb der Ruhezeit (07.00 Uhr bis 09.00 Uhr, 13.00 Uhr bis 15.00 Uhr und 20.00 Uhr bis 22.00 Uhr): $t_\mathrm{B} = 2$ Stunden
- Nachts, (0.00 Uhr bis 7.00 Uhr und 22.00 Uhr bis 24.00 Uhr): $t_\mathrm{B} = 1$ Stunde, lauteste Nachtstunde

Die Beurteilungszeiten spielen eine wesentliche Rolle in der Bildung des Beurteilungspegels, der letztendlich mit den Immissionsrichtwerten verglichen wird, vgl. Abschnitt 9.3. Die Zeit t_S in der die Veranstaltung tatsächlich stattfindet und Emissionen erzeugt werden kann nämlich auf den jeweiligen Beurteilungszeitraum t_B gemittelt und als Korrekturfaktor K_t berücksichtigt werden. Die zeitliche Korrektur K_t setzt also die tatsächliche Spielzeit t_S in Relation zum jeweiligen Beurteilungszeitraum t_B:

$$K_\mathrm{t} = 10 \cdot \log_{10}\left(\frac{t_\mathrm{S}}{t_\mathrm{B}}\right). \tag{1}$$

Tagsüber werktags außerhalb der Ruhezeit beträgt beispielsweise $t_\mathrm{B} = 12$ Stunden. Spielt nun eine Band von 18.00 Uhr bis 20.00 Uhr, so beträgt $t_\mathrm{S} = 2$ Stunden und $K_\mathrm{t} = -7,8$ dB. Die Spielzeit mittelt sich also auf die Beurteilungszeit. Dies ist tagsüber sehr hilfreich, da K_t den Beurteilungspegel entsprechend absenkt. Meistens finden Veranstaltungen jedoch abends statt. Hier liegen $t_\mathrm{B} = 2$ Stunden in der Ruhezeit von 20.00 Uhr bis 22.00 Uhr bzw. $t_\mathrm{B} = 1$ Stunde nach 22.00 Uhr in der Nachtzeit vor. Es bleibt also bei Abendveranstaltungen meist leider wenig Potential für eine Absenkung des Beurteilungspegels durch K_t.

Insbesondere zur Nachtzeit ist nicht nur die Beurteilungszeit sehr kurz, sondern es wird zudem auch nur die lauteste Nachtstunde betrachtet. Ist also in einer einzelnen Stunde der Immissionsrichtwert überschritten – was bei einer kurzen Beurteilungszeit von einer Stunde recht schnell passieren kann – so ist dies die formale Beurteilungsgrundlage für die gesamte Nachtzeit. Die anderen Nachtstunden sind formal nicht relevant.

Es kann insgesamt durchaus sinnvoll sein, die Programmplanung an die Beurteilungszeiten anzupassen und die Intervalle geschickt zu nutzen. Bei schalltechnisch kritischen Situationen wird dies auch so praktiziert.

9.3. Beurteilungspegel

Die Immissionsrichtwerte sind mit dem sogenannten Beurteilungspegel L_r zu vergleichen. Neben dem mit einem Schallpegelmesser erfassten oder mit der Prognosesoftware errechneten äquivalenten Dauerschallpegel L_Aeq gehen hier noch verschiedene Zuschläge und Korrekturfaktoren ein. Es wäre somit keinesfalls korrekt, lediglich den Pegel L_Aeq mit dem Schallpegelmesser am Immissionsort zu erfassen und ihn direkt mit dem Richtwert zu vergleichen. Zwischen dem

gemessenen L_{Aeq} und dem Beurteilungspegel L_{r} können im ungünstigen Falle 12 dB Differenz liegen. Statt der in der Tabelle angegebenen 55 dB(A) in einem allgemeinen Wohngebiet zur Tagzeit außerhalb der Ruhezeit dürften auf dem Schallpegelmesser somit beispielsweise nur $L_{\mathrm{Aeq}} = 43\,\mathrm{dB(A)}$ gemessen werden, um den Immissionsrichtwert einzuhalten.

Der Beurteilungspegel L_{r} wird (vereinfacht geschrieben) wie folgt gebildet:

$$L_{\mathrm{r}} = L_{\mathrm{Aeq}} + K_{\mathrm{inf}} + K_{\mathrm{imp}} + K_{\mathrm{t}} + K_{\mathrm{f}}. \qquad (2)$$

Mit:

L_{Aeq}	Anteiliger äquivalenter Dauerschallpegel innerhalb eines einzelnen Beurteilungszeitraums
K_{inf}	Zuschlag für Informationshaltigkeit / Tonhaltigkeit
K_{imp}	Zuschlag für Impulshaltigkeit
K_{t}	Zeitliche Korrektur, vgl. Abschnitt 9.2
K_{f}	ggf. Fremdgeräuschkorrektur

Vor dem Vergleich mit dem Immissionsrichtwert wird der Beurteilungspegel L_{r} stets auf volle dB gerundet. Der anteilige äquivalente Dauerschallpegel bezieht sich in der o. g. Gleichung darauf, dass auch nur während laufender Veranstaltung gemessen wird (ggf. mit einem geringen zeitlichen Vorlauf) und die Messung nicht genau am Beginn des Beurteilungszeitraums begonnen und genau am Ende des Beurteilungszeitraumes beendet wird. Die Differenz wird entsprechend durch K_t berücksichtigt, vgl. Abschnitt 9.2.

Der Zuschlag K_{inf} für Informationshaltigkeit / Tonhaltigkeit wird z. B. abhängig von der Sprachverständlichkeit nach Höreindruck seitens des Sachverständigen vergeben. Hier können Werte von $K_{\mathrm{inf}} = 0\,\mathrm{dB}$, $+3\,\mathrm{dB}$ oder $+6\,\mathrm{dB}$ angesetzt werden. Ist jede Textzeile am Immissionsort perfekt zu verstehen und klar zu hören, so fallen $K_{\mathrm{inf}} = +6\,\mathrm{dB}$ Zuschlag an. Sind zwischendurch Bruchteile zu verstehen oder ggf. bekannte Lieder gerade zu erkennen, so entspräche dies $K_{\mathrm{inf}} = +3\,\mathrm{dB}$. Die Grenzen und Interpretationsspielräume sind nicht klar definiert.

Der Zuschlag K_{imp} für Impulshaltigkeit wird entweder nach Höreindruck des Sachverständigen vergeben oder alternativ messtechnisch nach dem Taktmaximalverfahren bestimmt, vgl. TA Lärm [4]. Bei der Vergabe durch den Sachverständigen können Werte von $K_{\mathrm{imp}} = 0\,\mathrm{dB}$, $+3\,\mathrm{dB}$ oder $+6\,\mathrm{dB}$ angesetzt werden. Bei messtechnischer Ermittlung wird das tatsächliche Messsergebnis angesetzt. Hohe Impulshaltigkeit ist z. B. anzusetzen, wenn die Bassdrum immer deutlich als einzelner Impuls zu hören ist.

Die zeitliche Korrektur K_{t} wird in Abschnitt 9.2 erläutert.

Die Fremgeräuschkorrektur K_{f} kann bei einer Messung durch den Sachverständigen vergeben werden, sofern am Immissionsort Fremdgeräusche z. B. aufgrund von Verkehrslärm vorliegen, die nichts mit der Veranstaltung zu tun haben. Hier können jedoch in Anlehnung an die DIN 45645-1 [18] nur Werte bis maximal $K_{\mathrm{f\,max}} = -3\,\mathrm{dB}$ vergeben werden, was in in der Praxis in vielen Situationen der tatsächlichen Fremdgeräuschbelastung nicht gerecht wird. Bei deutlich höheren Fremdgeräuschanteilen können aber auch in der Regel keine aussagekräftigen Messungen durchgeführt werden. Die Behörde fordert üblicherweise nach der Veranstaltung einen eindeutigen Messwert bzw. Beurteilungspegel zum Vergleich mit der vorher genehmigten Immissionsschutzanforderung. Somit kann hier oft die Geräuschsituation nur abgeschätzt werden, wodurch in kritischen Situationen oder in Grenzfällen größere Herausforderungen oder Diskussionen zu erwarten sind.

10. Ausnahmegenehmigungen

Bei Veranstaltungen können die regulären Immissionsrichtwerte aus Abschnitt 9.1 sehr häufig nicht eingehalten werden. In diesen Fällen kann die Veranstaltung nur über eine entsprechende Ausnahmegenehmigung realisiert werden. Dabei gibt es verschiedene reguläre Ausnahmen, die bereits in den Regelwerken vorgesehen sind. Weiterführende Ausnahmen werden individuell verhandelt. Ein Veranstalter hat jedoch grundsätzlich keinen Rechtsanspruch auf die Erteilung einer Ausnahmegenehmigung. Die Entscheidung darüber obliegt der Ortsbehörde. Wesentlich für die Erteilung einer Ausnahmegenehmigung ist die Begründung. Nicht nur der Veranstalter muss der Behörde eine Begründung liefern, sondern die Behörde muss selbst ihre Entscheidung begründen und rechtfertigen können.

Gründe für eine Ausnahmegenehmigung liegen in der Regel dann vor, wenn besonderes öffentliches Interesse an der Durchführung der Veranstaltung besteht und/oder historische, kulturelle oder sonstige sozialgewichtige Umstände vorliegen. Insbesondere bei allgemeinen oder auch örtlichen Brauchtumsveranstaltungen wie Karneval, Schützenfest oder Oktoberfest liegen diese besonderen Umstände lokal vor. Bundesweit zählen dazu beispielsweise Festivitäten wie Silvester. Fragwürdig sind hingegen die tatsächlich historischen und kulturellen Umstände eines Oktoberfests in 500 km Entfernung zu München oder vergleichsweise neu importierte Festivitäten wie Halloween, u. s. w.

Es können aber mitunter auch große Konzerte und Tourneen unter eine solche Regelung fallen, sofern die Veranstaltung von besonders großem öffentlichen Interesse ist und ggf. die Entwicklung der Region in besonderem Maße fördert. Auch für Events im Rahmen von international relevanten Sportveranstaltungen wie Europa- oder Weltmeisterschaften im Fußball sind Ausnahmen gut begründbar. Die Ortsbehörde trifft in jedem einzelnen Falle eine individuelle Abwägung und entscheidet über die Erteilung einer Ausnahmegenehmigung.

Im Folgenden werden kurz zwei übliche reguläre Ausnahmen diskutiert, die bereits in den Regelwerken verankert sind und daher oft angewandt werden.

10.1. Seltene Ereignisse

Sowohl in den Freizeitlärmrichtlinien als auch in der TA Lärm werden sogenannte Seltene Ereignisse aufgeführt. Läuft eine Veranstaltung unter einem Seltenen Ereignis, so gelten gemäß Freizeitlärmrichtlinie LAI [9] oder TA Lärm [4] im Außenbereich Immissionsrichtwerte von 70 dB(A) zur Tagzeit und 55 dB(A) zur Nachtzeit und dies unabhängig von der Gebietsausweisung. In Industriegebieten gelten zur Nachtzeit

weiterhin immer 70 dB(A). Höhere Überschreitungen sind explizit zu begründen. In NRW gilt beispielsweise eine andere Regelung [10]. Hier dürfen die regulären Immissionsrichtwerte um jeweils 10 dB überschritten werden, jedoch darf die Höchstgrenze von 70 dB(A) zur Tagzeit und 55 dB(A) zur Nachtzeit dabei nicht überschritten werden.

Wie der Name bereits vermuten lässt, sollte ein Seltenes Ereignis ein seltenes Ereignis sein. In der TA Lärm [4] sind 10 Seltene Ereignisse pro Jahr festgesetzt. In der Freizeitlärmrichtlinie LAI [9] oder im Runderlass Freizeitlärm NRW [10] sind es immerhin 18 Tage pro Jahr, was eine der wesentlichen Abwandlungen der Freizeitlärmrichtlinien gegenüber der TA Lärm [4] darstellt.

Zudem beziehen sich die Seltenen Ereignisse immer auf ein gemeinsames akustisches Quartier. Liegen also mehrere Veranstaltungsstätten im Einwirkbereich gemeinsamer Immissionsorte, so hat nicht jede Veranstaltungsstätte 18 Seltene Ereignisse zur Verfügung, sondern die Anzahl bezieht sich auf alle Veranstaltungsstätten im gleichen akustischen Quartier zusammen. Somit haben kleinere Locations in der Nähe von Stadien, Arenen und großen Open Air Locations meistens das Nachsehen und bekommen keine Seltenen Ereignisse zugesprochen, da die Seltenen Ereignisse für die großen Locations mit Veranstaltungen von größerem öffentlichen Interesse benötigt werden.

10.2. Verschiebung der Nachtzeit

Die Freizeitlärmrichtlinien wie z. B. die Freizeitlärmrichtlinie LAI [9] sehen in besonders gelagerten Fällen die Möglichkeit einer Verschiebung der Nachtzeit um bis zu zwei Stunden vor. Somit können dann bis 23.00 Uhr oder maximal 24.00 Uhr die Tagesrichtwerte angesetzt werden. Zu beachten ist, dass weiterhin eine achtstündige Nachtruhe gewährleistet bleiben muss. Eine Verschiebung der Nachtzeit kommt also in der Regel nur vor Feiertagen und am Wochenende infrage und sollte nicht an mehreren Tagen oder Wochenenden hintereinander stattfinden.

10.3. Weitere Ausnahmen

Neben den üblichen Ausnahmen gibt es verschiedene bundesweite, länderbezogene und sogar städtebezogene Regelungen für besondere Ereignisse. Auch existieren z. B. Sehr Seltene Ereignisse, die nach den Landesimmissionsschutzgesetzen begründet werden. Darunter fallen ganz besondere Veranstaltungen wie z. B. Weltmeisterschaften und Europameisterschaften im Fußball. Sonderregelungen gibt es für Public Viewing. Beispielweise wurde in Köln für die Durchführung des ehemaligen Ringfestes das Ringfest-Kriterium entworfen. Die Ausnahmen und Sonderregelungen sind zahlreich.

In besonderen Situationen können in Abstimmungsgesprächen mit den Behörden auch ganz individuelle Lösungen für die Genehmigungssituation erarbeitet werden. Dies wird in der Praxis zunehmend zum Standardprozedere, da die individuellen Situationen in Ballungsräumen immer schwieriger aufzulösen sind. Bei detaillierter Begründung seitens der Behörde und ausführlicher schalltechnischer Prognose können der Geräuschsituation angepasste und sach- und fachgerecht erstellte Ausnahmegenehmigungen außerhalb der regulären Immissionschutzanforderungen im Streitfall vor Gericht standhalten.

Bei der Erteilung einer Ausnahmegenehmigung fordert der Ortsgesetzgeber jedoch in der Regel einen Nachweis über die Einhaltung der inhaltlichen Vorgaben in Form einer messtechnischen Begleitung der Veranstaltung durch einen anerkannten Sachverständigen oder eine benannte Messstelle.

10.4. Aufbau der Genehmigung

In der Genehmigung sind üblicherweise verschiedene Auflagen erteilt. Neben der Angabe des Datums der Veranstaltung und den Spielzeiten, die unbedingt einzuhalten sind, wird darauf verwiesen, auf Basis welcher Regelwerke die Genehmigung erteilt wird.

In der Genehmigung können Minderungsmaßnahmen (z. B. Schallschutzwände, Ausrichtung der Bühne, Beschallungskonzept) festgesetzt sein. Diese Minderungsmaßnahmen sind unbedingt sehr genau einzuhalten, da die Veranstaltung bei Verstoß gegen die ordnungsbehördliche Erlaubnis jederzeit sowohl bei der Bauabnahme im Vorfeld als auch bei laufender Veranstaltung seitens der Behörde stillgelegt werden kann.

Neben den o. g. Auflagen werden in der Genehmigung Immissionsorte (mindestens der maßgebliche Immissionsort, ggf. weitere Immissionsorte) benannt und dazugehörige Beurteilungs- und Spitzenpegel aufgeführt. Diese werden bei nicht seltenen Ereignissen und bei Seltenen Ereignissen auf die Beurteilungszeiträume (innerhalb/außerhalb der Ruhezeit, Nachts – lauteste Stunde) gemäß heranzuziehender Freizeitlärmrichtlinie bezogen. Bei Sehr Seltenen Ereignissen werden die Pegel oft auf eine volle Zeitstunde bezogen. Wie bereits diskutiert, werden teilweise auch Vorgaben für den Innenpegel der Veranstaltung (z. B. am F.o.H. oder in einem bestimmten Abstand zur Beschallungsanlage) auferlegt.

11. Messung

Sofern eine Ausnahmegenehmigung erteilt wird, muss die Veranstaltung in der Regel auch messtechnisch durch einen anerkannten Sachverständigen oder eine benannte Messstelle begleitet werden. Die Behörde muss einen Nachweis über die Einhaltung der inhaltlichen Vorgaben der Genehmigung führen. Die Auferlegung der Messung ist somit nicht als Schikane seitens der Behörde zu verstehen. Die Behörden haben selbst Nachweispflicht und müssen ihre Entscheidungen im Streitfall rechtfertigen können.

11.1. Stichproben- vs. Dauermessungen

Grundsätzlich gibt es bei Messungen zwei Vorgehensweisen. Entweder werden Dauermesspunkte errichtet, an denen die Pegel kontinuierlich während der gesamten Veranstaltung erfasst und aufgezeichnet werden oder es werden stichprobenartige Messungen durchgeführt. Dabei fährt der Gutachter mehrere Immissionsorte ab und misst dort jeweils für ein paar Minuten. Problematisch ist bei Stichproben, dass kein lückenloser Nachweis geführt wird und es schwer einzuschätzen ist, ob die erhobenen Stichproben tatsächlich repräsentativ für die gesamte Veranstaltung sind. Bei Veranstaltungen gibt es in der Regel verschiedene Emissionssze-

narien, z. B. Live-Band, Moderation, DJ, Umbaupausen, etc. Eine Stichprobe kann immer nur einen Teil der Szenarien erfassen. Zudem kann sich während der Veranstaltung die meteorologische Situation verändern. Eine Veränderung der Windrichtung oder der Temperaturschichtung kann zu maßgeblichen Veränderungen der Immissionssituation führen.

Bei Dauermessungen gibt es diesbezüglich keinerlei Unsicherheiten und Diskussionen, was letztendlich für alle Beteiligten von Vorteil ist. Ferner sollten alle Dauermesspunkte durchgehend mit Protokollführern besetzt sein, die Fremdgeräuscheinwirkungen, besondere Vorkommnisse und Höreindrücke mit genauer Uhrzeit und Dauer notieren oder entsprechende Marker im Messsystem setzen. So wird ausgeschlossen, dass die gemessenen Pegelwerte durch zufällige Fremdeinwirkung (Verkehr, Passanten, Hundegebell, Ehestreit auf dem Balkon, Rasenmäher u. s. w.) gestört oder gar vorsätzlich durch einen beschwerdeführenden Anwohner manipuliert werden.

Abb. 8: Dauermesspunkt bei einer Veranstaltung. Messwagen mit Mast und Protokollführer.

Die jeweiligen Protokollführer müssen zwar eingewiesen werden, es können dafür aber z. B. studentische Aushilfen ohne detaillierte Sachkenntnis eingesetzt werden. Jedoch muss dann der verantwortliche Sachverständige jeden Messpunkt persönlich besuchen, um einen Höreindruck zu erhalten und ggf. entsprechende Zuschläge zu vergeben, da die Zuschläge ausschließlich durch einen Sachverständigen vergeben werden dürfen.

In vielen Genehmigungen werden zunehmend explizit Dauermessungen an den Immissionsorten gefordert, da diese eine wesentlich zuverlässigere Einschätzung der Situation ermöglichen. Ggf. sind die Dauermesspunkte noch um stichprobenartige Messungen an zusätzlichen Immissionsorten zu ergänzen. Aus manchen Genehmigungen geht nicht eindeutig hervor, ob die Messungen stichprobenartig erfolgen sollen oder ob eine Dauermessung durchgeführt werden soll. Die erste Wahl sollte immer die Dauermessung sein, um spätere Diskussionen und Unsicherheiten zu vermeiden. Bei sehr großen Veranstaltungen kann es ggf. auch problematisch sein die Immissionsorte für Stichprobenmessungen überhaupt in realistischer Zeit zu erreichen. Im Umfeld großer Veranstaltungen sind oft Straßen gesperrt und Wege durch die Polizei blockiert, sodass die sequentielle Messung an mehreren Immissionsorte innerhalb einer verhältnismäßig kurzen verfügbaren Zeitspanne zu einer großen und vor allem unabsehbaren logistischen Herausforderung wird.

11.2. Messposition und -höhe

Gemäß TA Lärm [4] und den Freizeitlärmrichtlinien muss die Messung in einem Abstand $d = 0,5\,\mathrm{m}$ vor dem geöffnetem Fenster am Immissionsort stattfinden. Das Fenster soll dabei geöffnet sein, um Reflexionen zu vermeiden.

Abb. 9: Normgerechte Messung in $d = 0,5\,\mathrm{m}$ vor dem geöffnetem Fenster.

Diese Vorgaben sind im Kontext von Messungen im Bereich Gewerbe- und Industrielärm möglicherweise sinnvoll und durchführbar, da hier meist einmalig für einen überschaubaren Zeitraum eine bestimmte Betriebssituation gemessen wird. Im Kontext von Veranstaltungen ist diese Vorgabe problematisch. Der vor schädlichen Umwelteinflüssen zu schützende Anwohner müsste demnach während jeder Veranstaltung das Fenster seines Schlafzimmers geöffnet lassen. Diese Idee stößt bei weitem nicht in jedem Fall auf begeistertes Verständnis der Betroffenen. In der Praxis wird meistens mit den Behörden ein entsprechender Ersatzimmissionsort abgestimmt. Dieser Immissionsort liegt dann beispielsweise einige Meter vom Haus entfernt im Garten oder auf dem Gehweg, vgl. Abbildung 10. Messtechnisch macht dies in typischerweise größeren Entfernungen von teilweise mehreren hundert Metern ohnehin keinerlei relevanten Unterschied.

Wesentlich für eine aussagekräftige Messung ist jedoch die Wahl der korrekten Messhöhe. Meist steigen die Pegel zu größeren Höhen an, sodass eine Messung im Erdgeschoss zwar bequem ist, aber keine realistischen Werte liefert. Die Geschosshöhe wird in manchen Genehmigungen genau vorgegeben oder ist in der zugehörigen Immissionsprognose genauer spezifiziert. Je nach Bebauung reicht ein konventioneller Mast für die Messungen aus. In manchen Fällen

Abb. 10: Typischer Ersatzimmissionsort während einer Veranstaltung, der nach Absprache mit der Behörde einige Meter neben dem relevanten Gebäude lokalisiert ist, um dem Anwohner das Öffnen des Fensters zu ersparen.

muss aber beispielsweise auch mal auf dem Balkon eines Hochhauses im 25. Obergeschoss gemessen werden. Hier müssen im Vorfeld die entsprechenden Zugänge mit den Anwohnern abgeklärt werden.

11.3. Messtechnik und Messwerte

Für die Durchführung der Messungen bestehen besondere Anforderungen an die einzusetzende Messtechnik. Für Immissionsmessungen sind geeichte Schallpegelmesser der Klasse I nach DIN EN 61672-1 [19] einzusetzen. Vorsicht ist bei der Anschaffung der Geräte geboten, da nicht jeder Schallpegelmesser der Klasse I auch automatisch eichfähig ist. Benannte Messstellen müssen zusätzlich zu der regelmäßigen Eichung auch jeweils noch eine akkreditierte Kalibrierung ihrer Geräte durchführen lassen. Hier entsteht nach der initialen Anschaffung der Geräte ein wiederkehrender und nicht unerheblicher Kostenaufwand. Messgeräte sind vor und nach der Messung zu kalibrieren. Die genaue Durchführung der Messungen und die zu erfassenden Parameter werden in der TA Lärm [4] beschrieben. Es gibt hierzu keine eigenen Angaben in den Freizeitlärmrichtlinien.

