Florian Wendt

Modeling the Perception of Directional Sound Sources in Reverberant Environments

DOCTORAL THESIS

submitted to

University of Music and Performing Arts Graz

Supervisors

Univ.Prof. Dr.techn. Robert Höldrich

Univ.Prof. Dr.phil. Gerhard Eckel

Univ.Prof. Dr.rer.nat. Torsten Dau (DTU)

Graz, September 2020

This work was partly funded by the Austrian Science Fund (FWF) under grant AR 328-G21,

Orchestrating Space by Icosahedral Loudspeaker (OSIL), PI: Franz Zotter.

Abstract

The perception of sound in rooms is influenced by the room acoustics. Depending on geometrical properties and texture of the room, a direct sound is followed by multiple reflections. For standard surrounding audio reproduction systems, the influence of reflections on the perception is well studied. Recent developments allow more particular constellations and compact loudspeaker arrays with highly pronounced variable directivity patterns that excite wall reflections from a single point in the room to spatialize auditory events. However, their prediction in space mostly fails when standard localization models are used. This is because the underlying psychoacoustic principles are different from those known for standard spatialization systems. This doctoral thesis investigates perceptions elicited by the sound field of a directional sound source in a room. Starting from auditory events evoked by a few precisely controlled sound instances examined in the laboratory, the aim of this work is to understand what perceptions are formed by the interaction of direct sound and its reflections. This bottom-up approach allows the development of models of perception building upon the measurements from the different stages of experimental complexity.

Kurzfassung

Die Wahrnehmung von Schall im Raum wird durch die Raumakustik beeinflusst. Abhängig von den geometrischen Eigenschaften und der Beschaffenheit des Raumes folgen dem Direktschall eine Vielzahl an Reflexionen. Für konventionelle, umhüllende Wiedergabesysteme sind die Einflüsse dieser Reflexionen bereits hinreichend untersucht worden. Die Entwicklung neuartiger kompakter Lausprechersysteme erlaubt die Verräumlichung von Klang ausgehend von einem einzigen Punkt im Raum. Dabei wird das ausgeprägte und variable Richtmuster solcher Systeme dazu verwendet, um Raumreflexionen gezielt anzusprechen und so Hörereignisse im Raum zu erzeugen. Deren Vorhersage unter Verwendung bekannter Lokalisationsmodelle ist jedoch nicht möglich, da die Wahrnehmung solcher Hörereignisse auf anderen psychoakustischen Effekten basiert als jene konventionieller Lausprechersysteme. Diese Doktorarbeit untersucht Wahrnehmungen die sich aus der Interaktion von Direktschall und dessen Reflexionen ergeben. Ausgehend von einfachen Laborexperimenten, in denen die überlagerte Wahrnehmung präzise gesteuerter Schallinstanzen untersucht wird, werden die Hörereignisse in einer Reihe von aufeinander aufbauenden Hörversuchen behandelt. Diese Vorgehensweise erlaubt die Entwicklung von Wahrnehmungsmodellen, die Erkenntnisse aus den verschiedenen Stadien experimenteller Komplexität beinhalten.

Acknowledgments

Throughout the writing of this dissertation I have received a great deal of support and assistance.

I would first like to thank my supervisor, Robert Höldrich, for his patient guidance, enthusiastic encouragement, and useful critiques of this research work. I would also like to thank my external supervisor, Torsten Dau, for the excellent cooperation and for the opportunity to conduct my research and further my dissertation at DTU.

I would like to thank the PEEK/FWF research group OSIL, especially Franz Zotter and Matthias Frank, whose expertises were invaluable in the formulating of the research topic and methodology in particular. Also, a huge thanks to all my colleagues from the Institute of Electronic Music ans Acoustics, who I had so much fun with, in particular all the fantastic people I shared an office with.

To my family, thank you for your love, support, and unwavering belief in me. Without you, I would not be the person I am today.

Above all I would like to thank Gudrun for her love and constant support. Thank you for being my muse, editor, and sounding board. But most of all, thank you for being my best friend.

Erklärung

Hiermit bestätige ich, dass mir der *Leitfaden für schriftliche Arbeiten an der KUG* bekannt ist und ich diese Richtlinien eingehalten habe.

Graz, am	(Unterschrift)

Related Publications

Journal papers

- 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, no. 1, pp. 76–88, 2017.
- 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.
- F. Wendt and R. Höldrich, "Precedence Effect for specular and diffuse reflections," Manuscript in revision, *Acta Acustica*, 2020.

Conference papers

- F. Wendt, M. Frank, F. Zotter, and R. Höldrich, "Directivity patterns controlling the auditory source distance," in *Proceedings of the 19th International Conference on Digital Audio Effects (DAFx-16)*, Brno, 2016.
- F. Wendt, "Investigations on Perceptual Phenomena of the Precedence Effect using a Bessel Sequence," in *Proceedings of the 142th Convention of the Audio Engineering Society*, 2017.
- F. Wendt, M. Frank, and R. Höldrich, "The role of median plane reflections in the perception of vertical auditory movement," in *Proceedings of the 23th International Congress on Acoustics*, 2019.
- F. Wendt, M. Frank, F. Zotter, and R. Höldrich, "Influence of directivity pattern order on perceived distance," in *Fortschritte der Akustik*, 2016.
- F. Wendt, R. Höldrich, and M. Frank, "The Influence of the Floor Reflection on the Perception of Sound Elevation," in *Fortschritte der Akustik*, 2017.
- J. Linke, F. Wendt, F. Zotter, and M. Frank, "How Masking affects Auditory Objects of Beamformed Sounds," in *Fortschritte der Akusitk*, 2018.
- J. Linke, F. Wendt, M. Frank, and F. Zotter, "How the perception of moving sound beams is influenced by masking and reflector setup," in *Proceedings of the 30th Tonmeistertagung*. Köln: Verband Deutscher Tonmeister, 2018.
- F. Wendt and M. Frank, "On the localization of auditory objects created by directional sound sources in a virtual room," in *Proceedings of the 30th Tonmeistertagung*. Köln: Verband Deutscher Tonmeister, 2018.

- F. Wendt and R. Höldrich, "Reflection properties influencing the precedence effect," in *Fortschritte der Akusitk*, 2018.
- K. Wegler, F. Wendt, and R. Höldrich, "How level, delay, and spatial separation influence the echo threshold," in *Fortschritte der Akusitk*, Rostock, 2019.
- F. Wendt, R. Höldrich, and M. Marschall, "How binaural room impulse responses influence the externalization of speech," in *Fortschritte der Akusitk*, 2019.
- P. M. Giller, F. Wendt, and R. Höldrich, "The influence of different BRIR modification techniques on externalization and sound quality," in *Proceedings on the Spatial Audio Signal Processing Symposium*, Paris, 2019.

Contents

1 Introduction

Spatial audio theory offers today numerous different methodologies for producing immersive auditory environments. Methods like Ambisonics or the Wavefield Synthesis (WFS) aim at recreating an exact reconstruction of the physical sound field, whereas perceptually-optimized methods such as Vector-Base Amplitude panning (VBAP) provide a more practical solution to spatial sound rendering and reproduction. While both approaches can provide satisfying reproduction of a desired sound field, they suffer from practical constraints as they require a large number of loudspeakers regularly placed around the listening area which is impractical for many applications.

Over the last years, the use of beamforming systems as spatial audio technique has been given increased interest. Sound from various directions is reproduced by projection of so-called *sound beams* on reflective boundaries, e.g., walls, using their variable directivity. In this way, the perceived direction of the auditory event changes from the direction of the physical array to the direction of the reflection, i.e., the projected source. Possibilities of such application are as manifold as the number of realizations of such systems. In 2005 Yamaha [Tak05] introduced the so-called *Digital Sound Projector*, the first commercially available beamforming soundbar that aims at recreating realistic auditory space for the home cinema. Recently, compact smart speakers such as Amazon Echo, Google Home, or Apple HomePod became popular in households and use beamforming technology to spatialize the intelligent personal assistant in acoustic demanding scenarios.

Long before domestic purpose, sonic arts discovered beamforming to create immersive sound environments. The notion of adjustable directivity loudspeakers was introduced in the late 1980s by researchers at IRCAM. For the renowned concept study *La Timèe* [WDC97], a cube housing six separately controlled loudspeakers was built to achieve freely controllable directivity. Since then, many other research groups established compact spherical loudspeaker arrays and employed them for electroacoustic performances, cf. [ZZFK17].

Since 2006, Zotter from IEM reconsidered classical beamforming technology by applying Ambisonics on a compact loudspeaker array, called IKO, yielding controllable directivity patterns [Zot09]. The research project *Orchestrating Space by Icosahedral Loudspeaker Array* (OSIL) dealt with the IKO as a new electroacoustic musical instrument. Pierre Boulez' remark "Le haut-parleur anonymise la source réelle" inspired the working hypothesis stating that if loudspeakers with their own directivity alienate natural sounds, spherical beamforming trajectories of the

IKO can acousmatically naturalize alien sounds in the surrounding room. In fact, IKO's sound beams involve the surrounding space and evoke constructs consisting of several static or dynamic *auditory events* that appear to be projected holographically. According to Blauert [Bla97], auditory events are "caused, determined, or elicited by (physical) sound events". Hence, both auditory events and sound events are related to each other and under certain conditions they occur with one another. The mapping of the location of an auditory event to certain characteristics of a sound event is termed *localization*. The auditory system uses several cues for auditory event localization which are evaluated in comparison with the position of other objects of perception which might be auditory events or the objects of other senses–in particular those of vision.

The aim of this thesis is the psychoacoustic characterization and quantization of auditory events created by directional sound sources like the IKO. Their examination is approached as simple and consistent as possible starting from psychoacoustic measurements in the anechoic laboratory, where the perception of a few precisely controlled sound instances is studied. By examining the localization of auditory events elicited by a directional sound source in a room, it is investigated which knowledge from the laboratory can be applied to more complex scenarios, where many psychoacoustic effects co-occur, e.g., precedence, masking, auditory grouping, etc. Throughout the evaluation, conceptual models based on acoustic measures of the sound field are build to understand the spatial perception of a mixture of multiple, partly concurrent sounds. The knowledge from the models is regarded in the excitation of beamformers to control the spatialization of sound from a single beamforming source in the room. Moreover, the models are used to develop a perceptually motivated sound field virtualization that allows a plausible reproduction of directional sound sources. In electroacoustic music, in which beamformers like the IKO are used as instruments, the insights gained from the models help to further develop their artistic means of expression.

1.1 Perception of sound in space

Almost a century ago, Blumlein [Blu31] patented the first spatial audio recording and reproduction technique. *Stereophony* plays back the same signal over two loudspeakers placed symmetric to the median plane of the listener and yields an auditory event between the arrangement. Control of spatialization, i.e., panning the auditory event between the loudspeakers, is achieved either by the relation of the loudspeaker gains or the time delays. The psychoacoustic principle explaining the perception of a single static auditory event that depends on two (or more) coherent sound waves arriving at the ears within the time interval of about < 1 ms is termed *summing localization*. It assumes that the localization cues elicit by the superimposed sound field are similar to the cues produced by a single real source. Thus, the listener perceives a single auditory event at the location of this virtually equivalent single sound source [Wen63].

The above-mentioned static cues are interaural differences of level and time (ILD and ITD), which have been identified to be most relevant for localization of single sound sources in the horizontal plane and are encoded in the head-related transfer function (HRTF), cf. Figure 1.1. For the localization of a single elevated sound source these cues are ambiguous, as ILD and ITD values are largely the same as those of a sound source on the horizon. As auditory perception along a so-called *cone of confusion*, cf. Figure 1.2, a 2D-surface with constant binaural differences [Bla97], is still possible, other cues, predominantly spectral properties of the HRTFs, play an important role complementing the interaural differences.

Pairwise amplitude panning became the standard spatialization technique and adapted versions of stereophony have been developed by adding additional loudspeakers to enable more stable horizontal panning around the listeners, e.g., quadraphonic sound or 5.1. In the early 2000s different attempts were made to extend stereophony to three dimensions enabling the creation of elevated auditory events. Although the motivation of loudspeaker placement differs, all these spatialization systems like Vector-Base Amplitude Panning (VBAP) [Pul97], Dolby Atmos, AURO-3D, and 22.2 [HHO05] are based on the assumption that, similar to the horizontal plane, summing localization fuses sound instances of vertical loudspeaker constellations to clearly localized auditory events. Given that spectral cues of superimposed HRTFs of vertical loudspeaker directions need not resemble the HRTF of the single source in between, one might doubt

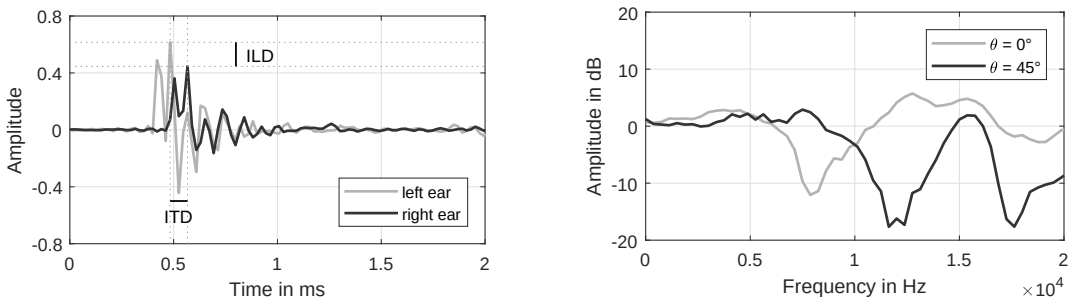


Figure 1.1: Localization cues encoded in the HRTF. Left: Interaural cues of a measurement taken at azimuth $\phi = 10°$ and elevation $\theta = 0°$. Right: Comparison of spectral information in the left ear for different elevation angles θ but the same azimuth angle $\phi = 10°$.

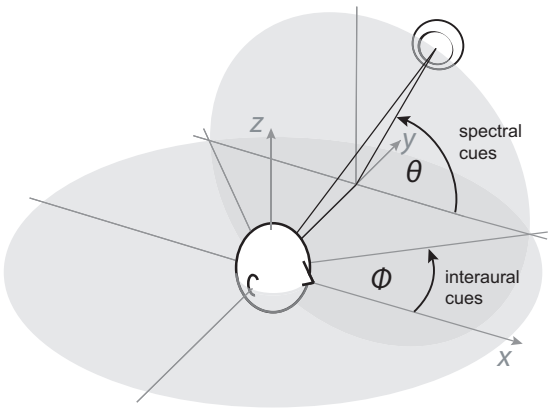


Figure 1.2: Schematic representation of the two localization concepts: interaural cues provide information on the horizontal location of the sound sources. The azimuth angle ϕ defines the cone of confusion, on which the difference in distance from both ears to any point on the cone is constant. For resolving this ambiguity of elevation θ, monaural spectral cues are processed. This definition of the coordinate system is used throughout the thesis.

this. However, studies found that the perception of a single fused auditory event is possible and can be described by weighted averaging of vertical loudspeaker directions [Pul01,BVV10,WFZ14].

A completely different approach to spatialize sound dates back to the early 1970s, were Gerzon [Ger73] came up with Ambisonics. This method differs from stereophony as it is based on sound field decomposition using spherical harmonics and aims at the reconstruction of a physical sound field at the listening point. The number of available loudspeakers for sound field reproduction and their arrangement define the Ambisonics order, which determines the number of used spherical harmonics.

A related spatial audio technique is WFS, initially formulated by Berkhout et al. [Ber88] in the late 1980s. It aims at recreating the sound field based on the Huygens–Fresnel principle, which states that any wavefront can be regarded as a superposition of elementary spherical waves. In theory with both an infinite number spherical harmonics for Ambisonics or infinite number of spherical waves for WFS the physical sound field can be perfectly recreated.

The drawback of all afore mentioned panning-based systems including Ambisonics is the limited sweet area size as they require the amplitude weighted loudspeakers signals arriving at the listening position to have a certain phase relation to spatialize auditory events.

Differing distances of loudspeakers to the ears at off-center listening positions distort the coherence. Small deviation from the central listening position still yield summing localization, perceived as slight distortions in the localization, as long as sounds are arriving with time lags $< 1\,\text{ms}$. However, for major deviations of the listening position, the auditory image collapses to the closest loudspeaker. The underlying psychoacoustic effect is known as the *law of the first wave front* [Cre48], an aspect of the precedence effect. It describes the process by which the spatial information of later arriving sound waves is suppressed and localization is dominated by the first-arriving wave front.

1.2 Beamformer: the icosahedral loudspeaker array

The directional sound source discussed in this thesis is a 20-sided, 20-channel loudspeaker array, in the from of a regular icosahedron. This compact spherical loudspeaker array, called IKO, is able to project strongly focused sound beams in any direction. In contrast to classical beamformers, which use different weights, delays, or filters to drive their transducers, the technology utilized for spherical loudspeaker arrays such as the IKO is called *spherical harmonic beamforming* [ZZFK17]. The technology is based on filtering to equalize different attenuation for spherical harmonics of different order yielding a uniform and direction-independent directivity.

Over the years, Zotter and colleagues worked on achieving a narrow beam that maintains its consistent shape over a broad frequency range [ZPS08]. The beamforming is based on the so-called max-$\boldsymbol{r}_\mathrm{E}$ weighting of the Ambisonics signals [DRP98], which exhibits sufficiently high side-lobe attenuation while maintaining a narrow main lobe. Directivities are described by:

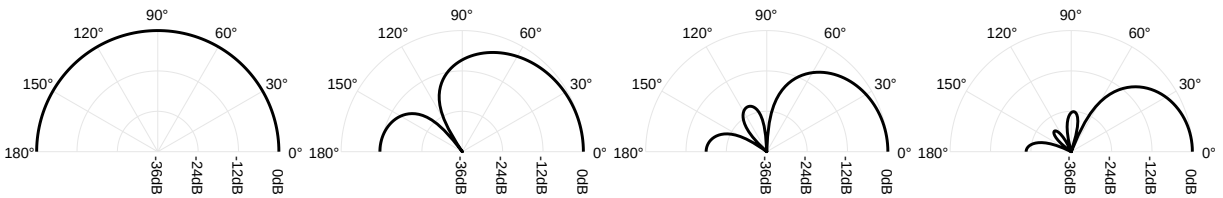
$$g_i(\boldsymbol{\theta}) = \sum_{N}^{n=0} \sum_{m=-n}^{n} Y_n^m(\boldsymbol{\theta}) w_{n,i} Y_n^m(\boldsymbol{\theta}_c), \tag{1.1}$$

with $Y_n^m(\boldsymbol{\theta})$ representing the fully orthonormal spherical harmonics and the two direction vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_c$ the direction of radiation and adjustable beam direction, respectively. The controllable Ambisonics order is considered by N with an upper limit given by the number of $(N+1)^2$ channels. The max-$\boldsymbol{r}_\mathrm{E}$ order weights are defined as:

$$w_{i,n} = \begin{cases} \dfrac{P_n[\cos(\frac{137.9^\circ}{i+1.151})]}{\sum_{n=0}^{i}(2n+1)\, P_n(\cos\frac{137.9^\circ}{i+1.51}))}, & \text{for } 0 \le n \le 1 \\ 0, & \text{elsewhere,} \end{cases} \tag{1.2}$$

with the nth-order Legendre polynomials $P_n(\cdot)$ [ZF12]. On-axis equalized max-$\boldsymbol{r}_\mathrm{E}$ beams of the orders $i \le N$ are shown in Figure 1.3.

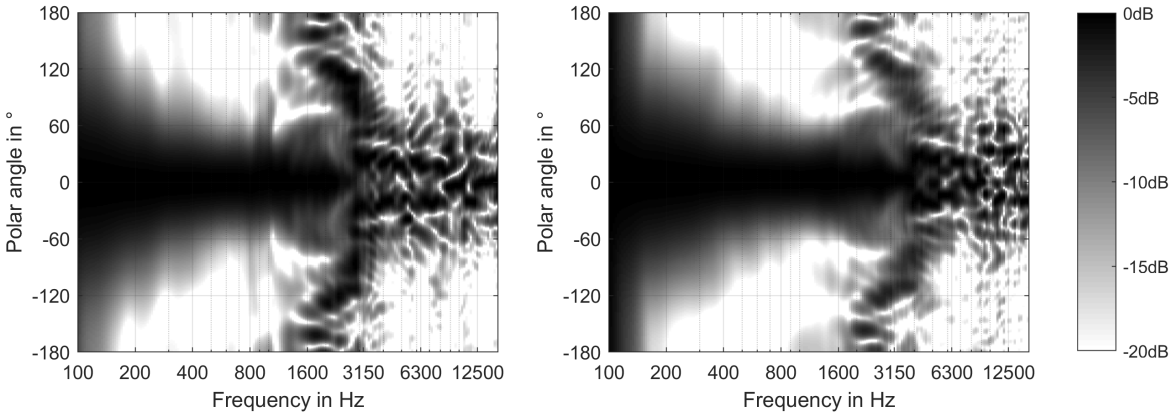
For the lower frequency range it has been shown that the higher-order directivity pattern synthesis is difficult to accomplish. More elaborated systems were developed, including radiation control filters, acoustic crosstalk cancellation, and a low-frequency amendment by crossing over to an omnidirectional subwoofer mode [Lös14, ZZFK17].



(a) $i = 0$ (b) $i = 1$ (c) $i = 2$ (d) $i = 3$

Figure 1.3: Spherical harmonic max-$\boldsymbol{r}_\mathrm{E}$ beam patterns plotted as $|g_i(\boldsymbol{\theta})|$ of orders i yielding rotationally symmetrical directivity patterns.

Figure 1.4: The icosahedral loudspeaker array consists of 20 loudspeakers, each placed on a equiangular face. Left: IKO_1, the prototype developed at the IEM with a triangular edge length of 34.5 cm; Right: commercially available IKO_2 with a triangular edge length of 28.8 cm, manufactured by Sonible.



(a) IKO_1 (b) IKO_2

Figure 1.5: Horizontal cross-sections through measured 3rd-order max-$\boldsymbol{r}_\mathrm{E}$ beampatterns of IKO_1 and IKO_2 normalized to its maximum over the polar angle. Magnitude levels indicated as levels of gray.

Throughout this thesis two different versions of the IKO are used. The IKO_1, shown in the left panel of Figure 1.4, is the prototype of the icosahedral loudspeaker array developed at the IEM in 2006. It consists of twenty 6.3-inch transducers driven by a multi-channel amplifier with 250W for each channel. In 2017 the technology was made commercial in a collaboration with the company Sonible[1]. The IKO by IEM and Sonible, hereafter called IKO_2 and shown in the right panel of Figure 1.4, was redesigned for easier transport and for easy integration with the MADI/Dante-capable sonible d:24 multichannel amplifier driving the slightly smaller 6-inch transducers. Measured 3rd-order beam patterns for IKO_1 and IKO_2 are given in Figure 1.5.

The directivity of the IKO can be controlled in real time and the 20 driving signals for the loudspeakers are generated using the ambiX plug-in suite [Kro14]. Firstly, a source signal is encoded using the ambiX encoder, then converted using ambiX converter, and lastly filtered according to [ZZFK17] using mcfx convolver. This allows one to create, modify, and decode higher-order Ambisonics on a DAW running on a desktop computer or laptop.

[1] `https://www.sonible.com`

1.3 Spatialization on beamforming loudspeaker arrays

In contrast to surrounding spatialization systems that play sound from the outside of the listening area into the audience, spatialization on beamforming loudspeaker arrays like the IKO is achieved from a single point in the room. Even if the beamforming is capable of uniform adjustment to all directions, the downside of this approach is that contiguous sound beam directions are not mapped to contiguous perceived directions, as reflection paths of a room are discrete. Thus, employing a directional sound source to "orchestrate" reflecting surfaces to spatialize auditory events at distinct locations in the room relies on the availability of reflectors and it is believed that spatialization is only precise if the reflections can be precisely controlled.

Moreover, due to physical limitations, the directivity of even highly focused beamformers is not pronounced enough to completely attenuate the sound radiated towards the direction of the listener. As a consequence the listeners percept is not only formed by reflections, but also by the direct sound of the sound source. Figure 1.6 compares gain relations, angles, and timing of direct sound and up to 3rd order specular reflections at the listeners position for a different sound source directivities.

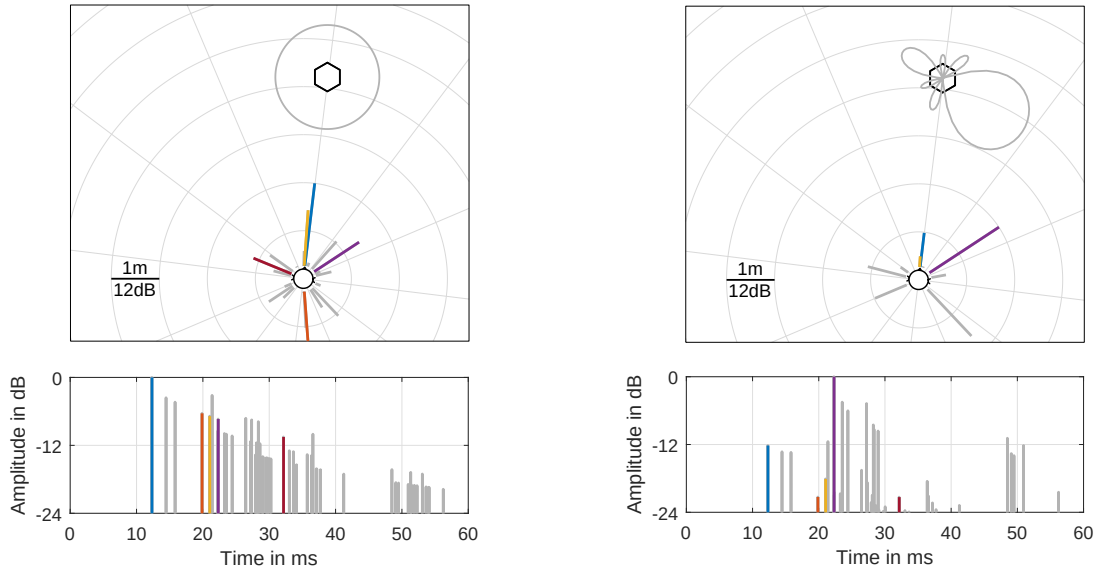


Figure 1.6: Angular and temporal representation of the reflectogram for a sound source with an omnidirectional directivity (left) and with an ideal 3rd-order directivity projecting its sound beams on right lateral wall (right). Direct sound and azimuthal first order reflections are color coded.

The impulse response obtained with an omnidirectional directivity shown in the left panel of Figure 1.6 represents a common reflection pattern. Compared to the direct sound, reflections are attenuated and delayed by more than 1 ms. Regarding horizontal localization, a property of auditory perception that resolves the competition for perception occurring between direct sound and a reflections is the *precedence effect* [WNR49]. Comprehensive studies could show that in many cases listeners localize a fused auditory event at the location cued by the first-arriving sound, e.g., [RH85, FCL91].

Projecting a higher-order directivity beam to a reflective wall (cf. right panel of Figure 1.6) yields a pronounced reflection containing much more intensity compared to the direct sound. Unlike environmental reflections, the perceived image for such a configuration will be shifted back in the direction of the more intense signal [Haa72]. Zotter and Frank [ZFFR14, ZF15] could show that the orientation of a directional sound source can be heard and it causes lateralizations deflected from the direct path. Their results indicate that, in most cases, the auditory event location is determined by the loudest unmasked acoustic reflection path. Recently, Wühle et al. [WMA19] investigated the localization dominance of a lagging reflection more closely by establishing level thresholds that beamformers should provide in isolating distinct reflection paths from the direct path.

Such level thresholds are often only accomplished for a limited set of the available reflection paths. Nevertheless, localization is not necessarily bimodal and studies could show that the auditory event can appear in the angular range spanned by the directions of direct sound and reflection [PB15, ZFFR14].

1.4 Organization of contents

Spatialization on beamforming systems is not necessarily restricted to directions of direct sound and the few prominent reflections but similar to standard amplitude panning techniques auditory events can be moved between these distinct directions. Chapter 2 of this thesis focuses on the lateralization of auditory events created by directional sound sources by examining psychoacoustic phenomena related to the precedence effect.

To complete for an exact location of auditory events in the two-dimensional space, spatial rendering systems should be able to produce distance effects. Although auditory distance is the most common localization-specific attribute [Mas17], it receives substantially less scientific attention compared to the horizontal direction. As a consequence, perception-based amplitude panning systems mostly use a simple gain scaling to create distance effects. In the free field loudness is the most prominent distance cue, but in closed environments other cues, such as the ratio of direct to reverberant sound energy are more important [ZBB05]. Chapter 3 studies how the directivity pattern of the directional sound source can be used to influences the distance perception.

In contrast to the horizontal plane, asynchrony effects in the perception of height play a minor role. Although some studies claim the existence of a vertical precedence effect, e.g., [LRYH97], recent findings suggest that fused auditory images, elicit if delayed reflections from the same sagittal plane superimpose with the direct sound, are due to backward masking [EOBW18]. For most playback situation there are only two distinct vertical reflection directions available, i.e., from floor and ceiling, which makes an exact positioning of auditory events difficult. Anyway,

in comparison to the sound's azimuth our performance in elevation localization is weak. Thus, requirements for three-dimensional spatialization systems regarding elevation are mostly limited to the perceived movement direction, i.e., does the auditory event move upwards or downwards, rather than an exact positioning. Chapter 4 focuses on possible asynchrony effects in the perception of height and how they can be incorporated in spatialization systems.

The setup of the directional sound source within a room is essential as its position and orientation defines the available reflections' paths. To allow the creation of similar auditory events in different acoustic environments, beam adjustments are mandatory which typically require a reasonable amount of time in the specific venue. Off-line adjustments are enabled by a sound field virtualization aiming at a synthesis of the main spatial properties of auditory events in three dimensions studied in previous chapters. Chapter 5 examines how these properties can be relayed from the original space to a binaural reproduction phase to allow a plausible or even authentic perception.

In course of the dissertation a total of twelve listening experiments were carried out. All experiments were published as conference or journal papers, and the subsequent chapters are based on these publications. For this thesis, however, most of the published findings have been adapted to put them in context to the other experiments. The dataset of each experiment was made available online and can be downloaded from `https://phaidra.kug.ac.at/o:107379`.