Während der Messung sind gemäß TA Lärm [4] mindestens folgende Werte zu erfassen:

- Äquivalenter Dauerschallpegel L_{AFeq}
- Taktmaximalpegel L_{AFTeq}
- Spitzenpegel L_{AFmax}
- 95%-Perzentilpegel $L_{\mathrm{AF\,95}}$

Alle Pegel werden mit A-Bewertung erfasst. Im Immissionsschutz ist in allen Regelwerken die Zeitkonstante *Fast* für die Messungen vorgeschrieben. Die Bedeutung der einzelnen Parameter aus der obigen Liste ist in der frei verfügbaren TA Lärm [4] genau beschrieben und eine detaillierte Erläuterung wird daher an dieser Stelle ausgespart. Neben den unbedingt notwendigen Parametern werden oft noch diverse weitere Parameter erfasst. Üblich ist auch die Erfassung des Spektrums in 1/3 Oktavbändern oder als kontinuierliches FFT-Spektrum. Aus den Spektren können die pegelbestimmenden Frequenzanteile identifiziert werden oder auch Rückschlüsse auf Fremdgeräusche gezogen werden.

Die Speicherung der Rohdaten erfolgt meist im gemittelten Sekundentakt, wobei die Erfassungsrate standardmäßig bei 125 ms liegt. Die entsprechenden Messwerte werden zur weiteren Aufbereitung auf die relevanten Zeitintervalle gemittelt. Zum Vergleich mit den Immissionsrichtwerten muss noch der Beurteilungspegel gebildet werden, vgl. Abschnitt 9.3.

11.4. Messbericht

Nach der Veranstaltung wird ein umfangreicher Messbericht in gutachterlicher Form erstellt. Hierin finden sich neben allgemeinen Informationen über die Veranstaltung und Lageplänen zur Verortung der Veranstaltung und der Immissionsorte die Messwerte und die Beurteilungspegel. Weiterhin können Höreindrücke oder besondere Vorkommnisse beschrieben sein. Ein üblicher Messbericht für eine einzelne Veranstaltung kann gut 20 bis 30 Seiten umfassen. Der Bericht wird dem Veranstalter (Auftraggeber) und in der Regel parallel der Behörde zugestellt.

Im Idealfall zeigt der Bericht auf, dass alle Immissionsrichtwerte entsprechend der Genehmigung eingehalten wurden und dass es keine besonderen Vorkommnisse gab. In diesem Falle sind alle Beteiligten zufrieden und das Festival kann im nächsten Jahr wieder stattfinden. In ungünstigeren Fällen zeigt der Bericht eine Überschreitung der Immissionsrichtwerte auf. Die Behörde wird daraus entsprechende Konsequenzen ziehen. Bei einer geringfügigen Überschreitung z. B. durch unvorhersehbare meteorologische Umstände kann es sein, dass die Behörde keine weiteren Schritte unternimmt. Gehen jedoch im schlimmsten Falle aus dem Bericht höhere Überschreitungen hervor, die zudem mit uneinsichtigem Verhalten des Veranstalters, Technikdienstleisters oder des Künstlers einhergehen, so werden relativ sicher entsprechende Konsequenzen folgen. Hier stehen beispielsweise Geldbußen an, die nicht nur den Veranstalter betreffen können, sondern bis herunter auf den einzelnen Tontechniker oder den Künstler greifen können. Manche Veranstalter rechnen diese Bußgelder in ihr Budget ein und nehmen diese bereits im Vorfeld der Veranstaltung in Kauf. Das wird von den Behörden selbstverständlich nicht gerne gesehen. Neben der Auferlegung von Bußgeldern hat die Behörde aber auch noch einen wesentlich stärkeren Hebel an der Hand. Sie kann nämlich die Genehmigungen für Folgeveranstaltungen verwehren, was die Veranstalter in der Regel noch deutlich härter trifft und ihnen ggf. sogar die Lebensgrundlage entzieht. Veranstalter sowie auch Techniker und Bands sollten die Auflagen in einer Genehmigung sowie die wohlgemeinten Ratschläge des Sachverständigen vor Ort somit ausgesprochen ernst nehmen.

11.5. Verteilte Messsysteme

Um einen optimalen Überblick über die Immissionssituation während der laufenden Veranstaltung zu erhalten, bietet sich der Einsatz verteilter und vernetzter Messsystemen mit Echtzeitauswertung an. Dabei werden die Daten der Messstationen an den Immissionsorten über eine öffentliche Datenverbindung, lokale Netzwerke oder über Funk auf das Veranstal-

tungsgelände übertragen und ggf. in ein Netzwerk eingespeist, das den Abgriff durch mehrere Clients am F.o.H., beim Systemtechniker, im Production Office, beim Sachverändigen etc. ermöglicht, vgl. Abbildung 11.

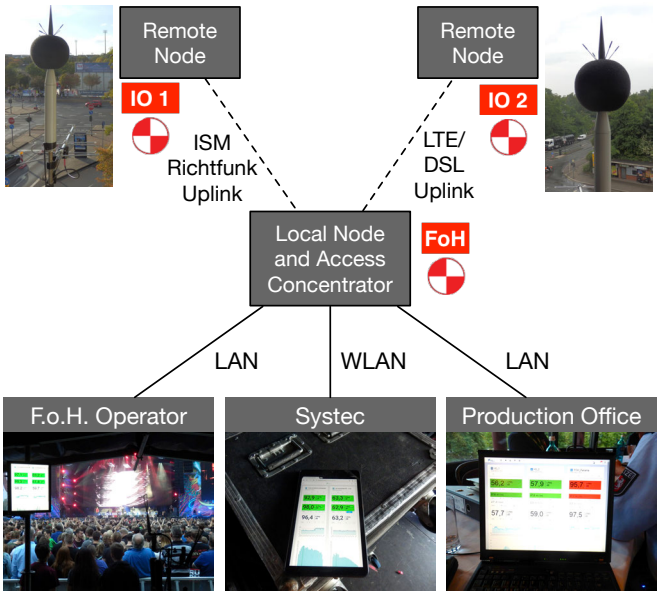


Abb. 11: Beispielkonfiguration des Konzertbegleiter Messsystems der Kramer Schalltechnik mit zwei Immissionsorten (IO) und Messung am F.o.H., bei dem die Messwerte der IOs übertragen, in einem Access Concentrator gesammelt und in einem lokalen Netzwerk auf verschiedene Clients (F.o.H. Operator, Systec, Production Office, u. s. w.) verteilt und dort dargestellt werden.

Derartige Systeme werden mittlerweile von den meisten größeren Herstellern von Schallpegelmessern angeboten. Jedoch setzen die Hersteller hier zumeist auf cloudbasierte Lösungen, bei denen die Messstationen ihre Werte über das Internet in die Cloud übertragen. Die Werte können dann aus der Cloud zur Darstellung und Sichtung wieder über das Internet von einem Client abgegriffen werden. Diese Systeme sind in der Regel für den Betrieb von festen Dauermessstationen z. B. für die Pegelüberwachung an Industrieanlagen, Flughäfen etc. entworfen und setzen eine gut funktionierende Verbindung zum Internet voraus. Bei dauerhaft installierten Systemen ist die Anbindung über DSL, Glasfaser oder auch über Mobilfunknetze kein großes Problem.

Die Situation bei Veranstaltungen unterscheidet sich aber durchaus. Hier werden in der Regel temporäre Stationen für die Dauer der Veranstaltung errichtet. Stationäre Internetverbindungen am Immissionsort sind nur in seltenen Ausnahmefällen verfügbar. Bei sehr großen Veranstaltungen sind zudem die öffentlichen Mobilfunknetze durch die hohe Anzahl an Devices der Besucher regelmäßig derart ausgelastet, dass keine stabile und kontinuierliche Datenverbindung mehr möglich ist. In diesen Fällen sind also alle Dienste ungeeignet, die eine stabile Internetanbindung erfordern. Diese Problematik wird sich vermutlich in Zukunft mit neuen Mobilfunktechnologien oder auf Basis lokaler mobiler Netzwerkinfrastrukturen entschärfen. Jedoch lässt sich auf dem aktuellem Stand der Technik häufig im Vorfeld nicht absehen, ob lokale Netzinfrastrukturen bei einer Veranstaltung verfügbar und zuverlässig sind oder ob eine Datenverbindung über öffentliche Mobilfunknetze bei großen Veranstaltungen ausreichend gut funktioniert.

Für dergestaltige Situationen ist es also momentan noch notwendig, lokal eine eigene Infrastruktur für die Übertragung der Messdaten und ggf. auch für die Kommunikation zu errichten. Hierfür bietet sich beispielsweise die Nutzung von klassischen Datenfunkmodems in den ISM-Bändern mit entsprechenden Richtantennen an. Ein Beispiel für einen Messpunkt mit einer Richtantenne ist in Abbildung 10 zu sehen. Es könnten aber durchaus auch alternative Übertragungswege und Funksysteme eingesetzt werden.

Verteilte und vernetzte Messsysteme bieten die Möglichkeit einer Echtzeitüberwachung der Immissionssituation. Sind neben der Datenfernübertragung auch noch entsprechende Mittelungs- und Auswerteroutinen im Messsystem vorhanden, so können direkt in Echtzeit der für die Beurteilung relevante intervallbezogene Immissionspegel und zusätzlich auch noch ein prognostischer Intervall-Restpegel berechnet werden. Fremdgeräusche an den Immissionsorten können teils automatisch erkannt oder manuell vom Messposten markiert werden. Die betroffenen Segmente können dann automatisch aus der Mittelung entfernt werden. Temporäre Fremdgeräusche am Immissionsort lassen sich zudem bei Vergleich der Pegel-Zeit-Verläufe von Immissionsort und Innenbereich der Veranstaltung identifizieren.

Der große Vorteil eines solchen Systems liegt darin, dass die Immissionsrichtwerte mithilfe der live ausgewerteten Information bis auf eine sehr geringe Restsicherheit von beispielsweise 0,5 dB vollständig ausgeschöpft werden können. Dabei werden auch automatisch mögliche meteorologische Veränderungen während des Veranstaltungsverlaufs berücksichtigt und es kann entsprechend unmittelbar auf Veränderungen jeder Art reagiert werden. Der vom System in Echtzeit ausgewertete und am F.o.H. angezeigte Pegel erscheint nach der Veranstaltung in exakt gleicher Höhe im Messbericht.

Dies gibt allen Beteiligten eine sehr hohe Handlungssicherheit während der Veranstaltung. Ohne derartige vernetzte Messtechnik wäre eine vergleichbare Annäherung mit genauer Punktlandung auf dem Immissionsrichtwert nicht möglich bzw. ein gefährliches Spiel mit dem Feuer. Ein seriöser schalltechnischer Gutachter würde hier immer einige dB Sicherheitsabstand zu den Immissionsrichtwerten halten, um keine versehentliche Überschreitung zu riskieren. Dieser Sicherheitsabstand geht entsprechend für den Innenpegel der Veranstaltung verloren, was in grenzwertigen Situationen genau den Unterschied zwischen guter oder eben kaum möglicher Durchführbarkeit der Veranstaltung ausmachen kann.

Ein weiterer wichtiger Aspekt liegt darin, dass mithilfe solcher Systeme und entsprechender grafischer Anzeige der Immissionssituation alle verantwortlich Beteiligten, wie der F.o.H. Operator, der Systemtechniker, der Veranstalter, der Sachverständige und ggf. auch Behörden, zu jedem Zeitpunkt völlig transparent über die aktuelle Lage informiert sind. Werden an einem Immissionsort die Richtwerte überschritten, so ist dieses allen Beteiligten sofort ersichtlich und es folgt keine große Überraschung am Ende des Abends oder gar

eine Woche später nach der Auswertung. An allen wichtigen Positionen sollten entsprechende grafische Anzeigen der Echtzeitdaten installiert werden. An den technischen Positionen sollten hier durchaus detaillierte Messwerte inklusive Immissionsspektren zur Verfügung gestellt werden, damit die Techniker entsprechend reagieren können. An anderen Positionen bietet sich oft auch eine einfache Ampeldarstellung (Grün: Immissionsrichtwert eingehalten und unterschritten, Gelb: Knapp unterhalb oder auf dem Richtwert, Rot: Richtwert überschritten) in Kombination mit einem gemittelten Immissionspegel an.

Im Bereich der verteilten Echtzeitmesssysteme besteht noch großes Entwicklungspotential. Anbieter, wie beispielsweise die Niederländische Firma Event Acoustics, führen in ihrem Messsystem METRAO [20] Audiodaten der verschiedenen Messpunkte auf einem serverbasierten Knoten zusammen und können durch Korrelation Fremdgeräusche erkennen oder die Anteile verschiedener Bühnen trennen. Eine entsprechend performante Netzwerkverbindung wird vorausgesetzt. Weiterhin wäre in Zukunft auch Musterekennung z. B. unter Einsatz von KI-Systemen zur Erkennung und Klassifizierung von Fremdgeräuschen denkbar. Das Problem der stabilen Netzwerkinfrastrukturen wird mittelfristig von dafür zuständigen Gewerken ohnehin gelöst werden.

Abb. 12: Mobile Messstationen des Konzertbegleiter Messsystems der Kramer Schalltechnik mit kalibrierten Klasse I Schallpegelmessern (hier Svantek SVAN 979) zur Erfassung der Messdaten. Die Daten werden auf einem Messkern (Rechner: Raspberry Pi, Implementierung: Node JS) aufbereitet, in Echtzeit gemittelt, ausgewertet und über ein im Case integriertes ISM Datenfunkmodem oder alternativ über einen LTE Router zum Veranstaltungsgelände übertragen. Die Messstationen verfügen über ein Terminal zur Konfiguration und eine integrierte unterbrechungsfreie Notstromversorgung für ca. 20 Stunden Dauerbetrieb. Die Messkerne spannen zudem ein lokales W-Lan Netzwerk auf, über das sich die Messhelfer per Browser einwählen können und so z. B. Marker setzen, Notizen machen oder in einem integrierten Chatsystem mit dem F.o.H. in Kontakt treten können. Letzteres ist bei Großerveranstaltungen sehr hilfreich, da die öffenlichen Mobilfunknetze oftmals überlastet sind und so keine Kommunikation möglich ist. Ferner ist die Konfiguration der Remote-Messkerne z. B. vom F.o.H. aus fernsteuerbar.

Bei aktuell kommerziell auf dem Markt verfügbaren Systemen der Messgerätehersteller sind es häufig eher Kleinigkeiten, die eine Nutzung im Veranstaltungsbereich etwas unpraktikabel machen. Dazu gehören z. B. derzeit noch die Voraussetzung von Internetverbindungen oder die Nutzung von Funktechnologien wie ZigBee, welche aufgrund der geringen Reichweiten die Notwendigkeit für die Errichtung von mehreren Funkzellen bzw. Repeatern auf dem Veranstaltungsgelände oder im Umfeld erfordern. Dies ist im Eventkontext zu aufwendig und somit unpraktikabel. Es beginnt aber oft mit noch wesentlich trivialeren Dingen. Viele Systeme und auch die üblichen Schallpegelmesser bilden zum Beispiel einen laufenden Stundenmittelwert L_{Aeq} direkt ab Messbeginn. Nach deutscher Regelwerkslage sind aber z. B. in der Nachtzeit Stundenintervalle der vollen Zeitstunden zu berücksichtigen. Auch im Tageskontext wird in der Praxis häufig auf volle Zeitstunden zurückgegriffen, die später auf die Beurteilungszeiten gemittelt werden. Mit einem stetig laufenden Mittelwert, der nicht exakt zur vollen Stunde zurückgesetzt wird, kann man vor Ort jedoch keine wirklich engmaschige und zuverlässige Abschätzung der Situation treffen, wie es bei Veranstaltungen in der Praxis notwendig ist. Der Veranstaltungsmarkt ist für die größeren Messgerätehersteller vermutlich zu klein und exotisch, als dass hier größerer Entwicklungsaufwand in eventspezifische Lösungen gesteckt wird, die für eine brauchbare Echzeitauswertung jeweils noch die Regelwerke der einzelnen Länder und sogar der Bundesländer berücksichtigen oder entsprechend flexibel konfigurierbar sein müssten. Aus diesem Grunde führen derzeit einige Messbüros Eigenentwicklungen durch, wie z. B. auch die Kramer Schalltechnik ein eigenes Messsystem namens Konzertbegleiter entwickelt hat und in Kooperation mit der Technischen Hochschule Mittelhessen stetig weiterentwickelt, vgl. Abbildungen 11 und 12.

12. Minderungsmaßnahmen

Bei Veranstaltungen können verschiedene Minderungsmaßnahmen eingesetzt werden. In vielen Fällen handelt es sich um relativ triviale Maßnahmen, die jedoch große Wirkung haben können. An erster Stelle steht hier eine sinnvolle Positionierung und Ausrichtung der Bühne, sofern es dafür Spielraum gibt. Die Bühnenöffnung sollte nach Möglichkeit nicht in Richtung sensibler Immissionsorte ausgerichtet werden.

Ausgesprochen wirksame Minderungsmaßnahmen lassen sich direkt an der Quelle vornehmen. Dabei geht es darum, die Beschallungsanlage derart aufzubauen und zu konfigurieren, dass der Schall gezielt auf die Publikumsfläche gerichtet wird, der Schallpegel auf der Publikumsfläche maximiert und im sonstigen Umfeld entsprechend abgesenkt wird.

Im Folgenden werden kurz einige Minderungsmaßnahmen diskutiert.

12.1. Schallschutzwände

Zur Minderung können, wie allgemein im Lärmschutz sehr üblich, klassische Schallschutzwände eingesetzt werden. Hier gibt es verschiedene mobile Varianten, die sich mit mehr oder weniger großem Aufwand temporär errichten lassen. Einige Beispiele finden sich in Abbildung 13.

In diesem Kontext ist es wichtig, dass die Wände auch bei niedrigen Frequenzen ein möglichst hohes Schalldämmmaß aufweisen. Die Schalldämmmaße werden in Datenblättern häufig nur bis zu 100 Hz angegeben, was mit den Normen

Abb. 13: Klassische Minderungsmaßnahmen bei Veranstaltungen in Form mobiler Schallschutzwände. Exemplarische Ausführungen als Seecontainerwand, Modulwand oder mit IBC-Containern.

aus der Bauakustik und dem bewerteten Bau-Schalldämm-Maß zusammenhängt. Im Eventbereich sind aber insbesondere auch die Frequenzen unterhalb von 100 Hz bis z. B. 30 Hz von besonderem Interesse. Der Trend geht generell zu immer tieferer Basswiedergabe und zu höherer Energie bei tiefen Frequenzen. Hier sei also große Vorsicht geboten bei Schallschutzfolien oder sonstigen leichten Konstruktionen, die ggf. einen attraktives mittleres Schalldämmmaß versprechen, jedoch bei tieferen Frequenzen oft kaum noch eine relevante Wirkung zeigen.

Weiterhin muss die Schallschutzwand eine ausreichende Höhe aufweisen. Zu niedrige Schallschutzwände haben so gut wie gar keinen Effekt. Insbesondere die Dämmung tieferer Frequenzen ist auch hier wieder ein wesentlicher Aspekt, da sich der Schall bei großen Wellenlängen besonders gut um Hindernisse wie eine Schallschutzwand beugt.

Schallschutzwände haben neben der eigentlichen Schalldämmung oft den zusätzlichen Vorteil, dass dahinter die Sprachverständlichkeit geringer wird und ggf. Impulse weniger intensiv zu vernehmen sind. Hier gehen also möglicherweise noch positive Effekte in Form geringerer Zuschläge einher, was die klassische Schallschutzwand zu einem zwar aufwendigen aber durchaus in vielen Situationen probaten und hilfreichen Mittel macht, das oftmals einen größeren Effekt auf den Beurteilungspegel hat als nur die Reduktion durch die rein technische Schalldämmung.

12.2. Line Arrays

Line Arrays können dazu eingesetzt werden, den Schall in der Vertikalen relativ gezielt auf die Publikumsfläche zu konzentrieren. Hier kommt es aber auf eine korrekte Konfiguration und einen entsprechenden Aufbau an. Das Line Array sollte dabei hoch hängen und von oben auf das Publikum nach unten abstrahlen, sodass das Publikum den Schall weitestmöglich absorbiert.

Ein klassisches Problem mit Bezug auf den Immissionsschutz sind zu tief hängende oder im Extremfall gar gestackte Line Arrays, vgl. Abbildung 14. Hier entsteht physikalisch bedingt eine ausgeprägte Richtkeule parallel zum Boden. Die größten Teile des Schalls werden dabei nicht auf das Publikum gerichtet und entsprechend in die weitere Umgebung emittiert. Längere Schallzeilen weisen ohnehin generell besonders in mittleren Frequenzbändern ein erweitertes Nahfeld auf, in dem der Schall nur mit 3 dB pro Entfernungsverdoppelung abfällt, was sich deutlich in der Nachbarschaft bemerkbar machen kann. Somit sollten gestackte oder sehr niedrig geflogene Arrays generell vermieden werden, damit weniger Schall in die Umgebung emittiert wird. Aber auch für eine gleichmäßige Pegelverteilung auf der Zuhörerfläche sind derartige Konfigurationen nicht besonders zuträglich. In der Regel sind dies eher budget- oder bauzeitenbedingte Notlösungen bei kleineren Veranstaltungen. Hier sollte aber nicht an der falschen Stelle gespart werden und das Line Array zumindest auf eine vernünftige Höhe gebracht werden. Diese Maßnahme kann ganz wesentliche Beiträge zur Reduktion der Immissionsbelastung liefern.

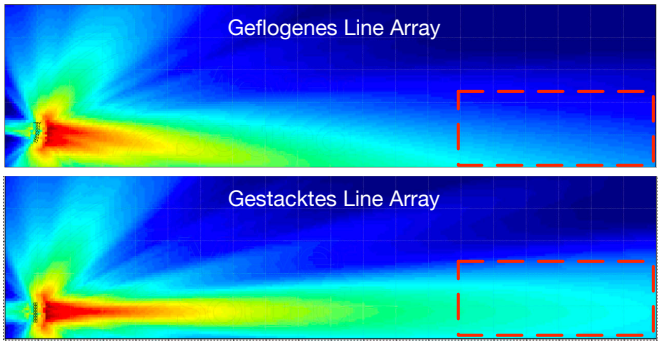


Abb. 14: Exemplarische und rein qualitative Darstellung der Richtwirkung eines geflogenen Line Arrays Systems gegenüber der gestackten Variante bei $f = 1$ KHz mit einer Oktave. Zu beachten ist jeweils der rot umrandete Bereich, der hinter der zu beschallenden Publikumsfläche liegt. Das geflogene Array ist gecurvt und geneigt. Die Energie wird auf die Publikumsfläche konzentriert und fällt dahinter schnell ab. Das gestackte Array hingegen weist eine ausgeprägte Richtkeule parallel zum Boden auf, die zu einem großen Problem bezüglich des Immissionspegels beim Anwohner werden kann. Simulation in MAPP XT [21].

12.3. Subwoofer Arrays

Bass lässt sich aufgrund der großen Wellenlänge schwieriger richten. Subwoofer ohne besondere zusätzliche Funktionalität strahlen nahezu kugelförmig ab. Jedoch lassen sich durch räumlich verteilte Anordnung und individuelle Ansteuerung mehrerer Subwoofer Arrays aufbauen, mit denen ganz erhebliche Richtwirkung auch bei tiefsten Frequenzen erzielt werden kann. Dazu gehören Standardanordnungen wie Straight Line, Delayed Straight Line, Arc, Cardioid oder Endfire. Es lassen sich aber auch von den Klassikern abweichende freie Konfigurationen erstellen, die einer spezifischen Situation möglicherweise noch besser gerecht werden. Hier ist natürlich entsprechende Kompetenz und Experimentierfreudigkeit gefragt. Für Standardanordnungen wie Cardioid Arrays bieten die Hersteller vorbereitete Presets an. Einige bieten auch Subwoofer mit einer bereits integrierten Cardioid-Funktionalität an, bei denen rückwärtig ein zweites Chassis verbaut ist und mit denen direkt eine hohe Rückwärtsdämpfung ohne explizite Arraybildung erzielbar ist.

Gut geplante, aufgebaute und eingemessene Subwoofer Arrays können ein maßgeblicher Schlüsselfaktor im Noise Management sein und die Immissionsbelastung in bestimmten Abstrahlrichtungen erheblich absenken. Wichtig ist dabei die kritischen Immissionsgebiete genau zu kennen und die Arrays entsprechend individuell auszulegen. Ein rein qualitatives und anschauliches Beispiel ohne nähere Erläuterung für eine Subwoofer-Optimierung auf einem Open Air Festival findet sich in Abbildung 15.

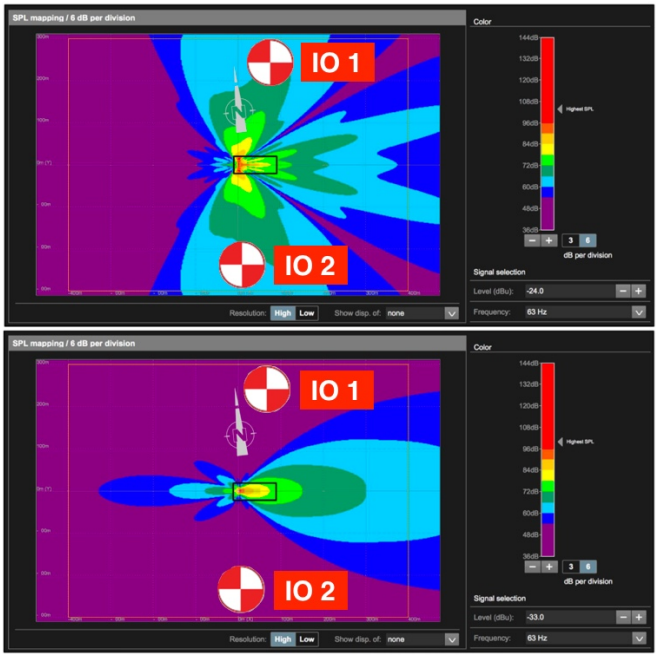


Abb. 15: Rein qualitatives Beispiel für eine Optimierung der Immissionssituation durch räumliche Umgruppierung und Processing der Subwoofer auf einem Festival. Das obere Bild zeigt das Abstrahlverhalten des ursprünglich vom Technikdienstleister geplanten Aufbaus für die exemplarische 63 Hz Terz. Das untere Bild zeigt das Abstahlverhalten nach der Optimierungsmaßnahme mit einer erheblichen Reduktion der Immissionsbelastung an den nördlich und südlich der Veranstaltungsfläche lokalisierten Immissionsorten. Simulation in ArrayCalc [16].

Diese Art der Optimierung ist in der Regel kein großer Aufwand und stellt für einen erfahrenen Systemtechniker keinerlei Schwierigkeit dar. Häufig wird beim Systemdesign aber das Thema Immissionsschutz einfach vergessen oder vernachlässigt. Das Nachsehen haben neben dem Veranstalter und den Besuchern mitunter der Systemtechniker und der F.o.H. Operator selbst, denn sie beschränken durch ungünstig konfigurierte Beschallungsanlagen ggf. deutlich den Innenpegel auf dem Platz, den Sie während der Show fahren dürfen.

12.4. Cardioid Low-Mids bei Line Arrays

Neben Subwoofern mit Richtwirkung erscheinen neuerdings auch Line Arrays mit Nierencharakteristik in den Low-Mids, vgl. SL-Serie von d&b audiotechnik [22]. Hier arbeiten zusätzliche Chassis im Gehäuse, die nur für die Unterdrückung der rückwärtigen Schallabstrahlung zuständig sind.

Der Low-Mid Bereich ist ansonsten nicht ganz unkritisch. Durch die im Vergleich zur Dimension der Gehäuse großen Wellenlängen werden Low-Mids relativ stark gebeugt und tendieren in Richtung omnidirektionaler Abstrahlung. Die rückseitige Abstrahlung der Low-Mids hat primär Nachteile für den Pegel auf der Bühne sowie auch auf die Raumakustik im Indoor-Bereich. Zusätzlich kommt es in diesem Frequenzbereich aber auch zu einer vergleichsweise großen Immissionsbelastung in Richtung der Rückseite der Bühne. Größere Line Arrays werden durchaus bei Trennfrequenzen von z. B. 60 Hz oder sogar fullrange betrieben, sodass hier wesentliche Bassanteile abgestrahlt werden, die stark um die Zeile gebeugt werden. Anders als bei den Subwoofern, mit denen entsprechende Arrays gebildet werden können, kann hier bei konventionellen Line Array Systemen meist nur wenig wirklich praktikable Abhilfe geschaffen werden. Das neue Konzept der kardioiden Abstrahlung in den Low-Mids bei Line Arrays löst somit verschiedene Probleme gleichzeitig und ist durchaus zukunftsweisend.

12.5. Dezentralisierung

Eine weitere mögliche Minderungsmaßnahme im Bereich der Beschallungstechnik besteht in der Dezentralisierung der Beschallungsanlage. Dabei werden mehrere Delaylautsprecher eingesetzt. Die Verteilung auf mehrere Quellen im Publikumsbereich, die entsprechend mit geringeren Pegeln und zielgerichteter Coverage arbeiten können, kann die Emissionen der Beschallung ins Umfeld deutlich verringern. Jedoch ist eine starke Dezentralisierung nicht für jedes Eventformat passend. Vor allem ist es bei starker Dezentralisierung nicht immer ganz einfach, überall eine stabile Lokalisation in Richtung der Bühne zu gewährleisten. In schalltechnisch kritischen Situationen kann die Dezentralisierung dennoch eine wichtige Minderungsmaßnahme darstellen. Oft lassen sich gut funktionierende Kompromisse finden.

12.6. Active Noise Cancelling

Active Noise Cancelling (ANC) ist bei Kopfhörern, im Automobil und auch im Bereich von Maschinenlärm Stand der Technik und es sind diverse kommerzielle Lösungen verfügbar. Eher im Stadium der Forschung und Entwicklung angesiedelt sind momentan noch die Verfahren rund um das Active Noise Cancelling im Kontext von Veranstaltungen bzw. Veranstaltungstechnik. Beim Active Noise Cancelling werden bei Veranstaltungen zusätzliche Lautsprechersysteme in der Peripherie der Veranstaltung aufgestellt und deren Signale mit passenden Treiberfunktionen gefiltert, sodass hier eine zu den entsprechenden Emissionsanteilen passende aber phaseninvertierte Version des Schalls abgestrahlt wird und sich die beiden Anteile im Idealfall gegeneinander auslöschen. So können die Immissionspegel in den hinter dieser aktiven Peripherie liegenden Zonen reduziert werden.

Solche Ansätze werden beispielsweise aktuell im Rahmen des MONICA Projekts [23] von der Arbeitsgruppe um Agerkvist an der DTU in Dänemark untersucht [24]. Zu erwähnen ist in diesem Kontext auch ein kleines Unternehmen namens rocket science in der Schweiz [25], das derartige Systeme bereits seit einigen Jahren aktiv bei Events einsetzt.

Die Verfahren funktionieren nur bei vergleichsweise tiefen Frequenzen bzw. großen Wellenlängen – wirken also auf den Bassbereich. Entsprechend werden üblicherweise auch nur

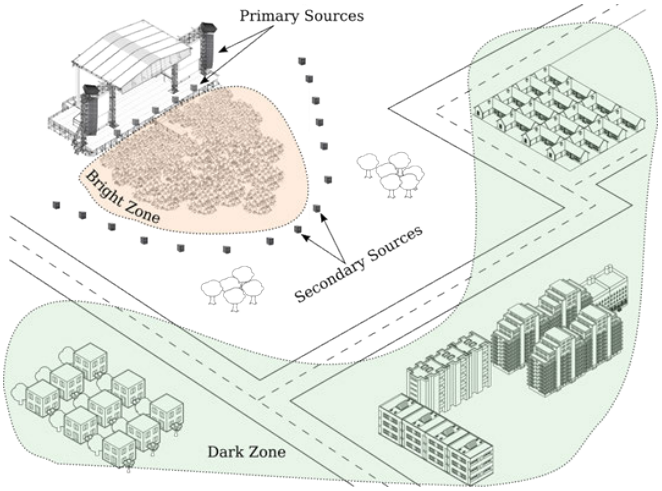


Abb. 16: Illustration von Active Noise Cancelling in Form eines Sound Field Control Systems. Die helle Zone (bright Zone) ist der zu beschallende Publikumsbereich und die dark Zone (dunkle Zone) ist der Bereich der umliegenden Nachbarschaft, in dem der Immissionspegel reduziert werden soll. Die Primary Sources (Primärquellen) sind das PA System des Events. Weiterhin werden Secondary Sources (Sekundärquellen) in der Peripherie der Veranstaltungsfläche aufgestellt, die durch passend phaseninvertierte Schallabstrahlung den entsprechenden Emissionsanteilen entgegenwirken sollen. [Quelle: MONICA Projekt [23]]

Subwoofer in der Peripherie eingesetzt. Der Einsatz von ANC in diesem Kontext erfordert momentan noch vergleichsweise großes Fachwissen und hohen Konfigurationsaufwand. In der Zukunft könnten hier möglicherweise fertige integrierte Module mit Messmikrofon und einem Eingang des F.o.H.-Signals, DSP Einheit und Subwoofer auf dem Markt erscheinen, so wie sie im Bereich des Maschinenlärms bereits verfügbar sind. Somit wäre kaum noch Expertenwissen zur Konfiguration notwendig und die Technologie könnte breitere Anwendung finden.

Bleibt jedoch immer ein grundsätzliches Problem im Kontext von Veranstaltungen: Die Cancelling-Systeme müssen hier außerhalb der zu beschallenden Publikumsfläche aufgestellt werden. Je weiter wir uns aber von der Emissionsquelle entfernen, desto größer ist die Fläche, auf die sich die Schallenergie bereits verteilt hat. Zur flächendeckenden Minderung müssten also relativ viele Subwoofer eingesetzt werden. Ferner ist das Problem dreidimensional. Der Schall wird auch nach oben emittiert und je nach Wetterlage ggf. in einiger Entfernung wieder nach unten gebeugt. Die ideale Anordnung für ein ANC System im Eventbereich wäre also eine umschließende Kuppel voller Subwoofer über dem Veranstaltungslände. Dies könnte jedoch durchaus an die Grenzen des Budgets, der Realisierbarkeit und des guten Geschmacks bei einer Open Air Veranstaltung stoßen.

Dennoch sind die Ansätze äußerst interessant und die Verfahren können für die Minderung besonders kritischer Sektoren oder ggf. bei einzelnen Beschwerdeführern durchaus realistisch eingesetzt werden.

Dieses Thema ist besonders deshalb interessant, da die Verfahren auf den kritischen Bereich der tieffrequenten Immissionsanteile einwirken, der zunehmend zu einem größeren Problemfeld heranwächst.

13. Tieffrequente Immissionen

Eine besondere und in jüngster Zeit an Brisanz gewinnende Herausforderung für Behörden, Sachverständige und Veranstalter sowie auch die Anwohner stellen tieffrequente Immissionen im Kontext von Freizeitlärm bzw. Veranstaltungen dar. Auch unter Einhaltung der Immissionsrichtwerte kann es zu wesentlicher Geräuschbelästigung im tieffrequenten Bereich kommen. Die kulturelle und hörästhetische Entwicklung ging in den letzten Jahren deutlich in Richtung steigender und tieffrequenterer Bassanteile auf Veranstaltungen. Durch die A-Bewertung sinkt die Gewichtung des tieffrequenten Bereichs bei einer Messung stark ab und mitunter bilden A-bewertete Messungen die tatsächliche tieffrequente Belastung am Immissionsort nicht hinreichend ab. Tieffrequente Anteile beugen sich zudem besonders gut um Hindernisse wie Schallschutzwände, verlieren kaum Energie durch Dissipation und werden auch vom umgebenden Baukörper des Anwohners meist nur in vergleichsweise geringem Maße gedämmt, sodass sie besonders stark in den Innenbereich einer schutzbedürftigen Nutzung propagieren.

Tieffrequente Immissionen sind auch im Gewerbe- und Industrielärm durchaus bekannt und in der TA Lärm [4] entsprechend durch Verweis auf die DIN 45680 [5] formal geregelt.

Die direkte Anwendung der DIN 45680 [5] auf den Bereich Freizeitlärm bzw. auf Veranstaltungen ist jedoch in der Praxis äußerst problematisch. Zum einen müssen die entsprechenden Messungen im Innenraum des Anwohners unter Ausschluss von hausinternen Fremdgeräuschen durchgeführt werden, was bei Veranstaltungen keineswegs praktikabel ist.

Zum anderen sind in der aktuell gültigen DIN 45680 [5] keine Seltenen Ereignisse erwähnt. Im Normentwurf der DIN 45680 [6] werden sie sogar explizit ausgeschlossen. Es ist in der Umsetzung nicht sinnvoll, einer Veranstaltung im Falle eines Seltenen Ereignisses einen erhöhten A-bewerteten Immissionsrichtwert zuzusprechen ohne dabei auch gleichartig die Immissionrichtwerte im tieffrequenten Bereich zu erhöhen. Andernfalls müssten die Bassanteile auf einer Veranstaltung im Bezug auf den Gesamtpegel vermindert werden. Der Bass müsste also ggf. deutlich abgesenkt werden. Es steht aber gegen den Charakter einer Veranstaltung das Spektrum der Beschallungsanlage derart zu modifizieren, da der Zuhörer eine ganz bestimmte spektrale Zusammensetzung von der Beschallung erwartet. Es handelt sich hier eben leider nicht um Industrieanlagen im Sinne der TA Lärm [4].

Eine etwas ausführlichere Diskussion des Themas inklusive eines praktiblen Vorschlags zur Messung und Beurteilung tieffrequenter Immissionen im Freizeitlärm unter starker Anlehnung an die Normen findet sich in [26]. Die Anwendung des diskutierten Verfahrens läuft aktuell als Pilotversuch im Raum Köln. Das Verfahren wird ggf. im weiteren Verlauf der Studie noch angepasst und modifiziert und soll im Anschluss ggf. als eine mögliche praxisorientierte Vorgehensweise publiziert werden. Abgesehen von dieser Experimentalstudie besteht hier generell grundlegender Handlungsbedarf, da die aktuelle Situation in der Handhabung sehr problematisch ist.

14. Schlussfolgerungen

Das Thema Schallimmissionsschutz bei Events gewinnt zunehmend an Relevanz, wobei davon auch durchaus nicht nur größere oder große Veranstaltungen betroffen sind. Die regulären Immissionsrichtwerte nach den gängigen Regelwerken können bei den meisten Veranstaltungen nicht eingehalten werden und es müssen entsprechende Ausnahmegenehmigungen seitens des Veranstalters beantragt und seitens der Behörde erteilt werden, damit die Veranstaltung überhaupt durchgeführt werden darf. So wie einerseits ein Veranstalter kein Recht auf die Zuteilung einer Ausnahmegenehmigung hat, so hat andererseits auch der Anwohner keinen Anspruch darauf „nichts“ zu hören. Der Veranstalter muss sich an die allgemeinen Richtlinien bzw. ersatzweise an die Auflagen der Ausnahmegenehmigung halten und der Anwohner hat die in diesem Rahmen auftretenden Geräuschimmissionen hinzunehmen. Die Ortsbehörden wägen für die Erteilung von Ausnahmegenehmigungen jeweils im Einzelfall die Interessenlagen sowie die Zumutbarkeit von Einwirkungen auf den Anwohner ab. Letztendlich geht es oftmals um die Erarbeitung eines allseitigen Kompromisses. Keiner der Beteiligten hat einen Rechtsanspruch auf vollständige Erfüllung seiner Belange. Das letzte Wort hat immer die örtliche Behörde, die auch über die Zukunft einer Veranstaltungsstätte oder eine spezifischen Veranstaltung entscheidet. Im ernsthaften Streitfall entscheiden dann weiterführend die Gerichte über die Situation.

Eine Veranstaltung funktioniert in der Regel nur, wenn sich alle Beteiligten kooperativ verhalten und aufeinander zugehen. Kommunikation im Vorfeld ist der maßgebliche Faktor. Es sollten jeweils frühzeitig enge Abstimmungsgespräche zwischen den Behörden, dem Veranstalter, dem schalltechnischen Sachverständigen und dem Technikdienstleister erfolgen. Vor allem sollte hier stets auf allen Positionen mit offenen Karten gespielt und die Situation sachlich vorgetragen werden. Mögliche Probleme sollten in jedem Falle lieber im Vorfeld ganz konkret angesprochen werden, statt ausgesessen oder verschwiegen zu werden. Der Veranstalter sollte den Anwohnern die Veranstaltung zudem rechtzeitig ankündigen, sodass die Anwohner zumindest über mögliche Einwirkungen informiert sind, entsprechend planen und ggf. ausweichen können. So kann in den meisten Fällen eine sachdienliche, praktikable und friedliche Lösung gefunden werden, die zudem Verbindlichkeit und Rechtssicherheit für den Veranstalter bietet.

Zu bedenken ist immer, dass die Ortsbehörden bei der Erteilung einer Ausnahmegenehmigung die Verantwortung übernehmen und im Streitfall ihre Entscheidung rechtfertigen müssen. Der Veranstalter ist somit entlastet, sofern er die Sachlage für den Abwägungsprozess der Behörden im Vorfeld vollständig und offen erläutert hat und die in der Genehmigung erteilten Auflagen vollständig erfüllt. Unter diesem Aspekt wird deutlich, dass die Behörden auf umfassende Informationen und die Kooperation aller Beteiligten angewiesen sind und entsprechende Auflagen und Anforderungen nicht aus Willkür erlassen.

Insbesondere die Technikdienstleister sollten sich mit diesem sehr interessanten Themenkomplex auseinandersetzen und den Immissionsschutz stärker in ihren Betrachtungshorizont aufnehmen, da die Beschallungsanlagen auf Veranstaltungen zumeist den maßgeblichen Emittenten darstellen und hier teils schon mit sehr einfachen Mitteln erhebliche Verbesserungen erzielt werden können. Eine Optimierung der Beschallungsanlage ist oftmals mit vergleichsweise wenig Budgetaufwand oder gar budgetneutral zu realisieren. Die Vorteile liegen sowohl in einer Verminderung der Immissionsbelastung als auch in einem höheren Innenpegel, der auf der Veranstaltungsfläche erzielt werden kann. Eine entsprechende Planung und Optimierung der Beschallungsanlage kann mitunter sogar klassische Minderungsmaßnahmen wie beispielsweise mobile Lärmschutzwände ersetzen und somit den Kostenaufwand für den Veranstalter senken.

In der täglichen Praxis führen meist kleine Unstimmigkeiten oder mangelnde Kommunikation zu großen Problemen. Sei es ein durch arrogantes Auftreten der Veranstalter verprellter Anwohner, der sich zur Revanche zum äußerst aktiven Beschwerdeführer entwickelt oder eine über die tatsächliche Sachlage nicht frühzeitig, wahrheitsgemäß oder ausreichend informierte Behörde. Oft sind es sogar die Sekundärfaktoren wie Verkehrschaos und falsch parkende oder pöbelnde Besucher, die ihren Müll in die Vorgärten werfen oder Blumen zertreten, die den großen Stein ins Rollen bringen. Die Anwohner kanalisieren ihren Unmut dann ggf. über das gut greifbare Thema Immissionsschutz. Veranstalter sollten also stets auch die Sekundärfaktoren im Blick haben und ggf. entsprechende Anwohnerschutzkonzepte entwickeln, um aufkeimenden Lärmkonflikten vorzubeugen, die eigentlich gar keine originären Lärmkonflikte sind. Ein wesentlicher Faktor ist die Schaffung der Akzeptanz der Veranstaltung in der Bevölkerung. Hier können ein kleiner Artikel in der lokalen Presse, ein paar Freikarten für direkte Anlieger, ein lokaler Künstler im Vorprogramm oder auch ein kurzes nettes Gespräch mit dem interessierten Anwohner am F.o.H. statt einer unfreundlichen Abfuhr wahre Wunder bewirken.

In allen Belangen gilt stets das Gebot der gegenseitigen Rücksichtnahme, das sogar explizit in den Regelwerken wie der TA Lärm [4] verankert ist.

Literatur und Quellverweise

[1] LärmVibrationsArbSchV (2016-11): Verordnung zum Schutz der Beschäftigten vor Gefährdungen durch Lärm und Vibrationen (Lärm- und Vibrations- Arbeitsschutzverordnung)

[2] DIN 15905-5 (2007-11): Veranstaltungstechnik - Tontechnik - Teil 5: Maßnahmen zum Vermeiden einer Gehörgefährdung des Publikums durch hohe Schallemissionen elektroakustischer Beschallungstechnik

[3] BImSchG (2017-07): Gesetz zum Schutz vor schädlichen Umwelteinwirkungen durch Luftverunreinigungen, Geräusche, Erschütterungen und ähnliche Vorgänge (Bundes-Immissionsschutzgesetz - BImSchG)

[4] TA Lärm (2017-09) Sechste Allgemeine Verwaltungsvorschrift zum Bundes-Immissionsschutzgesetz (Technische Anleitung zum Schutz gegen Lärm - TA Lärm)

[5] DIN 45680 (1997-03) Messung und Beurteilung tieffrequenter Geräuschimmissionen, unter Berücksichtigung des zugehörigen Beiblattes 1, gültige Fassung der Norm.

[6] DIN 45680 (2013-09) Messung und Beurteilung tieffrequenter Geräuschimmissionen, Normentwurf.

[7] LImschG (2016-09) Gesetz zum Schutz vor Luftverunreinigungen, Geräuschen und ähnlichen Umwelteinwirkungen (Landes-Immissionsschutzgesetz - LImschG) des Landes NRW

[8] 18. BImSchV (2017-06): Achtzehnte Verordnung zur Durchführung des Bundes-Immissionsschutzgesetzes (Sportanlagenlärmschutzverordnung - 18. BImSchV)

[9] Freizeitlärmrichtlinie LAI (2015-3) der Bund/Länder-Arbeitsgemeinschaft Immissionsschutz

[10] Freizeitlärmrichtlinie NRW (2016-04): Messung, Beurteilung und Verminderung von Geräuschimmissionen bei Freizeitanlagen, Runderlass des Ministeriums für Umwelt und Naturschutz, Landwirtschaft und Verbraucherschutz

[11] DIN ISO 9613-2 (1999-10): Akustik - Dämpfung des Schalls bei der Ausbreitung im Freien - Teil 2: Allgemeines Berechnungsverfahren

[12] MAPANDGIS (Modelling And Prediction of Air Pollution and Noise Data Management with Geographic Information System), Kramer Schalltechnik GmbH. URL: https://www.kramer-schalltechnik.de/cms/de/software/mapandgis/, Stand: 10.12.2018.

[13] Belcher, D. und Shabalina, E. (2018): Schallimmissionsprognosen mit Beschallungsanlagen; Import und komplexe Addition, in: Akustik Journal der Deutschen Gesellschaft für Akustik (DEGA), 01/2018

[14] NoizCalc, d&b audiotechnik GmbH. URL: https://www.dbaudio.com/global/de/produkte/software/noizcalc/, Stand: 10.12.2018.

[15] SoundPLAN noise, SoundPlan GmbH. URL: https://www.soundplan.eu/deutsch/soundplan-laerm/soundplan-noise/, Stand: 10.12.2018.

[16] ArrayCalc, d&b audiotechnik GmbH. URL: https://www.dbaudio.com/global/de/produkte/software/arraycalc/, Stand: 10.12.2018.

[17] VDI 3770 (2012-09): Emissionskennwerte von Schallquellen - Sport- und Freizeitanlagen

[18] DIN 45645-1 (1996-07): Ermittlung von Beurteilungspegeln aus Messungen, Teil 1: Geräuschimmissionen in der Nachbarschaft

[19] DIN EN 61672-1 (2014-07): Elektroakustik - Schallpegelmesser - Teil 1: Anforderungen

[20] METRAO, Event Acoustics. URL: http://metrao.com, Stand: 10.12.2018.

[21] MAPP XT, Meyer Sound Laboratories Inc., USA URL: https://meyersound.com/product/mapp-xt/, Stand: 10.12.2018.

[22] SL-Serie, d&b audiotechnik GmbH. URL: https://www.sl-series.com/global/de/, Stand: 10.12.2018.

[23] MONICA Projekt am Fraunhofer Institute for Applied Information Technology (Projektkoordinator) mit diversen angeschlossenen internationalen Projektpartnern. URL: https://www.monica-project.eu/, Stand: 10.12.2018.

[24] Heuchel, F. M.; Caviedes Nozal, D.; Brunskog, J.; Fernandez Grande, E. und Agerkvist, F. T. (2017): An Adaptive, Data Driven Sound Field Control Strategy for Outdoor Concerts, AES International Conference on Sound Reinforcement - Open Air Venues

[25] rocket science GmbH. URL: http://www.rocket-science.ch/, Stand: 10.12.2018.