2

The Precedence Effect affecting Lateralization

The precedence effect is thought to be involved in resolving competition for perception and localization between the direct sound and its reflections. Thus, it is important for our ability to guess the lateralization of sound sources in acoustically complex environments. At the sounds onset, localization cues for a static sound source are the same as for a source in the free field. However, as reflections join in, interaural cues start to vary and conflict with the cues carried by the onset. To maintain localization, the auditory systems differentially weights the spatial cues over the stimulus duration. Reliable onset cues are emphasized, while less reliable cues of the ongoing sound are de-emphasized.

Outline This chapter is structured as follows. Section 2.1 presents different phenomena of the precedence effect by introducing corresponding measures. It provides an overview of the relevant literature in the context of directional sound sources and outlines open topics, which are studied in 4 listening experiments in subsequent sections. Section 2.2 presents a listening experiment to study the relative importance of reflection's delay, attenuation, and angle on the precedence effect. Section 2.3 presents a model for simulating diffuse reflections and studies its perceptual influence on the precedence effect. Section 2.4 studies the influence of multiple emphasized reflections on lateralization. Section 2.5 applies the IKO for spatialization and models obtained lateralizations. Finally, Section 2.6 summarizes the chapter.

2.1 Measures and parameters defining the precedence effect

The precedence effect has been extensively studied for more than half a century, in literature. Respective studies typically utilize a two-source paradigm that is carried out as so-called lead-lag experiment, representing direct sound and specular reflection at the same intensity. In most precedence studies, the parameter of interest is the lead-lag delay ΔT. For very short delays between direct sound and reflection of < 1 ms, *summing localization* occurs and the two sounds appear as a single fused image located between the speakers. As the delay exceeds the limits for summing localization, the precedence effect becomes active. For delays just slightly longer than 1 ms *fusion* is still maintained and the lead dominates the perceived location of the single fused auditory event. This state of *echo suppression* is called *localization dominance*. Another increase of the delay yields widening of the auditory event until it breaks apart at the *echo threshold* (ET), and separate auditory events are heard at the locations of lead and lag. However, also beyond this critical threshold of the precedence effect, the lead dominates the overall auditory event and the lag is only barely audible. The lack of sensitivity to features of the lagging signal, most especially its location, is termed *discrimination suppression*. Figure 2.1 gives an overview of temporal dynamics of the precedence effect.

The active range of summing localization has been studied extensively by listening tests yielding an inter-subjective upper limit of around 1 ms, e.g., [WT02]. Although obtained localization curves vary with experiments, the inter-subjective variation in the perceived direction of the phantom source within an experiment is relatively low. Contrastingly, measures of precedence such as fusion, echo suppression, localization dominance, and discrimination suppression are found to be highly subjective and, in addition, appear to depend on the listening situation and under some conditions they diverge substantially.

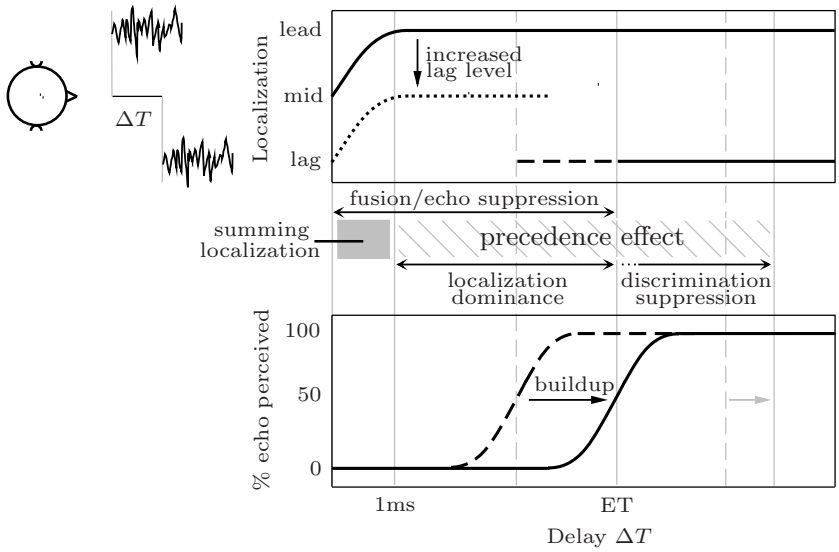


Figure 2.1: Schematic representation of temporal dynamics of the transient precedence effect found in lead-lag experiments as a function of the inter-stimulus delay ΔT. The upper panel shows the localization curve and the lower panel the corresponding psychometric function for echo perception. The effect of a repetitive representation, i.e., buildup of precedence, is shown by dashed lines and the effect of an increased lag level by the dotted line.

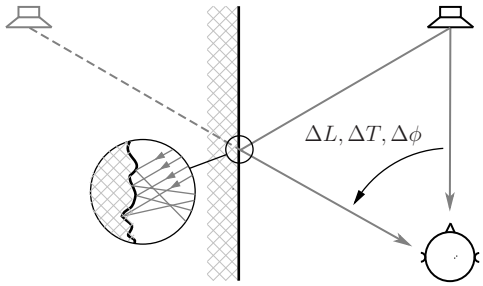


Figure 2.2: Illustration of a typical setup used to study the echo threshold. Compared to the direct sound, the specular reflection arriving at the listener is defined by the image source determining the delay ΔT, intensity difference ΔL, and the directional separation $\Delta\phi$. The zoomed wall area illustrates a possible scattering of the reflected sound.

Experiments studying the strength of the precedence effect typically measure the echo threshold as upper bound of both echo suppression and localization dominance. This is done by a variation of the lead-lag delay ΔT. Lowest echo threshold delays of 5 to 10 ms are found for transient clicks. For signals with longer onset times, such as running speech or music, the delay has to exceed 50 ms before the echo is perceived [Haa72]. Thus, the perception strongly depends on the excitation signal and depending on its onset time and duration, different types of precedence effect are identified in literature. Freyman et al. in [FGZ14] introduce a categorization of three different precedence effect types based on the lead and lags signal type. The *transient* precedence effect occurs if a leading brief transient dominates the lateralization of the lagging transient. Long duration stimuli with onset durations exceeding 100 ms on the other hand also elicit the precedence effect. In this case, it is the interaural cue of the ongoing lead component that controls lateralization, which is called *ongoing* precedence effect. A third precedence phenomenon is observed in lateralization studies, when the transient onset controls the perceived localization of the ongoing sound. This *onset capture* effect only occurs when interaural cues of the ongoing sound are ambiguous.

In addition to the signal envelope, the precedence effect depends on the number of times the lead-lag pair is presented. A repetitive representation yields *buildup* of precedence and results in a significant increase of the echo threshold delay. After presenting the two sound instances several times, the lag becomes less salient and perceptually fades away [Cli87, FCL91, DB01]. Similarly, albeit weaker, the lead-lag repetition increases the discrimination suppression threshold [YG97], cf. Figure 2.1.

Alternatively to varying the lead-lag delay ΔT, fewer studies on the precedence effect manipulate the level of either lead or lag. However, in a constellation of sound source, receiver, and reflective wall, cf. Figure 2.2, these two parameters are closely linked because the distance traveled by the reflection determines not just the delay ΔT, but also the inverse-square law intensity difference ΔL of the reflection compared to the direct sound. The echo threshold is then defined as the intensity difference ΔL for a fixed delay ΔT [DK86, RHH00]. Thus, a stronger

precedence effect, i.e., a stronger echo suppression, is implied by a longer echo threshold delay or a higher echo threshold level. Regarding localization, these studies found that lateralization shifts towards the lag as the level of the lag exceeds the lead [XSC11, PB15]. More specific, Pastore and Braasch [PB15] studied the localization of a fused auditory image as a function of lead-lag delay and intensity. In the context of spatialization on beamforming systems this is especially relevant, as a trade off between delay and level has to be solved in the setup of the beamformer: a longer delay ΔT decreases the perceptually needed level of the refection to be audible but, at the same time, decreases its physically achievable level due to the increased propagation attenuation. However, in the data of [PB15] indications are found that the intensity might be the more prominent parameter compared to the delay and a more pronounced but temporally closer reflection is preferred to a less a pronounced one that arrives later in time. The third parameter linked to the setup shown in Figure 2.2 which is known to influence the precedence effect is the directional separation $\Delta\phi$. Assuming a specular reflection, the direction of incidence is derived from the corresponding image source and respective studies could show that the echo suppression is higher if direct sound and reflection arise from similar directions than when they are spatially separated [Sep61, San76, SCZD93, RWFB13, GvdPT17]. Section 2.2 studies the contribution of the reflection's delay, level, and directional separation on the strength of the precedence effect by measuring the echo threshold in a listening experiment.

Except for the few studies that focus on the directional separation, most of the previously cited literature uses headphone ITD or ILD stimuli. These paradigms are rather artificial, as stimuli of the real word are always a combination of both. More realistic free-field conditions are created by filtering stimuli with respective HRTFs for the desired lead and lag virtual locations or by loudspeakers playback of the sound instances. In this way both ITD and ILD cues as well as spectral shape cues in their natural combinations are regarded. Studies employing real wall panels for creating the lagging sound instance are rare and apart from a handful of contributions, e.g., [RH85, RH86, Gus90], the reflection is mostly simulated as a synthesized copy of the direct sound, cf. [BST15]. Such a *specular* reflection occurs on an infinitely large, smooth, and rigid wall [Kut09]. Its impulse response is described by a Dirac delta distribution with the delay compared to the direct sound calculated from the distance between its image source and the receiver.

Obviously, the surface of a real-life wall is not completely smooth and sound is inevitably scattered into angles other than the specular reflection angle, cf. Figure 2.2. This diffusion yields a directional and temporal widening of the reflected sound. Regarding perception, the temporal diffusion yields spectral colorations of the reflection. According to Blauert and Dinveyi [BD88] suppression occurs as long as the spectrum of the lag contains exclusively those regions that are present in the spectrum of the lead. However, for reflections differing in magnitude spectrum and phase spectrum, the suppression was found to be generally weaker compared to a specular reflection, e.g., [PSM87, BD88, WRS13]. Section 2.3 presents a theoretical model for a diffuse reflection yielding a directional and temporal diffusion of the reflected sound and the influence of diffuse reflections on the precedence effect's strength is quantified in a listening experiment.

All previously cited literature offers comprehensive insight into how the precedence effect potentially influence the perception of directional sound sources, but it is limited to single reflections. And yet, the sound field of closed environments will always consist of many reflections, and with directional sound sources some of them might be stronger than the direct sound. Section 2.4 studies the influence of multiple reflections with increased level on the lateralization of sound with different frequency content and determines where knowledge from lead-lag experiments is applicable.

Finally, the last experiment in Section 2.5 applies the IKO with its highly focused directivity and studies the perception of lateralized auditory events. Knowledge obtained from previous sections is used to fit a localization model which considers the precedence effect.

2.2 The echo threshold as a function of constellation parameters

According to the definition introduced in the previous section, the echo threshold represents the transition from one to two auditory events. This transition is everything but distinctive; the presence of reflections is already audible before the transition of summing localization to precedence ends, in particular in terms of a change in timbre, blurriness, or width of the auditory event. Thus, in addition to the stimulus parameters, the echo threshold critically depends on the instructions and tasks given to listeners. Blauert in [Bla97] asked the listeners for "the shortest delay time at which a second auditory event becomes audible". Another echo threshold definition measures the delay for "a second sound heard at the vicinity of the lag speaker", e.g., [FCL91], yielding higher thresholds. Further increases of the echo threshold are obtained if the definition asks for a "reflection that is annoying", e.g., [DK86], or for a "reflection that is equally loud as the primary auditory event", e.g., [Haa72]. Hence, quantitative estimates of the echo threshold vary tremendously, depending on the definition. A common feature, on the other hand, is their inherent subjective nature and estimates of the echo threshold depend on individual subject factors. Conservative listeners for example tend to yield higher thresholds compared to more liberal listeners. At the lower bound of the range, spanned by different definitions of the echo threshold, the reflection is not heard at all. The *masked threshold* is the absolute threshold of perceptibility of the reflection, based on criterion of a change in the auditory event without specification of the type of change.

The constellation of the directional sound source and the listener within a room plays a significant effect on the sound propagation paths. The matter of the investigation in this section is to determine the influence of reflection parameters such as the reflection's delay time, intensity, and direction. The investigation studies the echo threshold of the transient precedence effect measured in a lead-lag experiment. Experiment 1 is confined to the direct sound, presented from the front followed by a single specular reflection, and consists of two consecutive parts. The author conceived and designed the experiments, which were conducted within the context of a master's thesis by Korbinian Wegler. The analysis of the data was done by the author and parts of it have already been published in [WWH19, Weg20]

2.2.1 The influence of directional separation

The first listening experiment was originally intended to examine if (similarly to delay and level) the echo threshold can be measured by varying solely the reflection's direction, while keeping constant all other parameters. This would extend Blauert's definition of the echo threshold towards "shortest delay, highest level difference, or smallest directional separation at which a second auditory event becomes audible". However, in an informal listening test where listeners varied solely the direction of the reflection with constant delay and level, it was not possible to measure the echo threshold as the echo was perceived either all the time or never.

Thus, Part 1 of Experiment 1 studies significant ranges of delay and directional separation for the echo threshold, defined as the level difference ΔL as a function of the delay and direction. Listeners are given control over the reflection's level starting from $\Delta L = -50\,\mathrm{dB}$ relative to the direct sound, with the task to adjust it until an echo is just barely audible as distinct auditory event. The direct sound is always simulated directly in front of the listener at $\phi = 0°$, whereas reflection angles and delays are varied for the investigated conditions. Fourteen equally sampled directional separations $\Delta\phi = 0° \ldots 90°$ (left) and $\Delta\phi = 0° \ldots 90°$ (right), each with three delays $\Delta T = (20, 40, 60)\,\mathrm{ms}$, are tested, yielding 42 conditions. An ongoing sequence of 10 ms-long white noise bursts (instant on- and offset) with a period of 300 ms is used as stimulus. For a precise controllability of the directions, the stimulus is filtered with the HRTF for the desired lead and lag virtual locations of the Neumann KU100 dummy head [Ber13] and presented over Beyerdynamic DT770 pro headphones in the anechoic laboratory of the IEM. The level of the direct sound is fixed with $L = 68\,\mathrm{dB(A)}$, except for 2 s at the beginning of each condition, when the level of the stimuli sequence is linearly faded in to support the buildup of the precedence effect.

Experimental results. As implied previously, the criterion for hearing an echo can be highly subjective. Accordingly, the range of individual echo thresholds collected from thirteen experienced listeners spreads remarkably and reaches up to 30 dB for the same condition. The primary aim of this experiment is to determine the influence of constellation parameters on the strength of the precedence effect rather than to estimate specific echo threshold values. Thus, for the analysis of the echo threshold levels, each individual threshold value ΔL_i in dB is corrected with respect to the individual mean threshold $\Delta\bar{L}_i$ over all conditions and the all-subject mean threshold $\Delta\bar{L}_{\mathrm{all}}$:

$$\Delta L_{i,\mathrm{corr}} = \Delta L_i - \Delta\bar{L}_i + \Delta\bar{L}_{\mathrm{all}}. \tag{2.1}$$

The data follows a normal distribution (Lilliefors, $p > 0.05$) and a statistical analysis reveals a symmetric response behavior between left and right, which is not a significant parameter[2] (t-test: $p > 0.05$), and the obtained levels ΔL_{corr}, pooled from left and right are given in Figure 2.3 as means and corresponding 95% confidence intervals. The echo thresholds ΔL progressively decrease with increasing delay ΔT and resemble those obtained by similar studies, e.g., [RHH00]. An analysis of variance (ANOVA) reveals the delay to be significant ($p \leq 0.05$) for all 7 reflection directions. The reflection direction $\Delta\phi$ on the other hand does not yield monotone trends. Although it is a significant parameter for $\Delta T = 40\,\mathrm{ms}$ and $\Delta T = 60\,\mathrm{ms}$ ($p \leq 0.05$), Tukey's HSD post hoc analysis reveals that there are only 2 significant groups: $\Delta\phi = 0°$ and $\Delta\phi \geq 30°$ with an effect size expressed as Cohen's $d \geq 0.46$, i.e., medium effect [Saw09]. For $\Delta T = 20\,\mathrm{ms}$, no significance is obtained. These results confirm the insights from the informal experiment and measuring the echo threshold by directly adjusting the direction $\Delta\phi$ is not possible.

[2] Statistical significance is determined if $p \leq 0.05$ and effect sizes inform about the magnitude of the significant effect.

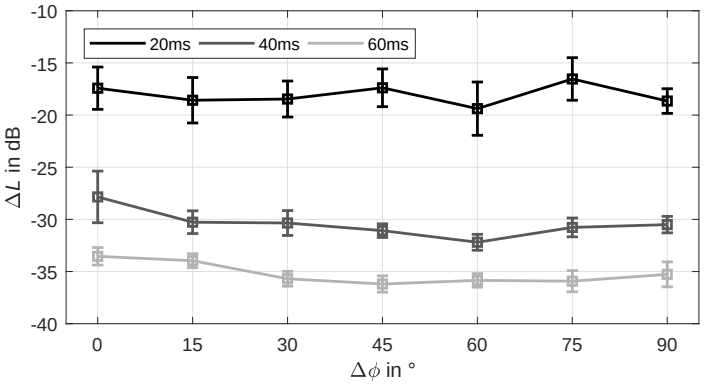


Figure 2.3: Results of the Part 1 of Experiment 1. Mean and 95% confidence intervals of echo threshold levels ΔL corrected with respect to the individual mean threshold and plotted over directional separations $\Delta\phi$ for delays ΔT.

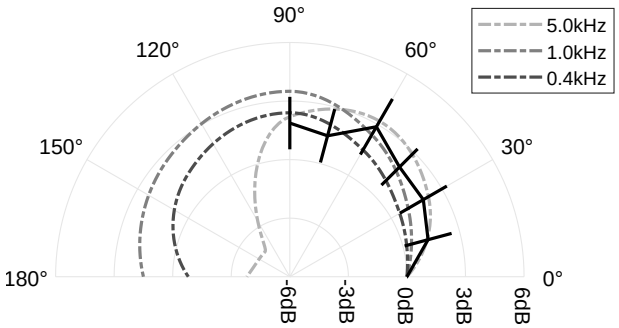


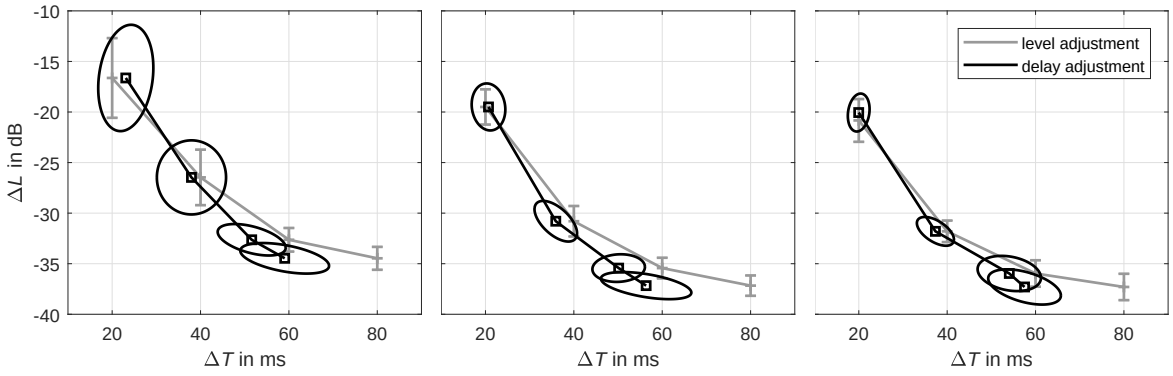
Figure 2.4: Directional mean echo thresholds and corresponding 95% confidence intervals normalized using Eq. (2.2) (solid) and mean directional loudness sensitivities from [SE06] for three different third-octave noise bands (dashed).

Modeling the echo threshold. The influence of directional separation is modeled by the direction-dependent sensitivity of the human ear. Sivonen and Ellermeir [SE06] measured the directivity of the ear by matching the loudness of sound sources at different angles with a frontal reference. To allow comparability of the data, for each delay ΔT and listener, direction-dependent echo threshold levels $\Delta L(\Delta\phi)$ in dB are normalized to the frontal direction with:

$$\Delta L_{\text{norm}}(\Delta\phi) = \Delta L(0°) - \Delta L(\Delta\phi). \tag{2.2}$$

Figure 2.4 compares mean values from [SE06] for third-octave noise bands centered at 0.4, 1.0, and 5.0 kHz with normalized echo thresholds given as means and 95% confidence intervals over all listeners and delays.

Almost all confidence intervals of the echo threshold intersect mean directional loudness sensitivities of all noise bands and it can be assumed that the influence of directional separation on the echo threshold is due to the directional sensitivity of the auditory system. However, in contrast to the rear hemisphere, there is just little frequency-variation in the directivity for the frontal hemisphere and further investigations are needed to prove this assumption.



(a) $\Delta\phi = 0°$ (b) $\Delta\phi = 30°$ (c) $\Delta\phi = 60°$

Figure 2.5: Results of Part 2 of Experiment 1 for directional separations $\Delta\phi$. Means and corresponding 95% confidence intervals of the adjusted level ΔL (first phase) are shown in gray; means and corresponding 95% confidence ellipses of the adjusted delay ΔT (second phase) are shown in black.

2.2.2 The influence of delay and level

Part 2 of Experiment 1 examines the relationship of delay and level in a two-phase experimental design. The first phase determines the echo threshold defined as the reflection's level ΔL for the delays $\Delta T = (20, 40, 60, 80)$ ms and the directions $\Delta\phi = (0°, 30°, 60°)$. Subsequently in the second phase, previously determined individual reflection levels are used to measure the echo threshold defined as the reflection delay. In doing so, the listeners' performance is monitored by comparing input delays ΔT_{in} with corresponding output delays ΔT_{out}. Subliminal conditions start with $\Delta L = -50$ dB (first phase) and $\Delta T = 5$ ms (second phase) relative to the direct sound. Binaural playback and excitation signal remain the same as used in Part 1.

Figure 2.5 gives the results of both phases separately for the investigated directions $\Delta\phi$ collected from fifteen listeners. Individual levels of the first phase are corrected using Eq. (2.1) and are given as means and 95% confidence intervals for fixed delays 20 ms, 40 ms, 60 ms, and 80 ms, cf. Figure 2.5. Individual delays (second phase) do not need any normalization as the individual criterion is considered in respective levels obtained from the first phase. They are shown as 2-dimensional mean values with the corresponding 95% confidence ellipses. Note that the deviating mean levels for $\Delta T_{\text{in}} = 20$ ms at $\Delta\phi = 60°$ is due to removing outliers. Semi axes of the ellipses are mostly parallel to the figures' axes, indicating the reflection's delay and level to be independent of each other. Generally higher answer spreads in Figure 2.5 (a) suggest that similar presentation directions of direct sound and reflection makes assessing the presence of separate auditory events more challenging. The ANOVA reveals the delay being a significant parameter ($p \leq 0.05$) for all 3 directional separations. Regarding directional separation, the results from Part 1 are confirmed and if the delay exceeds $\Delta T \geq 40$ ms, the direction of the reflection is significant with two significant groups ($\Delta\phi = 0°$ and $\Delta\phi \geq 30°$, Tukey's HSD $p \leq 0.05$).

Modeling the echo threshold. Regarding consistency, it was expected that the mean curves of delay adjustment and level adjustment would coincide. Interestingly, delays ΔT_{out} adjusted

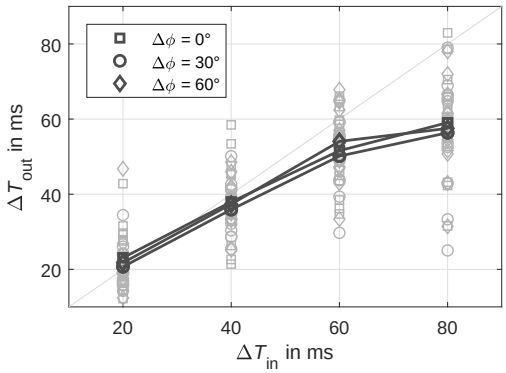


Figure 2.6: Individual output delays ΔT_{out} adjusted in the second phase over input delays ΔT_{in} and corresponding means over directional separations $\Delta\phi$

in the second phase are consistently lower than corresponding input delays ΔT_{in}. Learning effects due to listeners' previous exposure to the tasks in the first phase are suggested as possible cause. A similar rapid procedural learning effect was reported by [WF01] for ITD and ILD discrimination tasks. The variation of the individual criterion defining an echo throughout the experiment and the specific test design can be considered as alternative cause. Figure 2.6(a) directly compares input delays ΔT_{in} over individual and mean output delays ΔT_{out}. For all listeners, the adjusted output delay is found to be progressively shorter than the corresponding input delay with increasing input delay in the majority of conditions. Hence, it is assumed to be a learning effect that is independent of the directional separation.

The progressive decrease of the echo thresholds ΔL with increasing delay ΔT resembles the one found in literature. According to a review of echo threshold studies presented in [Bla97], slopes of the decrease vary with signals but are similar for different (individual) echo threshold definitions. The review includes the experimental data presented originally in [Dam71], which measured echo thresholds with the same definition and signal, and a similar setup (loudspeakers at $\phi = 0°$ for the direct sound and $\phi = 45°$ for the reflection) as used in Experiment 1. Interestingly, their threshold levels are almost 20 dB higher compared to the thresholds given in Figure 2.5, whereas agreeing with [Bla97] the slope is similar. Hence, the modeling focuses on the interdependence between delay and level, i.e., the slope of the echo threshold, rather than on absolute values of the echo threshold. In the interesting time range $\Delta T \leq 40$ ms slopes amount to $\beta_{\text{ET},1} = (-0.50, -0.57, -0.55)$ dB/ms (first phase) and $\beta_{\text{ET},2} = (-0.66, -0.74, -0.68)$ dB/ms (second phase) for $\Delta\phi = (0°, 30°, 60°)$, respectively. Accordingly, a linear fit of the experimental data of [Dam71] in this range yields $\beta = -0.70$ dB/ms.

2.2.3 Discussion

This section examined how the level, delay, and directional separation of a refection compared to the direct sound influence the echo threshold. Part 1 of Experiment 1 focused on the contribution of the reflection's direction on the echo perception. Although influences of the direction on the echo threshold were proven, the impact is negligible in the context of directional sound

sources. Significances were obtained only between $\Delta\phi = 0°$ and $\Delta\phi \geq 30°$ if the delay exceeded $\Delta T = 40\,\text{ms}$. However, successful modeling of the influence of direction is achieved by the sensitivity of the human ear.

Part 2 of Experiment 1 examined the relation of the parameters level and delay in a two-phase paradigm. In the first phase echo thresholds were measured by adjusting the reflection's level for a fixed delay. The results were then input to the second phase and listeners had to adjust the reflection's delay for highly individual levels obtained from the first phase. Ideally, both input and output delays would be identical. However, although level and delay were found to be independent parameters, the adjusted delays of the second phase were lower than the input delays of the first phase. These findings underline that the wide range of echo thresholds reported in literature is not only due to stimulus parameters, tasks employed, and echo threshold definitions, but also arises from individual subject factors such as the criterion. Further support thereof is found by the comparison of the echo threshold to the literature in which threshold levels of the reflection were reported to be 20 dB above the levels found here. However, there is a strong similarity in the threshold slopes within the interesting time range $\Delta T \leq 40\,\text{ms}$, which is used to model the interdependence of level ΔL and delay ΔT.

2.3 Reflection parameters influencing the echo suppression

While many psychoacoustic experiments have contributed to our understanding of the precedence effect with specular reflections, findings on the influence of spatio-temporal diffuse reflections are still sparse and partly contradictory. Robinson et al. [RWFB13] studied the influence of temporal diffusion on the echo threshold delay ΔT by modeling measured reflection responses. Resulting echo threshold delays for speech and music signals revealed mostly no difference between the suppression of compact and a temporal diffuse reflection. Grosse et al. [GvdPT17] studied the perceptual influence of directional diffusion of either direct sound or reflection by presenting the sound from a single loudspeaker or by presenting mutually uncorrelated versions of similar sounds from nine adjacent loudspeakers at the same distance. Although no pure specular lead/lag condition was tested in their experiments, there is evidence that the directional diffusion of the reflection might reinforce the echo suppression and for both speech and noise bursts, the echo threshold delay is longer if directional diffuse reflections are used. Lokki et al. [LPT$^+$11] studied the qualitative influence of the temporal diffusion of reflections and found significant perceptual differences. They describe that a more clear and open impression of the acoustics was reported for specular reflections as opposed to weak and muddy impressions reported for temporally diffuse reflections. In contradiction to [GvdPT17], [LPT$^+$11] suggested the stronger precedence to cause this impression.

This section presents a theoretical model for simulating Lambertian reflections and subsequently studies how reflection parameters influence the precedence effect in Experiment 2. Except for Section 2.3.5, the content of this section has been submitted to a journal [WH20].

2.3.1 Modeling diffuse reflections

The law of specular reflections states that the angle of incidence (ϕ_S, θ_S) equals the angle of reflection (ϕ_R, θ_R) and the incident, normal, and reflected directions are coplanar, cf. Figure 2.7. The delay T_r of the corresponding impulse response $\delta(t - T_r)$ is calculated from the distance between the corresponding image source and the receiver.

Specular reflections represent one of two extreme conditions that have been identified for rigid surfaces. The other extreme is the *diffuse* reflection which occurs when the reflected energy is scattered [CDD$^+$06, Kut09]. In room acoustics, reflections on rigid walls are typically approximated by dividing the reflected energy into two components, specular and diffuse, with the scattering coefficient s defining the ratio of the scattered energy to the total energy reflected by the surface. Thus, the scattering coefficient is defined in the interval $s = [0, 1]$ and a value of 1 corresponds to a directionally and temporally diffuse reflection in which there is no specular reflection component.

For simulating a diffuse reflection, suppose the point source S in Figure 2.7 emits a short power impulse at time $t = 0$, represented by a Dirac delta distribution with unity energy. The sound power $\mathrm{d}P_W(t)$ reaching the wall element $\mathrm{d}W$ by the angle of incidence ϕ_S and distance r_S

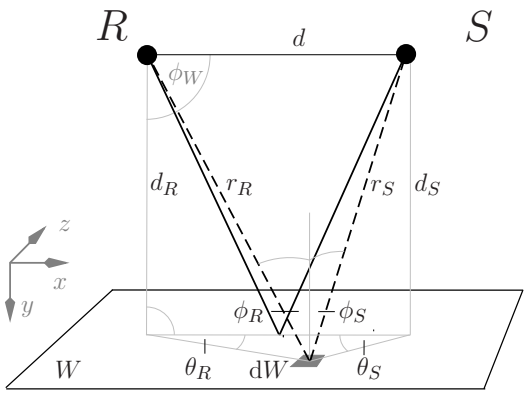


Figure 2.7: Schematic geometry of a reflection on the xz-plane. The basic constellation consists of a source S, a receiver R, and a wall W. The sound propagation path of the specular reflection is shown as solid line and the path of the wall element $\mathrm{d}W$ as dashed line.

is then defined by

$$\mathrm{d}P_W(t) = \frac{\cos\phi_S}{4\pi\, r_S^2}\,\delta(t - t_S)\,\mathrm{d}W, \tag{2.3}$$

with $r_S = \sqrt{(x + x_S)^2 + (z + z_S)^2 + d_S^2}$, time $t_S = r_S/c$, and speed of sound c. Diffuse reflections are typically interpreted by assuming Lambertian reflection in which the scattered sound power from the element $\mathrm{d}W$ is proportional to the cosine of the angle of reflection $\cos\phi_R$ [Kut09]. The intensity $\mathrm{d}I_W$ reflected by the wall element $\mathrm{d}W$ of a perfectly reflective wall without absorption that reaches the receiver R is defined by

$$\mathrm{d}I_W(t) = \frac{\cos\phi_S\cos\phi_R}{4\pi\, r_S^2\, r_R^2}\,\delta(t - t_{SR})\,\mathrm{d}W. \tag{2.4}$$

with $r_R = \sqrt{(x + x_R)^2 + (z + z_R)^2 + d_R^2}$ and $t_{SR} = (r_S + r_R)/c$. Assuming the reflected sounds of wall elements $\mathrm{d}W$ to be incoherent, the overall intensity I_{diff} reaching the receiver R is obtained by the integration over both dimensions of the wall

$$I_{\mathrm{diff}}(t) = \iint_W \frac{d_S\, d_R}{4\pi\, r_S^3\, r_R^3}\,\delta(t - t_{SR})\,\mathrm{d}x\mathrm{d}z. \tag{2.5}$$

The cosines are expressed by corresponding distances $r_{S/R}$ and $d_{S/R}$ yielding an elliptical integral, which we solve numerically.

Figure 2.8 shows the reflected intensities $\mathrm{d}I_W$ reaching the receiver R directionally spreading around the specular direction (ϕ_0, θ_0). It is easy to see that the size of the wall not only influences the amount of the reflected energy, but also the temporal diffusion of the reflected sound. The overall intensity I_{diff} for two different wall sizes is depicted in Figure 2.9 (setup A, cf. Table 2.1) and represents the temporal spread. The dimension of the rectangular finite-sized wall is chosen in a way that compared to an infinite wall approximately 90% of the energy is reflected with (ϕ_0, θ_0) at the center of the wall. For a better comparability intensity envelopes are shifted in time by $T_d = d/c$. In this way, the direct sound of all conditions reaches the receiver at $t = 0\,\mathrm{ms}$, whereas corresponding reflections start at $\Delta T = T_r - T_d$.

In [SLTS12] the temporal structure of diffuse reflections is modeled based on Biot's rough surface theory [Bio68]. It is assumed that the surface roughness Ra is small relative to the wavelength λ. The energy envelope of this model for a single diffuse reflection is assumed to produce an exponentially decaying tail of reflections after the specular reflection. The decay coefficient τ of the exponential function $e^{-t/\tau}$ is related to the local reverberation time and can be found by fitting a line to the Schroeder-integrated energy curve corresponding to the measured impulse response.
Fitting exponential functions to our model with decay coefficients based on the early decay yields an envelope decay of $e^{-t/\tau}$ wit $\tau = \Delta T/1.33$. Figure 2.9 shows a neat overlapping of the early parts of the reflective impulse responses and the exponential functions regard about 95% of the overall energy reflected by an infinite wall.

In room acoustics, diffusion of the incoming sound is typically achieved with so-called Schroeder diffusors [Sch79]. In contrast to the low surface roughness studied in [SLTS12], the surface of these diffusors has pronounced periodic wells of different depths yielding a considerable phase shift and, with this, the diffusion of the reflected sound. However, the wells do not only influence

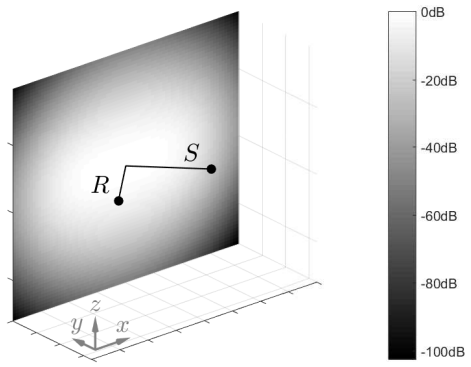


Figure 2.8: Directional spread of diffusely reflected intensities $\mathrm{d}I_W$ normalized to the maximum and coded in gray-scale. The sound propagation path of a specular reflection is indicated as solid line.

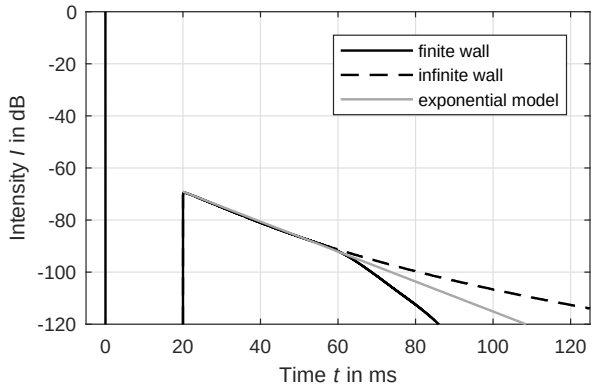


Figure 2.9: Intensity envelopes of the impulse response of condition A_{20} consisting of direct sound an diffuse reflection. The temporal spreads for diffusely reflected intensities I_{diff} for an infinite wall and the finite-sized wall W are normalized to the direct sound. The exponential model proposed in [SLTS12] is fitted to the early decay.

the diffusion but also the delay ΔT, as due to the surface structure the specular path is not necessarily the shortest. Consequently, measured reflections' envelopes of diffusers with prominent surface structure resemble a gamma function starting with a damped onset and – depending on the well depth and size–reaching their peak amplitude several milliseconds later [DT98, DT00].

2.3.2 The influence of scattering

The influence of reflection properties on the transient precedence effect is studied by measuring the echo threshold. The intensity envelopes of the reflection response obtained by the model were shown to vary with the setup of source, wall, and receiver–hence also with the delay. Thus, in contrast to the specular reflection, varying the delay is not straightforward which is why the echo threshold is studied by the level ΔL for fixed delays, i.e., fixed setups. A method of adjustment is used for the ongoing sequence of direct sound and reflection similar to Experiment 1 of Section 2.2, and the listener is given control over the level of the reflection. To ensure buildup, the amplitude of the ongoing sequence is linearly faded in for 2 seconds at the beginning of each condition. The excitation signal for all conditions consists of a 50-ms Gaussian noise burst (instant on- and offset) presented at a period of 250 ms.

Before the breakdown of fusion, listeners can detect the presence of the lag even when they do not perceive the lag as a separate auditory event [LSC01]. In addition the echo threshold, a second part of the experiment determines how much additional attenuation is required to eliminate all audible effects of the lagging sound. The *masked threshold* is defined as the absolute threshold of perceptibility of the lagging sound and thus no variation of the criterion is possible. It was measured by a two-interval two-alternative forced choice adaptive procedure (2I2A) in a 3-down 1-up rule [Lev71] estimating the 79.7% point on the psychometric function. In each interval, the direct sound consisting of four consecutive noise bursts was presented with a pause of 1 s between the intervals. In one randomly chosen interval the reflection was added, whereas in the other interval it was absent. The listeners' task was to specify the interval containing the reflection. Possible loudness cues were removed by roving the level of both intervals and feedback was given after each response. At the beginning of each run direct sound and reflection were equally loud with a step size of 10 dB. After two reversals the step size was decreased to 5 dB and set to a final value of 2 dB after another two reversals. Using this 2 dB step size, 6 more reversals were obtained and the threshold value is calculated by averaging over all levels obtained by final step size according to [Kle01].

Conditions. The directional and temporal distribution of diffuse reflection responses is influenced by the setup of the geometric constellation, i.e., source-to-receiver distance and respective distances to the reflecting wall. In the experiment three different setups A, B, C are tested, schematically shown in Figure 2.10. The wall W is simulated to be plane yielding a consistent definition of the delay ΔT for both specular and diffuse reflections. Respective conditions derived from the setups are listed in Table 2.1 with the receiver R facing always the point source S at

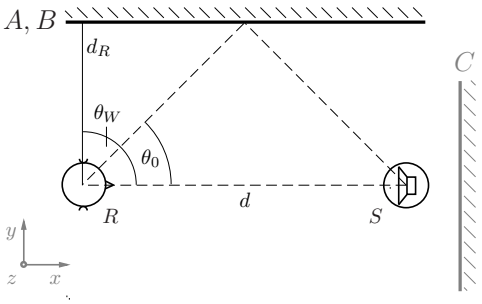


Figure 2.10: Schematic representation of the constellation of the reflecting wall for setuos that are examined in Experiment 2. For setups A and B (black) sound paths of direct sound and specular reflection, and reflection angle ϕ_0 and wall angle ϕ_W are indicated.

equal height. Note that the condition's subscript number inform about the delay ΔT in ms and the superscript indicates its scattering coefficient s. To study the influence of the delay ΔT on specular ($s = 0$) and diffuse ($s = 1$) reflections while keeping the angle $\phi_0 = 45°$ constant, conditions $A^s_{\Delta T}$ with $\Delta T = (5, 10, 20, 30)$ ms are tested for both suppression thresholds. The influence of the angle of reflection ϕ_0 on suppression thresholds is tested by conditions A^1, B^1, C^1 while keeping the delay $\Delta T = 20$ ms constant. Moreover, to study the influence of the scattering coefficient, condition A^s_{20} with $s = 0.5$ is included in the test. The condition $A^\star_{20}$ is a hybrid of specular and diffuse reflection, and examines the influence of the directional spread. Compared to fully-diffuse A^1_{20}, the diffuseness of condition $A^\star_{20}$ considers only the monaural-temporal spread but not the directional spread. Instead, similar to conditions of [RWFB13], the reflection is presented from a single direction, i.e., the specular angle ϕ_0.

For the numerical solution of Eq. (2.5), the size of the wall W is chosen in a way that approximately 90% of the energy reflected by an infinite wall is taken into amount to by a uniform grid consisting of quadratic patches w. Computed intensities $\mathrm{d}I_W$ were encoded into Ambisonics (order $N = 17$) and decoded on a spherical t-design (degree $t = 35$) using max-$\boldsymbol{r}_\mathrm{E}$ weighting [Dan01] resulting in 632 envelope signals describing the spatio-temporal energy distribution of a diffuse reflection. The corresponding impulse responses were obtained by multiplying white Gaussian noise with the square roots of intensity envelopes.

Agreeing with observations in [LPT$^+$11], the temporal diffusion yields a coloration and the respective reflections sound more muddy than specular ones. To avoid timbral influences, an iterative whitening procedure similar to the Hilbert transform approach described by [KKvdH$^+$97] is applied to each diffuse reflection waveform, and they are normalized thereafter in terms of their total energy. Impulse responses of specular reflections are simulated by a Dirac delta distribution and, thus, the specularly reflected sound is an exact copy of the direct sound.

The excitation signal is convolved with respective impulse responses and binaural stimuli are created by convolving them with corresponding HRTF measurements [Ber13]. Playback level and equipment are the same as for Experiment 1.

Table 2.1: Composition of conditions that are tested in the listening experiment. Condition letters indicate the wall and reflection angle and their indices indicate the delay ΔT, which is adjusted by scaling distances d, d_R and walls W, w.

condition	ϕ_0	ϕ_W	ΔT	d	d_R	W	w
A_5	45°	90°	5 ms	4.1 m	2.0 m	$10\times8.5\,\mathrm{m}^2$	$0.1\times0.1\,\mathrm{m}^2$
A_{10}	45°	90°	10 ms	8.3 m	4.1 m	$20\times17\,\mathrm{m}^2$	$0.2\times0.2\,\mathrm{m}^2$
A_{20}	45°	90°	20 ms	16.5 m	8.3 m	$40\times34\,\mathrm{m}^2$	$0.4\times0.4\,\mathrm{m}^2$
A_{30}	45°	90°	30 ms	24.8 m	16.6 m	$60\times51\,\mathrm{m}^2$	$0.8\times0.8\,\mathrm{m}^2$
B_{20}	70°	90°	20 ms	4.0 m	5.1 m	$20\times18\,\mathrm{m}^2$	$0.2\times0.2\,\mathrm{m}^2$
C_{20}	0°	0°	20 ms	4.0 m	7.4 m	$18\times18\,\mathrm{m}^2$	$0.2\times0.2\,\mathrm{m}^2$

2.3.3 Experimental results

Twelve experienced listeners participated in the experiment. To prevent listeners fatigue, the listening experiment was performed in three runs. In the first run all echo conditions $A^1_{(5,10,20,30)} + A^0_{(5,10,20,30)} + A^{0.5}_{20} + B^1_{20} + C^1_{20} + A^\star_{20} = 12$ were tested twice yielding 24 adjustment tasks. The second run included masked conditions with $\Delta T = 20\,\mathrm{ms}$, which were tested once yielding $A^0_{20} + A^{0.5}_{20} + A^1_{20} + B^1_{20} + C^1_{20} + A^\star_{20} = 5$ adaptive tasks. The last run consisted of masked conditions for specular and diffuse reflections $A^0_{(5,10,20,30)} + A^1_{(5,10,20,30)} = 8$. In this way masked conditions A^0_{20} and A^1_{20} were tested twice, once in the context of the refection properties (second run), and once in the context of the delay ΔT (third run). Conditions within each run were tested in individual random order.

The experimental results follow a normal distribution (Lilliefors, $p > 0.05$) and are depicted in Figure 2.11 as mean values by filled symbols for the echo threshold (ET) and open symbols for the masked threshold (MT) with corresponding 95% confidence intervals. Note that in comparison to Experiment 1 no correction of individual threshold was necessary. Although different procedures with a different number of runs were used to access the thresholds, the level gap between specular echo and specular masked threshold is constant for delays $\Delta T \geq 20\,\mathrm{ms}$ agreeing with the gap found in [RHH00] for speech. Note that masked threshold levels of conditions $A^{(0,1)}_{20}$ tested once in the context of the delay (Figure 2.11, left panel) and once in the context of the refection property (right panel) are slightly but not significantly different (paired sample t-tests, $p > 0.05$).

Influence of delay and scattering coefficient. The left panel of Figure 2.11 shows suppression threshold levels examined with setup A as a function of the delay ΔT. The ANOVA reveals the threshold type, the delay, and the reflection type to be significant parameters ($p \leq 0.05$). As expected, suppression decreases progressively with increasing ΔT, and masked thresholds are lower than the echo thresholds. For both reflection types masked thresholds decrease non-linearly with increasing the delay. The echo threshold on the other hand decreases linearly which is in contrast to the results from Experiment 1 in Section 2.2. However, delay ranges investigated differ substantially from Experiment 1, and a comparison of the small overlapping range $20\,\mathrm{ms} \leq \Delta T \leq 30\,\mathrm{ms}$ reveals a similar linear decay of $-0.5\,\mathrm{dB/ms}$ in both experiments.

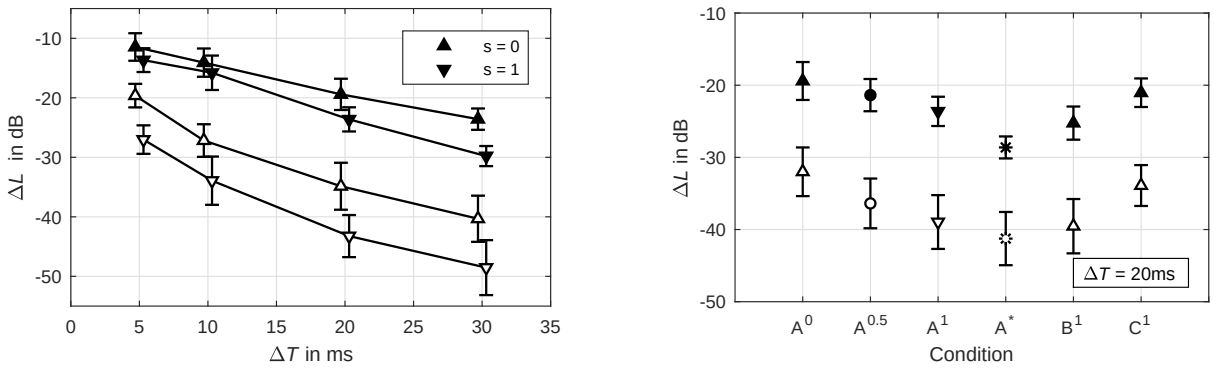


Figure 2.11: Means and 95% confidence intervals of echo threshold (filled symbols) and masked threshold (open symbols). Left: Specular reflection ($s = 0$) and diffuse reflection ($s = 1$) tested with conditions $A_{(5,10,20,30)}$. Right: The influence of the scattering coefficient s examined by conditions A^s_{20}, the influence of the angle ϕ_0 examined by conditions A^1, B^1, C^1 with delay $\Delta T = 20\,\text{ms}$, and the influence of the directional spread examined by hybrid condition A^*_{20}.

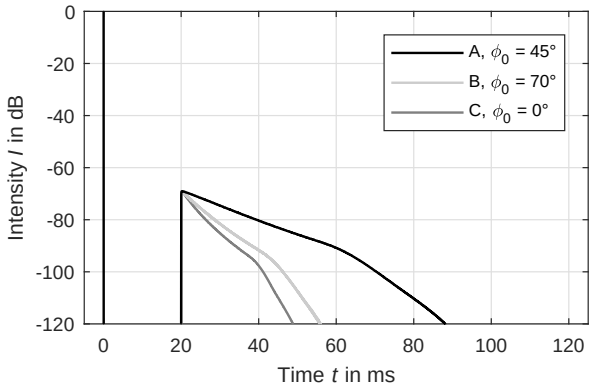


Figure 2.12: Diffusely reflected intensity I_{diff} arriving at the receiver R of conditions A^1_{20}, B^1_{20}, and C^1_{20} with the finite sized wall W, normalized to their maximums and shifted in time by T_d.

Interestingly, the diffuse reflection appears to weaken the echo suppression. Corresponding reflections are more easily detectable and in each condition the diffuse threshold level is lower than the corresponding specular threshold level. A Tukey HSD post hoc analysis of both reflection types shows conditions $A_{(5,20,30)}$ of the echo threshold and all conditions of the masked threshold to be significantly lower ($p \leq 0.05$) with effect sizes expressed as Cohen's d above $d \geq 0.9$, i.e., large effect [Saw09]. This finding is in contrast to results obtained by diffuse reflection responses modeled in [RWFB13] for speech and music signals, who found mostly no significance between echo thresholds of specular and temporally diffuse reflections.

In addition to specular ($s = 0$) and diffuse ($s = 1$) reflections, condition A_{20} was tested with a scattering coefficient of $s = 0.5$. Respective results are shown in the right panel of Figure 2.11. For both suppression types, means of $A^{0.5}$ are in between corresponding means of extreme conditions A^0 and A^1 and an ANOVA reveals the scattering coefficient to be a significant parameter ($p \leq 0.05$).

Influence of directional spread and directional separation. The influence of directional spread of the diffuse reflection on the echo suppression is examined with the hybrid condition $A^\star$, cf. right panel of Figure 2.11. This condition combines the temporal characteristics of a diffuse reflection with the directional characteristics of a specular reflection. Mean threshold levels are below mean levels of the corresponding condition A^1 that spreads also directionally and a paired sample t-test reveals significant differences of $A^\star$ and A^1 for the echo threshold ($p \leq 0.05$; Cohen's $d = 0.99$, large effect), whereas for the masked threshold no significant difference is obtained ($p > 0.05$). It thus can be concluded that the temporal and directional spread have a reverse effect on the echo suppression: compared to fully diffuse reflections, reflections spreading only temporally but not directionally are more easily detectable as an echo. Conversely, this means that the directional diffusion increases the suppression, making reflections more difficult to perceive as a separate auditory event. The directivity of the ear with an increased sensitive for $\phi = 45°$, cf. Figure 2.4, is suggested as possible cause. Thus, depending on the angle of the specular reflection, this finding might not apply for any constellations of sound source, wall, and receiver.

The influence of directional separation is examined by diffuse conditions A_{20}, B_{20}, C_{20}. Echo and masked threshold levels of condition B_{20} are not different from A_{20} ($p > 0.05$, Tukey HSD post hoc analysis) suggesting that the influence of directional separation is not applicable for directionally and temporally diffuse reflections. However, corresponding reflection responses depicted in Figure 2.12 reveal a higher temporal spread of condition A_{20}, which compensates any possible effect of directional separation. In contrast, conditions B_{20} and C_{20} exhibit a similar temporal spread and significant increases of echo and masked thresholds levels are seen for the decreased directional separation of direct sound and reflection of condition C_{20} ($p \leq 0.05$; Tukey HSD post hoc analysis). Thus, similar to specular reflections, the influence of directional separation also applies for directionally and temporally diffuse reflections.

2.3.4 Modeling the echo suppression

Echo threshold. Following the approach in [RHH00] a linear model is used to describe the echo threshold level $\Delta L_{\mathrm{ET}}(\Delta T)$

$$\Delta L_{\mathrm{ET}} = \alpha_{\mathrm{ET}} + \beta_{\mathrm{ET}} \cdot \Delta T, \tag{2.6}$$

with the intercept α_{ET} in dB and the slope β_{ET} in dB/ms. Modeling individual echo thresholds with the regression fit given in Eq. (2.6) reveals the differences of obtained slopes between both reflection types to be significantly different form zero (t-test: $p \leq 0.05$). Thus, the slope of the specular regression line is different from the slope of diffuse regression line. The differences of individual intercepts of specular and diffuse threshold on the other hand are not significant ($p > 0.05$). It can be conjectured that the temporal spread of diffuse reflections results in an increase of the perceptually effective delay. To account for different scattering coefficients s in a

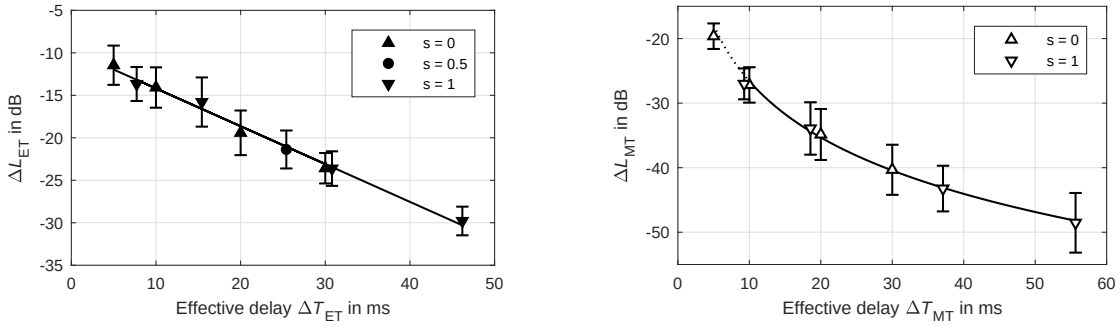


Figure 2.13: Means and 95% confidence intervals of the echo threshold (left panel) and the masked threshold (right panel) for different scattering coefficients s tested with setup A. All data is time-aligned to corresponding effective delays. For the echo threshold this is done by delays $\Delta T_{\mathrm{ET}}^{ec}$ calculated from the temporal energy centroid of corresponding reflection responses using Eq. (2.7) with $k_{\mathrm{ET}}^{ec} = 0.54$. The alignment of masked thresholds is achieved with Eq. (2.8) and effective delays ΔT_{MT} are fitted to the logarithmic Eq. (2.9) yielding $k_{\mathrm{MT}}^{opt} = 0.62$.

single model, a simple predictor of the effective delay ΔT_{ET} is introduced, which replaces ΔT in Eq. (2.6) by

$$\Delta T_{\mathrm{ET}} = \Delta T\,(1 + s \cdot k_{\mathrm{ET}}). \tag{2.7}$$

This yields a temporal alignment of the results with the alignment parameter k_{ET}. The optimal alignment parameter is obtained by pooling individual responses of setup A for $s = (0, 0.5, 1)$ with corresponding effective delays of Eq. (2.7). The maximum coefficient of determination $R^2 = 0.99$ of the model in Eq. (2.6) with corresponding means is found with $k_{\mathrm{ET}}^{opt} = 0.48$ yielding $\alpha_{\mathrm{ET}} = -9.5\,\mathrm{dB}$ and $\beta_{\mathrm{ET}} = -0.47\,\mathrm{dB/ms}$ with a 95% confidence interval of $[-0.53, -0.41]\,\mathrm{dB/ms}$.

In [RWFB13] the echo threshold of specular and temporally diffuse reflections is examined by a variation of the delay. With their delay defined as the time between the direct sound and the centroid of energy of the reflection response did not yield any perceptual differences between specular and temporally diffuse reflections for most conditions (5 out of 6) for speech and music signals. Thus, it could be seen as a perceptually effective delay.

Effective delays $\Delta T_{\mathrm{ET}}^{ec}$ based on the centroid of energy of diffuse reflections of setup A are calculated using Eq. (2.7) yielding $k_{\mathrm{ET}}^{ec} = 0.54$. This value falls within the 95% confidence interval of the optimal parameter k_{ET}^{opt}. The left panel of Figure 2.13 shows the temporally aligned data with $k_{\mathrm{ET}}^{ec} = 0.54$ and the corresponding regression line with $\beta_{\mathrm{ET}} = -0.45\,\mathrm{dB/ms}$ (95% confidence interval $[-0.50, -0.41]\,\mathrm{dB/ms}$). The latter resembles the slopes obtained for $\Delta T = 20 \ldots 40\,\mathrm{ms}$ of Experiment 1 presented in Section 2.2 with mean values $\beta_{\mathrm{ET}} = -0.55\,\mathrm{dB/ms}$ in the first phase and $\beta_{\mathrm{ET}} = -0.69\,\mathrm{dB/ms}$ in the second phase.

Masked threshold. The temporal masking effect of Experiment 2 can be seen as a mixture of *simultaneous masking* and *forward masking*. Fastl and Zwicker [FZ06] studied simultaneous masking of 50-ms noise bursts by uniform masking noise and found threshold levels not more than 10dB below the masker level. Conversely, masked threshold levels of Experiment 2 are sharply lower than the masker (i.e., the direct sound) and it is save to assume that obtained

thresholds are mainly caused by forward masking. Following temporal masking theory, a signal is masked as long as its excitation pattern is below the temporal masking pattern evoked by the masker. According to [FZ06] (p. 83, Fig. 4.22) temporal masking patterns for forward masking of noise bursts do not exhibit an exponential decay, but thresholds can be well approximated by a logarithmic function of the delay time, i.e., $\Delta L_{\mathrm{MT}} = \gamma - \epsilon \cdot \ln(\Delta T/\mathrm{ms})$ for $\Delta T = 10 \ldots 100\,\mathrm{ms}$.

The experimental results indicate that masked thresholds of diffuse reflections are shifted by a constant factor δL in dB compared to the thresholds obtained by specular reflections, cf. Figure 2.11. Thus, in order to model masked thresholds of arbitrary reflections, the logarithmic function is extended by the scattering coefficient yielding

$$\Delta L_{\mathrm{MT}} = \gamma + s \cdot \delta L - \epsilon \cdot \ln(\Delta T/\mathrm{ms})\,. \tag{2.8}$$

Figure 2.14 shows a schematic representation of the forward masking function from [FZ06] together with envelopes of specular and diffuse reflections. For a specular condition, the reflection amplitude exceeds the temporal masking pattern for the first time at the instant offset $t = \Delta T$. For a diffuse condition on the other hand, the decay of the reflection amplitude is flatter than the logarithmic function of Eq. (2.8). Accordingly, the diffuse reflection amplitude exceeds the masking pattern around the decay at $t = \Delta T_{\mathrm{MT}}$ at a lower level, cf. Figure 2.14. As the temporal masking pattern for $\Delta T = 10 \ldots 100\,\mathrm{ms}$ yields a straight line in a log-log representation, a delay ΔT_{MT} of the form $\Delta T_{\mathrm{MT}} = \Delta T \cdot f(s)$ is assumed, yielding Eq. (2.8) to be rewritten as

$$\Delta L_{\mathrm{MT}} = \gamma - \epsilon \cdot \ln(\Delta T_{\mathrm{MT}}/\mathrm{ms}) \tag{2.9}$$

with the effective delay $\Delta T_{\mathrm{MT}} = \Delta T \cdot e^{s\,k_{\mathrm{MT}}}$ and the parameter $k_{\mathrm{MT}} = \delta L/\epsilon$. Fitting parameters for masked thresholds of conditions $A_{(10,20,30)}$ are calculated using Eq. (2.9) and amount to $\gamma = 2.4\,\mathrm{dB}$, $\epsilon = 12.6\,\mathrm{dB}$, $k_{\mathrm{MT}}^{opt} = 0.62$ and therefore $\delta L = 7.8\,\mathrm{dB}$. This yields a coefficient of determination $R^2 = 0.99$ with mean masked thresholds.

The right panel of Figure 2.13 shows time-aligned masked threshold levels of specular and diffuse conditions of setup A with the corresponding logarithmic fit. As for $\Delta T < 10\,\mathrm{ms}$ the forward masking function shown in Figure 2.14 saturates, thresholds of conditions $A_5^{(0,1)}$ are overestimated by the logarithmic model of Eq. (2.9). In contrast to our model, a linear approach is used for masked thresholds of delays $20\,\mathrm{ms} \leq \Delta T \leq 80\,\mathrm{ms}$ in [RHH00]. However, in this time range the logarithmic function can be well approximated with a linear model.