[26] Bernschütz, B. und Latz, J. (2018): Leitfaden zur Messung und Beurteilung tieffrequenter Immissionen im Freizeitlärm, Experimentalstudie. URL: http://digdok.bib.thm.de/volltexte/2018/5231/

Planung und Bau eines Tonstudiokomplexes – Neubau der Studios 301 in Sydney

J. Veith[1]

[1] *jv.acoustics, Brunnthal, Germany, Email: info@jv-acoustics.de*

Abstract

Die Studios 301 in Sydney, Australien, sind eine der ältesten Studios der Welt und wurden im Jahr 1926 als Teil der Columbia Graphophone Company gegründet. Die traditionsreichen Studios zählten Künstler wie Bob Dylan, David Bowie, Prince, U2, Snoop Dogg, Metallica, Lady Gaga, Snoop Dog, Lana Del Rey und viele andere zu ihren Kunden. In 2018 wurde ein Neubau der Studios am Standort 3 Ellis Ave Alexandria eröffnet. Mit elf Regie- und Masteringräumen, einem 200m² Aufnahmesaal und zahlreichen Aufnahmekabinen ist es Australiens bekanntestes Studio und zählt zu den größten Musikstudiokomplexen der Welt.

Der folgende Erfahrungsbericht beschreibt die Planung und den Bau der neuen Studio 301 aus Sicht des Akustikers.

1. Einführung

Dieser Text soll eine Übersicht über die verschiedenen Aufgabenstellungen, Zielsetzungen und die Konstruktion der neuen Studios 301 geben. Es ist kein wissenschaftlicher Report, sondern vielmehr ein persönlicher Erfahrungsbericht. Es werden einige interessante Details herausgehoben und mit vielen Bildern soll ein Einblick vermittelt werden.

Im Folgenden wird beschrieben wie es zum Auftrag und zur Zielsetzung des Projektes kam und unter welchen Gesichtspunkten ein geeignetes Gebäude gefunden wurde. Es wird auf Details der bauakustischen und raumakustischen Planung eingegangen. Insbesondere Belange der Schalldämmung und Klimatechnik werden näher beleuchtet. Weiterhin werden raumakustische Messungen der fertiggestellten Räume gezeigt und Besonderheiten aus der Bauphase angesprochen.

Abb. 1: Harbour Bridge und Opera House zum Festival Vivid Sydney 2017.

2. Auftrag, Zielsetzung und Grundlagenermittlung (Teil 1)

Das Projekt begann Anfang 2016 mit dem Anruf von Tom Misner. Er erwarb die Studios 301 im Jahr 1998 und wollte nun, da der alte Studiokomplex geschlossen werden musste, die Studios an einem neuen Standort in bester und modernster Qualität neu aufbauen. Tom Misner hatte bereits sehr genaue Vorstellungen davon, was der Studiokomplex beinhalten sollte. Nach einigen Vorgesprächen und Interviews war der geplante Umfang der Räume definiert. Insgesamt sollten drei Recording beziehungsweise Mixing Räume, drei Mastering Studios, vier Pre-Production Räume, ein großer Aufnahmesaal, sieben Aufnahmekabinen sowie ein Empfangsbereich, Maschinenräume und Bürofläche in der Anlage integriert werden. Für alle Räume wurde der Platzbedarf ermittelt sowie die jeweiligen technischen Anforderungen in einem Whitebook beschrieben.

Nach ersten Kalkulationen war klar, dass dafür insgesamt, also Büro- und Studioflächen, knapp 2000°m² Grundfläche nötig waren. Es wurde ein erster Budgetrahmen festgelegt und die Aufgaben im Projekt klar verteilt. Tom Misner gestaltete das ‘Interior Design‘ der Räume, also Oberflächenmaterialien und Farben quasi im Alleingang. Mit den Planungen von Statik, Elektroanlagen, Klimatechnik und Studiotechnik wurden örtliche Ingenieure beauftragt. Im Verantwortungsbereich von jv.acoustics lag die raumakustische, bauakustische und raumbildende Planung, sowie die Beratung für die Klimatechnik. Das umfasste natürlich auch die Detailplanungen zu den einzelnen Gewerken. Für die Durchführung vor Ort zeichnete Herr Paul Oaks als Bauleiter verantwortlich.

ISBN
978-3-9812830-9-9

Der Autor hält die Urheberrechte für den gesamten Inhalt seines Manuskripts. Er überlässt dem VDT das Recht auf Veröffentlichung in gedruckter Form, online oder als PDF-Download von https://tonmeistertagung.com

Abb. 2: Leere Halle vor Beginn der Bauarbeiten.

3. Grundlagenermittlung (Teil 2)

Nach der Klärung des Bedarfs wurde eine für das Vorhaben passende Halle gesucht. Die Anforderungen an ein entsprechend geeignetes Gebäude sind vielseitig. Priorität wurde auf ausreichende Grundfläche und lichte Raumhöhe gesetzt. Die statischen Voraussetzungen müssen den Einbau der Studios gewährleisten oder einen entsprechenden Umbau zulassen. Eine weitere Rolle bei der Auswahl spielen die Beschaffenheit von Dach und Wänden sowie technische Gegebenheiten wie hier zum Beispiel die wichtige Frage der Dachentwässerung. Wenn alle diese Faktoren berücksichtigt sind, wird die Umgebung des Gebäudes betrachtet: kann es zu Beeinträchtigungen durch Immission oder Emission kommen? Ist eine gute Erreichbarkeit, sowohl im Bau, wie auch im Betrieb, gegeben?

In einem dritten Anlauf konnte eine geeignete Halle mit 1600m² Grundfläche für Studios und einem zusätzlichen Bürogebäudeteil gefunden werden. Die mittelbare Nähe zum Flughafen von Sydney machten Untersuchungen zum Fluglärm nötig. Wie in Abbildung 3 zu sehen, befindet sich die Halle in der Nähe von Start- und Landebahn.

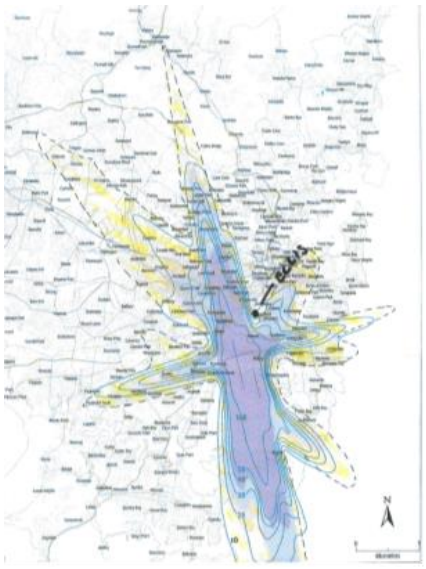


Abb. 3: Karte zur Verteilung des Fluglärms in Sydney. Der schwarze Punkt markiert die günstig gelegene Position des neuen Studio 301.

Ein wichtiger Fakt war für uns, dass der Hallenboden für eine Nutzlast von mehreren Tonnen pro Quadratmeter ausgelegt war.

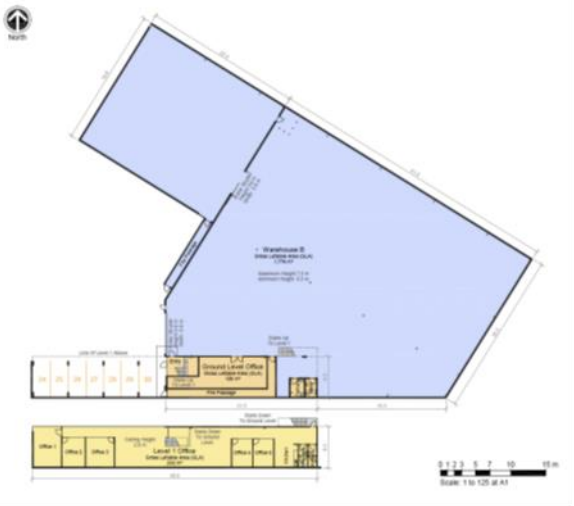


Abb. 4: Der Grundriss der Halle mit einer Grundfläche von etwa 1600m² + Büroeinheit.

4. Grundriss und Visualisierung

Nachdem die Voruntersuchungen abgeschlossen, sowie die Anforderungen und Zielsetzungen geklärt waren, konnte mit der Planung des Grundrisses begonnen werden. Trotz der Größe der Halle konnten die zahlreichen Räume nur untergebracht werden, weil die Büros in einem anderen Gebäudeteil Platz fanden. Es wurden mehrere Vorschläge erarbeitet und nach einigen Entwicklungsstufen wurde sich für die finale Ausführungsvariante entschieden, die in Abbildung 5 dargestellt ist. Im weiteren Verlauf des Projekts wurden virtuelle 3D Darstellungen fast aller Räume erstellt. In dieser Phase kam die Idee von Tom Misner, Jochen Veith‘s Photographien des Lichtinstallations-Festivals Vivid Sydney mit in das Design der Räume einzubinden. Die Visualisierungen sind in den Abbildungen 6 bis 8 zu sehen.

Abb. 5: Visualisierung der geplanten Studioräume im Grundriss.

Abb. 6 - 8: 3D Visualisierungen von Studio 1, Aufnahmesaal und Empfangsbereich. Im Studio 1 wurden Bilder des Festivals Vivid Sydney in das Design integriert.

5. Schalldämmung

Neben der Raumakustik spielt im Studiobau die Schalldämmung eine entscheidende Rolle. Es sollen möglichst alle Schallemissionen und -immissionen aus und in die Studios soweit unterdrückt werden, dass ein ungestörtes Arbeiten ermöglich ist. Hierbei spielen drei Faktoren eine grundsätzliche Rolle: die mechanische Entkopplung der Bauteile, die Masse der trennenden Bauteile und der Abstand zwischen einzelnen Schalen. Da es die hohe Nutzlast der Halle zuließ, haben wir uns dazu entschieden die einzelnen Räume baulich getrennt in Massivbauweise zu errichten. Diese Bauweise reduzierte die Kosten bei gleichzeitiger hoher Dämmung speziell im tieffrequenten Bereich und bietet Erleichterungen in der Statik. Zuerst wurden Ringanker gegossen, die mittels Bahnen aus Elastomeren vom Boden der Halle abgekoppelt sind. Die Abstimmung der Entkopplung wurde auf fres ≤ 15Hz ausgelegt, um eine möglichst gute Körperschallentkopplung zu gewährleisten. Innerhalb der Ringanker wurden schwimmende Estriche eingebaut, die nach unten zum Fundament und ebenso seitlich zum Ringanker entkoppelt sind. Die Lagerung erfolgte wieder auf Elastomerstreifen. Hier ist die Abstimmfrequenz fres ≤ 18Hz.

Die Wände wurden auf den Ringankern aus zum Teil zwei Reihen Betonhohlblocksteinen gemauert. Diese wurden mit Beton verfüllt, um die für die Schalldämmung benötigte Masse zu erreichen. Wir reden hier über einen Bereich von bis zu 900kg pro Quadratmeter Wandfläche, je nach Wandtyp. Auf die Wände aufgelegt wurden die Stahlträger für die Betondeckenkonstruktion. Diese müssen nicht nur die Last der Decken tragen können, sondern auch zusätzlich die der Klimakanäle und der raumakustischen Deckeneinbauten. Die Rohdecken sind zweischalig aufgebaut. Die obere Deckenlage ist eine 12,5cm Stahlbetondecke und die untere Decke eine 3-lagige Trockenbaudecke. Auch diese wurde mechanisch über Elastomerhänger entkoppelt. Nur so können die hohen Anforderungen an die Schalldämmung erreicht werden.

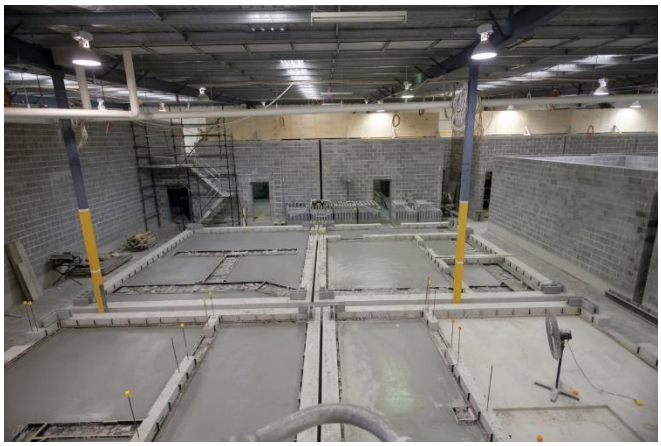


Abb. 9: In diesem Bild sind die verschiedenen Stadien des Ausbaus zu erkennen. Vorne im Bild sind die elastisch gelagerten Ringanker und Bodenplatten zu sehen, an beiden Seiten und hinten die bereits aufgemauerten freistehenden Räume.

Zu jedem Raum gehören auch Fenster und Türen. Diese Öffnungen in der Primärschale bieten einen Angriffspunkt für Schallemissionen und müssen besonders beachtet werden. Für typische Regiefensteranlagen gibt es vier die Schalldämmung bestimmende Größen: Koinzidenz-Effekte, Koppelresonanzen zwischen zwei Scheiben, Stehwellen zwischen den Scheiben und Kopplung über den Rahmen. Der Koinzidenz-Effekt ist ein schmalbandiger Einbruch in der Schalldämmung einer Scheibe um durchaus über 10dB. Er entsteht, wenn die projizierte Wellenlänge eines schräg auf die Scheibe einfallenden Schalls bestimmter Frequenz mit der Biegewellenlänge dieser Frequenz in der Scheibe übereinstimmt. Um der Schwächung durch diesen Effekt

entgegenzuwirken, wird eine Glasscheibe biegeweicher gemacht, indem sie als Verbundscheibe mit einem Schallschutzfolienkern gefertigt wird. Somit verschiebt sich die Koinzidenzfrequenz zu höheren Frequenzen hin. Da in einem Studiofenster immer zwei Scheiben verwendet werden, wird auch immer darauf geachtet, dass die zwei Scheiben jeweils eine unterschiedliche Koinzidenzfrequenz aufweisen. Die Schwächungen durch Stehwellen zwischen den Scheiben werden minimiert, indem die Scheiben schräg zueinander eingebaut werden. Durch einen möglichst großen Scheibenabstand zweier Scheiben rutscht die Koppelresonanz des Masse-Feder-Masse-Systems möglichst tief. Daraus resultiert dann wiederum eine entsprechende hohe Schalldämmung des gesamten Fenstersystems. Als viertes muss eine mögliche Kopplung über den Rahmen unbedingt vermieden werden. Das erreichen wir, indem wir zwei komplett getrennte und berührungsfreie Rahmen einsetzen. Etwas zusätzlicher Dämmstoff im Rahmenbereich vergrößert die Verluste im Scheibenzwischenraum. Mit einer solchen zweischaligen Konstruktion lassen sich unter einem vernünftigen Kosten-Nutzen-Aspekt Bau-Schalldämm-Maße im Bereich von 65dB im eingebauten Zustand erreichen.

Nur am Rande erwähnt sei, dass nicht nur in den Räumen auf eine mechanische Entkopplung der Bauteile geachtet wurde, auch die Böden und Decken der verbindenden Flure mussten entkoppelt ausgeführt werden, um keine Schallbrücken zwischen den Studioräumen zu erzeugen.

6. Klimatechnik

Um in den Studioräumen über lange Zeiträume entspannt arbeiten zu können, ist eine ausreichend dimensionierte Klimatisierung der Räume zwingend notwendig. Gleichzeitig sollen keine Geräusche durch Strömungsrauschen oder Ventilatorgeräusche im Studioraum wahrnehmbar sein. Auch darf die Schalldämmung zwischen den Studioräumen nicht durch die Klimainstallationen (z.B. Luftschächte) geschwächt werden.

Die elektronischen Geräte und Personen im Studio produzieren zum Teil eine erhebliche Wärmelast. Für ein großes Mischpult, wie sie in den Regieräumen 1 und 2 installiert sind, muss eine Wärmelast von 4 bis 6 Kilowatt angenommen werden. Hinzu kommen die Geräte-Racks, Licht, Personen und teilweise auch der Wärmeintrag durch Außenfenster. Zusätzlich zu den Anforderungen im Studio ist es notwendig die klimatischen Bedingungen der jeweiligen Region mit zu Bedenken. In Sydney zum Beispiel kann es zum Teil sehr heiß und feucht werden.

Die Verdichter für die Kältegeräte wurden auf dem Hallendach montiert. In dem Zwischenraum zwischen Hallendach und Studiodecken befinden sich Umluftkühlgeräte, Entfeuchter, Wärmetauscher sowie große Lüftungskanäle mit den Hauptschalldämpfern. Alle aktiven Geräte sind entkoppelt an der Dachkonstruktion aufgehängt. Die Kanäle und Schalldämpfer lasten auf den Studiodecken und sind von den Geräten entkoppelt.

Von den Studiodecken wird die Luft über Kanäle in den Zwischendeckenraum (also zwischen Betondecke und Trockenbaudecke) geführt. Dort befinden sich Verteiler und weitere Schalldämpfer. Erst danach durchdringen die Lüftungsschächte die Zwischendecke in den jeweiligen Studioraum hinein. Nur durch diese aufwendige Konstruktion konnte die Schalldämmung bei den Durchbrüchen aufrechterhalten und das geforderte Ruhegeräusch eingehalten werden.

Zusätzlich erfolgt natürlich eine Fischluft und Fortluft-Injektion in das jeweilige System im Hohlraum über den Studiodecken. Aufgrund der klimatischen Bedingungen musste die Frischluft vorkonditioniert, also gekühlt und entfeuchtet werden.

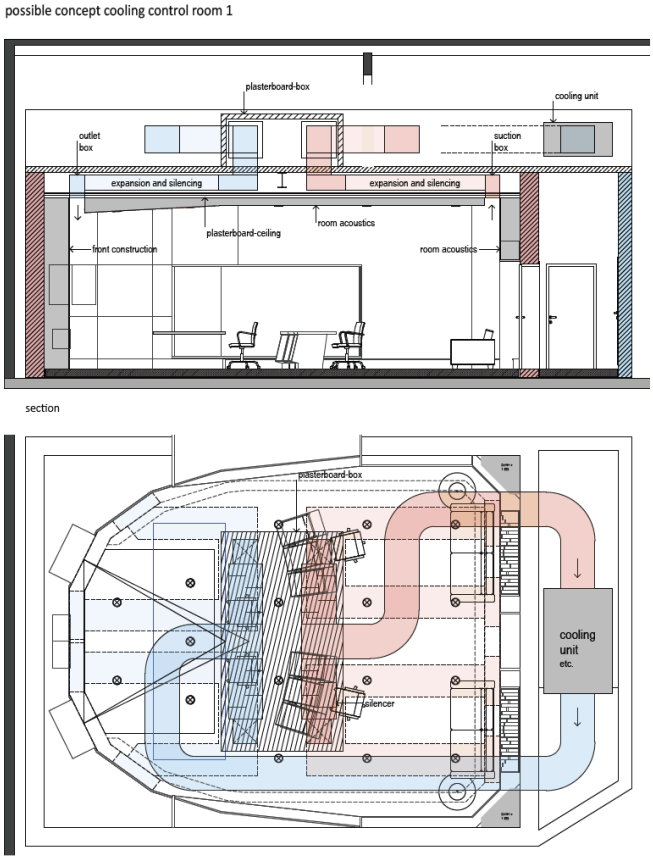


Abb. 10: Klimakonzept der Regie 1

7. Raumakustik

Neben der Schalldämmung und Schwingungsisolierung spielt natürlich die Raumakustik in den Studios die größte Rolle. Auch hier wurden von Seiten des Kunden die höchsten Anforderungen gestellt. In den Tonregien Eins, Zwei und Drei wurden die Lautsprecher in der Frontwand eingebaut, wobei Regie 3 als 5.1 Regie ausgelegt wurde. Die Mastering Studios sind in Stereo ausgelegt, mit den Lautsprechern in Freifeld-Aufstellung. Alle weiteren kleinen Tonregien arbeiten ebenfalls mit frei aufgestellten Lautsprechern. Im Folgenden möchte ich die raumakustischen Parameter der Räume, deren Simulation und die Gestaltungsparameter beschreiben.

Beispielhaft wird auf die Räume Regie 1, Mastering 3 und den großen Aufnahmeraum eingegangen. Alle gezeigten Messungen wurden mit dem Easera Messsystem durchgeführt.

Die wichtigsten raumakustischen Parameter in den Regieräumen sind neben der frequenzabhängigen Nachhallzeit und dem Reflexionsverhalten (Unterdrückung der frühen Reflexionen) das Verhalten bei tiefen Frequenzen und die Übertragungsfunktion zwischen Lautsprecher und Hörposition. Im Prognoseprozess kommen entsprechend drei Modelle zum Einsatz: die geometrische Raumakustik (Strahlenmodell), die statistische Raumakustik (z.B. Nachhallzeitberechnung) und der wellentheoretische Ansatz (z.B. Modenberechnung bei tiefen Frequenzen).

Mit dem Modell der geometrischen Akustik wird im Besonderen darauf geachtet, dass die ersten Reflexionen im Abhörbereich unterdrückt werden. Dies verbessert die Ortung, Klangverfärbung und vor allem die Raumerkennung der Aufnahme. Die geplanten Nachhallzeiten richten sich nach den bekannten Empfehlungen der ITU und EBU Richtlinien. Wobei wir bei solchen kurzen Nachhallzeiten gerne von Quasi-Nachhallzeiten sprechen. Im Bereich oberhalb der Schröder Frequenz erfolgt die Bedämpfung durch Dämmstoffe mit verschiedenen Strömungswiderständen und Dicken. In Australien wird interessanterweise kaum noch mit Mineralwolle, sondern meist mit Polyesterwolle gearbeitet.

Besondere Aufmerksamkeit wurde auf die Bedämpfung im tieffrequenten Bereich unter 200Hz gelegt. Hierfür wurden Schwerfolienschwinger und Helmholtzresonatoren genutzt. Die Helmholtzresonatoren wurden nach unseren Vorgaben gefertigt und dann im jeweiligen Raum später auf die entsprechende Frequenz gestimmt und bedämpft. Eine Abstimmung im Vorfeld ist zwar rechnerisch möglich, allerdings liefert eine genauere Abstimmung an der genutzten Position im Raum bessere Ergebnisse. Zusätzlich wurden an den Rückwänden von allen Regie- und Mastering- Räumen sogenannte Helmfusoren eingesetzt. Diese Kombination aus Diffusoren und Helmholtzresonatoren sind eine Eigenentwicklung. Diffusoren mit dieser Strukturtiefe arbeiten am besten in einem Frequenzbereich von ca. 900-5000Hz. Gleichzeitig bietet dieses Bauteil dringend benötigte Absorption bei tiefen Frequenzen unter 100Hz.

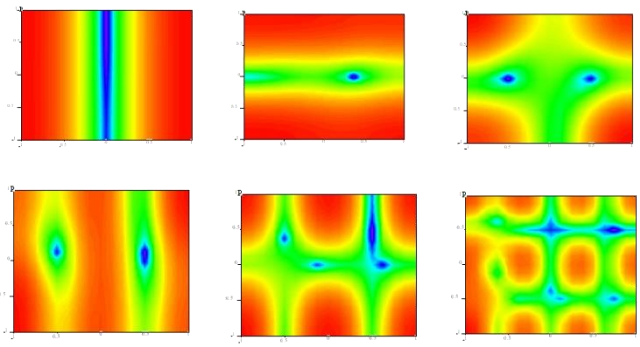


Abb. 11: Darstellung 2-dimensionaler Moden. Druckzonen werden in Rot dargestellt, Schnellezonen in Blau.