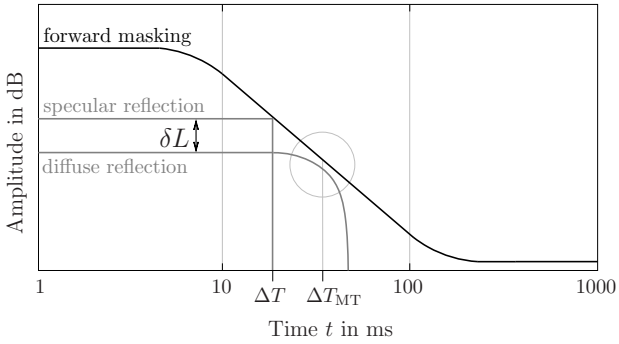


Figure 2.14: Schematic representation of the forward masking pattern from [FZ06] (p. 83, Fig. 4.22) as a function of the delay time $\log(\Delta T)$ together with nominal amplitudes of specular and diffuse reflection as a function of time t. The masker (direct sound) ends at $t = 0$. The masked level is achieved, if reflection's envelope peaks exceed the post masking pattern. For the specular reflection this when the reflection ends at ΔT, whereas envelope fluctuations of the diffuse reflection exceed the level of masking in the time range of ΔT_{MT}.

2.3.5 The wall's surface structure

To study the influence of the wall's surface structure on the suppression of diffuse reflections a consecutive listening experiment was conducted by Korbinian Wegler in the context of his Master's thesis [Weg20]. Relevant parts of it are presented in the following and discussed in the light of the previous results.

To consider a distinct surface structure of the wall, the diffuse model in Eq. (2.5) is used. In contrast to the plane wall model, the reflection response is not obtained by the envelope, but quadratic patches of size $w = 0.2 \times 0.2\,\mathrm{m}^2$, simulated at random height y_i, contribute directly an intensity $\mathrm{d}I_W$ to the reflection response. To avoid spectral coloration, an additional spatial filter is applied to the amplitude y_i of the patch. In this way, the structure resembles a 2D quadratic residue diffusor [Sch75] with an arithmetic average roughness $Ra = 0.4\,\mathrm{m}$, which is in the range of the wavelength λ.

The conditions under test are derived from a setup with fixed source-to-receiver distance $d = 2\,\mathrm{m}$ and specular direction $\phi_0 = 70°$. The fixed set of delays $\Delta T = (20, 40, 60)\,\mathrm{ms}$ is accomplished by consistent constellations of the wall's distance d_R and angel ϕ_W (cf. Figure 2.7) with the wall size chosen in a way that 95% of the overall energy is considered. The echo and masked threshold are measured in the same way as previously using a method of adjustment and a 2I2A forced choice adaptive procedure, respectively, with noise bursts as excitation signal. Figure 2.15 delineates the experimental results as means and 95% confidence intervals of twelve listeners.

Compared to the specular means obtained with setup A (shown as gray filled markers in Figure 2.15) there is a constant offset of the specular echo threshold of about $-8\,\mathrm{dB}$. This is explained by the different setup with a variation of directional separation ($\Delta\phi = 45°$ vs. $\Delta\phi = 70°$), but also by different listeners (with different individual criteria). The absolute masked thresholds of specular reflections on the other hand agree with the corresponding mean results of Experiment 2 (gray open markers) and the regression of the means yields a coefficient

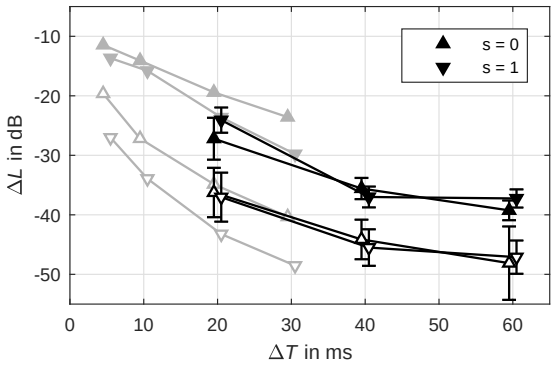


Figure 2.15: Means and 95% confidence intervals of echo threshold (black, filled symbols) and masked threshold (black, open symbols) for specular ($s = 0$) and diffuse ($s = 1$) reflections. In comparison to the plane wall panel studied previously (mean values shown in gray), the diffuse wall surface is simulated to resemble a Schroeder diffusor with $\phi_0 = 70°$ and $d = 4\,\mathrm{m}$.

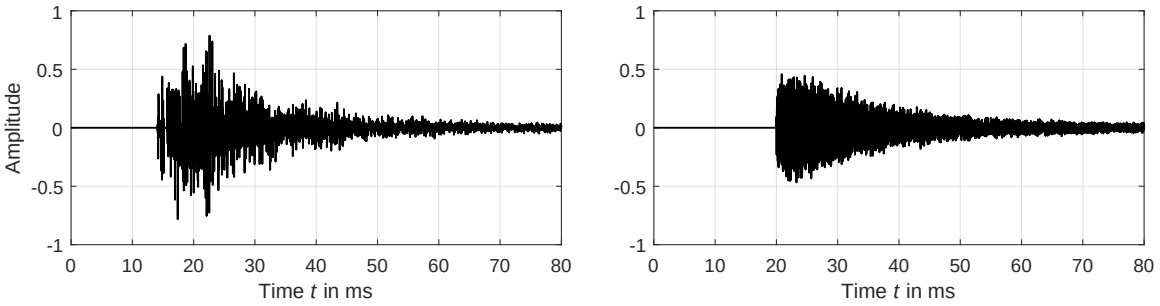


Figure 2.16: Comparison of reflection responses (left) directly obtained by Eq. (2.5) with finite-sized quadratic patches w and an arithmetic average roughness $Ra = 0.4\,\mathrm{m}$ as used in [Weg20] and (right) indirectly obtained by considering only the envelope calculated with Eq. (2.5) and no roughness. Both reflections contain the same amount of energy.

of determination of $R^2 = 0.99$. Interestingly, for both suppression types, diffuse and specular reflections yield similar results and the reflection type is not significant for any threshold ($p > 0.05$, ANOVA). At first, this seems to be in contrast to the findings from above according to which diffuse reflections yielded lower threshold levels compared to specular ones. However, examining the temporal shapes of reflection responses for $\phi_0 = 70°$ given in Figure 2.16 reveals that compared to a plane wall panel (right), the wall's surface structure yields gamma-shaped reflection responses resembling those obtained from diffusor measurements [DT00] and used in [RWFB13]. Besides the shape of the reflection response, the distinct surface structure influences the definition of the delay ΔT. The random height (and depth) of quadratic patches yields a ramped onset before the specular delay time ΔT and a steeper decrease after ΔT. In this way the temporal energy centroid is shifted towards the specular delay and for a nominal delay of $\Delta T = 20\,\mathrm{ms}$, the temporal energy centroid of the diffuse reflection almost coincides with $\Delta T^{ec} = 22\,\mathrm{ms}$. Thus, the effective delay for perceiving an echo ΔT_{ET}, cf. Eq. (2.7), is almost identical for 20 ms-conditions of both reflection types. For the masked threshold it is assumed that due to the steeper decrease of reflection response after the specular delay, higher reflection levels are needed to exceed the forward masking pattern, cf. Figure 2.14. Energy centroids of diffuse reflections with nominal

delays $\Delta T = (40, 60)$ ms amount to $\Delta T^{ec} = (47, 73)$ ms, respectively. This mismatch is not negligible. However, corresponding threshold levels approach their lower limits for such long delays and, thus, discrepancies of ΔT barely influence the suppression ΔL.

2.3.6 Discussion

This section introduced a model for simulating diffuse reflections based on Lambert's cosine law. The reflections modeled spread spatially yielding a directional and temporal diffusion of the reflected sound, which is influenced by the geometrical setup of sound source, receiver, and reflective plane wall. The early decay of the temporal diffusion resembles an exponential function and agrees with the model proposed in [SLTS12]. Control over diffusion is achieved by the scattering coefficient, which describes the ratio of the scattered energy to the total energy reflected by the surface.

The perception of different reflection responses was evaluated in Experiment 2 studying the echo suppression. Agreeing with results presented in Experiment 1, cf. Section 2.2, for large delays between direct and reflected sounds both the echo suppression levels are low for either type of reflections, specular and diffuse. A comparison of reflection types reveals the echo suppression to be weaker for spatio-temporal diffuse reflections than for specular reflections of the same total energy; if the reflected sound is scattered, the perception of it being present as well as of it being noticed as echo occurs at lower levels. This finding agrees with the observations made in [LPT$^+$11].

By removing the directional spread of diffuse reflections, the suppression further decreases and thresholds observed for temporally diffuse refections are below the corresponding levels of fully diffuse, i.e., also directionally diffuse, conditions. Thus, it is assumed that directional and temporal diffuseness have opposing effects on the echo suppression threshold. This finding dissolves the apparent contradiction in the available literature. In agreement to [GvdPT17], directional diffusion increases the suppression. However, reflections that spread also temporally remove this effect, cf. [LPT$^+$11].

The effect of directional separation, comprehensively investigated in Section 2.2 applies also to directionally and temporally diffuse reflections and the suppression is stronger when direct sound and reflection are directionally close together. Similarly, this effect is relatively weak and it becomes negligible as different reflection's directions yield different effective delays.

The apparent contradiction of the results with findings from Robinson et al. in [RWFB13] who reported mostly no significance between the reflection types is explained by the definition of the reflection's delay ΔT. Their envelope shapes are simulated by a gamma function which is based on diffusor measurements with distinct surface structure. In this way diffuse responses are detached from any geometry of the geometrical setup and the authors determine the delay from the corresponding temporal energy centroid. This suggests the existence of an perceptually effective delay and modeling of the echo thresholds presented in this section is achieved by temporally aligning the threshold levels to energy centroids of corresponding reflection responses.

Similarly, masked threshold levels are modeled by a temporal alignment of the data with an alignment parameter deduced from the forward masking pattern found in [FZ06] exhibiting a logarithmic relation between reflection's delay and level.

Lastly, it was shown that by considering a surface structure in the diffuse model, reflection responses strongly resemble those obtained from diffusor measurements presented in [DT00]. Perceptually, this removed any significances of the scattering in accordance with [RWFB13], and the suppression levels for diffuse reflections converge with the ones of specular reflections.

2.4 Lateralization with increased reflection level

Previous experiments studying the echo suppression offer a detailed insight into the precedence effect. However, their direct applicability on beamformers is limited as spatialized signals are seldomly transient clicks, and reflection patterns typically consist of more than a single reflection. Studies on the contribution of multiple reflections to the precedence effect are rare as a free number of sound instances complicates experimentation by increasing the number of potential conditions. Moreover, it is not clear if the categorization of precedence effect types introduced in [FGZ14] is applicable to more complex reflection patterns. The few studies directly assessing the precedence effect with multiple reflections focus on transient clicks and knowledge of the transient precedence effect was confirmed by their data [ESN68, TH99a, Yos07, GYL12]. Knowledge on precedence phenomena caused by long duration stimuli is mostly obtained from experiments that focus on room acoustics, e.g., [Bec98], or such describing recording and rendering techniques, e.g., [LR05, Fra13, Sti15, ZFKC14].

This section presents a study that covers the transient precedence effect, the ongoing precedence effect, and the onset capture effect, which are investigated using sounds of different envelope, frequency range, angular and temporal spread. Except for the part that studies lateral reflections, the content of this section already appears in [Wen17] and it is summarized here again and put in context with the previous experiments.

2.4.1 The influence of multiple reflections

Spatial all-pass: The Bessel sequence. With the number reflections, the possibilities to vary the reflection response is increased and an exhaustive systematic investigation appears infeasible. To limit the set of free parameters and at the same time attain a flat frequency response for all parameter settings a Bessel sequence is used as the temporal pattern of reflection. Keele in [Kee89] proposed the distribution of a signal to an equal spaced loudspeaker array using weights of a sequence of Bessel functions $J_m(\hat{\varphi})$ [OOL+20, Eq. 10.2.2] of order m and modulation depth $\hat{\varphi}$. This yields an all-pass impulse response

$$h(t) = \sum_{m=-\infty}^{\infty} J_m(\hat{\varphi})\delta(t - m\Delta T), \tag{2.10}$$

which can easily be truncated to $N \leq |m|$ whenever the modulation depth stays limited, e.g., $\hat{\varphi} \leq \pi/2$. Figure 2.17 shows the impulse, magnitude, and phase responses of the truncated system as defined in Eq. (2.10) with $\hat{\varphi} = 40°$ and $N = 2$. By assigning $J_m(\hat{\varphi})\delta(t - m\,\Delta T)$ to a ring of $L = 2N + 1$ evenly spaced loudspeakers, cf. Figure 2.18, the impulse response resembles the horizontal first order reflection pattern of a beamformer projecting its beam to a reflective wall (cf. Figure 1.6). The Bessel system response is all-pass, and the directional and temporal assignment allows to study the precedence effect with multiple sound instances.

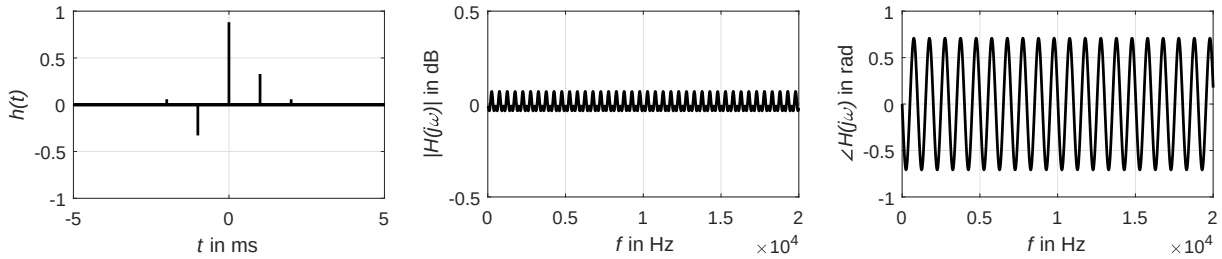


Figure 2.17: (from left to right) Impulse response h, magnitude $|H|$, and phase φ of a truncated FIR all-pass filter with order $N = 2$, modulation depth $\hat{\varphi} = 40°$, and delay $\Delta T = 1\,\text{ms}$.

Conditions. Conditions are created by the variation of the reflection pattern's

- delay: $\Delta T = (1, 2, 4, 8)\,\text{ms}$;
- directional separation: $\Delta\phi = (15°, 30°, 45°)$;
- playback direction: clockwise or counterclockwise;
- angle of the prominent reflection: $\phi_0 = (0°, 90°)$;
- negative weight $J_{-1}(\hat{\varphi})$: leading or lagging.

The Bessel order $N = 2$, modulation depth $\hat{\varphi} = 40°$, and source-receiver distance $r = 1.5\,\text{m}$ are kept constant yielding 48 reflection-pattern conditions.

The experiment consists of 2 parts which are tested in separate listening sessions. Part 1 studies the influence of the direction of reflections. First, sound instances are presented from the front of the listeners with the angle of the prominent reflection $\phi_0 = 0°$ and the results are compared to lateral reflections with $\phi_0 = 90°$. Signals S are composed of 4 identical linearly faded pulses defined by their length t_{pulse} and onset and release times $t_{\text{in/out}}$ followed by a pause of length t_{pause}, cf. Figure 2.18. In contrast to the randomly created pink noise used in previous experiments, the basis $x(t)$ of pulses in this experiment is a deterministic pink complex tone with

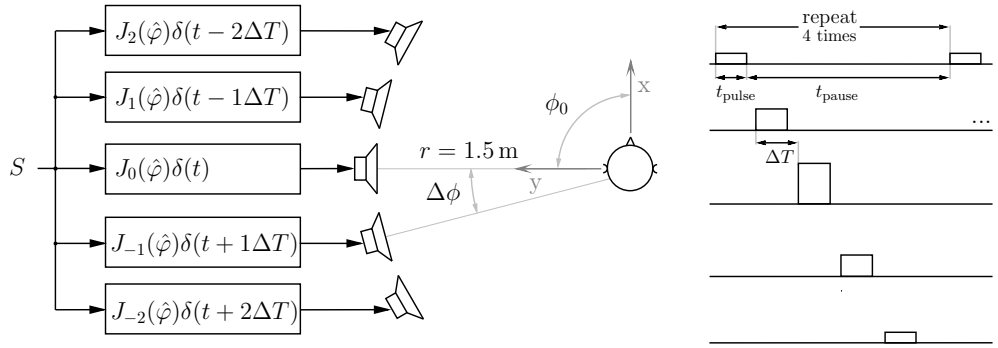


Figure 2.18: Block diagram of the experimental setup (left) and noise burst train (right) used in Experiment 3. Each pulse, i.e., reflection, of a condition with the same length t_{pulse}, onset and release times $t_{\text{in/out}}$ is defined by its delay $m \cdot \Delta T$ and weight J_m.

a fundamental frequency of 40 Hz and harmonic components between 120 Hz and 20.4 kHz, each of which using random phase offset φ:

$$x(t) = \sum_{k=3}^{510} \frac{\sin(2\pi\, 40\, k\, t + \varphi_k)}{\sqrt{k}}. \tag{2.11}$$

To maintain determinacy the phase is calculated pseudo-randomly for each component k with $\varphi_k = 2\pi\ \mathrm{sinc}(1000\, k\, \pi)$.

Part 2 studies the influence of the signal spectrum of the excitation signal. The lateralizations of the full band pink complex tone x of Eq. (2.11) is compared to high- and low-pass versions of x. Both parts of Experiment 3 are conducted at IEM's anechoic laboratory using a circle of level- and delay-compensated Genelec 8020A loudspeakers with a radius of $r = 1.5$ m adjusted at ear height. Five-channel audio files are created by convolving the impulse response of the five active loudspeakers for each condition with the corresponding signal S. During the listening experiment, the listeners were requested to keep their heads immobile and to face the central loudspeaker at $\phi = 0°$. Their task was to determine the lateralization of up to two auditory events by pointing with a motion-tracked toy gun [FMSZ10]. The corresponding azimuth angles were stored when the listener pressed a key on a keyboard. In contrast to the previous experiments, the stimuli were presented only after pressing a key but listeners were given control over the playback and could repeat pressing the respective key as often as they wanted. This is because informal listening of the author indicated that by build-up of precedence (presenting conditions in a loop), no differences might be perceivable. Accordingly, lateralization is then always determined by the sound instance containing the most energy.

2.4.2 Experimental results

The lateralization of three different signals S delineated by the envelopes of their pulses is studied: short pulses with instantaneous onsets (S_1), long duration pulses with long onsets (S_3), and pulses with medium onset and duration times (S_2). Respective pause times were chosen in a way that each train consisting of four pulse-pause combinations had a length of 1 s. Table 2.2 lists the pulse composition of the conditions that were investigated.

Table 2.2: Signal parameters of the conditions under investigation in Part 1 of Experiment 3.

signal	t_{pulse}	t_{in}	t_{out}	t_{pause}
S_1	1 ms	0 ms	0 ms	249 ms
S_2	15 ms	10 ms	5 ms	235 ms
S_3	250 ms	200 ms	50 ms	0 ms

The negative filter weight $J_{-1}(\hat{\varphi})$ is always assigned to the chronologically second active loudspeaker. For frontal playback ($\phi_0 = 0°$), all conditions were tested in full permutation with two repetition, whereas for lateral playback ($\phi_0 = 90°$) the loudspeaker separation is restricted

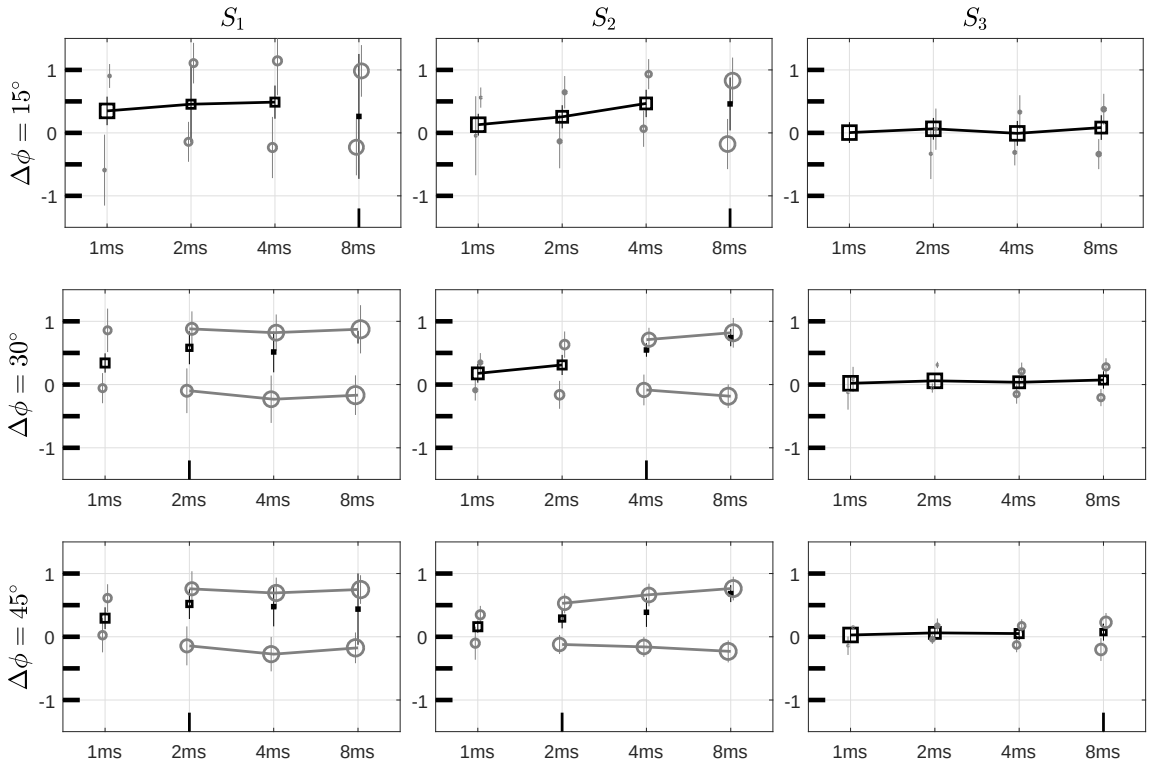


Figure 2.19: Normalized lateralizations over delays ΔT for signals S and loudspeaker spacings $\Delta\phi$ with the most prominent loudspeaker at $\phi_0 = 0°$. Squares indicate the mean result across all listeners and representations for single percepts and circles for left/right (primary/secondary) percepts; vertical bars indicate corresponding ± 1 standard deviations. The markers are scaled with regard to relative response frequency across all listeners and representations. Loudspeaker positions (1 = left … -1 = right) are indicated at the left of each plot and the ET (threshold criteria 50%) at the bottom. Along the delay, neighboring means below and above the ET are connected.

to $\Delta\phi = (15°, 45°)$ with two repetitions. This leads to 240 conditions in total, that are equalized in level and presented randomly at $58\,\mathrm{dB(A)_{eq}}$ to each of the twelve listeners.

Frontal reflections. In the analysis of the listeners' angular responses for $\phi_0 = 0°$, perceived directions are either classified as *single* direction, obtained by one-fold responses, or as the *left* and the *right* direction of a two-fold response.

An ANOVA of single, left, and right directions (normal distribution according to Lilliefors test, $p > 0.05$) revealed that the side of the leading sound is not significant if lateralizations are mirrored accordingly. Thus, further analysis is done with a combined set. Figure 2.19 shows means and ± 1 standard deviations of collected answers over delays normalized to the separation $2 \cdot \Delta\phi$. In this way the lateralization and chronology of pulses is $(1, 0.5, 0, -0.5, -1)$, i.e., left to right. Mean values for responses containing one perceived direction are represented as squares and means for responses containing two perceived directions are represented as circles. The respective marker size informs about the relative frequency of the corresponding responses.

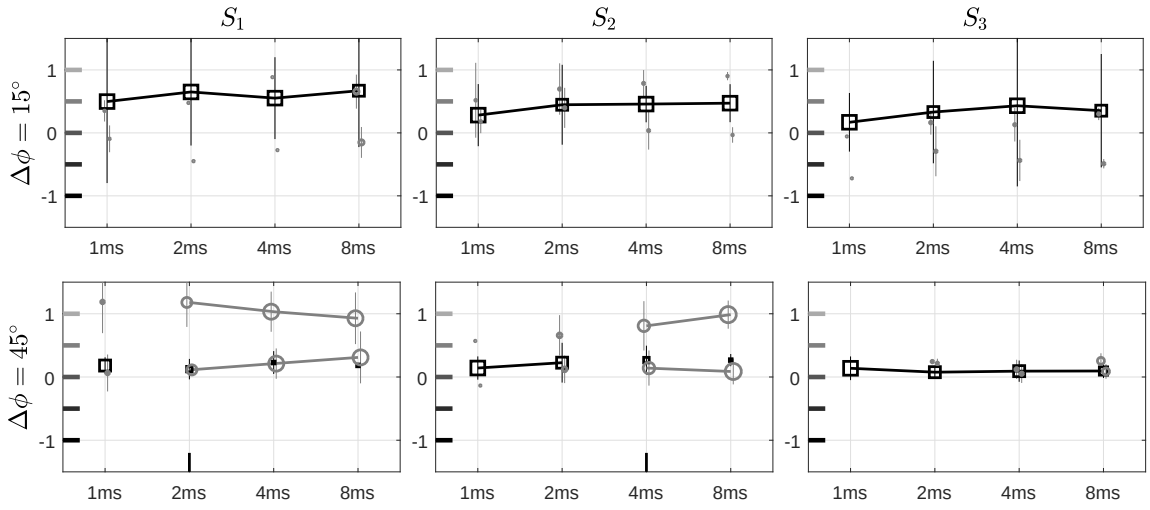
For all signals S and directional separations $\Delta\phi$, at the shortest delay $\Delta T = 1\,\mathrm{ms}$ mainly one fused sound image was heard. As the delay increases a second image emerges. The shortest delay at which more than 50% of collected answers consist of two directions is defined as echo

threshold (ET), which is marked by a vertical bar at the bottom of panels in Figure 2.19. For all signals and in accordance with the results of lead-lag experiments presented in Experiment 1, cf. Section 2.2, the ET varies with directional separation and is longer if sound instances arise from adjacent directions. Moreover, the critical effect of signal onset is seen in the data and ETs are shorter for transient sounds compared to sound with long onset times. The ANOVA of individual ET reveals the signal ($p \leq 0.05$) and directional separation ($p \leq 0.05$) to be significant parameters. Moreover, in accordance with Experiment 1 the ET is found to be subjective: while individual ETs of most listeners (8/12) correlate with modal ETs ($R > 0.80$), for some listeners the ET is independent of signal or loudspeaker spacing, and for others the ET is not reached by any condition.

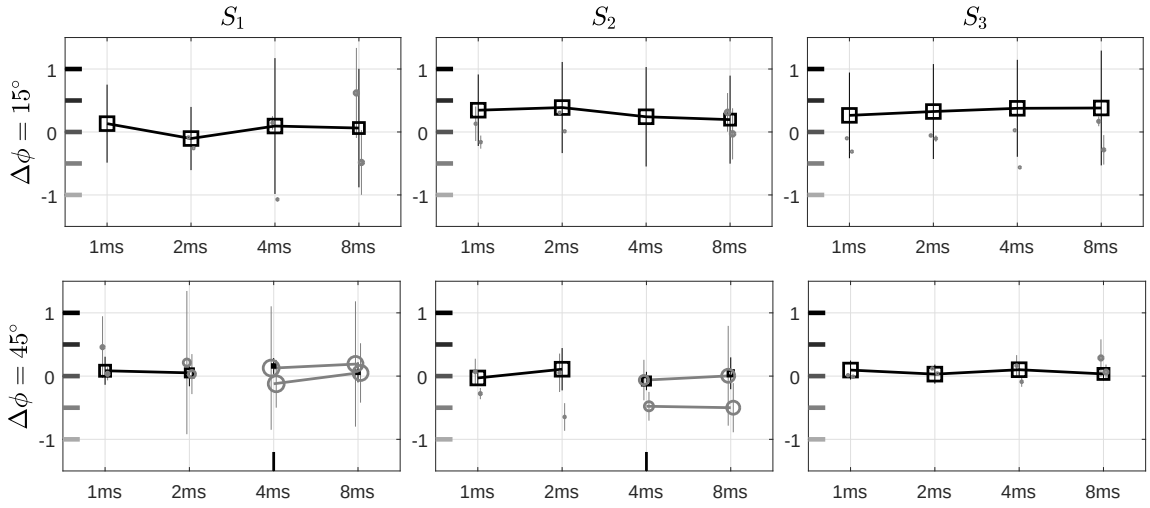
One-fold responses for conditions above the individual ET are significantly closer to the individual left mean, revealing the left direction to be the more prominent one: single directions merge into corresponding left directions above the individual ET. This is underlined by the standard deviations of respective answers, showing the variability of right directions to be significantly higher than the standard deviations of more lateralized left directions ($p \leq 0.05$, t-test). Therefore directions are classified as primary (left) and secondary (right).

For all $\Delta\phi$, mean values of primary directions for S_1 are fully lateralized indicating the transient precedence effect. The slight offset with regard to the outermost loudspeaker for $\Delta\phi = (15°, 45°)$ is in accordance with the typical direction-dependent over- and underestimation for lateral sound sources [Bla97]. Single direction means of S_1 are indicating precedence similarly. For the shortest delay $\Delta T = 1\,\text{ms}$ the fused image is perceived near the direction of the most prominent loudspeaker. By increasing ΔT, localization dominance becomes active and the lateralization of fused sounds is determined by the directional information of the leading loudspeaker. The mean values of single and primary directions for S_2 are less lateralized compared to S_1, and reflect a weakened transient precedence effect with increased ETs. Signal S_3 is mainly localized at the direction of the most prominent loudspeaker. Although the ET is reached for $\Delta\phi = 45°$, the left and right lateralization means are perceived symmetric with regard to the single mean. This is in contrast to the asymmetry obtained for $S_{1,2}$ and it is assumed that more than 50% of the answers characterize the width of the fused image rather that the lateralization of two distinct images. Thus, there is no (or only a weak) precedence effect and localization is dominated by the loudest loudspeaker which is further underlined by the standard deviations of S_3 being significantly smaller than for S_1 and S_2 ($p \leq 0.05$, t-test).

Right or secondary images are perceived mainly at the direction of the most prominent loudspeaker. Tan et al. in [TTY00] studied the precedence effect with two lagging source directions. They argue that secondary auditory images are due to a summing localization interaction between lagging sources as the secondary image was perceived only when the delay between the two lagging sounds came within the summing localization range. However, this assumption cannot be proven with the results of Experiment 3 because delays required to provoke a second image are beyond this range.



(a) counterclockwise: front to rear



(b) clockwise: rear to front

Figure 2.20: Shifted and normalized lateralizations over delays ΔT for signals S and loudspeaker spacings $\Delta\phi$ with the most prominent loudspeaker at $\phi_0 = 90°$ plotted separately for both playback directions. Squares indicate the mean result across for single percepts and circles indicate mean front/rear percepts with corresponding ± 1 standard deviations. The playback chronology of loudspeakers (1 = front ... -1 = rear) is coded in gray by corresponding markers at the left of each plot (gray = start; black = end).

Lateral reflections. For reflections presented from the lateral hemisphere the playback direction of pulses significantly influences the lateralization. Figure 2.20 shows mean lateralizations across all listeners and representations for single and double percepts with corresponding ± 1 standard deviations separately for both playback directions. For representation, the direction of the prominent loudspeaker is compensated and the lateralization is normalized to the separation $2{\cdot}\Delta\phi$ similar to Part 1. In this way the normalized lateralization of 0 equals ϕ_0 and the chronology of pulses is either $(1, 0.5, 0, -0.5, -1)$, i.e., clockwise front to rear, or $(-1, -0.5, 0, 0.5, 1)$, i.e., counterclockwise rear to front. Perceived directions are either classified as *single* direction, obtained by one-fold responses, or as *primary* and *secondary* direction of a two-fold response.

Regarding ETs, for lateral reflections with $\Delta\phi = 45°$ an overall increase is observed. Contrastingly, with a separation of solely $\Delta\phi = 15°$ the ET is not reached at all. The increase of

the localization uncertainty for lateral sound sources is likely the cause for not reaching the ET. According to Blauert [Bla97] a noise burst presented from $\phi = 90°$ is underestimated by 10° with a smallest perceivable change, i.e., uncertainty, of lateralization of $\pm 10°$. Thus, the uncertainty is in the range of the smallest loudspeaker spacing $\Delta\phi = 15°$, yielding high answer variations of single-fold answers. Several listeners reported that although they perceived more than one image, for some conditions it was hard to assign corresponding lateralizations.

Distributing pulses from front to rear with $\Delta\phi = 45°$, cf. Figure 2.20 (a), yields the precedence effect, and mean lateralizations resemble means obtained for frontal playback, cf. Figure 2.19. However, although primary directions are lateralized all the way to the first active loudspeaker, the localization dominance is weakened and single directions are hardly lateralized. By reversing the playback direction (rear to front), the precedence effect is further weakened and primary directions for S_2 are less lateralized. Surprisingly, primary directions for the more transient S_1 are less lateralized compared to S_2. Taking a closer look to individual responses reveals a bimodal answer distribution, and it is assumed that for the salient signal S_1 front-back confusion occurs more often, yielding a mean lateralization at (or near) the interaural axis.

Spectral content. Part 2 studies the lateralization of 400 ms-long pulses and varies onset and release times. Thus, the leading pulse temporally overlaps with all lagging pulses. The composition of pulses delineating signals $S_{4\ldots8}$ is listed in Table 2.3. The influence of the signal spectrum is evaluated by filtered versions of signals $S_{5\ldots7}$. Both low-pass and high-pass filtering is attained by restricting the component $k = 3\ldots25$ of the signal x in Eq. (2.11) for low-pass signals, and $k = 25\ldots510$ for high-pass signals. The playback employs a fixed loudspeaker spacing of $\Delta\phi = 30°$ with $\phi_0 = 0°$. Other system parameters are tested in full permutation leading to 80 full-band conditions and 96 filtered conditions. Full-band conditions are evaluated twice and filtered conditions are evaluated once in random order by each of the eight listeners.

Table 2.3: Signal parameters of the conditions under investigation in Part 2 of Experiment 3.

signal	t_{pulse}	t_{in}	t_{out}	t_{pause}
S_4	400 ms	0 ms	200 ms	100 ms
S_5	400 ms	10 ms	200 ms	100 ms
S_6	400 ms	200 ms	200 ms	100 ms
S_7	400 ms	200 ms	10 ms	100 ms
S_8	400 ms	200 ms	0 ms	100 ms

Responses are pooled by mirroring responses of conditions with the right-most loudspeaker leading ($p > 0.05$). An ANOVA of resulting means of single, primary (left), and secondary (right) directions revealed, that the assignment of the negative weight $J_{-1}(\hat{\varphi})$ is not significant, independent of the signal spectrum. Although most studies state that a reflection need not be identical to the direct sound to be suppressed, e.g., [BD88, Zur80], a deviating lag tends to weaken the precedence effect. Divenyi [Div92] for example found that a single opposite-phase reflection of a narrow band pulse centered at 2 kHz is not suppressed. Further evaluation of

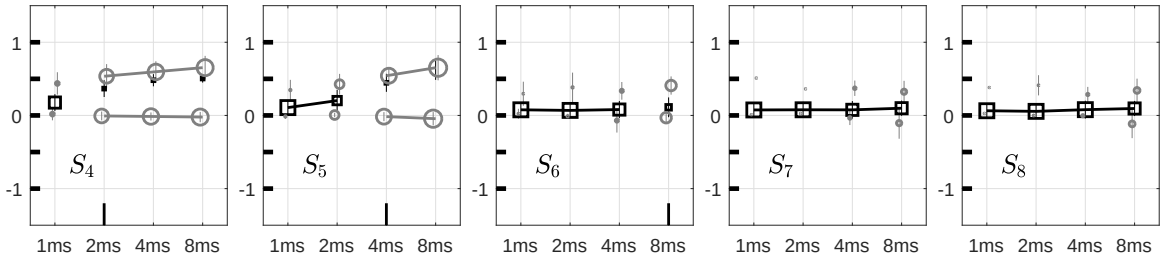


Figure 2.21: Normalized lateralizations of full-band signals $S_{4...8}$ over ΔT with $\Delta\phi = 30°$. The data is pooled of all listeners, both playback directions (mirrored, starting at 1), sides of the negative weight (leading/lagging), and repetitions. Squares represent means for single percepts and circles for primary/secondary percepts with corresponding ± 1 standard deviations.

the data presented here is done using a combined data set containing both leading and lagging negative weights.

Responses for full-band signals $S_{4...8}$ are given as mean values for single and primary/secondary directions with corresponding ± 1 standard deviations in Figure 2.21. Signal conditions $S_{4,5}$ strongly resemble their corresponding conditions $S_{1,2}$ at $\Delta\phi = 30°$ and indicate the onset capture effect. The ETs of $S_{4,5}$ converge to the same delays ΔT as $S_{1,2}$ at $\Delta\phi = 30°$ and localization dominance is shown for single directions. However, the lateralization of single/primary directions of $S_{4,5}$ is less distinct and precedence is weakened by the temporal overlap of the lead with its lags. Moreover, in contrast to secondary directions of not temporally-overlapping signals $S_{1,2}$, which are slightly lateralized to the right, the secondary directions of signals $S_{4,5}$ are collocated with the direction of the prominent reflection. Signals $S_{7,8}$ resemble the no-precedence condition S_3. Interestingly, for S_6 the echo threshold is reached at $\Delta T = 8\,\text{ms}$ and similar to S_3 at $\Delta\phi = 45°$ primary/secondary directions are assumed to characterize the apparent source width.

The influence of the spectral content on signals $S_{5...7}$ is shown in Figure 2.22. All means of low-pass and high-pass signals strongly resemble respective full-band conditions. While for the more transient signal S_5 the onset capture is obtained also for the filtered versions, long onset times of $S_{6,7}$ do not yield precedence. Interestingly, all individual ETs of the high-pass S_5 are within the range of delays that are tested and the mean ET agrees with full-band S_5, whereas individual ETs of the low-pass S_5 are significantly increased and three listeners reported to hear only fused sounds. This finding is in accordance with [ESN68] showing an influence of signal spectrum on the ET.

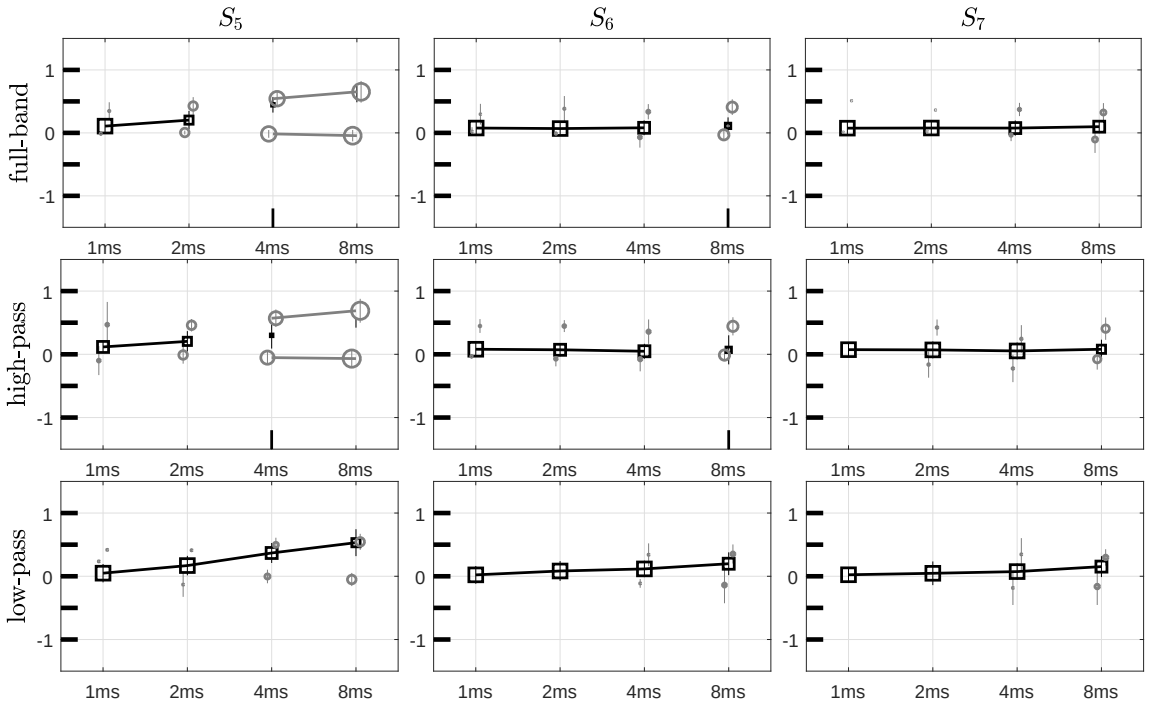


Figure 2.22: Normalized lateralizations over ΔT for $S_{5...7}$ with different signal spectrum at $\Delta\phi = 30°$. Squares represent means for single percepts and circles for primary/secondary percepts with corresponding ± 1 standard deviations.

2.4.3 Discussion

This section systematically evaluated three precedence effect types known from two-source experiments for multiple reflections with increased lag levels. The onset capture effect, the ongoing precedence effect, and the onset capture effect were examined over different inter-signal intervals and angles of incidence, and directional separations. An all-pass array consisting of five active horizontally arranged loudspeakers was used and lateralizations of differing signal envelopes were acquired in Experiment 3.

In Part 1 of Experiment 3 the transient precedence effect is shown for complex reflection patterns with a critical effect of signal onset of signals $S_{1,2}$ on lateralization. Findings from two-source Experiments 1 and 2 could been applied to multiple sources and the suppression increases with decreasing directional separation. Similarly, for lateral presentations precedence is shown for widely spread reflections with $\Delta\phi = 45°$. However, the playback direction was found to be critical and starting the reflection patterns from the front of the listener ($\phi = 0°$) yields more consistent lateralizations compared to reflection patterns that start from the rear ($\phi = 180°$). The ongoing precedence effect expected for smooth-onset signal S_3 could not been obtained by any condition as the ongoing leading pulse with $-24\,\mathrm{dB}$ compared to the loudest pulse was not prominent enough to dominate the localization. Instead, the ambiguous ongoing sound, with incoming reflection from different directions, induced a single fused percept of S_3 at a weighted-average position, with weights depending on relative sound intensities.

Part 2 of Experiment 3 quantifies the onset capture effect with multiple lags. The onset capture effect was shown to be robust against modifications of phase of the second sound instance, which

emphasizes the importance of the signal onset. Regarding signal spectrum, the echo threshold significantly increased for signals without high frequency components, while lateralizations remain unaffected. Overall the transient precedence effect was found to be the strongest for both fusion and localization dominance, followed by the onset capture effect. The ongoing precedence effect could not be observed with any condition of this experiment as fusion and localization dominance were not sufficiently pronounced. Last but not least, evidence is found that the perception of secondary sounds is determined by a level-weighted average of the source locations.

2.5 Perception of static and dynamic directivities in a room

The knowledge of the localization of a real sound source projecting its sound beams on reflective boundaries is still rough. However, literature offers extensive knowledge from laboratory experiments studying the perception of direct sound and a single reflection. Lateralization of fused images for lead-lag stimuli with delays above summing localization show a high degree of inter- and intra-subjective variability. Interaction effects between sound instances yield an oscillation in the lateralization over delay [BC78, TH99b, DC06], which becomes even worse, if the reflection level is increased [PB15]. Similarly, lateralizations of stimuli consisting of multiple reflections studied in the previous section are highly variable if presented from the side of the listener, whereas a frontal presentation yielded relatively consistent perceptions. Accordingly, Zotter and Frank [ZF15] found relatively little variation in the lateralization of auditory events elicit by simulated reflection patterns of a directional sound source in front of the listener.

To study if a beamformer like the IKO is able to produce inter-subjective perceptions, this section investigates the spatial perception of auditory events for different sound projections in a room. Experiment 4 is structured in 3 parts presenting listening experiments, which were carried out in the IEM lecture room, a $8.3\,\mathrm{m} \times 7.0\,\mathrm{m} \times 3.0\,\mathrm{m}$ large shoebox-shaped room with $T_{60} = 0.57\,\mathrm{s}$. Part 1 and Part 2 investigate static and dynamic auditory events, respectively. These parts are excerpts of a series of listening experiments conceived, designed, and conducted by Franz Zotter and Matthias Frank with Gerriet K. Sharma composing the stimuli and the author analyzing the data. The comprehensive results of the listening experiments are published in [WSF+17] and appear also in [Sha16, ZZFK17, ZF19]. The last part examines two approaches for increasing the lateralization of auditory events. Corresponding listening experiments were conducted within the context of a student's project by Julian Linke which was supervised by the author and results are published in [LWZF18, LWFZ18].

Figure 2.23: Experimental setup of Part 1 and 2 with the IKO_1 in the IEM lecture room. Answers were obtained by the marker placement on the graphical user interface shown in the foreground.

2.5.1 Static sound beams

Sound projection of static sound beams studied in Part 1 of Experiment 4 is achieved by four different third-order spherical directivity patterns using IKO_1. Three of them are perpendicular to the front, left, and rear wall, respectively, whereas $\phi_\bigcirc = 235°$ is in line with the specular direction of the right wall. The setup is delineated in the left panel of Figure 2.24. The hexagon and the head symbol indicate position of the IKO and the listener. Respective beam angles $\phi_\bigcirc = (0°, 90°, 180°, 235°)$ are indicated around the hexagon.

Conditions. To study how the precedence effect influences lateralization, the perception of transient signals is compared to signals with smooth onsets. The excitation signal S_1 consists of a sequence of irregular transient bursts and signal S_2 consists of a 1.5 s-long pink noise sample, with a smooth linear fade-in and fade-out of 0.5 s, followed by a 0.5 s-long silent gap. A condition consisted of signal S played back in loop and spatialized with the directivity direction $\phi_\bigcirc$. Each randomly presented condition is examined twice and listeners have to determine the location of the auditory event by placing a marker on a graphical user interface showing the two-dimensional outline of the setup. During the experiment listeners are free to move their heads while seated and can listen to each condition as long as they want.

Results. For a compact representation of individual data points obtained by 15 experienced listeners, two-dimensional mean values and the corresponding 95% confidence ellipses are computed. Beam directions are color coded and indicated around the IKO. The signal onset, i.e., transient or smooth, is coded by the chroma. With this setup, the delay between direct sound and prominent reflections is in the range of $8\,\text{ms} < \Delta T < 20\,\text{ms}$. Thus, lateralization is determined by the build-up of precedence rather than by summing localization. Half-axes of the 95%-confidence ellipses are mostly tangential/perpendicular to the listener-centric circular grid, indicating the

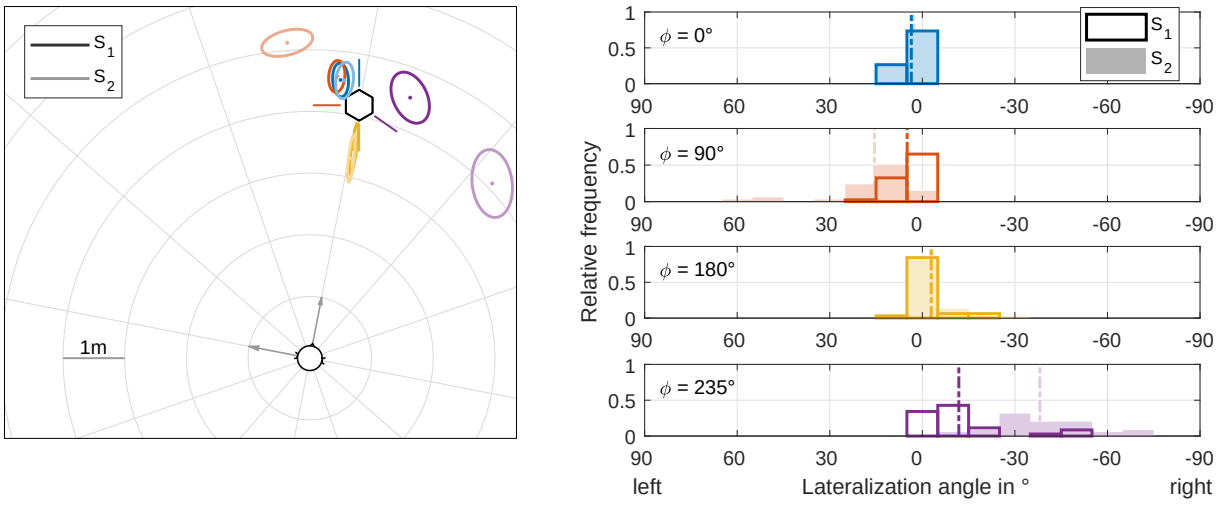


Figure 2.24: Left: 2D-means and corresponding 95% confidence ellipses of auditory events elicit by directivity directions $\phi_\bigcirc = (0°, 90°, 180°, 235°)$ for both signals types. Right: Mean and histogram of lateralizations for both signals and plotted separately for each beam direction.

lateralization and auditory distance to be independent from each other. The right panel of Figure 2.24 gives the histogram of obtained lateralizations for different beam directions separately for both signals. All beam orientations yield a shift away from the directional sound source that mostly agrees with the beam orientation. Lateral orientations $\phi_{\bigcirc} = (90°, 235°)$ mainly yield lateral shifts, whereas median-plane orientations $\phi_{\bigcirc} = (0°, 180°)$ shift the auditory events along the distance. For lateral orientations, the critical effect of signal onset indicates the precedence effect. Especially for the specular beam direction $\phi_{\bigcirc} = 235°$, smooth onset events almost stick at the reflective wall, whereas transient bursts are perceived in the direction of the IKO with a difference of mean lateralizations of 26°. Similarly but slightly less distinct, the reflected energy of the beam at $\phi_{\bigcirc} = 90°$, is sufficient to pan the smooth-onset noise away from the direction of the direct sound and the difference of mean lateralizations is 10°.

Comparing the size of 95% confidence ellipses reveals small variations in lateralizations for median-plane orientations. The directional alignment of direct sound and prominent reflection can be considered as possible reason. In contrast, for directivity directions, where the prominent lag deviates from the direction of the direct sound, the spread of perceived lateralization increases. Interestingly, the histogram given in the right panel of Figure 2.24 shows bimodal lateralizations of the smooth signal S_2 at $\phi_{\bigcirc} = 90°$ and of the transient signal S_1 at $\phi_{\bigcirc} = 235°$. While most listeners perceived the respective auditory event near the directional sound source, for some of them level or delay of the prominent reflection may have exceeded the individual echo threshold yielding a prominent auditory event at the direction of the prominent reflection. However, except for these few outliers the results exhibit the lateralization perception to be consistent.

2.5.2 The extended energy vector

Shinn-Cunningham et al. in [SCZD93] model the perceived lateralization in a precedence situation by a weighted sum of interaural delays and quantify the localization dominance of the lead over a single lag by the weight w. Reformulating their model as two dimensional vector base with base vectors $\boldsymbol{\theta}_{1/2} = [\cos(\phi), \sin(\phi)]^T$ of unit length pointing at the direction of lead (index 1) and lag (index 2) yields:

$$\boldsymbol{r} = w\,\boldsymbol{\theta}_1 + (1 - w)\,\boldsymbol{\theta}_2. \tag{2.12}$$

where the vector $\boldsymbol{r}$ points at the perceived direction and w is related to the strength of the precedence effect, i.e., the amount of the stationary signal. The weight w is defined in the interval $[0, 1]$; precedence results give values of $w > 0.5$ and with transient stimuli in the localization dominance region w is typically close to 1. Interestingly, this formulation corresponds to the *velocity vector* $\boldsymbol{r}_V$, which has been proposed by Gerzon in [Ger92] and referred to as the *Makita theory of localisation* [Mak62]. The velocity vector is intended to predict the perceived direction of a phantom source utilizing amplitude panning on 2 loudspeakers at directions $\boldsymbol{\theta}_{1,2}$ and gains

$g_1 = w$ and $g_2 = 1 - w$. Its general form for an arbitrary number of L loudspeakers in 3D is

$$\boldsymbol{r}_\gamma = \frac{\sum_{l=1}^{L} g_l^\gamma \boldsymbol{\theta}_l}{\sum_{l=1}^{L} g_l^\gamma}, \tag{2.13}$$

with the exponent $\gamma = 1$ yielding the $\boldsymbol{r}_V$ vector. The direction of this vector is assumed to correspond to the localization of low-frequency signals where a constructive interference of sound instances is expected and the formulation is equal to the tangent law in stereophony [Pul97].

In line with the energy interpretation of the perception of diffuse reflections in Section 2.3.4, recent studies assume an energetic superposition of the sound instances and use the *energy vector* $\boldsymbol{r}_E$ with $\gamma = 2$ to successfully model the horizontal localization. Frank [Fra13] used it to predict localization for various spatial audio techniques. At off-center listening positions, the distances to the loudspeakers, i.e., sound instances, are not equal anymore, resulting in additional attenuation and delay for each loudspeaker depending on the position. This effect can be incorporated into the energy vector by additional weights $w_{r,i}$ and $w_{\tau,i}$ yielding

$$\boldsymbol{r}_E = \frac{\sum_{l=1}^{L} (w_{r,l}\, w_{\tau,l}\, g_l)^2 \boldsymbol{\theta}_l}{\sum_{l=1}^{L} (w_{\mathrm{r},l}\, w_{\tau,l}\, g_l)^2}. \tag{2.14}$$

Assuming a point-source-like propagation results in $w_{\mathrm{r},l} = 1/r_l$.

The modeling of suppressed contributions of delayed signals, i.e., precedence, requires a weight w_τ that attenuates the lagging signals in order to reduce their influence on the predicted directions. This transformation from lth reflection's delay τ_l to level is achieved by:

$$w_{\tau,l} = 10^{\frac{\beta}{20}\tau_l}, \tag{2.15}$$

with the attenuation β in dB/ms.

In an early approach [ZFFR14] the attenuation is modeled by a linear slope of $\beta = -0.25\,\mathrm{dB/ms}$. Similar to Eq. (2.6) in Section 2.3, this value is based on the modeling of the echo threshold in [RHH00]. Recently, Kurz [Kur18] studied the localization at off-center listening positions with surrounding loudspeaker systems and found the attenuations to be signal-dependent by minimizing the model prediction error. According to [Bla97] and supported by findings in Section 2.2, echo threshold slopes are similar for a specific signal but independent of the echo thresholds definition. Hence, similar to [ZFFR14], the attenuation β is obtained from measures of the echo threshold. In the interesting range $\Delta T \leq 40\,\mathrm{ms}$ echo thresholds are approximated with linear fit, cf. Eq. (2.6). Mean fitting parameters of Experiment 1 studying the perception of 10-ms transient noise bursts are found with $\beta_{\mathrm{ET}} = -0.55\,\mathrm{dB/ms}$ for the first stage and $\beta_{\mathrm{ET}} = -0.69\,\mathrm{dB/ms}$ for the second stage, cf. Section 2.2. For 50-ms transient bursts of Experiment 2 the slope is $\beta_{\mathrm{ET}} = -0.47\,\mathrm{dB/ms}$, cf. Section 2.3. Similarly, Damaske [Dam71] found the echo threshold to vary with burst length with steeper shapes for shorter pulses. However, a more prominent parameter influencing the echo thresholds slope β_{ET} is the transient time amount of the signal.

For naturally produced sounds such as speech, echo thresholds are found to vary systematically among syllables [MLK09]. Linearly fitting experimental data examining speech yields slopes ranging from $\beta_{\mathrm{ET}} = -0.20\,\mathrm{dB/ms}$ to $-0.35\,\mathrm{dB/ms}$ [LB58, San76, RHH00]. Similarly for music, fitted slopes are content dependent and resemble those of speech [DK86]. Hence, the slope β_{ET} depends on the non-stationariness of the signal and more transient signals yield steeper slopes β_{ET} compared to smooth-onset signals. At the other end, if no temporal cue is provided, e.g., pure tones or stationary decorrelated noise without onset, the precedence effect vanishes, resulting in a purely energy-averaged percept with $\beta = 0\,\mathrm{dB/ms}$ and $w_\tau = 1$, cf. *Franssen* illusion [HR89].

Stitt [Sti15] introduced further weights in order to model the angle-dependency of the precedence effect for surrounding loudspeaker systems. However, listening test results of Kurz and Frank [KF17] modeled by different extensions of the energy vector revealed that the simple weighting with $w_{\mathrm{r},i}$ and $w_{\tau,i}$ is sufficient for predicting the perceived direction. This agrees with the findings of Experiment 1 in Section 2.2 for which the parameter spatial separation $\Delta\phi$ was found not to be significant within the delay $\Delta T \leq 40\,\mathrm{ms}$.

2.5.3 Modeling the lateralization of static sound beams

For modeling the lateralization of auditory events the extended energy vector is used. Instead of an elaborate room impulse measurement for finding exact model parameters, a simplified multi-path propagation model, i.e., image-source model [AB79], is used as a first approach. Directions and timing of sound instances are calculated from the room geometry, whereas for calculating the levels an ideal frequency-independent source directivity and a frequency-independent absorption coefficient $\alpha_0 = 0.3$ are considered. The latter results from Sabine's reverberation equation:

$$\alpha_0 = 0.16\frac{V}{T_{60}\,S}. \tag{2.16}$$

with measured volume V, surface S and reverberation time T_{60} of the IEM lecture room. To account for the reflecting blackboard on the front wall and absorbers on the rear wall (cf. Figure 2.23) corresponding absorption coefficients are subsequently adjusted to $\alpha_{\mathrm{front}} = 0.1$ and $\alpha_{\mathrm{rear}} = 0.9$.

The signal-dependent slope β, which incorporates the precedence effect in the the energy-vector in Eq. (2.14), is found by minimizing the cost function

$$J_{\mathrm{RMS}}(\beta) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\phi_{\boldsymbol{r}_E}(\beta) - \bar{\phi}_n)^2}, \tag{2.17}$$

with the lateralization angle $\phi_{\boldsymbol{r}_E}$ calculated form the extended-energy vector and mean perceived lateralizations $\bar{\phi}$ elicit by $N = 4$ beam directions $\phi_{\bigcirc}$.

Corresponding directions of the extended energy vector are represented as dashed lines in Figure 2.25 for transient S_1 with $\beta = -1.51\,\mathrm{dB/ms}$ (left panel) and smooth S_2 with $\beta = -0.44\,\mathrm{dB/ms}$

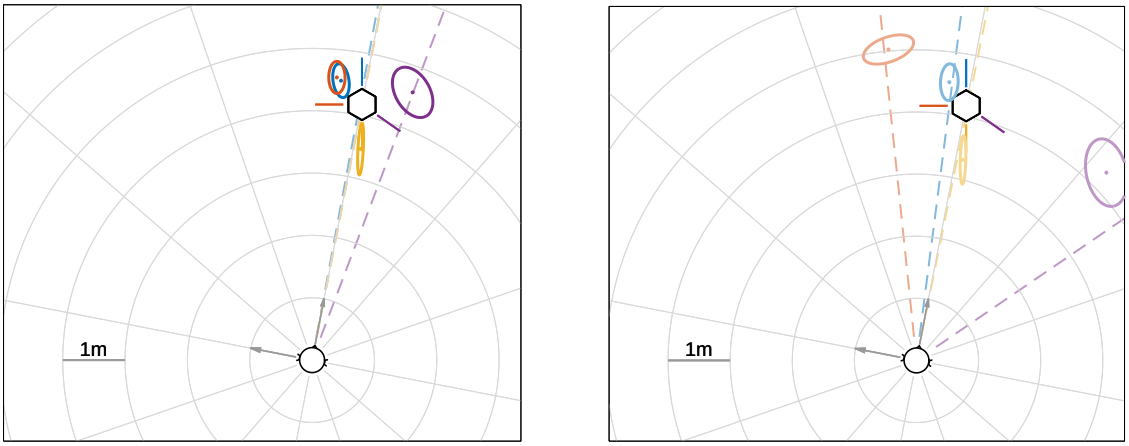


Figure 2.25: Modeling the results of the transient S_1 (left) and the smooth S_2 (right) using the extended energy vector $\boldsymbol{r}_E$ assuming an ideal 3rd-order directivity with frequency-independent absorption coefficient.