Bei akustisch gesehen kleinen Räumen, wie den Tonregien, wird der Schwerpunkt der Untersuchungen bei tiefen Frequenzen und somit beim wellentheoretischen Ansatz liegen. Hier geht es vor allem um die Übertragungsfunktion zwischen Lautsprecher und Hörplatz. Zur Prognose des tieffrequenten Verhaltens wurden Simulationen zur Raumgeometrie durchgeführt, um die Ergebnisse zu optimieren. Wir betrachten Volumen, Verhältnisse der Raumdimensionen, Lautsprecherpositionen und Hörbereich, um eine ausgeglichene Übertragungsfunktion im unteren Frequenzbereich zu erreichen. Als Berechnungsverfahren wurde für den wellentheoretischer Ansatz die analytische Modensynthese gewählt um die Übertragungsfunktionen im Tieftonbereich zu berechnen. Dies ist auch mit FEM/BEM Simulationen möglich, allerdings sind diese noch zu rechenaufwendig, um damit gezielt nach den besten Verhältnissen zu suchen und somit die Übertragungsfunktionen zu optimieren. Es sind meist einfach zu viele Freiheitsgrade.

Es ist bekannt, dass Raummoden im tieffrequenten Bereich bis etwa 200Hz die Übertragungsfunktion Lautsprecher zu Hörposition dominieren. Im Frequenzbereich darüber ist die Modendichte so hoch und die Wellenlängen entsprechend klein, dass sich kaum noch negative Auswirkungen zeigen. Bei jeder einzelnen Raummode bildet sich im Raum ein eigenes, spezifisches Muster aus Druck- und Schnelle-Zonen aus. Da unsere Ohren Druckempfänger sind, können wir die Frequenz der jeweiligen Mode nur in den Druckzonen wahrnehmen. Ähnlich können Drucksender, welche fast alle Lautsprecher sind, nur in Druckzonen Energie in das System (Raum),einkoppeln'. Somit hängt die Übertragungsfunktion sehr stark von der Raumgeometrie, der Lautsprecherposition und der Hörposition ab. Dies gilt es zu untersuchen.

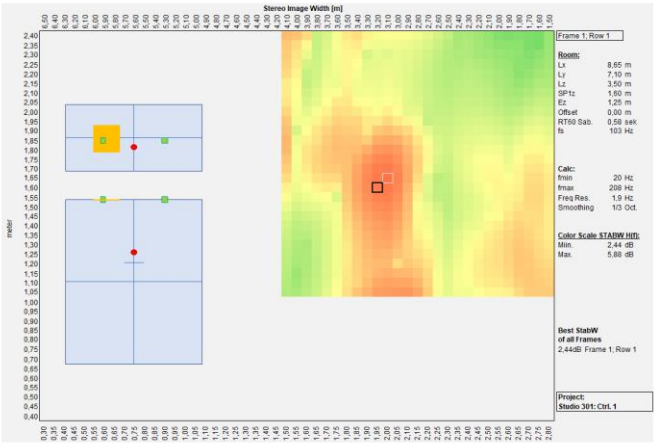


Abb. 12: Berechnung möglicher Lautsprecherpositionen an der Frontwand. Rote Bereiche bezeichnen günstige Positionen.

In Abbildung 12 ist ein Beispiel einer Untersuchung zur Positionierung von Lautsprechern in einer Frontwand zu sehen. In der linken Hälfte der Abbildung sind die Frontwand und der Grundriss eines rechteckigen Raumes zu erkennen. Der orange Bereich kennzeichnet den Bereich möglicher

Aufstellungsorte des linken Lautsprechers. Dieser ergibt sich aus einer jeweils sinnvollen minimalen und maximalen Basisbreite, sowie einer sinnvollen möglichen Aufstellungshöhe. Der rote Punkt und die grünen Rechtecke stellen die Lautsprecher und die zugehörige Abhörposition dar. In der rechten Hälfte der Abbildung sind die Bewertungen der jeweiligen Frequenzgänge an den verschiedenen Positionen über eine Farbskala dargestellt. Diese Grafik entspricht der orangen Fläche in der Raumdarstellung. Hier bezeichnen rote Bereiche günstige Aufstellpositionen (mit entsprechendem Hörplatz) mit möglichst ausgeglichenen Frequenzgängen, und grüne entsprechend ungünstige. Die Raumgeometrie entspricht hier recht gut dem akustischen Verhalten der Tonregie 1. Die weiße Markierung bezeichnet die ideale Position, die schwarze die tatsächlich genutzte. Bereits einige 10cm Veränderung in der Position der Lautsprecher würde massive Veränderungen im Frequenzgang auf der Abhörposition nach sich ziehen.

Abb. 13: Tonregie 1 im Studio 301 nach der Fertigstellung.

Als Ergebnis und Beleg ist in Abbildung 14 die nicht entzerrte Übertragungsfunktion des linken und rechten Lautsprechers in Regie 1, gemessen auf der Abhörposition, zu sehen.

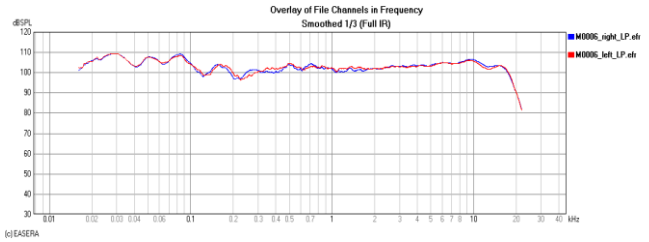


Abb. 14: Übertragungsfunktion im Regieraum 1 auf Abhörposition ohne Entzerrung.

Aufgrund der Geometrie, der raumakustischen Maßnahmen, aber auch auf Grund der Größe der Regie 1 ergibt sich im der gesamten Regie ein recht ausgeglichenes Verhalten der Übertragungsfunktionen über einen großen Bereich. In Abbildung 15 sind die gemessenen Kurven auf der Abhörposition, 40cm dahinter und 1m dahinter jeweils vor einer Entzerrung zu sehen.

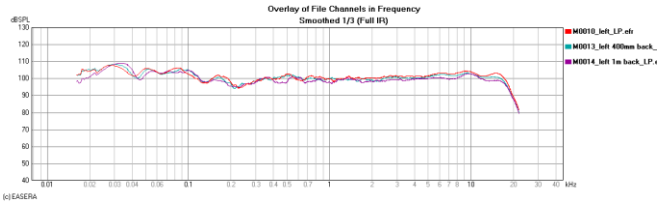


Abb. 15: Ü-Funktion auf Abhörposition – ohne Entzerrung; Auf Hörposition, 400m zurück, 1000mm zurück

In Abbildung 16 ist die Kurvenschar von mehreren Messungen auf und um die Abhörposition herum nach einer leichten Entzerrung dargestellt (40cm rechts, 40cm links, 40cm zurück).

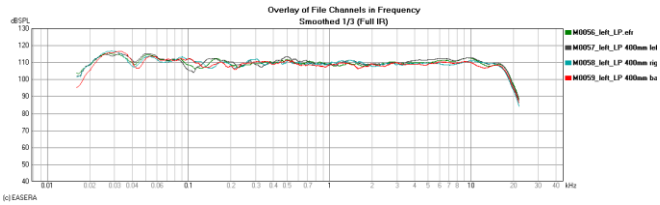


Abb. 16: Ü-Funktion auf Abhörposition – mit Entzerrung; Auf Hörposition, 400mm links, 400mm rechts, 400mm zurück.

Die gemessene Nachhallzeit ist in Abbildung 17 zu sehen. Nachhallzeiten unter 100Hz lassen sich vor allem auf Grund der Modendominanz in diesen Räumen nicht mehr zuverlässig messen und darstellen.

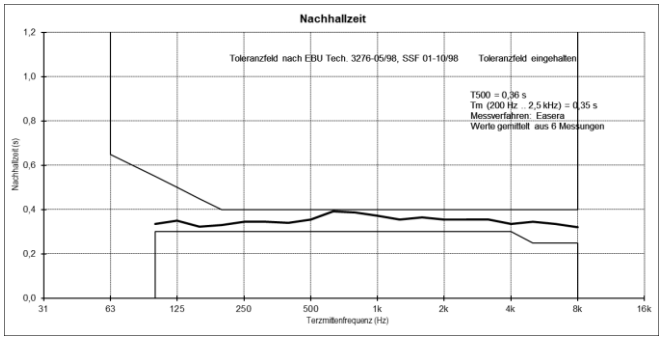


Abb. 17: Gemessene Nachhallzeit der Regie 1.

Abbildung 18 zeigt das gemessene Energie-Zeitdiagramm ETC (Energy-Time-Curve). Es sind keine dominanten frühen Reflexionen nach dem Direktsignal zu erkennen. Die RFZ (Reflexion Free Zone) funktioniert.

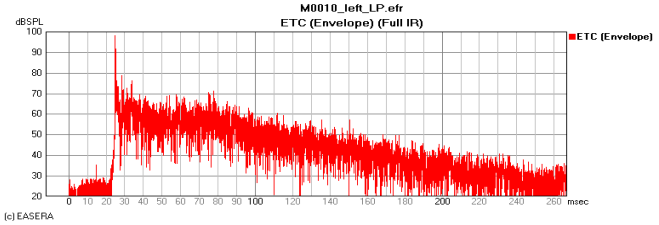


Abb. 18: ETC (Energie-Zeit-Kurve) auf der Abhörposition des linken Lautsprechers.

In den Abbildungen 19 und 20 sind jeweils ein Wasserfalldiagramm vor, beziehungsweise nach Abstimmen der Resonatoren in den hinteren Kanten zu sehen.

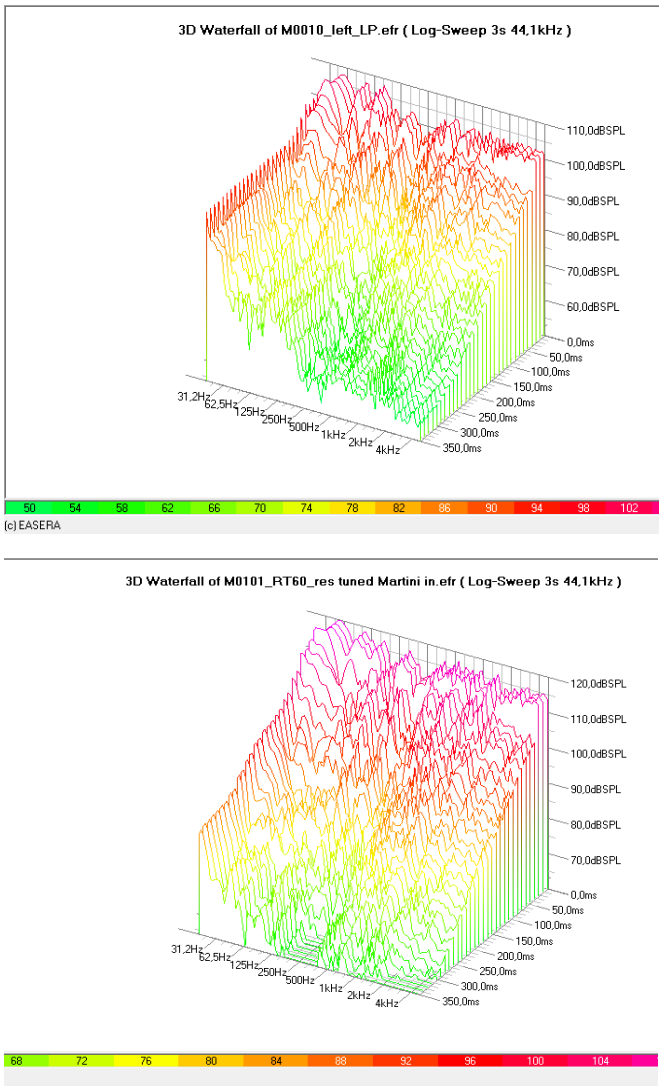


Abb. 19 und 20: Wasserfalldiagramm vor und nach dem Abstimmen der Resonatoren.

Bei den Mastering-Studios handelt es sich, wie gesagt, um 3 identische Räume mit frei aufgestellten Lautsprechern. Generell lässt sich sagen, dass es mit frei aufgestellten Lautsprechern schwieriger ist eine ausgeglichene Übertragungsfunktion im tieffrequenten Bereich auf der Abhörposition zu erreichen. Im Gegenzug jedoch kann bei falsch eingebauten Lautsprechern keine Korrektur mehr vorgenommen werden. In den Abbildungen 22 bis 25 wird ein Ausschnitt aus den Messergebnissen gezeigt.

Abb. 21: Mastering Studio 3.

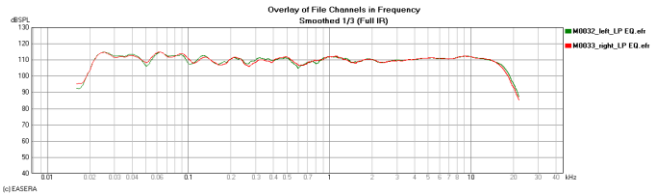


Abb. 22: Mastering Studio 3: Übertragungsfunktionen des linken und rechten Lautsprechers, gemessen auf der Abhörposition, leicht entzerrt.

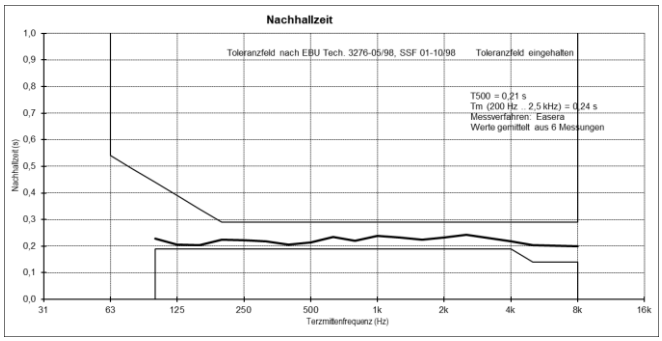


Abb. 23: Messung der Nachhallzeit des Mastering Studios 3.

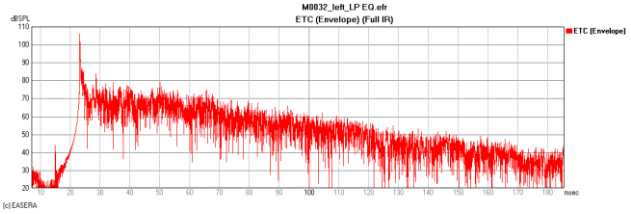


Abb. 24: ETC (Energie-Zeit-Kurve) auf der Abhörposition des linken Lautsprechers.

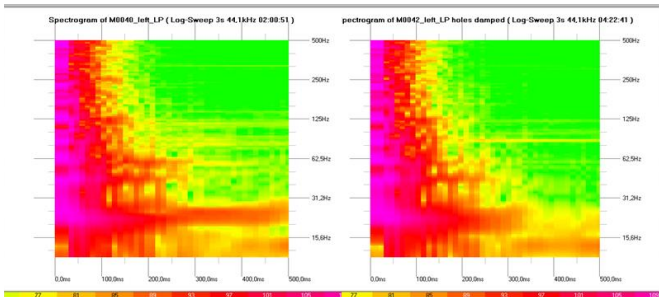


Abb. 25: Spectrogramme vor und nach dem Stimmen der Helmholtzresonatoren.

Abb. 26: Aufnahmesaal des Studio 1.

Als Letztes soll noch ein kurzer Einblick in den großen Aufnahmeraum 1 gegeben werden. Er ist direkt über eine große Studioscheibe an Regie 1 angekoppelt. Des Weiteren sind 2 Aufnahmekabinen mit einem Fenster, beziehungsweise einer zweiflügligen Tür mit großer Glasfüllung angeschlossen. Sein Zweck sollte multifunktional sein. Das heißt, in ihm sollen Aufnahmen einzelner Musiker (z.B. Popmusik), kleine Gruppen (z.B. Jazz) oder kleine bis mittlere Orchester aufgenommen werden. Das akustisch wirksame Volumen ist ca. 660m³ und die akustisch wirksame Höhe 4,3m.

Bei der Konzeptionierung wurde Wert darauf gelegt, dass keine dominanten frühen Reflexionen auftreten, gleichzeitig sollten sich Musiker untereinander gut hören können. Auch war uns eine gute Bedämpfung der Tiefen wichtig und der Frequenzbereich von ca. 250-350Hz sollte in der Nachhallzeit eine leichte Senke aufweisen, um unseren Klangvorstellungen zu entsprechen. Eine hohe Diffusität ist eine der Grundvoraussetzungen für einen gut funktionierenden Aufnahmeraum.

Wir haben uns für ein modulares Design entschlossen. An den Seitenwänden sind Schlitzplattenresonatoren zu erkennen, um die gewünschte Senke im Bereich 250-350Hz zu erreichen. Des Weiteren bestehen die Elemente an den Wänden aus reflektiven Teilen, Tiefenabsorbern und Breitbandabsorbern. Leider wurden die Deckensegel nicht nach unseren Plänen verwirklicht, sondern aus Zeitgründen verändert und stark vereinfacht. Messungen (mit umlaufendem Polstergestühl) zeigten daher, dass das Nachallzeit-Niveau niedriger als unser angestrebtes Niveau war. Aus diesem Grunde wurde es notwendig hier noch einmal nachzubessern. Die nun vorhandenen Segel wurden leicht im Winkel verändert, höher gehängt und teilweise Dämmstoff entfernt. Neuere Messungen liegen uns zum jetzigen Zeitpunkt leider noch nicht vor.

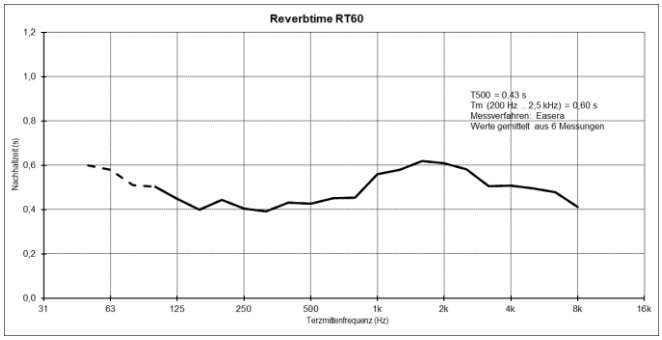


Abb. 27: Nachhallzeitmessung im großen Aufnahmeraum mit umlaufenden Polstergestühl vor der Nachbesserung der Segel.

8. Zusammenfassung

In diesem Erfahrungsbericht wurde ein Einblick in die großen Herausforderungen bei Planung und Neubau der Studios 301 in Sydney gegeben. Es wurde der Projektablauf zur Fertigstellung eines der größten Musikstudiokomplexe weltweit skizziert. Nach Auftrag und Zielsetzung des Projektes wurde auf die Standortsuche und Vorplanung eingegangen. Weiterhin wurden Details zu Klimatechnik, Schalldämmung und Raumakustik behandelt.

Von den ersten Gesprächen bis zur Inbetriebnahme sind knapp 2 Jahre vergangen. Das Gesamtbudget belief sich auf fast 12 Millionen Australische Dollar. Es ist nicht sehr häufig, dass Musikstudioprojekte dieser Größe entstehen und ich empfinde es daher immer als Privileg an solchen Vorhaben beteiligt sein zu dürfen. Daher möchte ich mich an dieser Stelle auch für das große Vertrauen und die angenehme und kreative Zusammenarbeit bei dem Bauherrn, dem Team vor Ort und auch meinem Team bedanken.

Abb. 28: Impression vom Festival Vivid Sydney 2017.

Abb. 29: Ja, die Bodenplatte hält das Gewicht aus

Abb. 30: Klimatechnik zwischen Studios und Hallendecke

Abb. 31: Einbau einer Studioscheibe

Abb. 32: Mastering 2

Abb. 33: Studio 3

Abb. 34: Studio 1

Abb. 35: Aufnahmesaal des Studio 1

The Duett – the first official audio drama in Dolby Atmos

Markus Rebholz, Volker Armbruster

Tonmeister, Boomoperator & Sounddesigner Email: markus.rebholz@tonmeister.de

Geräuschfabrik Audio Post-Produktion für Film und Medien Email: volker@geraeuschfabrik.de

Abstract

„Das Duett“ is officially the first audio drama in Dolby Atmos. It happened to be a students’ experiment and offers a whole new Dolby Atmos experience for cinema. The story was specifically designed for the immersive audio format Dolby Atmos to bring the experience to another level of sound. Volker Armbruster and Markus Rebholz give detailed insights about the making-of and invite everyone to dive into the immersive world of “Das Duett”. It is alternative content for the cinema and a thrilling piece of applied research, a re-birth of the audio drama.

1. Introduction

As student of film sound and sound design at Filmakademie Baden-Württemberg one creates a downright flood of sounds for Film and TV, Shorts, Commercials, Feature Films, Documentaries, Soaps etc.

More often than not the current work load is so heavy one can forget to actually study: What worked out well, what didn’t? What has to be done differently, where are weaknesses, where strengths.

At one point during their studies both students Markus and Volker were absolutely certain: The time to experiment is now. They were not satisfied with having hands-on film experience – they wanted to explore, research, try out, test, play, fail and broaden their horizon with their experiences in other words: Create something that later (in the real world) no one would pay for, no production would care for, no one would stand for.

“How about turning the known around and not create sound for picture? Let’s make picture for sound.

- The idea was born.

2. The story

The story evolved when Markus stumbled across some interesting facts about a mystery theft back in the 1970s in Stuttgart. He constructed a short story around them and built a so called faction story – a story that combines facts and fiction.

During the process writing both Markus and Volker agreed on writing the piece specifically for a Dolby Atmos – an immersive cinema audio format. So every character, every scene had “3D-Sound” in the back of the author’s head.

Synopsis:

“You know how it is! You had to get rid of some witnesses”

- “Wrong! I WAS the witness!”

- “Alright, alright, alright,...well, then YOU try again. Maybe today you’re playing better than me.”

- “Everything began…randomly…in a streetcar…”

The case had not been solved for 40 years: Theft of rare coins in Stuttgart City. The intermezzo of defendant Victor Bethke and culprit Heinrich Dasler was everything but random – that’s at least what inspector Hartmann believes. Why else only the two of them survived the streetcar accident on that Saturday of May 1974?

On the other hand Bethke is wondering how he ended up a suspect. If he wasn’t the accomplice – who was? Hartmann…?

An interrogation, a storm and an exploding historical Stuttgart streetcar build a faction-story full of action. A dance searching for the truth.

Alright then: Da capo!

3. Field recording

The project is a student project without budget. Luckily one day the two sound engineers had the chance to get a clean recording of the historical streetcar they needed. A Schoeps ORTF Surround takes care of the Surround bed in Atmos. ORTF is ideal because of a homogenous stereo spread and combines the advantages of time-of-arrival stereophony and intensity stereophony. [1]

For the height layer the IRT-cross is being chosen since both ORTF Surround and IRT-cross are very similar in terms of sound characteristic.[2]

The IRT-cross is built of 4 DPA 4011 which in the required frequency range (150Hz-10kHz) sound very similar to the Schoeps MK4 capsules, although their noise floor is a little bit higher.[3]

The IRT-cross is ideal for ambience recordings. It has a good spatial representation of the real acoustic environment. [4].

The microphones are tilted upwards in order to receive more of the sound from above. Vertical and horizontal recording angles are adjusted so they are balanced out to each other. The IRT cross ends up being mounted about a foot above the ORTF Surround.

A Sennheiser 418 MS microphone as well as a MKH 40 are used on extra booms for all other sound effects that the old

ISBN 978-3-9812830-9-9

The authors hold the copyrights of all of content in this manuscript. They have granted the VDT the right to publish it in print, online and as PDF download on *https://tonmeistertagung.com*

Fig. 1: ORTF Surround in streetcar

Fig. 1: IRT-cross for height layer recordings

Fig. 3: Extra SFX of the train

4. Studio Recordings

Unfortunately, the three main characters must be x-ed, but an elegant dialogue edit brings back the authenticity that is developed when they "act" together in the studio. The Neumann U87 is the main microphone for all voice recordings.

Additional foleys that were not recorded inside the streetcar wagon are now added. But also important sounds for the "typical" sound design that can't be found in available sound libraries at the Filmakademie have to be thought of and recorded. Depending on the sound a Schoeps MK4 (mainly foleys), the Neumann U87 (bass phattener) and a Microtech Gefell M 930 (room microphone, very low self-noise, hi-boost at 6kHz) are mixed most of the time.
All sounds are recorded at 192kHz to have the chance to pitch them down at full quality. A lot of sounds have to be recorded in order to "fill the room" a lot of coverage is needed.