(right panel). For most beam directions the vector based on the image-source model coincides with corresponding mean angles $\bar{\phi}$ yielding residual prediction errors $J_{\mathrm{RMS}}(\beta) = (3.9°, 5.2°)$ for signals $S_{1,2}$, respectively. However, slopes β are remarkably steeper than those found in other modeling approaches with the extended energy vector and do not comply with previously cited signal-dependent echo threshold slopes β_{ET}.

A more elaborated version of the image-source model considers the measured directivity of the IKO_1 by a weighted energy average for different frequencies, cf. Figure 1.5. The weight accounts for the respective signal spectrum and for the relative loudness (A-weighting). Figure 2.26 shows corresponding results for the advanced image-source model yielding slightly reduced prediction errors $J_{\mathrm{RMS}}(\beta) = (3.6°, 2.0°)$ for S_1 (left panel) and S_2 (right panel), respectively. More important than the decrease of prediction error are obtained slopes β, which agree with literature. For the transient signal S_1 the slope $\beta = -0.57\,\mathrm{dB/ms}$ agrees with values found in Experiment 1 and is between respective mean echo threshold slopes (first stage: $\beta_{\mathrm{ET}} = -0.55\,\mathrm{dB/ms}$; second stage $\beta_{\mathrm{ET}} = -0.69\,\mathrm{dB/ms}$) and is slightly above the value found in Experiment 2 ($\beta_{\mathrm{ET}} = -0.47\,\mathrm{dB/ms}$). The smooth-onset S_2 exhibits only weak precedence with $\beta = -0.12\,\mathrm{dB/ms}$ yielding fused percepts at energy-weighted average positions.

By incorporating the IKO's measured directivity in the image-source model, the extended energy vector is also able to assess the answer spread for different beam direction. Figure 2.27 compares energy vectors calculated for third-octave bands from 20 Hz to 16 kHz with individual answers for different beam orientations. The energy vector directions of lateral orientations $\phi_{\bigcirc} = (90°, 235°)$ exhibit a higher spread compared to median-plane orientation $\phi_{\bigcirc} = (0°, 180°)$, as do individual lateralizations.

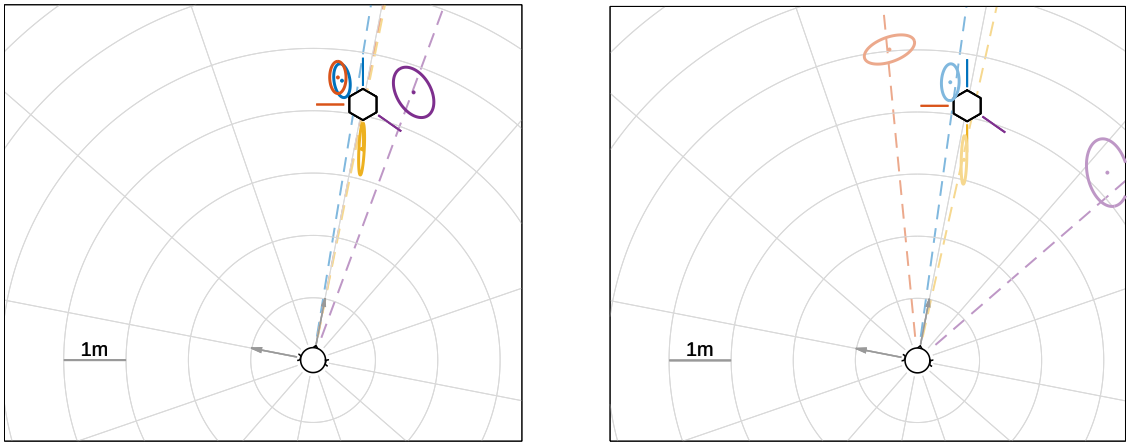


Figure 2.26: Modeling the results of the transient S_1 (left) and the smooth S_2 (right) using the extended energy vector $\boldsymbol{r}_E$ and considering the IKO's measured directivity.

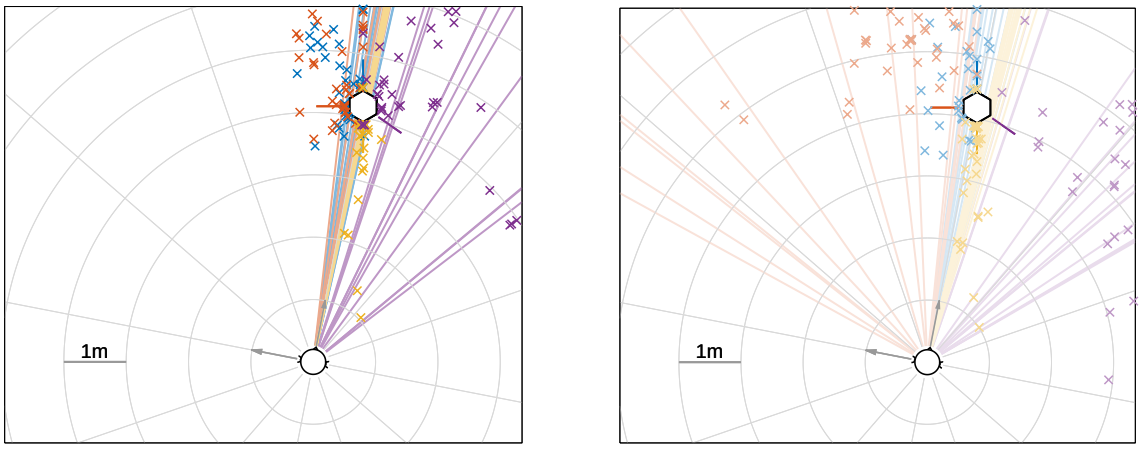


Figure 2.27: Modeling the individual lateralizations of the transient S_1 (left) and the smooth S_2 (right) using the extended energy vector $\boldsymbol{r}_E$ calculated separately for third octave bands (solid lines). The directional spread of the vector correlates with the spread of individual answers (×) for different beam orientations.

2.5.4 Dynamic sound beams

An auditory event spans a succession of acoustic events over time. The auditory system groups these separate stimuli together into a temporal sequence forming an auditory "stream", which can itself be termed as moving event. Surrounding amplitude-panning loudspeaker systems use dynamic gain modifications of adjacent loudspeakers to create an auditory stream, i.e., a moving auditory event between the loudspeakers. Beamformers offer an elegant way for controlling acoustic events; dynamic changes of the spatial projection by varying the beam direction or order, or by blending between multiple beams. Part 2 of Experiment 4 evaluates the perception of trajectories of sound beams.

Conditions. A circular trajectory consist of a single beam rotating counterclockwise by $\Delta\phi_\bigcirc = 360°$ in the horizontal plane around the IKO_1 starting at $\phi_\bigcirc = 90°$. For a pendulum trajectory, the amplitude of a beam at $\phi_\bigcirc = 90°$ is linearly faded out and at the same time the amplitude of a beam at $\phi_\bigcirc = 270°$ is faded in. Another pendulum trajectory reduces the effective

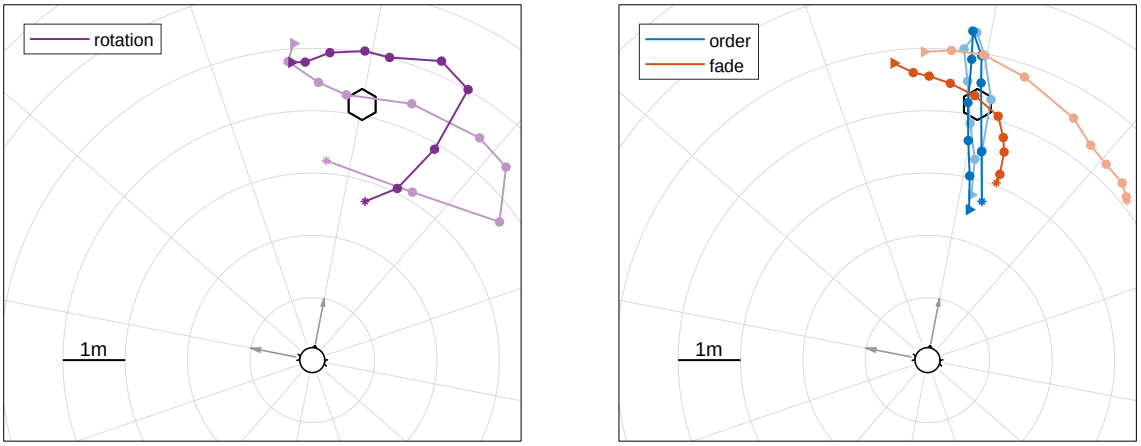


Figure 2.28: 2D-means of auditory events elicit by time variant beam trajectories. Trajectories were created by a full rotation of a 3rd-order beam (violet, left), by variation of the effective beam order (blue, right), and by a linear fade between two opposing 3rd order beams (red, right). Excitation signals were noise bursts (dark) and uniform pink noise (light).

beam order of a 3rd-order beam at $\phi_{\bigcirc} = 180°$ until 0th-order, i.e., becomes omnidirectional, and increases it again back to order 3. All movement trajectories (rotation/order/fade) are continuous, last for 5 s, and are played back in loop. In this way, the rotation trajectory and the order trajectory are closed in their movement, whereas the directivity of the fading trajectory jumps from $\phi_{\bigcirc} = 270°$ to $\phi_{\bigcirc} = 90°$ at its end. Listeners are asked to localize the auditory events and adjust ten markers in 2D Cartesian coordinate system to the perceived location in successive half-second intervals during playback in loop. Markers are flashing successively at the associated playback time, and can be moved using a mouse on a graphical interface showing the floor plan of the test setup. Excitation signals are a sequence of irregular transient bursts, known from Part 1 of the experiment, and uniform pink noise.

Results. For representation of the experimental data of beam trajectories obtained by 15 listeners two-dimensional means of the positions for each time interval of a condition are shown in Figure 2.28. Markers $\triangleright$ and $*$ indicate starting and ending position of the trajectory, respectively. Beam trajectories are color coded and the signal type is coded by the chroma.

A review of literature on dynamic spatial discrimination of a moving source by Stecker and Gallun [SG12] yielded the conclusion that the perception of sound source motion involves similar processes as the localization of steady sources. Accordingly, mean results for the circular trajectory resemble obtained means for the static beams as the course of the moving auditory event roughly meets static auditory events given in Figure 2.24. Interestingly, the critical effect of signal envelope on lateralization is seen only for the more prominent right reflection, especially when the beam orientation meets the direction of the specular reflection (7th and 8th marker). A similar tendency is seen for the fading trajectory at its end ($*$); although the corresponding beam orientation ($\phi_{\bigcirc} = 270°$) is not coincident with the specular reflection ($\phi_{\bigcirc} = 235°$), achieved lateralizations are comparable to the circular trajectory. This effect, which cannot be seen for the closed rotating movement, is explained by the *auditory representational momentum* [GL07]

yielding a displacement of the final position of a moving sound source in the direction of motion.

Agreeing with median-plane directions of the first experiment, varying the order yields mainly changes of distance impression of both sounds. This is explained by the variation of the direct-to-reverberant energy ratio (DRR), a cue for distance perception [ZBB05], which is influenced by the source directivity. This effect does not show any signal-dependency and perceived distances at the start and end of the closed trajectories are almost the same. Note that the auditory event elicit by an omnidirectional directivity (0th order) is not perceived at the physical position of the loudspeaker but behind it. The 3rd order beam facing the listener on the other hand is perceived closer compared to its static version in the first experiment.

Overall, trajectories including lateral beam directions shift the auditory events towards the more prominent lateral reflection, i.e., right wall. Nevertheless, continuous beam trajectories equally sampled in time, yield mostly equally distributed answers in space. In accordance with the *auditory saltation illusion* [BPWJ77], this suggests that the auditory system interprets consecutive "snapshots" of locations, i.e., discretely localizable directions as gradually moving auditory event. Thus, in accordance with studies on auditory motion elicit by a moving real sound source [CG92], it does not require continuous movement cues between these directions.

2.5.5 Increasing the lateralization

The experimental results presented in the previous sections revealed a difficulty in the detachment of transient auditory events from the physical location of the IKO. Part 3 of Experiment 4 examines two approaches to overcome the precedence effect to increase lateralization.

Adding reflectors. In the trade-off between delay and level, Part 1 (cf. Section 2.5.1) showed that more pronounced but temporally closer reflections are preferred. In line with this finding, reflectors are used to increase the reflected energy and the number of discretely localizable directions. Their influence on the of auditory movement is studied in a listening experiment [LWFZ18].

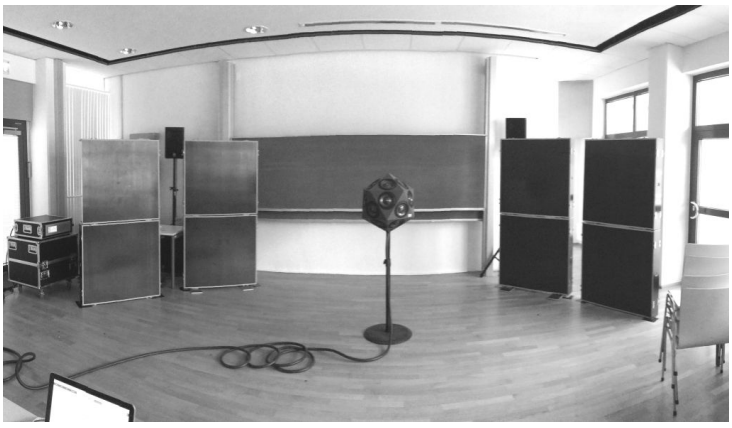


Figure 2.29: Experimental setup of Part 3 with the IKO_2 and reflectors at both sides in the IEM lecture room. In comparison to the setup tested in Part 1 and 2, listener and IKO are shifted slightly to the left.

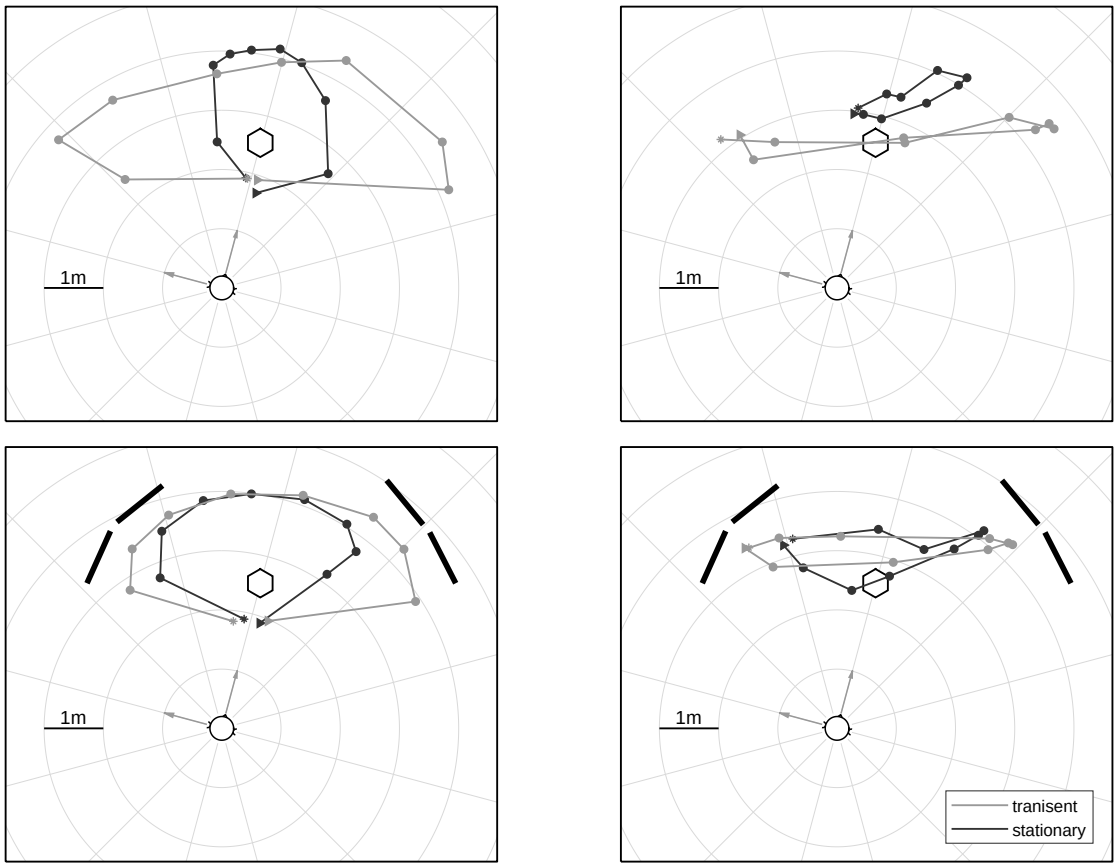


Figure 2.30: Mean values for each time interval coded in grayscale for transient and stationary sounds. Upper row: without reflectors. Lower row: with additional reflectors indicated as bold lines. Left column: single beam rotating counterclockwise by 360° starting from $\phi_{\bigcirc} = 180°$. Right column: linear fade between 2 opposing beams at $\phi_{\bigcirc} = \pm 90°$.

The excitation signals and test interface are similar to previous parts of Experiment 4 and spatial projections include the circular trajectory of a single beam starting at $\phi_{\bigcirc} = 180°$ and the fading trajectory, both lasting 5 s. In contrast to the previous parts, the fading between two opposed beams is done from left ($\phi_{\bigcirc} = 90°$) to right ($\phi_{\bigcirc} = 270°$) and back to left. In this way both movement trajectories are closed. To test the influence of reflectors on the perception, the session is consecutively conducted by each listener once with and once without acoustically reflecting baffles with the IKO_2 placed more in the center of the room. Ten experienced listeners participated in the experiment, half of them started the experiments with reflectors, the other half started without reflectors.

The upper row of Figure 2.30 shows mean values for both signal types and each movement trajectory without additional reflectors. The comparison to the previous results given in Figure 2.28 demonstrates how important the beamformer positioning can be when it comes to lateralization: Increasing the IKOs distance to lateral walls obviously opens space for auditory event positioning but at the same time decreases the reflected energy due to the propagation attenuation. However, the more centered setup of the IKO in the IEM lecture room yields more lateralized auditory event at both sides of the IKO than in Figure 2.28. Stationary auditory events are lateralized almost all the way to the lateral walls, whereas transient sounds elicit the

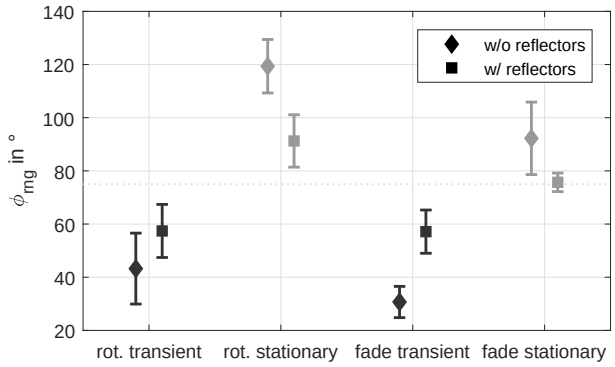


Figure 2.31: Means and 95% confidence intervals of lateralization range ϕ_{rng} computed individually for both trajectories, both signal types, and both reflector setups. The horizontal dotted line marks the medium aperture angle spanned by the reflector pairs.

precedence effect and tend to stay close to the IKO. Regarding auditory distance, findings from the previous experiments are confirmed and beams aligned to the median plane, i.e., to/off the listener, vary the distance impression, whereas lateral beams of the centered IKO are perceived mostly at the physical distance of the beamformer.

Adding reflectors (lower row of Figure 2.30) increases the lateralization of transient sounds and at the same time limits the lateralization of stationary sounds. A measure for lateralization is found by calculating the maximum lateralization range $\phi_{\mathrm{rng}} = |\phi_{\mathrm{max}}| - |\phi_{\mathrm{min}}|$ for each listener. It is given as means and corresponding 95% confidence intervals in Figure 2.31 for both trajectories, both signal types, and both reflector setups.

The ANOVA of the lateralization range ϕ_{rng} (normal distribution according to Lilliefors test, $p > 0.05$) reveals the trajectory type to be a significant parameter for both reflector setups ($p \leq 0.05$). For the fade trajectory, most lateralized sounds are achieved with a single 3rd order beam projected to $\pm\phi_{\bigcirc} = 90°$. The circular trajectory on the other hand yields slightly higher lateralizations as the beam meets directions of specular reflections of both lateral walls. A more prominent parameter is the signal type. By adding reflectors, however, the difference between transient and stationary sound is not as pronounced as without reflectors. Although it is still significant, the increased laterally reflected energy for transient sounds weakens the precedence effect and at the same time stationary sounds are laterally limited by the reflector positions. Thus, the difference of corresponding lateralizations of both signal types is reduced.

Adding ambient noise. The second approach aiming at increasing the lateralization of transient sounds is inspired by a study of Chiang and Freyman [CF98], who discovered substantial weakening of precedence using background noise in lead-lag experiments. Their effect was found to be especially prominent, when the level of the lag was increased. A listening experiment is performed to determine the influence of maskers on the perception of transient auditory events, i.e., the target, created by the IKO_1 [LWZF18]. A similar setup as tested with the previous approach is used with two reflectors pairs arranged on each side of the IKO, cf. lower panels of Figure 2.30, but with the listener located directly in front of the IKO. The target is the same

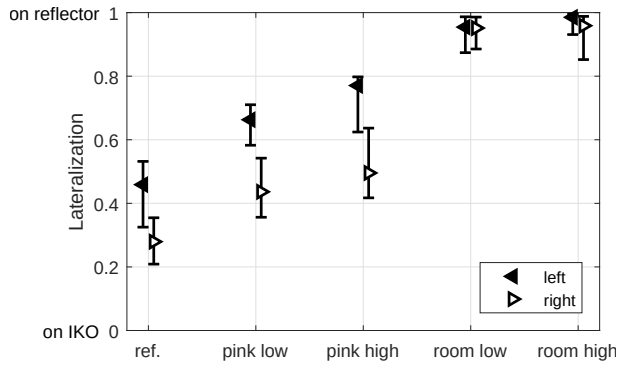


Figure 2.32: Medians and 95% confidence intervals of lateralizations for different masker types (pink/room) and target-to-masker levels (low/high).

transient sound as in the previous experiment. For each condition this signal is spatialized using the IKO with a stationary directivity aiming at the middle of either the left or right reflector pair with $\phi_{\bigcirc} = \pm 60°$. Two masker types are added, each tested at two different target-to-masker levels (low/high). For pink masker conditions, the transient target is played back at comfortable loudness and include an omnidirectional broadband pink masking noise at two levels below the target. For ambient masker conditions, the noise floor of the listening room present in quite is used for masking the direct sound of the target. Hence, the target level is tested once just above hearing threshold (low target-to-masker level) and once slightly louder (high target-to-masker level). Additionally, a reference condition without masker tests the lateralization of the spatialized target at comfortable loudness.

Eleven experienced listeners participated in the experiment. Lateralization of the transient auditory event had to be adjusted on a graphical user interface displaying a continuous slider for each condition in a multi-stimulus set including all conditions ([2 room masker + 2 pink masker + 1 reference] × 2 directions). The left-most position of the sliders represent the lateralization on the leftmost reflector and the rightmost position on the rightmost reflector.

Results violate the normality test assumptions (Lilliefors, $p \leq 0.05$) and Figure 2.32 shows medians and corresponding 95% confidence intervals of the conditions mirrored accordingly for left/right conditions. For both beam directions, the lateralization of the transient auditory event significantly increases if a masker is added ($p \leq 0.05$, nonparametric Kruskal-Wallis test with Bonferroni-Holm post hoc analysis). While the parameter masker type is significant, the target-to-masker level is not a significant, and there is no difference between high/low conditions of the same masker type. The comparison of the left and right results reveals significant differences of the reference and pink masking conditions, whereas there is no significance found for left and right ambient masker conditions. The slight asymmetry in the positioning of the reflectors is suggested as possible reason as listeners reported to hear reflections from the left to be louder than conditions from the right.

2.5.6 Discussion

Section 2.5 studied the perception of static and dynamic sound projections. In Part 1 of Experiment 4 it has been shown that strongly focused lateral sound beams can trigger lateralized auditory events in space. The results revealed where knowledge from literature is applicable: timing of direct sound and prominent lateral reflection yields the precedence effect which is seen in the onset-dependency of lateralization. Transient signals tend to be localized closer to the IKO than smooth signals. Beams aligned to the median plane influence mainly auditory distance and add spatial depth to the scene.

Successful modeling of lateralizations was achieved by the extended energy vector with reflection parameters derived from a simple image source model. Based on the constellation of source and receiver within the room, it was able to coarsely predict perceived lateralization. Refinement of the model included the consideration of the IKOs measured directivity. In this way the echo threshold slopes β_{ET} established in Section 2.2 and Section 2.3 agree with the attenuation parameter incorporated in the extended energy vector to consider propagation paths preceded by others contributing to the perceived direction.

Part 2 of Experiment 4 showed that the perception of moving sound beams, i.e., beam trajectories, cannot be fully explained by extrapolating the perception of static sound beams as it additionally involves perceptual properties indicated in studies on auditory motion. Like the perception of a real moving sound sources, where the detection of the stimuli displacement yields the perception of auditory motion, the discretely localizable directions of the beamforming trajectory are heard as gradually moving auditory event.

In Part 3 of Experiment 4 two different approaches for increasing lateralization of transient sounds were studied, one with stationary and the other with dynamic sound beams. The addition of omnidirectional pink masking noise weakens the precedence effect and increases the lateralization of the static transient sounds. Alternatively, target transient sounds can also be softened until they are just above the level of the ambient room noise acting as masker. Lateralization of respective conditions was highest with the drawback that the transient sounds were sometimes barely audible. Another approach is the use of reflectors at the sides of the IKO. In this way, the lateralization of dynamic precedence-effect affected transient sounds is increased, which is explained by the additional lateral energy from the reflectors. At the same time, the position of reflectors limits the lateralization of smooth-onset sounds and, thus, a more signal-independent perception is achieved.

2.6 Summary

This chapter examined the precedence effect and how it affects the lateralization of sounds. Experiment 1 focused on setup-specific reflection parameters and their influence on precedence. The listening experiment studied the relative importance of the reflection's delay, attenuation, and angle compared to the direct sound by measuring the echo threshold. In the first part of the experiment the directional separation between direct sound and reflection was found to be the least significant parameter. Nevertheless, although highly individual echo threshold levels needed a correction to account for the individual echo criteria, the experiment showed significant differences between the echo thresholds of different reflection angles. In numbers, the echo threshold level decreased by almost 3 dB over all examined delays, if instead of a spatially coincident presentation of direct sound and reflection from the front, the refection angle is changed to $\phi = 60°$. A possible explanation of this effect is found by the direction-dependent sensitivity of the human ear. Successful modeling is achieved by comparing normalized echo threshold levels of different reflection angles with ear directivity measures from literature.

The second part of Experiment 1 examined the more prominent reflection parameters delay and level. Obtained echo threshold levels ΔL and echo threshold delays ΔT were not only found to be highly individual, but also time-variant; individually adjusted delays resulting in the perception of an echo became shorter throughout the experiment. This is assumed to be due to an effect of learning. Moreover, in accordance with individual differences in absolute values ΔL, the comparison of echo threshold levels to results of the literature showed tremendous differences. However, the shape of the echo threshold curves was found to be similar which is why the interrelation of level and delay was modeled by a linear approximation in the interesting time range yielding a signal-dependent slope $\beta_{\mathrm{ET}} = \Delta L/\Delta T$. In the context of spatialization on beamforming systems, obtained slopes β_{ET} describing the trade-off between delay and level, revealed that for shifting the auditory image away from the physical sound source, strong but temporally closer reflections are preferred over weaker ones that arrive later in time.

Experiment 2 studied how reflection-specific parameters influence the precedence effect. The section introduced a model for diffuse reflections on a plane wall panel based on Lambert's cosine law. Resulting reflection responses yield a directional and temporal smearing of the reflected sound. The perception of different reflections was studied by measuring echo threshold and masked threshold. The experimental results revealed that diffusion makes reflections more easily perceivable as masked threshold levels are below the corresponding levels of specular reflections. Accordingly, diffuse reflections similarly weaken the precedence effect and less level is needed to perceive a diffuse echo compared to specular echoes. The results suggest that this is mainly due to the temporal diffusion, whereas similar to the influence of the directional separation tested in Experiment 1, the effect of spatial diffusion is a result of the direction-dependent sensitivity of the ear. Consequently, the modeling of echo threshold concentrated on the temporal properties, and temporally aligning the threshold levels to energy centroids of corresponding reflection responses

yielded highly correlated curves. Successful modeling of the masked threshold on the other hand was achieved by considering the temporal masking patterns for forward masking of noise bursts from literature. Lastly, Experiment 2 examined the influence of the surface structure on the reflection response. Considering a surface structure of a Schroeder diffusor in the model, resulted in reflection's responses that resemble diffusor measurements exhibiting a gradual onset of the reflected energy.

Experiment 3 studied different types of the precedence effect known from two-source experiments using multiple reflections with increased reflection levels. The onset capture effect, the ongoing precedence effect, and the onset capture effect were examined by presenting noise bursts with different envelopes. An all-pass array consisting of five active horizontally arranged loudspeakers with weights derived from a Bessel sequence was used as setup. The experiment tested different directional separations and angles of incidence. In contrast to previous experiments, all effects were studied without buildup of precedence and listeners were asked to determine perceived lateralizations of a single fused percept and a potential echo for different inter-signal intervals. In this way echo threshold delays were determined indirectly.