Fig.4: SFX recordings studio

Fig. 5: Foley recordings studio

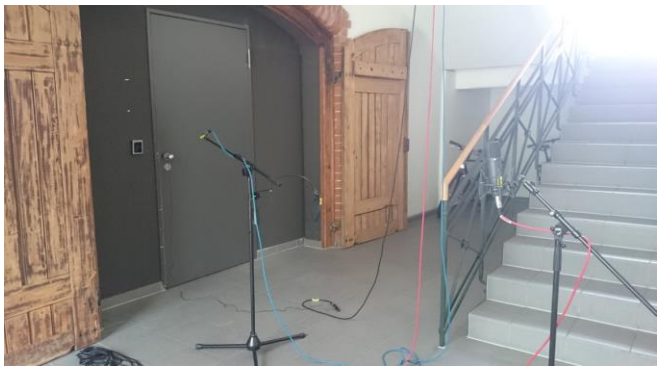


Fig. 6: Hard FX recording

5. Sound Design

5.1 Concept:

The audio drama has two main sets: The streetcar and an interrogation room. The interrogation room (where it all starts) shows how single sounds can be panned around – this shows the true effect of immersive sound because one can really determine were a single effect is happening in spatial scape. Then the height layer is intentionally being used to bring in flashbacks while staying in the same room. This way one doesn't have to leave the "surround" to enter a new situation – the new situation extends the surround.
Also this way the listeners realize that this is something more than 5.1 surround. Tests show that people even have the impression of being grilled themselves while inspector Hartmann is walking around them – a planned effect by the two sound designers.

Once entering the streetcar situation, the sound field gets fuller and fuller. The listeners are surrounded by ambience beds and single objects that fill the room on different spots.

By changing up the two main situations using half flashbacks or full flashbacks the story and its dramaturgy become wavelike but with a rising tension. This allows the listeners to enjoy the full energy of an object filled room vs. a quiet room with only few objects that can be traced precisely – but again enriched by a flavour of the power of Atmos sound (flashbacks).

As said earlier, it wasn't clear that the piece was going to be an audio drama. But as more focus was laid on the details (slight movements etc.) the character of an audio drama could already be presumed.

5.2 Dialogue:

The x-ed recordings need need more time for editing because the characters have to match perfectly to each other although they didn't act together in the studio. The loop groups are recorded in three different layers – so a bit of the sound design already happens while recording:
General walla: Beds. Unintelligible phantasy language, general emotions
Single voices: Objects. For concentrated listening one can understand what they are saying, but they are still put in the background
Single peaks: Objects. Intelligible voices important to help explain the plot.

5.3 Movements & Steps:

Both sound very film-like: A little bit larger than life. One mustn't forget: The listeners are still in a movie theatre. They expect movie theatre-like sound. Also it helps locate where e.g. inspector Hartmann is at while his walk-and-talk-moments.

5.4 Ambiences:

Sound designer Louis Molinas hits the nail on the head: For the series Rabbids invasion he had to build silence but used up to seven layers for it to come alive [5]. So we listen to a mix of room tone, a/c ventilation, an apart police office, a barking police dog and sirens whenever Bethke is under distress. Water going down the sewer in the ceiling and an obnoxious ticking of the clock on the wall communicates stress and silence at the same time.
Inside the streetcar the original ambience does it for a general setting (bed) and is enhanced by a lot of objects to create a streetcar with a lively character.

5.5 Relation of Power:

Hartmann, a powerful inspector standing, walking through his territory. Bethke, weak, sitting, not moving much – this is the beginning. As the story unfolds the relationship starts to change. In scene 5 Bethke now uses Hartmann's words. He is now powerful, Hartmann starts mumbling and stuttering. The sound design uses this interesting change and shows it with intelligent sound effects:

Scene 1 - Hartman flicks the coin vs. Scene 5 – Bethke flicks the coin (start and turning point)

Scene 2 – Hartmann throws his chair through the room vs. Scene 5 – Bethke stands up and throws it in the corner

Hartmann hits the table vs. Bethke it's the table

In the first half Hartmann has more movements, Bethke's movements are soft. In the second half it is the other way round.

Scene 1 – stands, Bethke sits. Later (scene 5) Bethke stands up as well, now they are "even".

6. Mix

Fig. 7: Mixing Stage

"Das Duett" uses every opportunity to fill the room. And we had enough space to place all the sound effects next to music without masking problems.
In Scene 3 there is a "funny" back and forth of the two main characters that the music and its panning picks up.
It is nice to see how different the sound is on different spots inside the cinema. The mix sounds different, some elements are accentuated or extenuated depending on where the listener is at. But the mix never falls apart because some sound elements are like strong posts in the room where we can orientate ourselves. Das Duett uses reverb as 7.1 beds and a 2.0 bed for the top layer. Different objects get a mono reverb if needed. The problem with reverb in an Atmos mix is a different story and not covered here but the two sound mixers make it work very realistic by sometimes use a little set of tricks. For example, when Bethke starts telling the story of the streetcar he slowly moves to the center channel in order to have enough power ad stability over all other sound effects. This way the listers themselves are now brought into the position of being in the train – of being…Bethke.
In order to know what they are actually talking about and of how to set up the spaces, storyboards are created for each scene to have an exact overview of where to place the sounds in the mix.

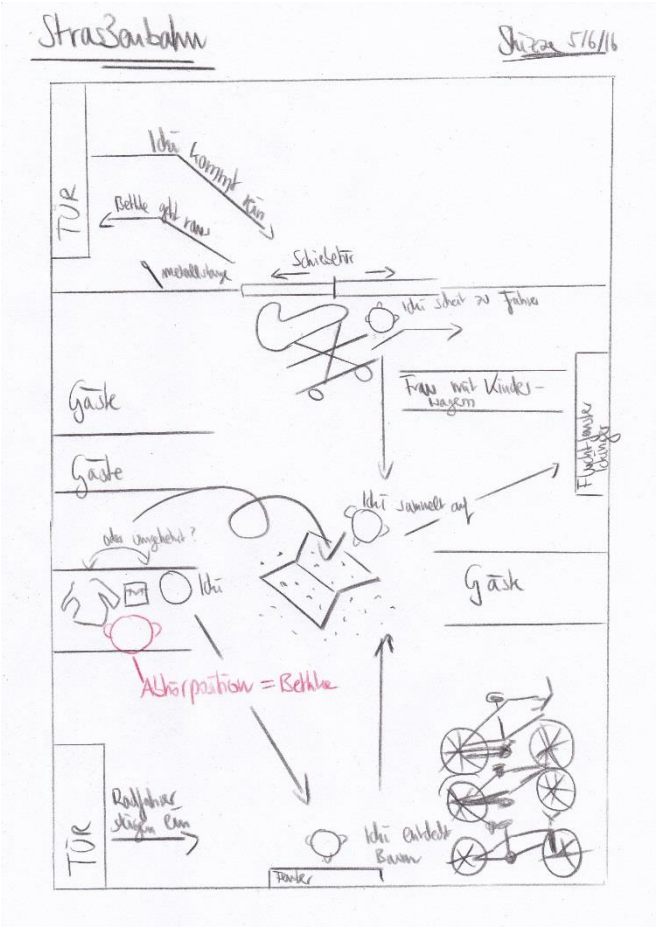


Fig. 8: Storyboard streetcar

Fig. 9: Poster

7. Visuals

The more test viewers and listeners they had the more they came to the point of deciding not to have any picture after all. The sound made it all work, the pictures where already there – inside the heads of the listeners.
Another visual would probably destroy what sound started to build from scratch
But Markus and Volker agreed on bringing a picture onto the screen after all: Complete Black.
This was the birth of a short film with sound and a black screen. Or: The world's first Dolby Atmos audio drama.

8. References

[1] see "Sengpiel 2017a"; http://www.sengpielaudio.com/DasStereo-MikrofonsystemORTF.pdf

[2] see "Plug-and-Play-Anordnung Für Surround Ambience-Aufnahmen ORTF Surround Outdoor Set - Übersicht - SCH OEPS.de" 2017; http://www.schoeps.de/de/products/ortf-surround-outdoor-set

[3] see "Microphone Data - 2017a"; http://microphone-data.com/microphones/4011/
"Microphone Data - 2017b"; http://microphone-data.com/microphones/ccm4/

[4] IRT-Mikrofonkreuze CB 250, CB 200, CB 140 Übersicht - SCH OEPS.de" 2017; http://www.schoeps.de/de/products/cb/overview

[5] At FMX 2016 in Stuttgart

vdt Verband Deutscher Tonmeister e.V.

AUDIO AES

AES Reviewed Paper at Tonmeistertagung 2018

Presented* by VDT in cooperation with the

Central European Region of the Audio Engineering Society (AES).

A Multi-Speaker Immersive Audio Middleware Solution for VR and Gaming Applications

K. Marais[1], R. Foss[2]

[1] *Rhodes University, South Africa, Email: g13m5942@ru.ac.za*

[2] *Rhodes University, South Africa, Email: r.foss@ru.ac.za*

Abstract

There is a need for precise and accurate audio spatialization in gaming and Virtual Reality applications. If the sounds of a virtual environment are localized exactly, the user will feel immersed in the environment. Certain software, Audio Middleware, works in cooperation with game engine software to facilitate the audio requirements for virtual environment and game developers. This paper describes the implementation of an enhanced Audio Middleware sound capability for Unity game engine developers. The capability leverages the functionality of a popular Audio Middleware software suite, FMOD, and a speaker-based spatialization system, ImmerGo. Additionally, a sample VR application that employs the capability has been developed. The application provides users with 3D sound control in real-time. Via the use of the Leap Motion controller and Oculus Rift Development Kit 2 Head-Mounted Display (HMD), users are able to select an audio source with a gesture, and then subsequently pan the audio around them in three-dimensional space.

1. Introduction

Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR) application development has blossomed over the past few years, and there is no indication that this development will slow down. [1] While most proponents of the technology would have predicted a massive influence on the entertainment industry, it is clear that VR and AR development is also being used to simulate environments for educational and training purposes. [2] In 2016, the global market value of AR and VR was estimated to be 6.1 billion dollars. By 2017, the global market value had increased to 9.4 billion dollars, and estimates indicate that it will grow to 215 billion dollars by 2021. [3] Apart from the VR and AR industry, the gaming industry has become the biggest and most lucrative component of the entertainment industry. [4] In parallel with the development of VR/AR/MR applications and high-fidelity games, more and more emphasis is being placed on realistic audio in both the gaming and VR sectors. Users and developers have recognized that a truly immersive experience requires immersive audio as much as it requires immersive visuals. [1] It follows that the importance of immersive sound in game and VR development cannot be overstated. Some research that has been conducted suggests that sound plays a crucial role in enabling users to be affected by whatever it is they are experiencing. [5] According to Schutze, “For the new realities (VR/AR/MR), the quality of the audio is not simply beneficial to your end result, or ‘nice to have,’ it is absolutely critical.” [1]

This paper outlines and analyzes the functionality and underlying mechanisms provided by an immersive audio capability that operates in tandem with ImmerGo, an immersive audio speaker spatialization system. [6] Figure 1 provides an outline of the ImmerGo system. It is a client-server system, in which the sound sources are tracks of a Digital Audio Workstation (DAW), residing on the server. A user uses the client, typically a mobile device, to request 3D localization of the tracks. The client communicates changes in the 3D coordinates of the audio tracks to the server. It is this client-server architecture that enables the ImmerGo system to be utilized by VR/gaming audio middleware.

*: Please note that the AES Reviewed Papers at Tonmeistertagung can be published by both, AES and VDT, in print, online and as PDF download.

ISBN 978-3-9812830-9-9

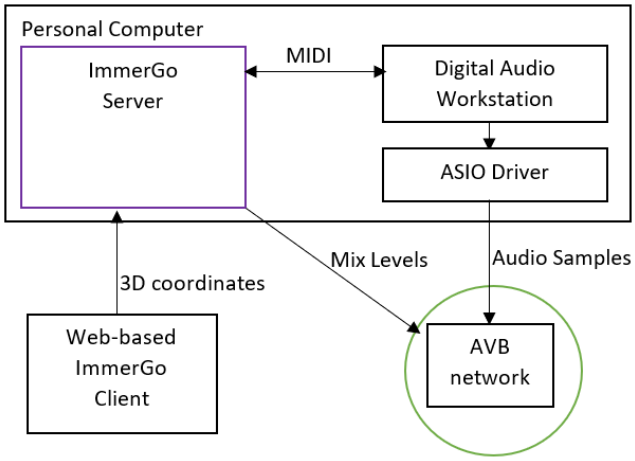


Fig. 1: A diagram representing the native ImmerGo spatialization system framework and components.

The audio middleware solution described in this paper controls sound sources in an FMOD[1] sound bank housed within the Unity game engine[2] (see Figure 2). The 3D coordinates are derived from Unity game objects that move in 3D space and provide access to their locations.

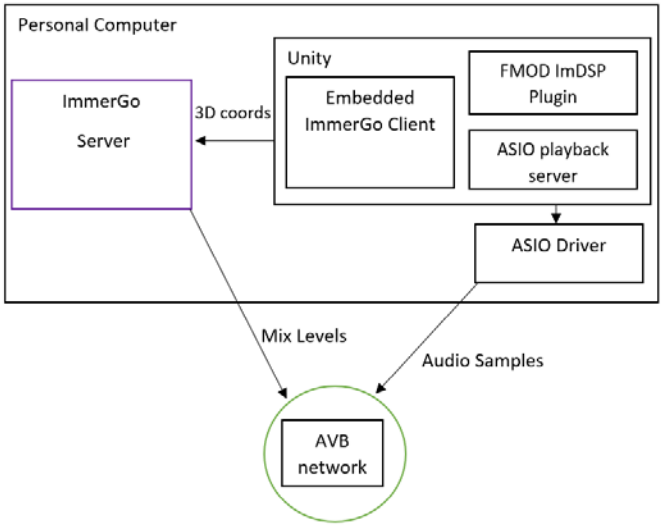


Fig. 2: An illustration of how the capability models the functionality of the native ImmerGo spatialization system.

Three components make up the ImmerGo-based middleware capability, and these communicate and function in parallel. As can be seen in Figure 3, the components of the capability have all been developed to be used from within a Unity application or game.

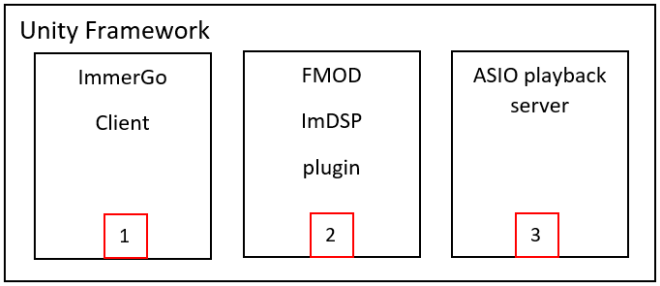


Fig. 3: A simple diagram indicating the three components that comprise the implemented capability.

The first component of the system keeps track of an array of game objects' positions. Should the position have changed for any of the audio sources, the 3D coordinates for the particular game object (audio source) values are sent to a modified version of the ImmerGo server, which subsequently applies spatial audio rendering algorithms using the coordinate values as input.

The second component of the system consists of instances of an FMOD DSP (Digital Signal Processing) plugin, entitled "ImDSP". The plugins function by transmitting audio samples from triggered FMOD events to the third component of the capability, the ASIO playback server.

The ASIO playback server is a socket server that can receive audio samples from the "ImDSP" plugins (see label "3" in Figure 3). The server functions by receiving samples and playing the samples back via an ASIO driver.

A sample Oculus Rift VR application that employs this middleware capability has been developed. The user is placed at the centre of a virtual room. A Leap Motion controller[3] attached to the HMD renders a digital representation of the user's hands and enables the use of a three-dimensional cursor. The user is surrounded by five audio sources (that are represented by different coloured spheres). After a brief pause, the audio sources all begin to play back, and the user can select (using the cursor) each of the audio sources and position them in 3D space at locations of their choice.

In the following sections, some of the literature and development platforms that are relevant to the project will be reviewed. The implementation of the system will also be discussed, and the sample VR application that has been developed will be described.

2. Context and Related Work

This section of the paper will analyze some of the research, platforms, and frameworks that are relevant to the project. After describing the ImmerGo spatialization system, Audio Middleware will be defined and the particular Audio

[1] https://www.fmod.com/

[2] https://unity3d.com/

[3] https://www.leapmotion.com/

Middleware, FMOD, will be introduced. The section will conclude with a brief discussion of immersive audio and its desirability for game/VR applications.

2.1. Overview of the ImmerGo spatialization system

2.1.1. Description of the framework

ImmerGo is a speaker-based audio spatialization system that can position audio sources in three dimensions. As has already been noted, the framework is implemented using a client-server based model. Importantly, the system also employs Object Based Audio (OBA), which attributes metadata to an audio source. This metadata can be used to convey positional information amongst other parameters. The system has been developed to function in parallel with audio software that can output multichannel audio and was initially envisioned to integrate with a user's DAW and enable the three-dimensional mixing of the audio tracks in the project in real time. (Review Figure 1 for an illustration of how the system functions.)

2.1.2. ImmerGo Server

The ImmerGo server was developed using the NodeJS framework,[4] and functions by connecting to and communicating with speaker end-points that are connected to an Ethernet Audio Video Bridging (AVB) network.[5] The server receives positional data from a client, and then employs spatial audio rendering algorithms to determine the mix-levels of the AVB end-points (speakers). The exact location and distance of each speaker is required for the server to accurately perform its rendering calculations: to ensure that the virtual audio object is correctly positioned relative to the speakers within the speaker array. In the current implementation of the framework, the system employs two rendering techniques: Distance-Based Amplitude Panning (DBAP) and Vector-Based Amplitude Panning (VBAP). [7] [8] The positional data that the server expects is provided from a web-based client that communicates to the server using a socket.io NPM module.[6]

In addition to rendering mix levels for the AVB end-points, the server also controls audio playback. The client sends a transport request to the ImmerGo server, which then translates the request to a Musical Instrument Digital Interface (MIDI) command that is received by the relevant DAW housing the audio tracks. To maintain synchronization with audio playback, the ImmerGo server also receives MIDI Timecode (MTC) information from the DAW, which it uses to timestamp a given track (information which is viewable from the web-based client).

[4] https://nodejs.org/en/

[5] http://www.electronicdesign.com/communications/understanding-audio-video-bridging

[6] https://www.npmjs.com/

2.1.3. ImmerGo Client

The traditional client is a web-based one (generally a mobile device) with a GUI that enables a user to pan an audio source in three dimensions. The interface also enables users to control audio playback from the DAW and record a three-dimensional mix for an audio track. As mentioned, while the user positions the audio, the client forwards these changes to the ImmerGo server in real time. Figure 4 below provides a schematic representation of the components of the ImmerGo system.

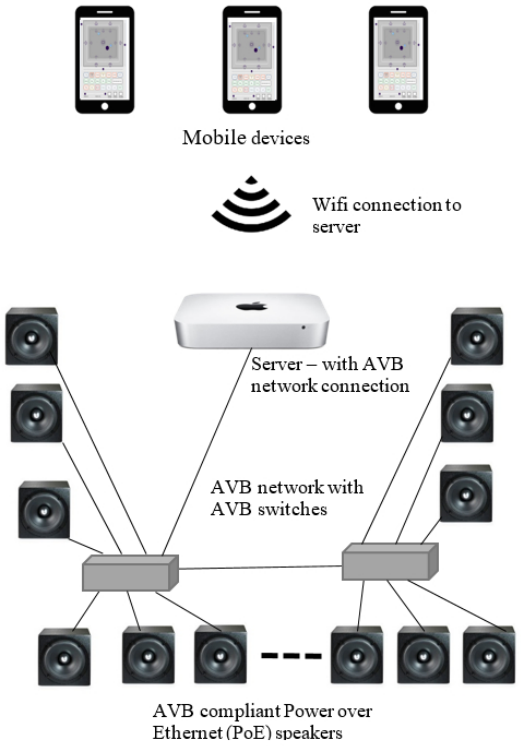


Fig. 4: A diagrammatic representation of the ImmerGo system.

2.1.4. Modifications to Framework

The system described in this paper features a modified client that is embedded in Unity (see the Implementation section below), and which imitates the behaviour of the web-based client, while transmitting coordinates using socket.io. Rather than controlling audio playback, the client is now used merely to transmit positional data to an instance of the ImmerGo server. The modified ImmerGo client represents the "coordinate transmission" component of the developed capability. Due to the modifications of the client, the server was also altered to no longer require MTC information, and to simply expect messages containing positional data for a relevant audio source in a Unity game engine scene.

2.2. Audio Middleware

Audio Middleware enables developers to process, create, and spatialize audio from a game development perspective. Since traditional audio processing in game engines is limited, Audio Middleware provides developers with tools

similar to those with which a DAW provides to sound engineers. [9] They greatly increase the audio processing functionality that a standard game engine provides and enable audio engineers to seamlessly integrate their work into a project in a way that does not interfere with other aspects of the game's development. Audio engineers can author audio events and build the events into a sound bank via the middleware software, which is then accessed by the game developers from within the engine. This separation of work flow means that the audio engineers are not required to understand the game engine's functionality, while the game developers do not need to understand how the Audio Middleware software functions. The middleware functions alongside the core game engine libraries and its main objective is to provide designers with more intuitive control over how sounds are used within the game. [10]

2.2.1. FMOD

FMOD is an Audio Middleware software suite that is developed by Firelight Technologies[7], and provides integration capabilities to both Unity and the Unreal Game Engine.[8] It consists of two components: FMOD Studio, and the FMOD Programmer API.[9] FMOD Studio is an audio content creation tool with a Graphical User Interface that emulates most DAWs, and it is the component that the sound engineer will work with. The programmer API is the set of functions that a game developer utilizes in their project and consists of two elements: The FMOD Studio API and the FMOD Studio Low Level API. The Studio API allows programmers to use FMOD-generated sound banks within their projects. It also enables developers to trigger events that have been created by the sound designers from within their project. The FMOD Studio Low Level API does not have any dependencies on events and sound banks and is used to trigger simple sounds from within a game or project.

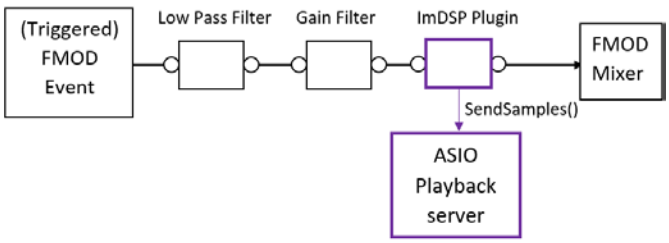


Fig. 5: An example of a signal processing chain of FMOD DSP objects.

Figure 5 provides an example of a signal processing chain (of FMOD DSP plugins) that could be applied to an FMOD audio event. There would be multiple chains for multiple audio events in an actual project that uses FMOD for signal processing.

2.3. Immersive audio and its desirability for game/VR applications

As the gaming industry has developed, increasing emphasis has been placed on hyper-realistic high-fidelity graphics. One of the reasons for this trend is the continual increase in the processing power of computers that has enabled developers to create resource intensive projects. Another is that the tools that developers use have also evolved with the industry (a game engine of today is a highly optimized piece of software). [4]

Spatial audio has been crucial in rendering immersion and realism in games for some time now. For example, in the First-Person Shooter (FPS) genre of games, if a player is being shot at from behind, the audio should immediately indicate the location of the opponent to them. However, realistic spatial audio in a Virtual Reality application is far more important than it is in a game with a traditional display. If a sound is spatialized incorrectly, there is a strong possibility that a user's sense of immersion will be broken. [1]

3. Implementation and System Design

This section will provide readers with a clear and concise understanding of the mechanisms that inform the functionality of the audio middleware system that was developed. Before discussing these mechanisms, it is worthwhile to reflect on the reasons behind the creation of the system and how the latter overcomes certain limitations.

3.1. The Limitations of FMOD with regard to Immersive Sound

The FMOD software suite provides basic 3D panning capabilities that function optimally with predetermined speaker configurations (mono, stereo, quad, surround, 5.1, 7.1, 7.1.4). While the software does allow for irregular speaker configurations for developers, it requires them to manually configure a mix matrix, which involves mapping speakers to outputs. Users of an immersive sound application developed using FMOD would have to find a way to configure their mix matrix according to their speaker configuration: the number of speakers and their positions. However, this is not possible without access to the source-code or precompiled version of the application—things that most users are unable to access. The default configurations limit the audio output channel count to eight (for eight speakers). As such, the middleware does not natively support immersive sound development for irregular speaker configurations with high (more than eight) output channels and large speaker numbers.

When using predetermined speaker configurations, the default FMOD 3D panner utilizes the FMOD mixer to determine the correct amplitude gains for each of the speakers in whichever predetermined configuration.

[7] https://www.fmod.com/

[8] https://www.unrealengine.com/en-US/what-is-unreal-engine-4

[9] https://www.fmod.com/resources/documentation-api

3.2. Enhancements to FMOD to overcome these limitations

The ImmerGo spatialization system has the ability to support thirty-two output channels and an unlimited number of speakers. To employ this functionality, while retaining the DSP power that FMOD possesses, an FMOD compliant DSP plugin was developed that is used to stream out audio from an audio event to an external audio driver. Each audio event has an instance of the DSP plugin bound to it, and an identifier that enables the audio driver to allocate the event to a specific output channel. By doing this, the eight-channel limitation mentioned in the preceding section is overcome. If there are enough output channels to provide one-to-one relationships with audio events, the system will be able to stream out audio on each available output channel. This solution bypasses the FMOD mixer, and therefore only provides playback functionality and does not enable spatialization from the onset. The spatialization of the audio sources is enabled by the ImmerGo system.

3.3. Mechanisms of the System

As mentioned, the capability outlined in this paper was developed for the Unity game engine. More specifically, the personal edition of Unity game development platform (version 2017.1.0f3) was used for the project and implementation. Unity is a community driven game engine that has been used to develop a number of games and VR applications. [11] When developing functionality for a Unity project or game, developers may implement their own logic and functions using imported libraries (known as “plugins”) and C# scripts. The engine also enables the integration of codebases and libraries in the form of plugins. As mentioned in the introduction of this paper, the capability functions through the interaction of three components. The following sections elaborate on the mechanisms of each component.

3.3.1. Coordinate Transmission

To enable spatialization of the audio sources, the coordinates of the sources need to be transmitted to the ImmerGo server. Each audio source is assigned an identifier (an integer value) that is also transmitted to the ImmerGo server along with the updated coordinates (see Figure 6). This identifier is the same one that is used when transmitting the audio samples to the external audio driver. Accordingly, each audio event is assigned an identifier that is used both to assign it a specific output channel (via the audio driver) and to notify ImmerGo about which event’s 3D position has changed.

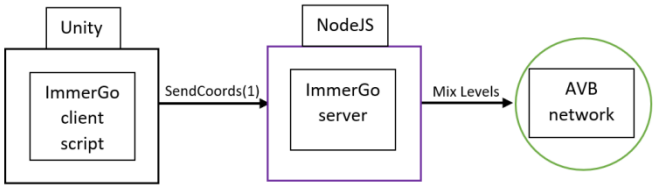


Fig. 6: A simple diagram indicating the transmission of 3D coordinates from a Unity application to the ImmerGo server, followed by the transmission of the speaker mix levels to the AVB speaker network.

To facilitate this socket communication from within Unity to the ImmerGo server, a socket.io Unity asset[10] was acquired, which enables developers to create, maintain, and close socket connections with applications that are external to Unity. A script was developed to act as the client to the ImmerGo server. It is launched at run-time with a list of the positions of each audio source in the scene. Once an audio source has moved, the previous positional values are compared against the current positional values of the specific audio source. If there is a difference (i.e. the object has moved), a socket message is sent to the ImmerGo server with the updated coordinates of the particular audio source (Figure 6 indicates the coordinates of an audio source being transmitted to the server).

3.3.2. FMOD ImDSP Plugin

This section outlines the procedure to be followed when constructing an FMOD DSP plugin. A developer needs to construct a C++ Dynamically Linked Library (DLL). This DLL exports a plugin description that enables the FMOD Studio application, as well as Unity, to register the DSP plugin, and use it in the studio application. The authored DSP is added to the signal chain of a FMOD audio event. The plugin description also enables the FMOD Programmer API to make callbacks to the plugin to perform a specific function or to report what state the plugin is in. Developers will bind an instance of the “ImDSP” plugin to their FMOD audio event. The “ImDSP” plugin supports Windows socket (Winsock) functionality and is used to act as a client to transmit audio samples from a particular audio event to the audio playback server. Essentially, the plugin retrieves FMOD event samples and transmits these samples across to the ASIO playback server. In its turn, the playback server consumes the samples and then sends requests to the clients for more audio samples. Each FMOD audio event has an instance of the ImDSP plugin bound to it, and a unique client identifier which the server then uses to assign an output channel.

[10] https://www.assetstore.unity3d.com/en/#!/content/21721

3.3.3. ASIO Playback Server

The final component of the system is the ASIO Playback Server, which utilizes the Windows implementation of sockets to receive samples sent to it by the ImDSP plugins. The server also utilizes the ASIO SDK to enable audio playback functionality. The SDK provides "low-latency and solves synchronization issues between the different input and output channels".[11]

The server operates using two threads that execute in parallel. After initialization, the first thread enters a receiver loop that constantly requests for more samples from all the FMOD audio events that have an ImDSP plugin attached. Upon receipt of a packet from an audio event (client), the server determines which audio event has sent the samples and then populates a corresponding circular buffer for the audio event. For example, if client number one sends a packet to the server, it analyzes the packet, recognizes it is from client number one, and then populates the circular buffer reserved for client number one. Figure 7 depicts this interaction. An instance of the ImDSP plugin, client number one, sends across samples to the server, which then populates the respective circular buffer–the red circle in the diagram. Should client number two also send a packet, the server will perform the same process, although the circular buffer that will be populated will be reserved for client two.

The second thread is used by the ASIO system to handle queries from the Streamware ASIO driver, which works in tandem with the Echo Streamware AVB compliant Network Interface Card (NIC)[12]. When the driver requires more data for audio playback, it is forwarded samples for a client and thus ASIO output. Figure 7 also represents this process. As can be seen, the ASIO driver requests samples from the server, which then retrieves the samples from the appropriate circular buffer, and provides them to the driver, which then streams out audio across the AVB network.

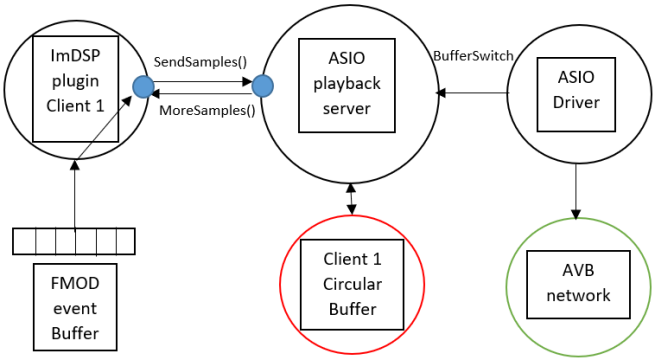


Fig. 7: A diagrammatic representation of how the ImDSP plugin sends audio samples to the ASIO playback server, which in turn supplies the incoming samples to an ASIO driver.

Figure 8 schematically represents the communication between two ImDSP plugins (Winsock clients) and the ASIO playback server. As the driver requests more samples for a specific output channel, the receiver thread in the server application requests more samples from the clients embedded in Unity. The clients then send their respective samples across to the server, which then places the samples in a corresponding circular buffer (represented as green circles in Figure 8). The ASIO thread is then able to pull samples from the circular buffer and populate the respective output channel's buffer. The parameters in the SendSamples() functions in Figure 8 represent the specific client identifier value.

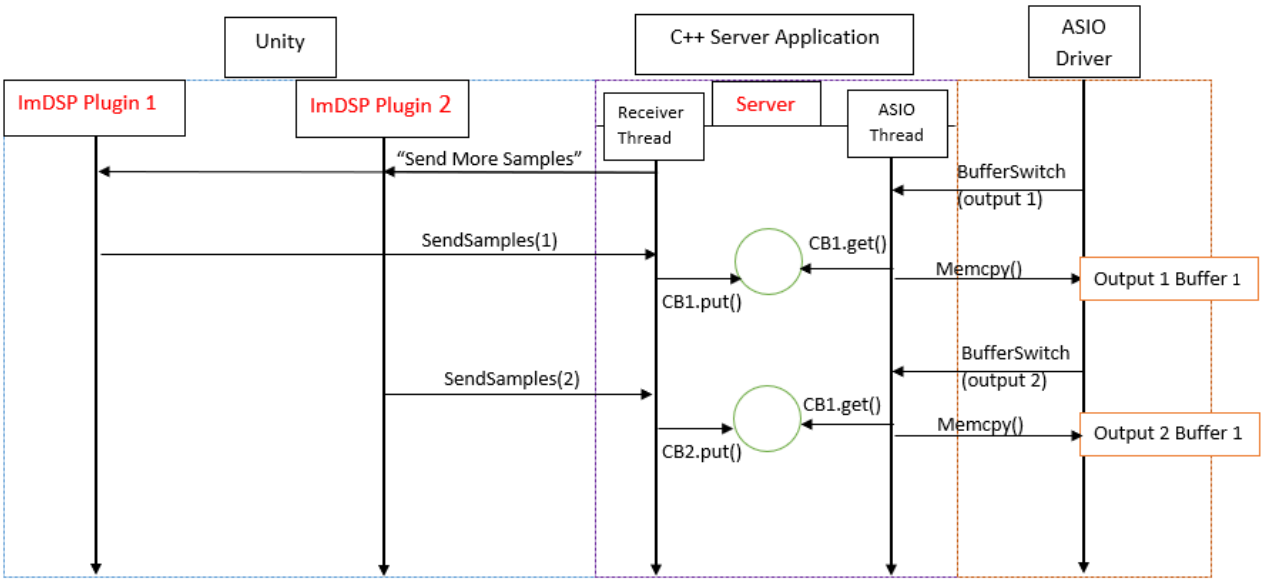


Fig. 8: A schematic diagram indicating the threads of execution and communication between the FMOD ImDSP plugin and the ASIO playback server.

[11] https://sdk.steinberg.net/viewtopic.php?f=1&t=9

[12] https://echoavb.com/products/streamware-nic-1

A synchronization issue between the ASIO playback server and the ImDSP clients resulted in a scenario whereby the ImDSP clients were occasionally filling up the circular buffers with samples faster than the ASIO driver was consuming the samples. This lead to audible distortion from the audio sources. The solution to the issue involved dropping one set of samples for each second of audio playback. This meant that the consumption of audio samples adequately matched the rate at which the circular buffers were filled, while maintaining perceived audio fidelity

3.4. Sample VR Application

Having described both the frameworks and platforms in which the implemented system belongs, a brief overview of the sample application will be provided.

The application's primary function is to confirm that the implemented capability performs as it should, to enhance the FMOD Audio Middleware by providing efficient, yet intuitive immersive sound functionality and accurate spatialization of audio sources in a Unity game engine application. Another important reason for the sample application's development was to indicate the potential behind the developed system from the perspective of a Virtual Reality application.

3.4.1. Utilized Hardware and Software

The sample application was developed using the Unity game engine. Since it is a Virtual Reality application, the Oculus Rift Development Kit 2 was utilized for the HMD. To enable user interaction, the Leap Motion Controller[13] was used, which provides a means to accurately track and digitally render a user's hands via the use of LED, infrared and camera sensors. The Leap Motion is not strictly a VR input device but can be used as one, if mounted atop the HMD. The SDK that the Leap Motion team provides, enables the recognition of specific gestures that a user may perform. Of course, the developer is then required to program an event after the gesture has been recognized. As mentioned previously, a three-dimensional cursor was utilized to enable the panning of the audio sources in the scene.

3.4.2. The Scene

The user is placed in the centre of the speaker array and then given the HMD. Once the application has been launched, the user is transported to the centre of a virtual room, whose dimensions are modelled on the physical room in which the speaker array is located. In front of the user, lie five spheres, each of which has an audio source attached. After a short delay, each of the audio sources will begin playback.

3.4.3. User Interaction

To select and pan an audio source, the user is required to locate the 3D cursor (which is a visible grey sphere that extends from the user's hand). To select an audio source, the user should ensure that the cursor has collided with the specific sphere (audio source). Once the cursor is positioned correctly, the user should simply pinch (a gesture that connects the index finger with the thumb). If the cursor is visible, the user should notice that it turned green, which indicates that the pinch gesture has been recognized by the controller. While holding the pinch, the user will notice that the Audio Source is "attached" to the cursor and will move wherever the cursor does (see Figure 9 below). After having positioned the audio source, the user should simply release their pinch, and the sphere will no longer be attached to the cursor— it will remain in the position it was last attached.

As soon as an audio source has been selected by a user and its position has changed, the aforementioned SendCoords function will register that the game object and thus audio source has moved and will transmit the new coordinates across to the instance of the modified ImmerGo server, which subsequently updates the mix levels of the AVB speaker end-points in real time.

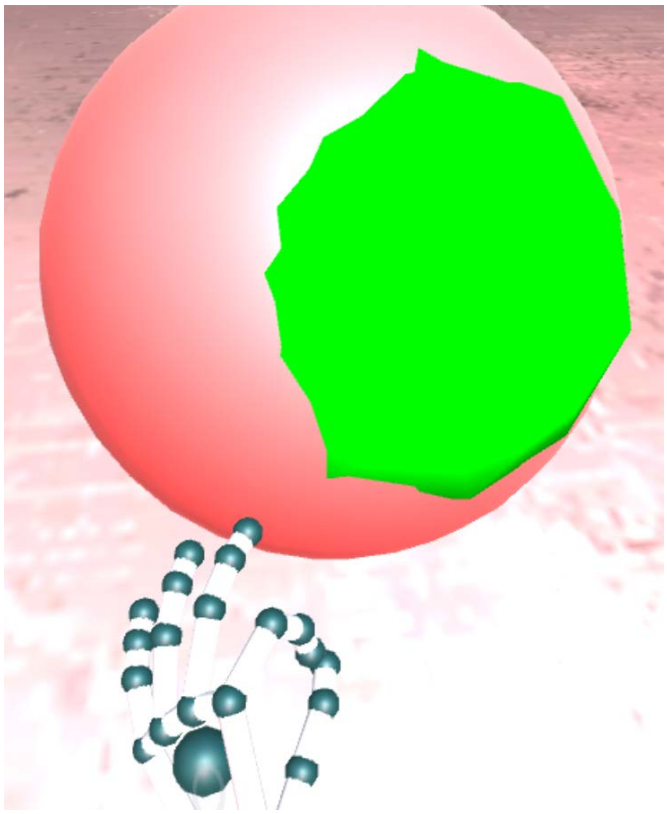


Fig. 9: A screenshot of the sample VR application that shows a user's hand rendered into the environment. The 3D cursor is the green sphere projecting out of the red sphere—an audio source. The cursor is currently "attached" to the source.

4. Conclusion

The main goal of the research was to enable an immersive speaker-based VR audio capability within the Unity game development platform. To this end, the research has been successful, as a functional system has been developed. Another goal of the project was to incorporate Digital Signal Processing functionality, which the Unity audio engine does not provide. Audio Middleware was used to provide this

[13] https://www.leapmotion.com/

functionality. FMOD was selected among other solutions[14] [15] because of its close integration with Unity, together with its extensive API. Due to certain limitations of the functionality that FMOD provides (see preceding section), a suitable means to bypass the FMOD mixer and playback audio samples had to be established. As mentioned, the FMOD middleware only provides eight audio output channels, whilst the ImmerGo system is able to utilize a maximum of thirty-two output channels. The developed capability therefore integrates and enhances both the middleware and the spatialization system. To enable audio playback functionality, the ASIO SDK was used and provided the tools for a lightweight and scalable solution. The processing functionality of FMOD has therefore been used to extend the capabilities of the ImmerGo audio spatialization system, and visa versa.

A functional immersive sound capability has been developed that enables Unity developers to utilize a scalable speaker-based spatialization system. This capability has the potential to be deployed by Unity virtual environment and game developers should they wish to employ immersive sound in their respective projects, and therefore deploy large, unconventional arrays of speakers. Importantly, in this regard, previous research has suggested that spatialization using an array of speakers (versus headphones) renders greater immersion. In conclusion, this research thus indicates that the capability described in this paper could greatly enhance Unity VR and game development that is dependent on immersive audio. [12]

5. References

[1] S. Schütze and A. Irwin-Schütze, New Realities in Audio: A Practical Guide for VR, AR, MR and 360 Video, Boca Raton: CRC Press, 2018.

[2] L. Freina and M. Ott, "A Literature Review on Immersive Virtual Reality in Education: State Of The Art and Perspectives," in *eLearning and Software for Education*, Bucharest, 2015.

[3] Statistica, "Forecast augmented (ar) and virtual reality (vr) market size worldwide from 2016 to 2021 (in billion u.s. dollars)," 2018. [Online]. Available: https://www.statista.com/statistics/591181/global-augmented-virtual-reality-market-size/. [Accessed 15 May 2018].

[4] P. Zackariasson and T. Wilson, The video game industry: Formation, present state, and future, New York: Routledge, 2012.

[5] L. E. Nacke and M. Grimshaw, "Player-Game Interaction Through Affective Sound," in *Games Computing and Creative Technologies: Book Chapters*, Bolton, University of Bolton Institutional Repository, 2011, pp. 264-285.

[6] R. Foss and A. Rouget, "Immersive audio content creation using mobile devices and ethernet avb," *Audio Engineering Society*, vol. Convention 139, 2015.

[7] T. Lossius, P. Baltazar and T. de la Hogue, "DBAP - Distance-Based Amplitude Panning," 2009.

[8] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *Audio Engineering Society*, vol. 45, no. 6, pp. 456-466, 1997.

[9] G. Somberg, Game Audio Programming: Principles and Practices, Boca Raton: CRC Press, 2016.

[10] S. Horowitz and S. Looney, "Masterclass: Using Game Audio Middleware," 11 August 2014. [Online]. Available: https://www.emusician.com/how-to/masterclass-using-game-audio-middleware. [Accessed 15 March 2017].

[11] Unity, "Made with Unity," 2018. [Online]. Available: https://unity.com/madewith. [Accessed 10 January 2017].

[12] K. D. Marais and R. Foss, "Gesture Controlled Immersive Audio for the Oculus Rift Head-Mounted Display," in *South African Telecommunication Networks and Applications Conference*, 2017.

[13] Oculus, "Oculus rift pre-orders now open, first shipments March 28," 6 January 2016. [Online]. Available: https://www.oculus.com/blog/oculus-rift-pre-orders-now-open-first-shipments-march-28/. [Accessed 25 May 2018].

[14] https://www.audiokinetic.com/products/wwise/

[15] http://www.tazman-audio.co.uk/fabric

Investigations on the effect of three-dimensionally produced background music in virtual reality environments

D. Wefers

University of Applied Sciences Düsseldorf, Germany, Email: dominik.wefers@study.hs-duesseldorf.de

Abstract

Binaural reproduction of 3D audio has become increasingly important due to the rise of virtual reality applications. For example, 3D audio plug-ins for game engines like Unity and Unreal provide HRTF filtering to place sounds anywhere in the virtual space. The aim of this work was to find out whether the feeling of immersion can be strengthened if background music is produced in 3D and which advantages and disadvantages this brings along compared to a normal stereo mix. In a listening test 18 subjects were asked to compare two three-dimensional mixes with a stereo version. The results indicate that the stronger the 3D effect is, the more the subjects are polarized when judging immersion.

1. Introduction

Interactive applications such as video games are increasingly supporting 3D audio formats. For example, commercially successful games such as "Overwatch" (2016) and "Assassin's Creed Origins" (2017) use Dolby Atmos for three-dimensional playback. [1] For VR applications, binaural technology is particularly useful as headphones are used for sound output. Real-time HRTF filtering of sound sources enables accurate localization, especially in combination with head-tracking. Currently there are various 3D audio engines that can be integrated into game development environments with comparatively little effort. The Ambisonics format, which is mainly used in 360-degree videos, can also be integrated for immersive sound design.

Traditionally, music in video games is implemented in stereo. The present work explores the question of whether music benefits from three-dimensional audio design in VR environments. Can this increase the sense of immersion, an integral part of VR?

Scientific studies have shown that subjects tend to prefer music productions in 3D audio formats to a regular stereo mix. For example, a listening test at the Technical University of Berlin showed that the spatial representation quality of a concert recording was rated better when binaural sound output was used. The emotional reactions to the music were also stronger than in the stereo version. [2] In cooperation with the Technical University of Ilmenau, advantages of a three-dimensional production of music could also be determined when using loudspeaker systems. [3]

2. Immersion

In connection with media, the term "immersion" is used to describe the phenomenon that users lose their awareness of the technical characteristics of the medium. For example, during a movie, moviegoers "forget" that images and sound are produced by projectors or speakers. Instead, they direct their attention to the content such as storyline, characters or environments. In interactive media like video games, necessary inputs are subconsciously executed in the case of a strong sense of immersion. [4] According to Christopher Dede of Harvard University Cambridge, three factors are key to strong immersion in interactive media: first, the human senses must be compellingly stimulated by technical devices (for example, high screen resolution and color fidelity, good sound quality etc.). Second, the application should motivate the user to perform actions in the virtual environment, for example, by enabling actions that are novel to the user or not possible for him in the real world. Third, the user should be involved on a psychological level, using his emotions, values, and beliefs as the basis for the content of the application. In this way, it is possible that strong emotional responses are evoked, though the user is aware that he is only part of a simulated environment. [5] The philosopher Samuel T. Coleridge shaped the term "Willing suspension of disbelief". That means people are willing to accept fictional and unrealistic representations in art, if they feel entertained (e.g., actors with superhuman powers, talking animals, stories taking place in the future). [6]

The immersive effect of VR has been proven in several studies. For example, it could be shown that subjects perceive a virtual body as their own, even if real and virtual body are not similar. In one experiment, subjects responded more intensely to simulated stress situations than subjects experiencing the scene from an uninvolved person's point of view. [7] In medicine, this knowledge is used as well, for example, to treat anxiety patients. Stress-inducing situations are simulated in VR to teach the handling of real situations. [8]

ISBN
978-3-9812830-9-9

The authors hold the copyrights of all of content in this manuscript. They have granted the VDT the right to publish it in print, online and as PDF download on *https://tonmeistertagung.com*

3. Music in interactive media

Music in interactive media such as video games has similar functions as it has in movies. In both cases, it is used to create a certain mood, represent time and place, describe characters and announce upcoming events in the story. [9]

However, unlike film music, game music is mostly non-linear. Although it may have linear components (e.g., cutscenes), the music often needs to be able to dynamically adapt to the gameplay. There are basically two types of dynamic music: interactive and adaptive music. Interactive music reacts to the player's input (e.g., pressing a certain button triggers a defined piece of music). In contrast, adaptive music responds to the gameplay. For example, the tempo of the music is increased during an action scene or during a countdown in the game. However, depending on the desired behavior, both principles can also be combined. [10]

There are various techniques to implement interactive and adaptive music. An essential approach is the creation of loops. The loops can be played back infinitely if the length of a scene depends on the player's actions. Additional variety can be created by dividing the loops into its instrument tracks (stems) and fade the instruments in and out randomly during gameplay. Stems can also be used to realize adaptivity. For example, by fading in additional instruments, the increasing intensity of a scene can be supported. Moreover, the subdivision of loops using special markers allows transitions to take place in the rhythm of the music. It can also be defined that only certain harmonically matching loops are triggered, depending on the current key. Such complex logics are often implemented using special audio middleware such as *Wwise* or *FMOD*. Middleware offer a (sometimes DAW-like) graphical interface and allow the linking of game parameters and audio files. The logics created in the middleware can then be called up by an audio engine that was previously integrated into the game's code.

4. Listening test

4.1. General setup

For the listening test a virtual environment was used, representing a seaside castle (see Figure 1). The *Unreal Engine* (version 4.20.1) was utilized, including the plug-in *Oculus Audio* which enables HRTF filtering of sound sources. Its filter set is a standard HRTF set. The integration of individualized HRTF is not possible. For the creation of the soundscape, a library with WAV files in mono and stereo format was used. Mono sources served to represent point sound sources, which were placed in the three-dimensional environment and should be made exactly localizable with the help of HRTF filtering (for example, individual bird sounds and insects). Stereo files, on the other hand, were mainly utilized for ambient sounds without HRTF filtering (e.g., general sea noise).