Part 1 of Experiment 3 studied the transient precedence effect for complex reflection patterns with the leading brief transient burst dominating the lateralization. In accordance with literature, increasing the onset length of the bursts weakened precedence as echo threshold delays increased and lateralization decreased. An ongoing precedence effect, where the ongoing lead component of smooth onset signals determines lateralizations, was not obtained in the results. Instead, reflection patterns induced a single fused percept of at a weighted-average position. However, considering that the weak direct sound is overlayed with four partly prominent reflections from different angles, this finding was not surprising.

The third precedence type that has been identified in complex sound fields is the onset capture effect. It implies that the transient onset controls the lateralization of the ambiguous ongoing sound and Part 2 of Experiment 3 quantified it for multiple reflections. Similarly, but not as pronounced as for the transient precedence effect, the experimental results identified a critical effect of signal onset in the lateralization of sound for the onset capture effect. Moreover, agreeing knowledge from two-source experiments, the influence of the spectral content was proven for this effect and echo threshold delays significantly increased for signals without high frequency components, without affecting lateralization.

The last section of this chapter applied the IKO for spatialization and studied the perception of auditory events with the focus on lateralization. Part 1 of Experiment 4 examined the perception of static directivities. Pronounced lateral beams directed to a reflective wall move the auditory event away from the physical position of the sound source. Agreeing with knowledge gained from previous sections the displacement showed a critical effect of signal onset. While smooth signals were almost fully lateralized to the reflective wall, the direct sound of transient signals dominated localization yielding fused percepts at the vicinity of the IKO. Successful modeling

of the perception was achieved by the extended energy vector which quantifies the perceived lateralization by an energetic summation of the incoming sound instances. The model's input are gain, delay, and angle of direct sound and its reflections simulated by a simple image source model. Additionally, to account for the precedence effect, reflections are weighted by a signal-dependent weight derived from the corresponding echo threshold slopes.

Part 2 of the experiment studied the perception of beam trajectories that change the directivity of the IKO dynamically over time. Although time-variant transient sounds tended to be perceived closer to the IKO than signals without transient character, auditory events were not fully explicable by the findings from static sound beams. It has been shown that the perception of such a complex time-variant sound field involves additional higher-order perceptual effects that are known from the perception of physically moving sound sources, e.g., auditory representational momentum and auditory saltation illusion.

Finally, Part 3 of Experiment 4 presented two approaches for weakening precedence and aiming at achieving a signal-independent lateralization. The first approach consisted of adding noise in order to mask the direct sound. Results of the listening experiment showed that masking noise effectively weakens the precedence effect as the lateralization of static transient auditory events increased. A different way to reduce precedence is the insertion of acoustically reflecting baffles. The experimental results showed a weakening of localization dominance for dynamic sound beams and the prominent reflection prevailed the transient direct sound yielding increased lateralizations. Moreover, the addition of reflectors decreased the lateralization of smooth-onset sounds as they physically restricted the space for auditory event positioning.

3

Auditory Distance Control by the Sound Source Directivity

Studies on sound localization mainly focus on the directional aspect, e.g., lateralization, and auditory distance perception receives substantially less scientific attention. A review of localization studies showed that when listeners are asked to describe the location of auditory events, the most common attribute used is distance [Mas17]. Studies regarding distance perception have proven the existence of multiple cues, which are assumed to be processed and combined in order to construct a distance percept.

Outline This chapter is structured as follows. Section 3.1 introduces the most relevant cues for distance perception in rooms. It gives an overview on how different rendering systems achieve distance impressions. Based on the literature, it presents a method for controlling the auditory distance. Section 3.2 introduces three different beampattern designs of directional sound sources to control the auditory distance. Their perception is studied in Experiment 5 using simulated acoustics and the section presents detailed results with discussions of the influence of room, signal, and reverberation. It establishes models of the measured results using simple acoustic measures. Section 3.3 applies the IKO to synthesize the previously established beampatterns in a room. Experiment 6 studies the correlation of auditory distance and the apparent source width and confirms the acoustic measures from the laboratory for the modeling. Subsequently, it takes up the results of Experiment 4 of the previous chapter and completes the modeling of the perception by applying the distance models established throughout this chapter. Finally, Section 3.4 summarizes the chapter.

3.1 Relevant distance cues and how they are incorporated in rendering systems

The ability to localize sound sources with regard to distance is generally much less accurate than it is with direction. A number of acoustical and non-acoustical cues are involved in the perceptual process, which is assumed to combine and weight the cues to produce a stable distance perception. The most studied non-acoustic cue is vision. The presence of visual cues increases the auditory distance accuracy [Zah01]. However, if the visual event does not comply with the auditory event, vision can also distort the auditory perception. Similar to the *ventriloquism effect* known from lateralization studies, the auditory event is shifted to match the distance of a plausible visual event [WAKW93].

Regarding acoustic cues the intensity has been the most frequently studied cue. In the free field, the relation between distance and intensity can be approximated by an inverse-square law. Under more natural conditions, where reflections join in, this simple intensity/distance relationship is destroyed. Instead, the loudness of the sound source may remain essentially constant with changes in source distance and, hence, a source that is perceived louder than another one is not automatically closer. Nevertheless, listeners can differentiate between intensity changes that result from changes in the source distance and those that arise from changes in the acoustic power of an unfamiliar source [Zah01]. A cue that is thought to dissolve the ambiguity in loudness perception is the ratio of energy of the direct sound to the energy reaching the listener via refections. This is because later-arriving reflections produce a diffuse sound field, which is defined to have uniform energy over varying source positions. Hence, independent of the acoustic power of the sound source, the source produces a greater proportion of direct-sound energy compared to the relative amount of reverberant energy, if it is closer to the listener. This makes the direct-to-reverberant energy ratio (DRR) to be inversely related to the distance of the sound source.

Other acoustic distance cues available to the auditory system are the spectrum of the signal and binaural information. However, in room environments, intensity and DRR are thought to prevail the process subserving distance perception [Zah02a]. Studies could show that when listeners are asked to guess the distance of a single sound source, they are more accurate in reverberant environments compared to anechoic ones, were the DRR-cue is not available [MK75, BH99]. This finding and the review of other relevant literature by Zahorik et al. [ZBB05] lead to the assumption that the DRR provides coarse but absolute distance information, whereas the intensity permits fine relative distance discriminations.

For spatial reproduction systems aiming at recreating the physical sound field such as scene-based Ambisonics or WFS, distance cues are inherently encoded in the signal. Perceptually-optimized amplitude-panning system such as stereo or VBAP are restricted to the intensity cue and spatial depth is added by scaling gains of corresponding loudspeakers based on the psychophysical function of loudness versus perceived distance. On the contrary, controlling

the DRR is not straightforward as the amount of reverberant energy as a function of time is determined principally by the size of the playback room and the acoustic properties of the reflecting surfaces. Hence, direct control of the direct-to-reverberant energy ratio in numbers is difficult to achieve for undetermined rooms. Nevertheless, the results for the previous chapter hint that spatialization on beamformers allows distance rendering achieved by the variation of the sound source directivity controlling DRR cues. Accordingly, Laitinen et al. [LPHP15] showed that directivity variations of compact spherical loudspeaker array in a relatively dry and small room result in significant changes of the auditory distance.

This chapter extends the approach introduced in [LPHP15] by applying it to a highly directional sound source. The first section evaluates the effect of a variable-order directivity on the auditory distance in a simulated sound field and studies the influence of sound source directivity, room, excitation signal, and reverberation. Subsequently, the practical applicability of the simulated directivity patterns to the directivity synthesis by the IKO is studied in a real room. The auditory distance and apparent source width are examined and the influence of visual cues is discussed. The descriptions of the two listening experiments presented in this chapter are published in [WFZH16a, WFZH16b, WZFH17]. Corresponding statistical evaluations of the results are revised.

3.2 Directivity-controlled auditory distance in auralized rooms

Experiment 5 studies the distance perception of 3 different directivity designs in reverberant rooms. Each design consists of 7 frequency-independent directivity pattern constellations that are a combination of spherical harmonic directivities up to the order $i = 3$, cf. Eq. (1.1). Proposed directivity designs vary:

A the beam order i from three to zero;

B the ratio a/b of two opposing 3rd-order beams;

C the angle $\phi_\bigcirc$ of a 3rd-order beam pair.

Table 3.1 lists all examined directivity designs in particular, which differently modify the amount of diffuse, lateral, and direct energy, thus the DRR. Each directivity indicated by the index 1 and 7 corresponds to a 3rd-order beam facing towards ($\phi_\bigcirc = 180°$: $A_1 = B_1 = C_1$) and away from the listening position ($\phi_\bigcirc = 0°$: $A_7 = B_7 = C_7$). Furthermore, directivity pairs indicated by indices 1/7, 2/6, and 3/5 of each design are identical in their shape but horizontally rotated by $\Delta\phi_\bigcirc = 180°$. Figure 3.1 shows the directivities $A_{1...4}$, $B_{1...4}$, and $C_{1...4}$ normalized to constant energy.

Table 3.1: Properties of directivity designs A, B, and C.

A	$A_{1/7}$	3rd-order max-$\boldsymbol{r}_\mathrm{E}$ beam to/off listener
	$A_{2/6}$	2nd-order max-$\boldsymbol{r}_\mathrm{E}$ beam to/off listener
	$A_{3/5}$	1st-order max-$\boldsymbol{r}_\mathrm{E}$ beam to/off listener
	A_4	omnidirectional directivity
B	$B_{1...7}$	3rd-order max-$\boldsymbol{r}_\mathrm{E}$ beams to and off listener linearly blended at $[\infty, 6, 3, 0, -3, -6, -\infty]$ dB
C	$C_{1...7}$	two 3rd-order max-$\boldsymbol{r}_\mathrm{E}$ beams horizontally arranged at $\phi_\bigcirc = 180° \pm 30° \cdot [0, 1, \ldots 6]$

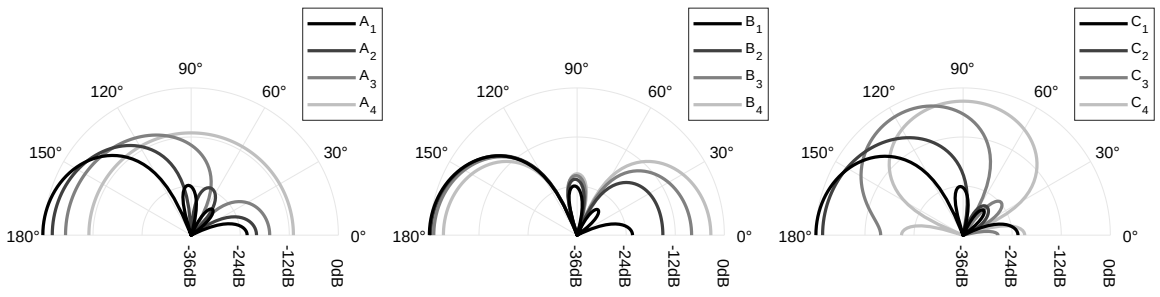


Figure 3.1: Directivity designs $A_{1...4}$, $B_{1...4}$, $C_{1...4}$ controlling the DRR. Corresponding directivities indicated by indices 5,6,7 are identical in their shape as 3,2,1 but horizontally rotated by $\Delta\phi_\bigcirc = 180°$.

3.2.1 The influence of directivity design, room, signal, and reverberation

The sound field of the variable-directivity source in a shoebox-shaped room is simulated. For the direct sound and specular reflections up to 3rd order an image-source model is used. The model considers a frequency-independent absorption coefficient derived from Sabine's formula using the rooms volume V, surface S and reverberation time T_{60}, cf. Eq. (2.16). For simplicity, the diffuse reverberation of an omni-directional excitation is considered using the software tool MCRoomSim [WEJV10]. Playback employed a ring of 24 equally-distributed Genelec 8020A loudspeakers with a radius of $r = 1.5\,\mathrm{m}$ placed in the anechoic laboratory of the IEM, cf. Figure 3.2. Listeners are sitting in the center of the arrangement with ear height adjusted to the loudspeaker ring.

On this circular setup each specular reflection is auralized by the loudspeaker with the closest azimuth angle. This avoids timbral effects of amplitude panning [TPKL13]. Elevated specular reflections are attenuated in the auralization by the cosine of their elevation θ. Diffuse reverberation is played back in Ambisonics format. The impulse response $h_l(t)$ of the l-th loudspeaker is obtained after superimposing specular and diffuse reflections. Obviously, a two-dimensional representation of a three-dimensional sound field is not optimal, but findings in [Gus90] indicate that reflections from floor and ceiling do not have a significant influence on the auditory distance. Impulse responses are convolved with signals $S_{1\ldots3}$, cf. Table 3.2, yielding a 24-channel audio file for each condition. To monitor the influence of room acoustics, three different layouts were tested, including two rooms and two source-listener distances, cf. $R_{1\ldots3}$ in Table 3.2. Geometry and reverberation time of the auralized rooms are based on the IEM CUBE, a $10.3\,\mathrm{m} \times 12\,\mathrm{m} \times 4.8\,\mathrm{m}$ large room with $T_{60} = 0.70\,\mathrm{s}$, and the IEM lecture room known from Experiment 4, cf. Figure 2.23.

The directional source is simulated near the corners of the room at a distance of 2 m and 3 m (IEM CUBE) and 1 m and 2 m (IEM lecture room). The listening position is chosen at a virtual distance of $d = 1.7\,\mathrm{m}$ to the sound source, which already lies outside of the loudspeaker ring. This value is in good agreement to the estimated specific distance tendency of the power function approaching the underlying distance psychophysical function [ZBB05]. Additionally, for the IEM CUBE an increased source-listener distance of $d = 2.9\,\mathrm{m}$ is tested. The listener is

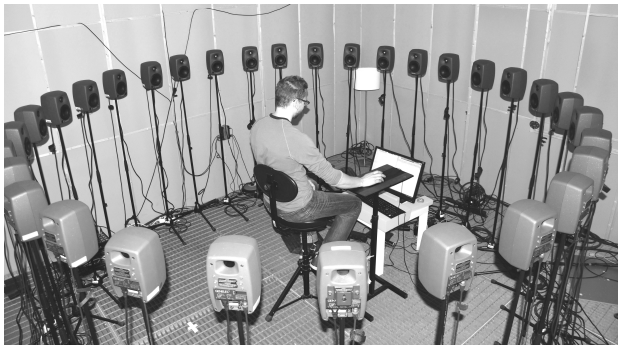


Figure 3.2: Experimental setup consisting of a horizontal loudspeaker ring in the anechoic laboratory. The sound filed of the frequency-independent directivity is simulated by an image source model.

Table 3.2: Properties of rooms R and signals S examine din Experiment 5.

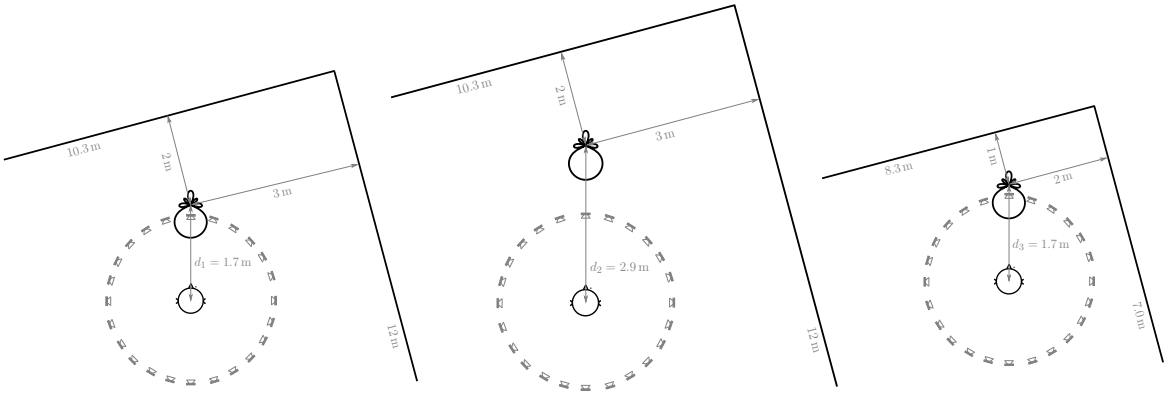
room	R_1	IEM CUBE,	$10.3\,\mathrm{m} \times 12.0\,\mathrm{m} \times 4.8\,\mathrm{m}$,	$T_{60} = 0.70\,\mathrm{s}$,	$d_1 = 1.7\,\mathrm{m}$;
	R_2	IEM CUBE,	$10.3\,\mathrm{m} \times 12.0\,\mathrm{m} \times 4.8\,\mathrm{m}$,	$T_{60} = 0.70\,\mathrm{s}$,	$d_2 = 2.9\,\mathrm{m}$;
	R_3	IEM lecture room,	$7.0\,\mathrm{m} \times 8.3\,\mathrm{m} \times 3.0\,\mathrm{m}$,	$T_{60} = 0.57\,\mathrm{s}$,	$d_3 = 1.7\,\mathrm{m}$;
signal	S_1	female speech[3]			
	S_2	sequence of irregular artificial bursts			
	S_3	speech-spectrum noise with increased kurtosis			

facing the sound source simulated at height of 1.8 m above the floor with an angular offset of $\Delta\phi = 15°$ with regard to the sidewalls. Figure 3.3 shows the setup of the auralized room using the 24-channel loudspeaker ring, and Table 3.2 lists rooms and source-listener distances tested in the experiment.

Studies suggest that the cue weighting process for creation of a distance percept is signal-dependent and the DRR is weighted relatively less than intensity for speech signals than for noise signals [Zah02a]. Hence, signals fed into auralization are chosen to investigate the influence of speech versus noise, noise spectrum, and noise envelope to the effect: anechoic female speech S_1, a sequence of irregular artificial bursts S_2 known from the previous Experiment 4, and Gaussian white noise shaped to speech spectrum S_3 as listed in Table 3.2. For S_3, envelope fluctuations are slightly accentuated by multiplying the noise with its Hilbert envelope and by restriction to its original bandwidth, cf. [KKvdH$^+$97]. By this procedure, S_1 and S_3 have similar spectra and kurtosis, whereas S_2 is more transient with more energy at frequency above 1 kHz. All signals are normalized to their RMS value for level equalization. The above signals are anechoic. To monitor potential influence of additional reverb, for some conditions signal S_1 is reverberated before auralization. Two levels of reverberation are tested, of which level 1 corresponds to a room impulse response with a reverberation time of $T_{60} = 0.5\,\mathrm{s}$, level 2 to one of $T_{60} = 1.0\,\mathrm{s}$, and level 0 to the anechoic signal.

The listening experiment is carried out as a multi-stimulus test where listeners have to comparatively rate multiple conditions, denoted as sets. Their task is to indicate the distance of auditory events on a graphical user interface displaying a continuous slider for each condition of a set along the ordinal scale *very close* (vc), *close* (c), *moderate* (m), *distant* (d), and *very distant* (vd). The listeners are allowed to repeat each stimulus at will, and audio files are played back in loop. Fifteen listeners participated in the experiment, all of them are experienced listeners in 3D audio and experienced participants in psychophysical studies of hearing.

Conditions. Examined sets (set 1 to 12, see Table 3.3) comprise 7 conditions, each representing a directivity, room, signal, and reverberation level. Under a varied directivity design, e.g., $A_{1\ldots7}$, the influence of room (set 1, 10, 11), signal (set 1, 2, 3), and reverberation level (set 1, 12) was only examined separately, yielding responses $x^{\mathrm{I}}_{1\ldots7}$ for each listener. These separate multi-stimulus sets do not yet permit cross comparison due to the absence of a common reference. As a solution, maintaining a limited testing time, the additional 9-stimulus comparison sets (13. . . 15) were



(a) R_1 (b) R_2 (c) R_3

Figure 3.3: Room and source configuration for R_1, R_2, and R_3 together with loudspeaker ring used for auralization. R_1 and R_2 are based on the IEM CUBE differing in the source-listener distance and room R_3 is based on the IEM lecture room. The setup is slightly offset compared to the lateral walls with the listener directly facing the IKO ($\phi = 0°$). The indicated directivity corresponds to $A_1/B_1/C_1$ with $\phi_\circ = 180°$.

tested with fewer directivities $A_{1,4,7}$ and instead involving cross-comparisons with regard to signal (13), room (14), and reverberation level (15). They yield cross-comparison responses $x^{\text{II}}_{1,4,7}$ that enable a comparison involving a fine-grained directivity variation in Figures 3.6–3.8.

In these Figures, responses $x^{\text{I}}_{2,3}$ and $x^{\text{I}}_{4,5}$ were re-mapped for each listener by linear scaling and shifting to match $x^{\text{I}}_{1,4}$ with $x^{\text{II}}_{1,4}$, and $x^{\text{I}}_{4,7}$ with $x^{\text{II}}_{4,7}$, respectively:

$$x_i = \begin{cases} x^{\text{II}}_i & \text{for } i \in \{1,4,7\}, \\ \dfrac{x^{\text{II}}_4 - x^{\text{II}}_1}{x^{\text{I}}_4 - x^{\text{I}}_1}(x^{\text{I}}_i - x^{\text{I}}_1) + x^{\text{II}}_1 & \text{for } i \in \{2,3\}, \\ \dfrac{x^{\text{II}}_7 - x^{\text{II}}_4}{x^{\text{I}}_7 - x^{\text{I}}_4}(x^{\text{I}}_i - x^{\text{I}}_4) + x^{\text{II}}_4 & \text{for } i \in \{5,6\}, \end{cases} \tag{3.1}$$

i.e., a complete response set $x_{1\ldots7}$ per listener, signal, room, and reverberation level.

During the listening session, the listeners were requested to face $\phi = 0°$ which corresponds to the direction of the auralized sound source, cf. Figure 3.3. At the beginning of the experiment, each listener was given a short training to familiarize with the evaluation scale. The training set included expected extreme values with regard to the auditory distance. Listeners were asked to rate along the whole scale and use extremes as an internal reference for further evaluations and ratings are assumed to be quantitative. After the training phase, multi-stimulus tasks were presented. Each time a multi-stimulus set was displayed, the arrangement of its stimuli was an individual random permutation. The listener could have the stimuli sorted by own ratings to facilitate comparative rating. The first part of the experiment consisted of the sets with 7 stimuli (set 1 to 12) in an individual random permutation, and the second part of the sets consisting of 9 conditions (set 13 to 15) in an individual random permutation. None of the listeners reported that they perceived the auralization as implausible or confusing; some emphasized the naturalness of the auralization.

Table 3.3: Composition of examined sets in Experiment 5, each consisting of 7 (set 1...12) or 9 samples (set 13...15).

set no.	design	index	signal	room	reverb. level
1	A	$1\ldots7$	S_1	R_1	0
2	A	$1\ldots7$	S_2	R_1	0
3	A	$1\ldots7$	S_3	R_1	0
4	B	$1\ldots7$	S_1	R_1	0
5	B	$1\ldots7$	S_2	R_1	0
6	B	$1\ldots7$	S_3	R_1	0
7	C	$1\ldots7$	S_1	R_1	0
8	C	$1\ldots7$	S_2	R_1	0
9	C	$1\ldots7$	S_3	R_1	0
10	A	$1\ldots7$	S_1	R_2	0
11	A	$1\ldots7$	S_1	R_3	0
12	A	$1\ldots7$	S_1	R_1	1
13	A	$1,4,7$	$S_{1\ldots3}$	R_1	0
14	A	$1,4,7$	S_1	$R_{1\ldots3}$	0
15	A	$1,4,7$	S_1	R_1	$0,1,2$

3.2.2 Experimental results

Influence of directivity design. Figure 3.6 shows the means and corresponding 95% confidence intervals of auditory distance for directivity designs A, B, C evaluated in room R_1 with signals $S_{1,2,3}$ evaluated with sets $1\ldots9$, cf. Table 3.3. For the analysis, conditions are pooled across signals $S_{1,2,3}$ maintaining normality (Lilliefors, $p > 0.05$). Resulting significance levels of the ANOVA corrected using Tukey's HSD and corresponding effect sizes are given in Table 3.4 for each directivity design. Design B yields a monotonic mapping to auditory distance for all signals with significant conditions $B_{1,4,5,6}$. For design A, the statistical analysis exhibits directivities $A_{1,3,4,5}$ to be significant. By contrast, directivities $A_{5\ldots7}$ do not yield significant changes, despite continuously reducing the DRR, cf. Table 3.5. This seems to comply with the *auditory horizon effect*, a general tendency to underestimate the physical distance of far away sources [BH99]. Directivity variations within these sub-sets of designs A and B have at least a very large effect on the distance perception (Cohen's $d \geq 1.2$). A comparison of the curve obtained for $A_{2,4,6}$ to the results of Laitinen et. al [LPHP15], who employed an omnidirectional pattern and two second-order cardioid patterns steering to and away from the listener, reveals a similar linear mapping to auditory distance

By contrast, the curve obtained for $C_{1\ldots7}$ is not monotonic in the proposed sequence. Comparing the strength and angle of direct sound and specular reflections arriving at the listener for directivities C_4 and C_7, cf. Figure 3.5, reveals more energy coming from lateral directions for C_4. The more diffuse sound field explains the significantly greater auditory distance for $C_{4,5}$ compared to C_7. However, a subset of 3 directivities $C_{1,3,4}$ can be derived to significantly influence the auditory distance with a reduced effect size (Cohen's $d \geq 0.6$).

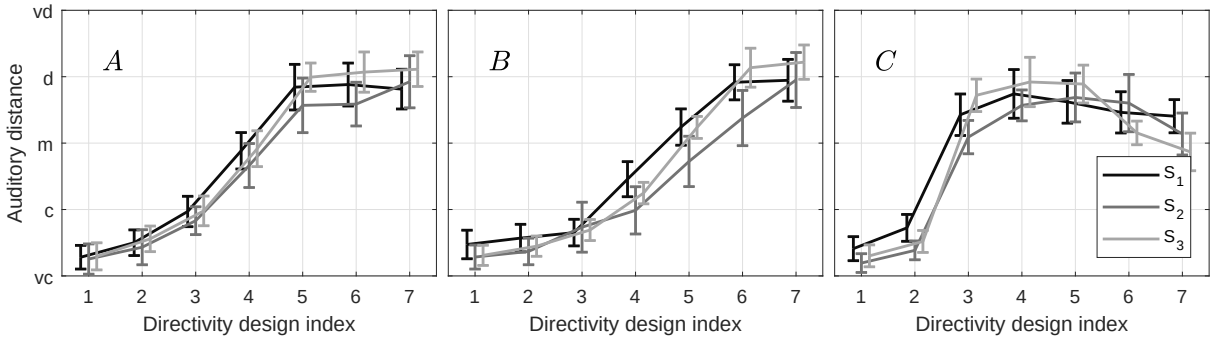


Figure 3.4: Means and corresponding 95% confidence intervals for designs A, B, and C (from left to right), plotted for signals $S_{1,2,3}$.

Table 3.4: Significance levels p (lower triangluar part) and absolute effect sizes as Cohen's d (upper triangluar part) obtained for directivity designs A, B, C for conditions pooled across signals $S_{1,2,3}$. Significances are highlighted in gray ($p \leq 0.05$). Effect sizes are coded in gray-scale according to the classification of [Saw09] with $|d| > 2$ defining a huge, $|d| = 1.2$ a very large, and $|d| = 0.8$ defining a large effect.

	A							*B*							*C*						
idx.	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	–	0.6	1.7	> 2	> 2	> 2	> 2	–	0.3	0.8	> 2	> 2	> 2	> 2	–	0.8	> 2	> 2	> 2	> 2	> 2
2	.36	–	1.1	> 2	> 2	> 2	> 2	.95	–	0.6	1.7	> 2	> 2	> 2	.25	–	> 2	> 2	> 2	> 2	> 2
3	.00	.00	–	1.9	> 2	> 2	> 2	.02	.28	–	1.1	> 2	> 2	> 2	.00	.00	–	0.6	0.6	0.0	0.5
4	.00	.00	.00	–	1.7	1.8	> 2	.00	.00	.00	–	1.5	> 2	> 2	.00	.00	.03	–	0.0	0.6	1.1
5	.00	.00	.00	.00	–	0.1	0.2	.00	.00	.00	.00	–	1.2	1.7	.00	.00	.04	.99	–	0.5	1.0
6	.00	.00	.00	.00	.99	–	0.2	.00	.00	.00	.00	.00	–	0.4	.00	.00	.99	.02	.03	–	0.5
7	.00	.00	.00	.00	.80	.96	–	.00	.00	.00	.00	.00	.29	–	.00	.00	.12	.00	.00	.13	–

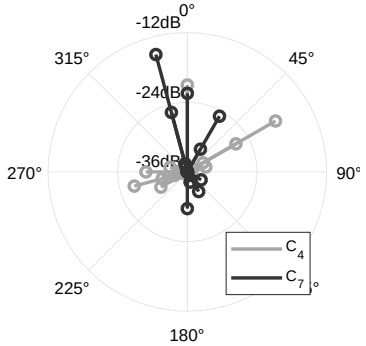


Figure 3.5: Amplitude levels of direct sound and specular reflections arriving at the listening position for C_4 and C_7, normalized with respect to C_1.

Influence of the signal. The influence of signals $S_{1...3}$ on the auditory distance of design A in room R_1 is evaluated by stimulus set 13, cf. Table 3.3. As the directivity indices $1, 4, 7$ of set 13 appear in the more detailed stimulus sets 1 to 3, a detailed analysis can be given by supplementing responses for indices $1, 4$ and 7 (set 13) with linearly re-mapped responses for $2, 3, 5, 6$ (set 1 to 3) for each listener. In this way the ranges between mean values for the indices 1, 4 and 4, 7, are filled out. Figure 3.6 shows the means and corresponding 95% confidence

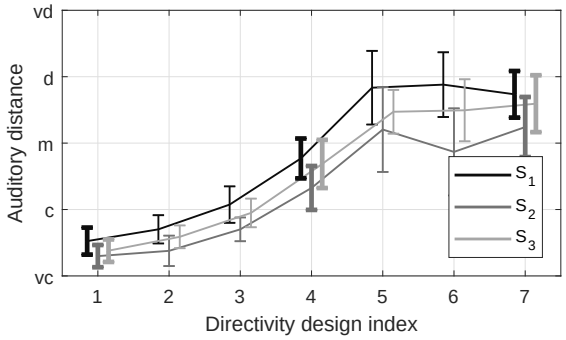


Figure 3.6: Means and 95% confidence intervals for signals $S_{1...3}$ in room R_1 with directivity design A. Individual responses for directivities indicated by indices $2, 3, 5, 6$ (thin lines) are linearly re-mapped to fill out the ranges between directivity indices $1, 4$ and $4, 7$ (bold lines) using Eq. (3.1).

intervals of the auditory distance for room R_1 and directivity design A. Along the indices, the distance impression exhibits a similar monotonic increase for all signals until A_5.