The piece of music for the test includes both natural and electronically processed orchestral instruments and is

Figure 1: View of the virtual environment

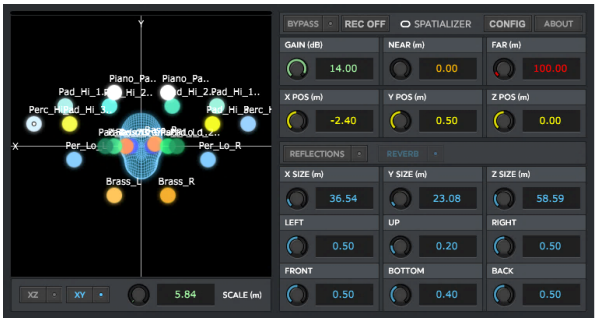


Figure 2: VST-Plug-in Oculus Spatializer

designed as a loop. First, a stereo version was mixed using the DAW *Reaper*. For the 3D mixes, the VST-Plug-in *Oculus Spatializer* allowed positioning instruments in 3D space and evaluating the acoustic effects of the Oculus HRTF set on sound sources within the DAW. Figure 2 shows the user interface of the plug-in, which must be integrated as an insert effect in every track to be positioned. Subsequently, in all instances of the plug-in, the tracks are represented by colored dots and can be arranged in space using X, Y and Z coordinates. In addition, the parameters *Near* and *Far* can be used to define the distances between sound source and listening position at which the simulated attenuation begins and reaches its maximum, respectively. In connection with an *Oculus Rift* VR headset it is possible to mix inside a VR environment, using the handheld controllers to move the sound sources and edit their properties. In order to ensure the most uniform sound possible in every gaze direction, the instruments were distributed spherically around the origin. It was decided to leave the lowest instruments out of the HRTF processing, since instruments like bass and drums suffered from the filtering tonally. Because low frequency components can hardly be localized by humans anyway, these instruments were integrated as stereo files instead. Finally, the following three mixes were created for the listening test:

A. Instruments in the low frequency range in stereo, mid & high frequency instruments in 3D (HRTF)

B. Instruments in the low & middle frequency range in stereo, high frequency instruments in 3D (HRTF)

C. All instruments in stereo

Figure 3 shows how the instruments were arranged in 3D space for mixes A and B. The center of the arrangement is linked to the center of the usable area in the real space. Therefore, if the user moves in the real environment, he can approach individual instruments. Various settings regarding the distance of the sources from the center and regarding the distance-dependent volume attenuation were tested. A compromise had to be found between not too strong volume differences while moving and a convincing, three-dimensional impression. Finally, the arrangement of the instruments approximately matched the dimensions of the usable area (1,5 x 2 meters) and an audible volume attenuation was implemented. The overall levels of the three mixes were adjusted to a comparable loudness level in order to minimize the influence of the psychoacoustic effect that people tend to rate louder music mixes better than quiet mixes.

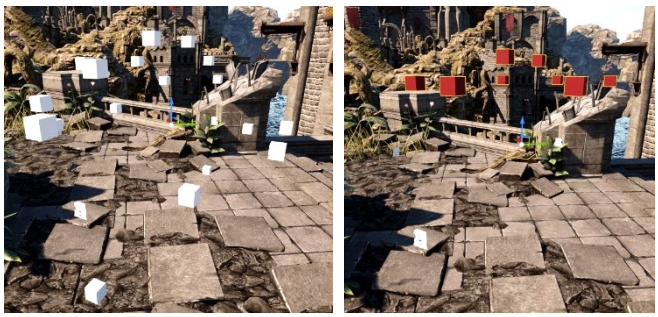


Figure 3: 3D arrangement of instruments: Mix A (left) & B (right)

4.2. Subjects

Eighteen subjects took part in the listening test, mainly employees and students of the university. No special selection criteria were set. Most participants had no (5) or little experience (10) with VR applications. Most of the participants were between 16 and 24 years old (10). The remaining subjects were in the age groups 25-33 (5), 34-42 (2) and 52-60 (1).

4.3. Procedure

In advance, the subjects were informed about the general topic of the experiment (three-dimensionally produced music in VR environments). Moreover, they did not know any further details about the three mixes. After an introduction to the test procedure, the subjects were able to familiarize themselves with the navigation (standard teleportation) and the switching of the mixes by pressing buttons. Subsequently, they always compared two mixes (A vs. C and B vs. C) according to following criteria:

1. Influence of movement on the perception of music
2. Externalization
3. Overall sound quality
4. Intelligibility of environmental sounds
5. Immersion
6. Overall impression

The participants were able to use three rating levels: "equal", "slightly stronger / better" and "much stronger / better". For criteria 1-3, all ambient sounds were muted and only the music mixes were rated. For criteria 4-6, subjects rated music and sounds in combination. For criterion 4 (intelligibility of environmental sounds), there was a predetermined location in the virtual environment to which the subjects were directed. For all other tests, they should move freely around the level. Subjects were given the opportunity to provide reasons for their scores on each criterion or to make further comments.

4.4. Results

Since always one of the 3D mixes (A or B) was compared with the stereo mix (C) and no direct comparison between A and B took place, only the respective 3D mix is mentioned below for reasons of readability.

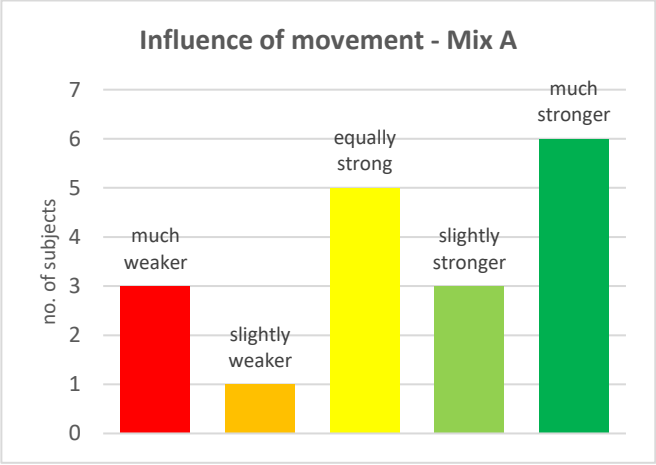


Figure 4: Results “Influence of movement” – Mix A

Half of the subjects experienced the influence of their movements in mix A much (6) or slightly stronger (3) than in the stereo mix. Five subjects noticed no difference. Four said that the influence seemed greater in the stereo mix.

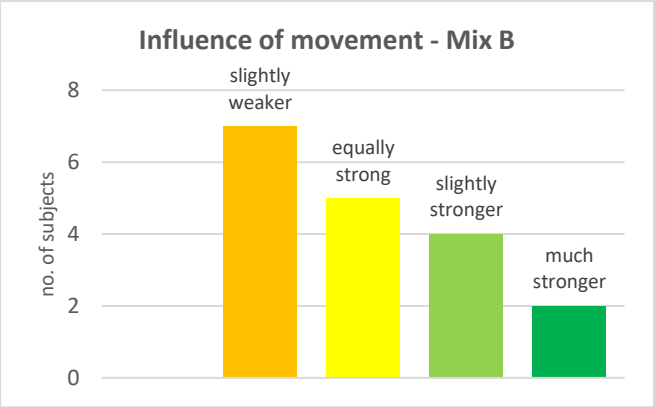


Figure 5: Results "Influence of movement" – Mix B

When listening to mix B, six participants perceived a stronger impact compared to the stereo mix. About the same number of participants stated that they did not notice any difference (5) or rated it slightly weaker (7).

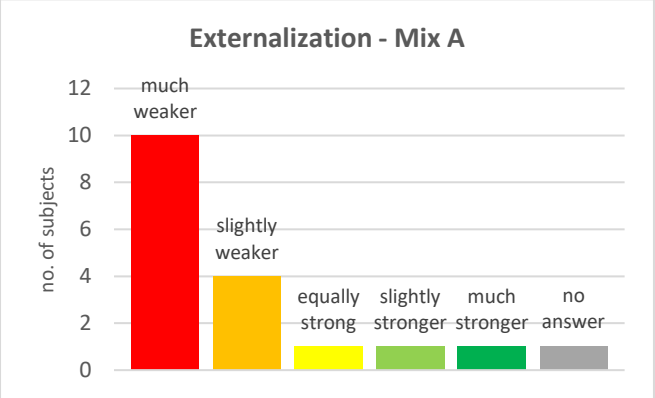


Figure 6: Results "Externalization" - Mix A

A clear majority perceived externalization to be weaker. Only two subjects rated it higher than the stereo version.

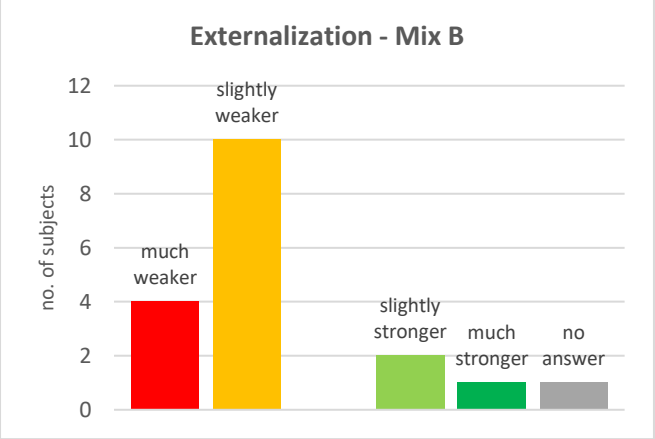


Figure 7: Results "Externalization" - Mix B

Externalization was also perceived weaker than the stereo mix, but less clearly compared to A. Three subjects thought it was stronger.

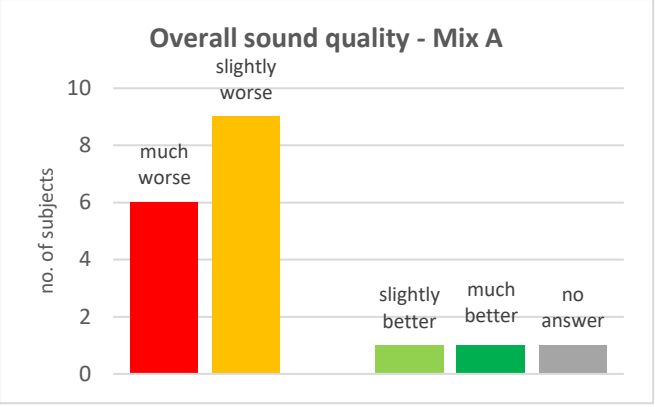


Figure 8: Results "Overall sound quality" – Mix A

Fifteen participants preferred the sound quality of the stereo mix, six of them clearly. The main reasons given were stronger externalization (4) and more pronounced high frequencies (4).

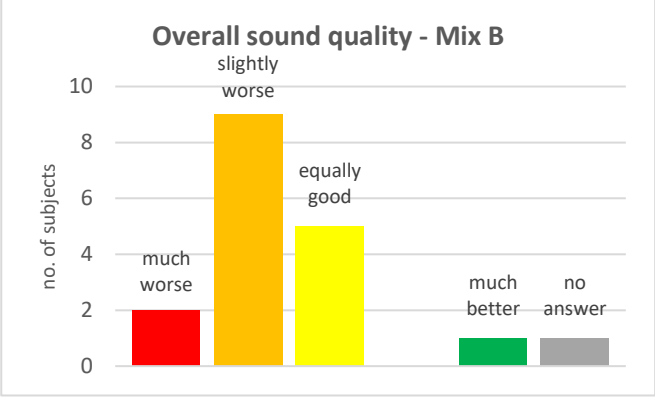


Figure 9: Results "Overall sound quality" – Mix B

Most subjects preferred the stereo mix, but less clearly compared to A. The main reason given was stronger high frequencies. Two subjects said the mix was more balanced.

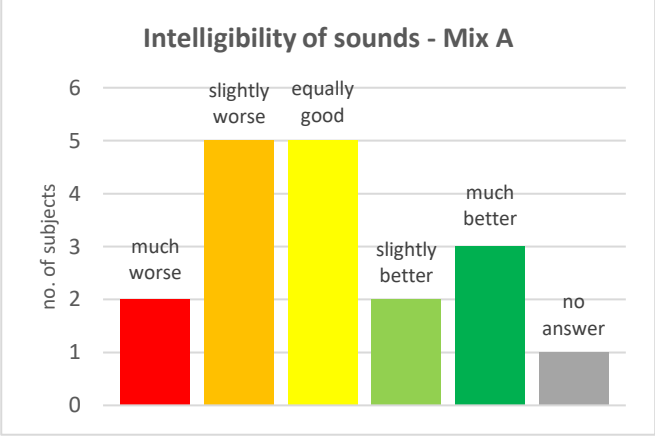


Figure 10: Results "Intelligibility of sounds" – Mix A

The result is mixed, with no clear preference for any of the mixes. The three groups (better / worse / equally good) are similar in size.

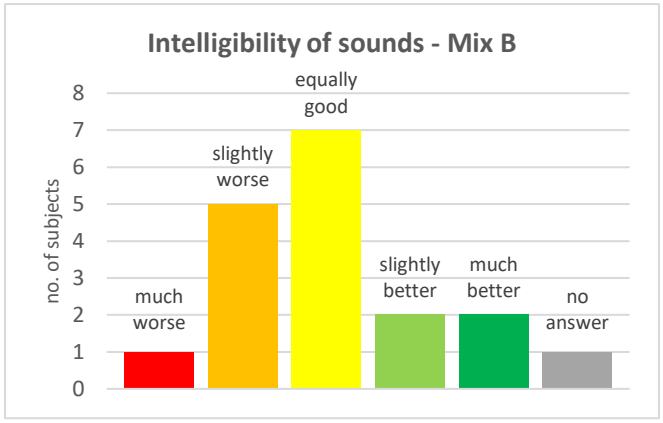


Figure 11: Results "Intelligibility of sounds" – Mix B

For B, the picture is mixed as well. However, the group that rates both mixes as "equally good" is the largest here. The mix does not seem to polarize as much as A.

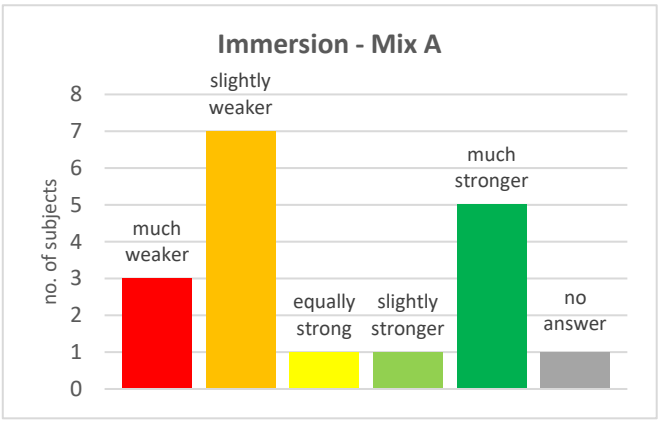


Figure 12: Results "Immersion" – Mix A

The majority find the stereo mix more immersive. However, five subjects clearly preferred the 3D mix. Two of them said that the better localization of sounds was the main reason.

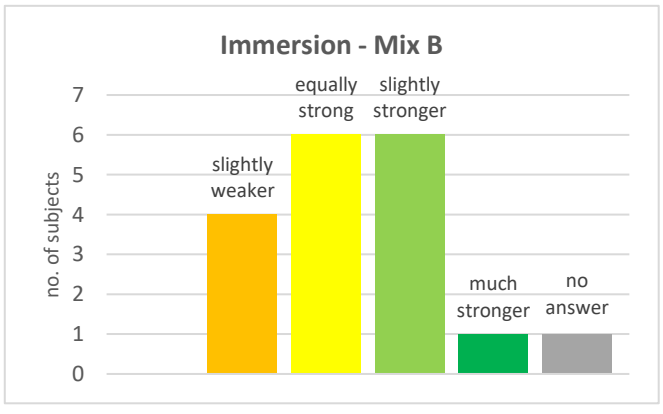


Figure 13: Results "Immersion" – Mix B

The 3D mix is slightly favored overall. Seven subjects thought it was more immersive, four preferred the stereo mix. In the comments, the interactivity of the 3D mix was evaluated as positive as well as negative ("unbalanced").

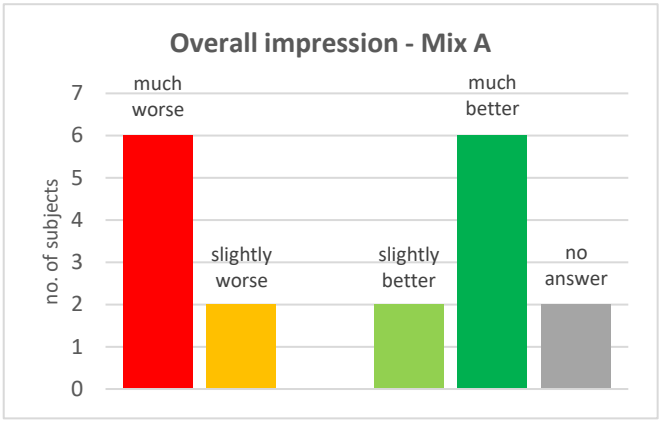


Figure 14: Results "Overall impression" – Mix A

The opinions are divided. Six people find the 3D mix "much better", mainly because of a better localization of environmental sounds. However, just as many find it "much worse", especially because of a "narrower" sound.

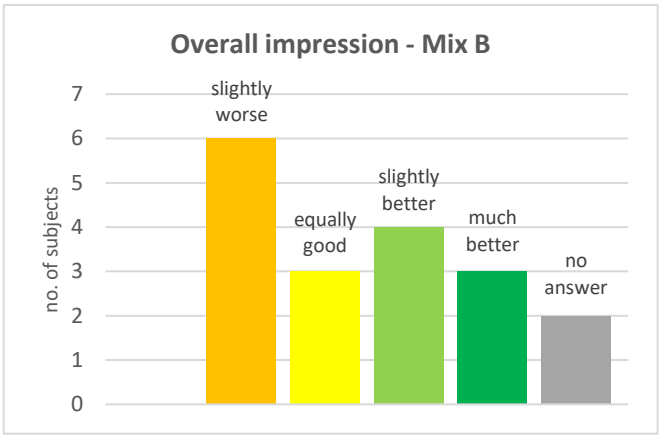


Figure 15: Results "Overall impression" – Mix B

Overall, the 3D mix is slightly preferred. Seven participants find it better, three of them clearly. Mostly they mention a stronger immersion and interactivity as a reason. Three subjects rate it as "equally good", six as "slightly worse". As with mix A, the latter mention the "narrower" sound as the main reason.

4.5. Evaluation of the results

Apparently, the externalization is weakened mainly by the 3D mix. Especially mix A with a lot of HRTF-filtered instruments is called "narrower" by the subjects. The reason for this is probably the use of the non-personalized HRTF dataset. The frequently mentioned lack of high frequencies in the 3D mixes seems to be due to the spatial arrangement of the instruments in the high frequency range. Since these were placed rather centrally in the usable area, their level dropped as the participants moved to the edge of the area. Apparently, a larger distribution of sources would have been beneficial. Alternatively, disabling the distance-dependent level attenuation could have helped. This would have improved the partially criticized imbalance of the 3D mixes as well. However, it is questionable whether this would have

had negative impacts on the ratings of the subjects who praised just this "interactivity" of the music in the 3D mixes.

In conjunction with ambient sounds, the influence of the overall poorer sound quality of the 3D mixes seems less relevant. Some subjects indicated that the sounds are more intelligible in return and that this is important for the feeling of immersion. However, since no clear majority preferred the 3D mixes during the tests for the intelligibility of sounds, it is difficult to draw clear conclusions from this. It must also be noted that during the test regarding immersion and overall preference, the participants moved freely in the virtual environment and were therefore likely to be exposed to slightly different environment sounds.

It is noteworthy that not more test persons perceived the motion-dependent sound changes in the 3D mixes. In part, they even rated the static stereo mix as "more interactive". This probably has several reasons. On the one hand, some participants stated that they did not understand the question right away. Therefore, there may have been uncertainties here, even with other subjects. In addition, the playing style of the instruments did not seem to be ideal for the listening test (mainly individual, sustained notes of the string instruments, with pauses in between). Two subjects stated that the pauses made it difficult to compare the mixes.

Furthermore, the feedback from some participants indicates that it would have made sense to randomize the played back mixes. Some of the subjects made the correct assumption that each letter shown was always assigned to the same mix. With a random assignment the danger of a possible "learning effect" could have been reduced. To what extent this ultimately had an impact on the results is difficult to estimate.

5. Summary and outlook

It turns out that interactive applications are increasingly supporting 3D audio formats. Especially for VR applications binaural sound makes sense, as headphones are used. Background music supports the desired mood and adapts itself dynamically to the behavior of the user or to the virtual environment through interactive and adaptive design techniques. The aim of the listening test was to clarify whether background music can benefit from a three-dimensional mix in VR. For this purpose, selected instrument tracks of a piece of music were included as mono signals with individual HRTF filtering. The remaining tracks, mostly instruments in the low-frequency range, were excluded from the filtering to avoid negative effects on the sound. The test results show that in the case of the 3D mixes, the externalization was weak and most of the subjects preferred the stereo mix sound-wise. Likewise, the fixed positioning of the instruments relative to the real usable space and also the distance-dependent level attenuation implemented here seems to be disadvantageous. Although some testers praised the "interactivity" of the music, others complained about an imbalanced mix. In future tests, consideration should be given to choosing either a purely perspective-based implementation or an implementation without distance-dependent level attenuation. Presumably, the Ambisonics format would be the most viable solution, as it would require only a single HRTF filtering, and therefore processor load would be lower. To exclude low-frequency content from the filtering, a combination of Ambisonics and stereo could be useful. Streaming services for 360-degree videos already use such mixed formats to integrate head-locked stereo content.

In order to increase the perceived sound quality and thus the general acceptance of 3D mixes, personalized HRTF datasets still seem to be important. Above all, externalization could be improved in this way. New solutions that allow for easy customization (such as image-based matching of the head anatomy using an HRTF database) will probably be helpful here. The fact, that at least the mix with a small amount of HRTF filtering was slightly preferred overall, is an indication that a three-dimensional music production may be advantageous. Further tests with personalized HRTF datasets and other music styles could show if a clear majority prefers three-dimensional mixes.

6. References

[1] Games in Dolby, URL: https://www.dolby.com/us/en/categories/games.html

[2] Lepa, et al.: Emotional Impact of Different Forms of Spatialization in Everyday Mediatized Music Listening: Placebo or Technology Effects? AES 136th Convention, Berlin, Audio Engineering Society, 2014

[3] Hold, C., Wierstorf, H., Raake, A.: The Difference Between Stereophony and Wave Field Synthesis in the Context of Popular Music. AES 140th Convention, Paris, Audio Engineering Society, 2016

[4] Phillips, W.: A Composer's Guide To Game Music. Massachusetts Institute of Technology, 2014

[5] Dede, C.: Immersive Interfaces for Engagement and Learning. Science (January 2009), 66, DOI: 10.1126/science.1167311

[6] Dörner, et al.: Virtual und Augmented Reality (VR/AR) - Grundlagen und Methoden der Virtuellen und Augmentierten Realität. Springer Vieweg, Heidelberg, 2013

[7] Slater, et al.: First Person Experience of Body Transfer in Virtual Reality. PLoS ONE 5(5): e10564. doi:10.1371/journal.pone.0010564, 12 May 2010

[8] Oxford VR, University of Oxford: Study provides first evidence that psychological therapy can be successfully delivered automatically in virtual reality (VR). 12 July 2018 (cited 2 August 2018). URL: http://oxfordvr.org/study-provides-first-evidence-that-psychological-therapy-can-be-successfully-delivered-automatically-in-virtual-reality-vr/

[9] Wardrobe, K.: The Guide To Composing Music For Video Games. 9 June 2016 (cited 9 August 2018). URL: https://midnightmusic.com.au/2016/06/the-guide-to-composing-music-for-video-games/

[10] Horowitz, S., Looney, S.: The Essential Guide To Game Audio - The Theory And Practice Of Sound For Games. Focal Press, Burlington, 2014

tmt30
expertise in audio media

30. Tonmeistertagung
VDT International Convention

Impressum

Veröffentlicht vom Verband Deutscher Tonmeister e.V.
Bergisch Gladbach, Januar 2019

Redaktion: Carsten Bänfer, Wolfgang Hoeg
Layout: Andrea Versteegh

© 2019 Verband Deutscher Tonmeister e.V., Bergisch Gladbach
Alle Rechte vorbehalten
Made in Germany

ISBN 978-3-9812830-9-9