For responses of each condition, the Lilliefors test failed to reject the null-hypothesis of normality at the 5% significance level. Corrected significance values of the ANOVA reveal the signal to be significant. While for speech-spectrum noise S_3 no significance is obtained, auditory distances of noise bursts S_2 elicit with indices 3 and 6 are perceived closer than corresponding conditions of speech S_1 ($p \leq 0.05$, corrected using Tukey's HSD). This medium effect (Cohens's $d < 0.8$) agrees with findings in [Col68, LMC92], according to which the auditory distance of broadband signals decreases with the relative amount of high-frequency energy. The increased distance range of the speech S_1 compared to noise burst S_3 contradicts the findings in [Zah02a], according to which the DRR is weighted relatively less for speech signals than for pulsed white noise.

Influence of the auralized room. The influence of a room size and source-listener distance on the auditory distance for design A and signal S_1 is evaluated by the sets 1,10 and 11, cf. Table 3.3. The statistical analysis of responses with normal distribution (Lilliefors, $p > 0.05$) reveals the same significant sub-set for all rooms $R_{1...3}$ consisting of conditions $A_{1,3,4,5}$ ($p \leq 0.05$, corrected using Tukey's HSD) and at least large effects (Cohen's $d \geq 1.4$). For cross-comparisons of the rooms $R_{1...3}$ the data of set 14 is used and supplemented by the corresponding responses from above using Eq. (3.1). According to the Lilliefors test, this transformation violates the assumption of normality for one third of the conditions at a significance of 5%. Figure 3.7 shows medians and corresponding 95% confidence intervals of re-mapped distances. A smaller room with shorter T_{60} and sound source closer to adjacent walls but with the same source-listener distance (R_3) leads to a flatter curve. This has a large effect on conditions $A_{1...3}$ and corresponding distances of R_3 significantly increase (Kruskal-Wallis with Tukey's HSD, $p \leq 0.05$).

Similar flattening accompanied by an additional offset to bigger auditory distances is achieved by extending the source-listener distance (R_2). Significance changes of auditory distance are found for conditions $A_{1...4}$ with a very large effect (nonparametric Kruskal-Wallis test with Tukey's HSD $p \leq 0.05$).

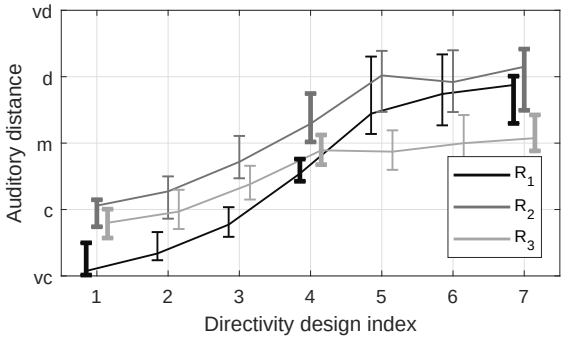


Figure 3.7: Medians and corresponding 95% confidence intervals for rooms $R_{1\ldots3}$ with design A and signal S_1. Individual responses for directivities indicated by indices $2, 3, 5, 6$ (thin lines) are linearly re-mapped to fill out the ranges between directivity indices $1, 4$ and $4, 7$ (bold lines).

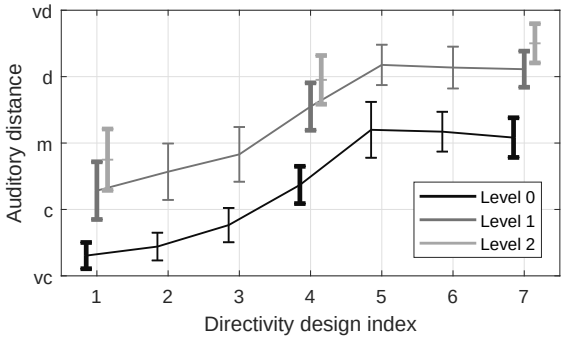


Figure 3.8: Means and corresponding 95% confidence intervals for reverberation levels 0, 1, 2 in R_1 with S_1 and directivity design A. Individual responses for directivities indicated by indices $2, 3, 5, 6$ with reverberation levels 1 and 2 (thin lines) are linearly re-mapped.

Influence of single-channel reverberation. Reverberation can provide listeners with a cue for sound source distance, e.g., [MBL+89]. In audio playback reverberation effects are often used to control the auditory distance. To get an idea how single-channel reverberation contributes to the proposed effect, the excitation signal S_1 is moderately reverberated and tested with design A in room R_1, cf. Table 3.3 set 12. A statistical analysis reveals that reverberation does not change the significant conditions, which still consist of directivities $A_{1,3,4,5}$ with the effect sizes interpreted to be very large (Cohen's $d \geq 1.5$).

Individually and linearly re-mapped responses from the sets 1 and 12 were used supplementing the responses from set 15 to provide a more detailed analysis for the reverberation levels in terms progression over the 7 design indices. Figure 3.8 shows respective means of the data following a normal distribution (Lilliefors, $p > 0.05$) together with corresponding 95% confidence intervals. Reverberation levels $1, 2$ yield a similar progression with the known saturation for $A_{>5}$. With the use of single channel-reverberation the perceived distances are significantly increased (ANOVA using Tukey's HSD $p \leq 0.05$) with corresponding effect sizes thought to be very large (Cohen's $d \geq 1.4$). However, the distance increases between reverberation level 1 and 2 are not significant (ANOVA using Tukey's HSD $p > 0.05$).

Table 3.5: Change of the DRR when increasing the directivity index, calculated for all directivity designs. Significant distance perception are indicated in gray.

idx.	$1 \to 2$	$2 \to 3$	$3 \to 4$	$4 \to 5$	$5 \to 6$	$6 \to 7$
A	−1.3 dB	−1.9 dB	−4.8 dB	−6.4 dB	−3.8 dB	−2.3 dB
B	−0.3 dB	−0.8 dB	−2.4 dB	−4.7 dB	−6.7 dB	−5.5 dB
C	−1.7 dB	−13.8 dB	−3.9 dB	−3.1 dB	−1.8 dB	3.8 dB

3.2.3 Modeling the auditory distance

This section discusses linear auditory distance models for the presented effect, based on characteristic metrics of the spatial sound field and their regression to the experimental data determined for designs A, B, and C. For the modeling, responses gathered with sets $1 \dots 9$ listed in Table 3.3 are pooled across signals. The direct comparability of all curves is feasible as all designs were determined to include reference patterns corresponding to a 3rd-order beam facing to ($A_1 = B_1 = C_1$) and off ($A_7 = B_7 = C_7$) the listening position, respectively. This allowed to linearly re-map the responses to fill out the entire interval $[0; 1]$ for each listener.

Direct-to-reverberant energy ratio. The most obvious predictor in this context is the DRR, defined as

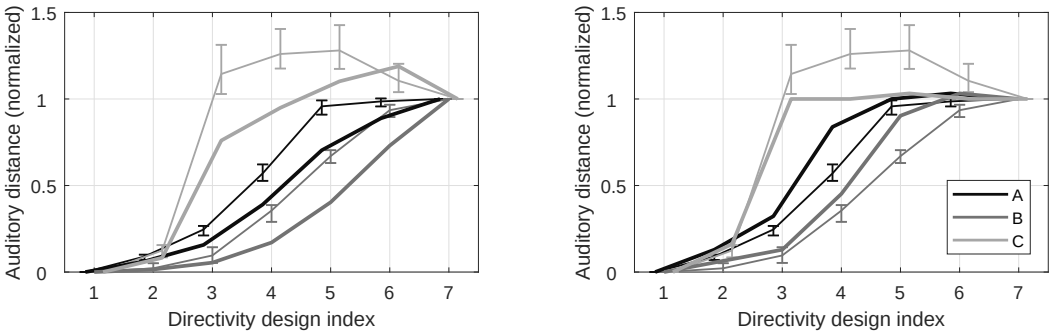
$$\mathrm{DRR} = 10 \log_{10} \frac{\int_{0\mathrm{ms}}^{T} s^2(t)\mathrm{d}t}{\int_{T}^{\infty} s^2(t)\mathrm{d}t}. \tag{3.2}$$

By using $s(t) = \sum_l h_l(t)$, the DRR can be calculated based on the loudspeaker impulse responses h_l, with a time constant T regarding only direct sound. Table 3.5 gives an overview on how the DRR changes if the directivity index is varied.

Zahorik [Zah02a] approximates the relation of DRR and auditory distance by the function $f(\mathrm{DRR}) = 2^{k\,\mathrm{DRR}+d}$. For the data of Experiment 5 the model $f(\mathrm{DRR}) = k\,\mathrm{DRR} + d$ with $k = -0.049$ and $d = 0.11$ yields a better approximation with the correlation to median values $R^2 = 0.93$. Figure 3.9 (a) shows the pooled data compared to model. Despite the high correlation, the models progression along the directivity indices tends to underestimate the distance.

Regarding sensitivity of changes in the DRR, Zahorik [Zah02b] found that increments need to succeed more than 6 dB to be just-noticeable. He concludes, that the DRR cue itself is a poor relative distance cue as this values corresponds to more than a doubling of sound distance for his acoustic environment. However, in Experiment 5 significant distance perceptions are achieved already at DRR changes around 2 dB, cf. Table 3.5. This finding indicates that a directional sound source might address additional cues that complete the construction of a distance percept.

Binaural spectral magnitude difference standard deviation. In [GMVM13] a feature is introduced related to the standard deviation of the magnitude spectrum of the room transfer function. Similar to the DRR, this feature, noted as BSMD STD, represents a distance-dependent behavior and is implemented to model the source-listener distance within the freely available



(a) DRR (b) BSMD STD

Figure 3.9: Comparison of medians and 95% confidence intervals for all conditions (thin lines) with distance predictors (bold lines).

Auditory Modeling Toolbox[4]. For calculating the BSMD STD, any binaural signal is sufficient.

Binaural input signals are generated by firstly convolving the signal of each propagation path arriving at the listener with respective HRTF measurements of dummy head and then summing up the signals for each ear respectively. The linear regression with $f(\text{BSMD STD}) = k\,\text{BSMD STD} + d$ yields the same correlation as the DRR ($R^2 = 0.93$ with $k = 0.32$ and $d = -1.52$), although their progression along the directivity index is qualitatively different, cf. Figure 3.9 (b).

Inter-aural cross correlation coefficient. As reverberation caused by the room simulation introduces binaural cues by altering the sound attributes at the two ears differentially, the inter-aural cross correlation coefficient (IACC) is used as an additional measure for auditory distance. The IACC is based on the inter-aural cross correlation function (IACF):

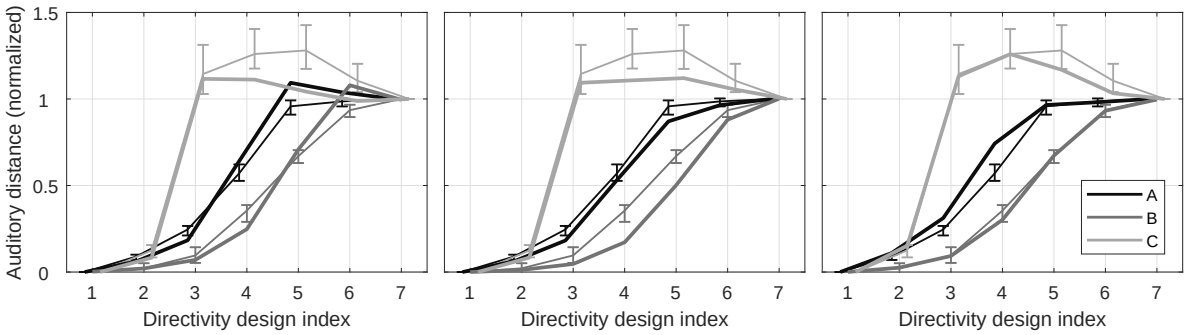
$$\text{IACF}(\tau) = \frac{\int_{t_1}^{t_2} s_{\text{left}}(t) s_{\text{right}}(t+\tau)\mathrm{d}t}{\sqrt{\int_{t_1}^{t_2} s_{\text{left}}^2(t)\mathrm{d}t \int_{t_1}^{t_2} s_{\text{right}}^2(t)\mathrm{d}t}}, \tag{3.3}$$

with $s_{\text{left}}(t) = h_{\text{left}}(t) * s(t)$ and $s_{\text{right}}(t) = h_{\text{right}}(t) * s(t)$. The binaural room impulse response $h(t)$ corresponds to responses for left and right ear at $\phi = 0°$. The IACC is defined as the maximum absolute value within $\tau = \pm 1\,\text{ms}$:

$$\text{IACC} = \max_{\forall\tau\in[-1\text{ms};1\text{ms}]} |\text{IAFC}(\tau)|. \tag{3.4}$$

The early IACC, considering a time window of $t_1 = 0\,\text{ms}$ to $t_2 = 80\,\text{ms}$, is commonly used in room acoustics as an objective measure for apparent source width (ASW), e.g., [HBO95]. It is widely accepted that a lower IACC value leads to a bigger ASW, and therefore $[1 - \text{IACC}]$ is positively correlated with the magnitude of perceived width. With the IACC binaurally measured in the experimental setup, linear regression yields $f(1 - \text{IACC}) = 1.52\,(1 - \text{IACC}) - 0.20$ to model the experimental data ($R^2 = 0.97$), cf. Figure 3.10 (a).

[4] `http://amtoolbox.sourceforge.net/`



(a) $[1 - \mathrm{IACC}]$ (b) $\boldsymbol{r}_\mathrm{E}$ (c) LF

Figure 3.10: Comparison of medians and 95% confidence intervals of distance ratings (thin lines) with models known from ASW prediction (bold lines).

Energy vector. Frank [FMS11] found a strong correlation of the IACC and the magnitude of the energy vector and uses it as alternative for predicting the ASW in surrounding loudspeaker systems. In contrast to lateralization, distance perception does not include asynchrony such as precedence. Hence, for distance modeling the simple energy vector introduced in the previous chapter, cf. Eq. (2.13), is sufficient. Like the IACC, it is negatively correlated with the auditory distance and therefore $1 - \|\boldsymbol{r}_\mathrm{E}\|$ is used for the modeling. Linear regression yields $f(\mathrm{LF}) = 0.32\,(1 - \|\boldsymbol{r}_\mathrm{E}\|) - 1.52$ with $R^2 = 0.98$, cf. Figure 3.10 (b).

Lateral energy fraction. The lateral energy fraction (LF) is another acoustic measure quantifying the spatial impression. Considering a time window up to 80 ms, it has been accepted as a measure of the effect of source broadening [Mar67, BM81] and describes the ratio of the sum of the early lateral energy to the sum of the early total energy:

$$\mathrm{LF} = \frac{\int_{5\mathrm{ms}}^{80\mathrm{ms}} s_\mathrm{lat}^2(t)\mathrm{d}t}{\int_{0\mathrm{ms}}^{80\mathrm{ms}} s^2(t)\mathrm{d}t}, \tag{3.5}$$

with $s_\mathrm{lat}(t) = \sum_l h_l(t)\sin(\phi_l)$ and ϕ_l as azimuthal angle of the l-th loudspeaker. Linear regression yields $f(\mathrm{LF}) = 7.3\,\mathrm{LF} - 0.54$, cf. Figure 3.10 (c). This LF-based linear model delivers the best matching results underlined by a sublime correlation to medians $R^2 = 0.99$.

3.2.4 Discussion

The mapping of the directivity designs $A_{1\ldots7}$ and $B_{1\ldots7}$ to auditory distance curves is sigmoid-shaped. It resembles the compressive power functions described in [Zah02a] characterizing the relation between physical and perceived distance. Moreover, agreeing with [Col68, LMC92], signals with an increased relative amount of high-frequency energy appeared to be closer in the study. Both decreasing the auralized room size and increasing the source-to- receiver distance yield a more compressed curve, which is slight offset in case of the increased source-to-receiver distance. Despite this, the range of discriminability is persistent.

The use of single-channel reverberation is also effective at increasing the auditory distance, however, it narrows the directivity-controllable range of distinct distance impressions. Studied perceptual models highly correlated with the experimental data. Interestingly, spatial measures used to quantify the apparent source width provide very accurate predictions. In a room, the physical distance to a source typically increases the amount of reflected sound in relation to the direct sound. Consequently, this affects the measures 1−IACC and LF for the apparent source width, as the measurements in [Lee13] showed. Experiment 5 only asked for distance ratings and further research is required to determine to what extent the auditory distance and width are separable.

3.3 Auditory distance control using the IKO

The experiment presented in this section studies the directivity designs introduced in the previous experiment in a real environment with the IKO. Considering the good performance of spatial parameters that were actually developed to predict the apparent source width (ASW), Experiment 6 evaluates the ASW in addition to the auditory distance to determine the interrelation of the two attributes. The ASW is defined as "the perceived width of a sound image fused temporally and spatially with the direct sound image" [MM88]. Although there has been much research in comparing the ASW between venues at the same physical source distance, only little is known on how spatial impressions vary in dependence of physical and auditory distance within the same venue.

3.3.1 Auditory distance and apparent source width

The experiment is conducted in the IEM CUBE with the same positions of the IKO and the listener as in condition R_2 at a distance of 2.9 m, cf. Table 3.2. The directivity of the IKO_2 is controlled using the ambiX plug-in suite [Kro14] with Reaper as DAW. The size controller in the ambiX encoder allows to increase the beam width from third to zeroth order ($\text{SIZE} = 0 \ldots 1$). However, this variation does not exactly meet the first and second max-$\boldsymbol{r}_\text{E}$ beam, cf. Table 3.1. Therefore, size values for directivity with index 2,3 (and 4,5) are determined by informal listening yielding design A^* as modified version of design A. Directivities $A^*_{1\ldots4}$ are facing to the listener ($\phi_\bigcirc = 180°$) with values of $\text{SIZE} = (0, 0.27, 0.47, 1)$ and conditions $A^*_{5\ldots7}$ are rotated by $\Delta\phi_\bigcirc = 180°$ with $\text{SIZE} = (0.47, 0.27, 0)$. Figure 3.11 compares both beampattern patterns A and A^*, normalized to constant energy. While main lobes are very similar, corresponding rear lobes of $A^*_{(2,3)}$ differ from $A_{(2,3)}$. The other designs B and C evaluated in the experiment are known from the previous section, cf. Table 3.1.

Listeners have to rate the distance of the auditory event on a graphical user interface. Contrastingly to the simulated-room experiment, they have to determine the absolute position on a screen showing the sketch of the setup. Seven randomly sorted markers are shown simultaneously, each representing a beampattern of the design under test, and can either be moved directly

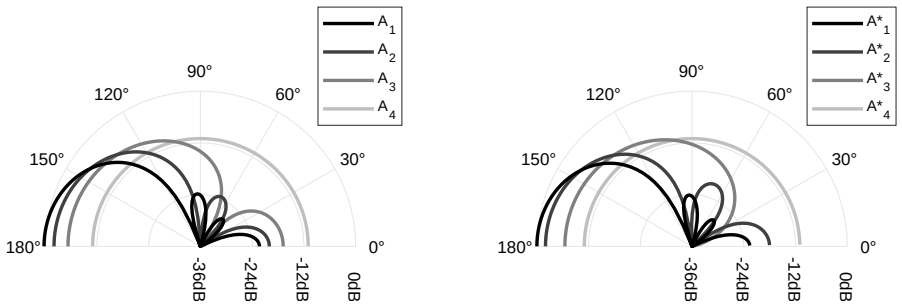


Figure 3.11: Comparison of the beampattern design A^* (right) tested in Experiment 6 with the original design A (left) varying the beam order and tested in Experiment 5.

Figure 3.12: Auditory distance control using the IKO_2. Left: Experimental setup in the IEM CUBE. Right: Sketch of the setup as represented in the GUI. Square and diamond represent listener and IKO, respectively.

(drag and drop) or, for fine adjustments, steered with a slider. Condition can be repeated at will until listeners are satisfied with the match between marker placement and what they heard. To facilitate the task a fine grid indicating distances of 0.5 m is displayed on the screen. In the room microphone stands mark distances of $(1, 2, 4)$ m, cf. Figure 3.12.

Listeners are asked to provide an honest report of what they actually perceive. This instruction has to do with the fact that there is no time limit to provide answers. It aims specifically at asking listeners to avoid developing theories about which condition they are presented, as some listeners are aware of results from the previous experiment. Moreover, this is for preventing any visual bias due to the limited physical distance between the listeners and the wall behind the IKO.

Additional to distance, the experiment examines the apparent source width (ASW) of auditory events in a separate task. The procedure is the same as the distance rating of the previous experiment, so that rating is done on a graphical user interface displaying a continuous slider for each condition of a set to permit comparative rating. Listeners are asked to rate using the whole scale *very narrow* (vn), *narrow* (n), *moderate* (m), *broad* (b), and *very broad* (vb). The signal feed into auralization is anechoic female speech[5] S_1. All conditions are normalized in loudness and are played back in loop at comfortable level.

During the listening session, listeners are sitting on a chair with ear height adjusted to the IKO. While listening to conditions, they are requested to face the IKO. Both tasks are performed consecutively with a short break in between. Half of the listeners start with the distance rating task and the other half with the rating of the ASW. Ten listeners participated in this experiment, nine of them performed already Experiment 5.

3.3.2 Experimental results

Auditory distance. Figure 3.13 shows results (normal distribution according to Lilliefors test, $p > 0.05$) for the distance rating task of Experiment 6. Significance levels of the ANOVA

[5] *Music for Archimedes*, CD Bang and Olufsen 101 (1992)

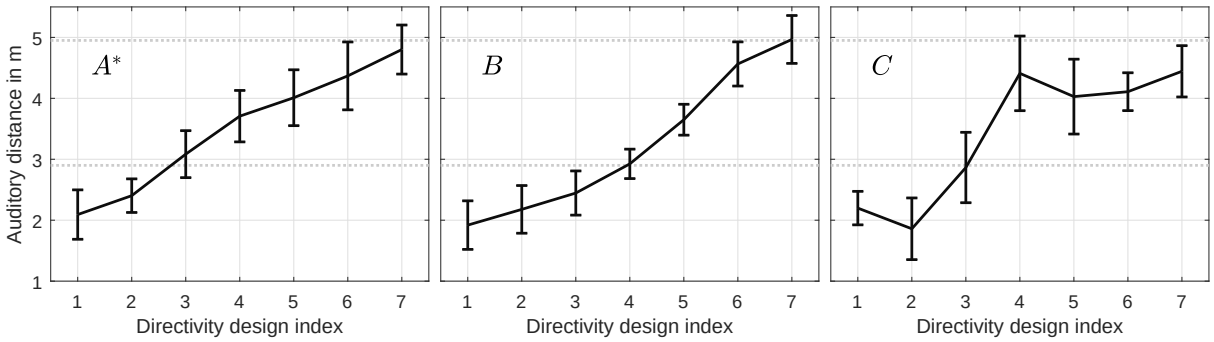


Figure 3.13: Means and corresponding 95% confidence intervals of the distance task with designs A^*, B, and C in the IEM CUBE using the IKO_2. Physical distances of IKO and front wall are indicated by dotted lines.

corrected using Tukey's HSD and corresponding effect sizes are given in Table 3.6.

Table 3.6: Significance levels p (lower triangluar part) and Cohen's d (upper triangluar part) determined with the IKO in the IEM CUBE for directivity designs A^*, B, C with speech as excitation signal. Effect sizes are coded in gray-scale according to the classification of [Saw09] with $|d| > 2$ defining a huge, $|d| = 1.2$ a very large, and $|d| = 0.8$ defining a large effect.

	A^*							B							C						
idx.	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	–	0.6	1.8	>2	>2	>2	>2	–	0.5	1.0	>2	>2	>2	>2	–	0.6	1.1	>2	>2	>2	>2
2	.9	–	1.5	>2	>2	>2	>2	.9	–	0.5	1.6	>2	>2	>2	.92	–	1.3	>2	>2	>2	>2
3	.01	.15	–	1.1	1.6	1.9	>2	.21	.88	–	1.1	>2	>2	>2	.33	.03	–	1.9	1.4	1.9	>2
4	.00	.00	.23	–	0.5	1.0	1.9	.00	.02	.31	–	>2	>2	>2	.00	.00	.00	–	0.4	0.4	0.0
5	.00	.00	.01	.91	–	0.5	1.3	.00	.00	.00	.02	–	>2	>2	.00	.00	.01	.87	–	0.1	0.6
6	.00	.00	.00	.18	.82	–	0.6	.00	.00	.00	.00	.00	–	0.8	.00	.00	.00	.96	.99	–	0.6
7	.00	.00	.00	.00	.06	.66	–	.00	.00	.00	.00	.00	.52	–	.00	.00	.00	.99	.82	.93	–

The comparison of the results with significances of the previous experiment should be interpreted with caution, as the source-listener distance is increased by 1.2 m (R_1 vs. R_2). However, indications are found that the IKO yields similar pronounced distance impressions as the ideal beamformer in the simulation. For design B and C the significant sub-set remains same with $B_{1,4,5,6}$ and $C_{1,3,4}$, and a very large effect [Saw09]. Only the significant set of the modified design A^* is reduced by the omnidirectional directivity (index 4) and consists of $A^*_{1,3,5}$. The mapping to auditory distance resembles the curves of the simulated sound field, cf. Figure 3.4. While for the modified design A^* the mapping is almost linear without saturation of the distance perception, means of design B remain sigmoid-shaped as they are in the simulated sound field. The mapping of design C shows major differences and conditions $C_{4,5}$ are no longer localized more distant compared to C_7 when auralized with the IKO. In accordance with Experiment 5, these conditions yield major intersubjective differences as indicated by the size of 95% confidence intervals.

Informal notes of listeners indicate that the spectral coloration of some conditions led to an impression as if the auditory event is right behind the IKO and the incoming sound is filtered due

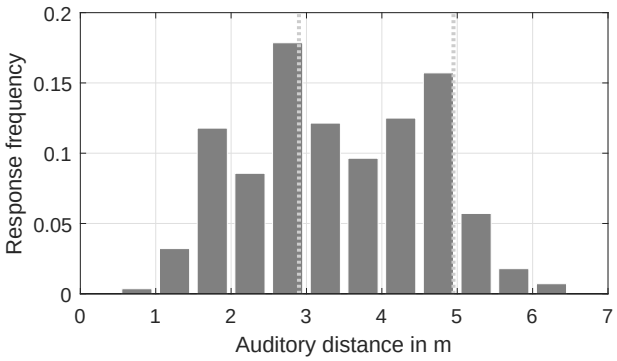


Figure 3.14: Histogram of all responses of the distance task. Distances of IKO and wall are indicated by dotted lines.

to acoustic shadowing. The availability of visual cues generally improves the distance judgments. Thus, seeing only one possible sound source biases the perceived distance towards the IKO [AZ14]. This explains the high frequency of responses within the interval of the IKO shown in Figure 3.14. Moreover, visual cues are thought to cause less pronounced ratings for large distances, i.e., conditions $C_{4,5}$. Responses of most listeners (7/10) are within the feasible space limited by the wall at approximately 5 m, leading to a high response frequency in the interval right in front of the wall.

Interestingly, in the fist experiment visual cues were available similarly as listeners saw the loudspeaker ring around them, but no influence thereof was obtained. Therefore it can be concluded that in the laboratory environment, in which visual cues do not comply with auditory cues, the former play a minor role. This agrees with findings in [EB02, MMP16] showing that sensory interactions, e.g., vision vs. audition, include a weighting process where the most reliable cue contributes the most to the multi-sensory percept.

Apparent source width. Figure 3.15 shows the results for the ASW rating task as medians and 95% confidence intervals. Normalized mean values of the distance task (highlighted in gray) strongly resemble corresponding source-width curves, underlined by a high correlation of central tendencies with $R^2 = (0.98, 0.96, 0.92)$ for designs A, B, C, respectively.

The correlation of ASW to auditory distance is not surprising, considering the model predictions of Experiment 5 in Section 3.2.3. Two of best predicting models for distance, $[1 - \text{IACC}]$ and LF, are measurements that are typically used in room acoustics to quantify the ASW. This is in contrast to the inverse relation between the physical source-listener distance and the ASW reported by Lee in [Lee13]. His data shows that the ASW almost linearly decreases as the distance is doubled. Regarding acoustic measures, Lee found the $[1 - \text{IACC}]$ and LF to be positively correlated to the distance, but negatively correlated to the ASW. This latter relation contradicts most other literature and the author concludes that it "seems difficult to establish a regular relationship between $[1 - \text{IACC}]$ or LF and ASW".

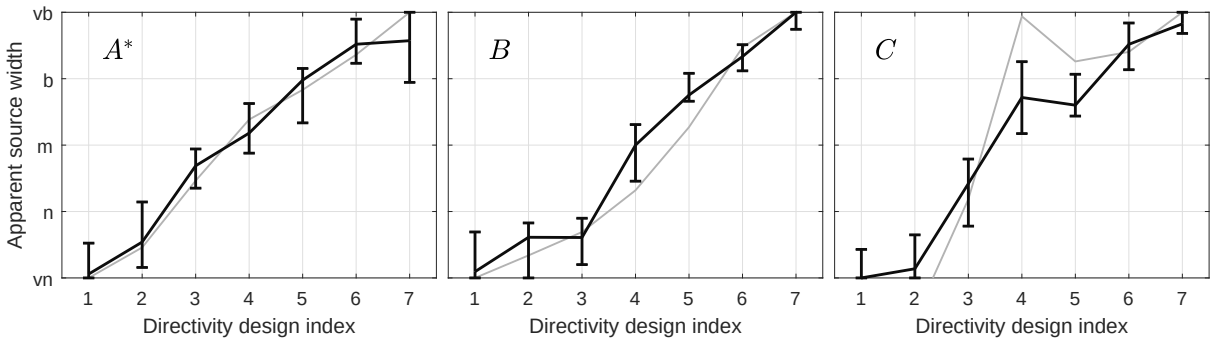


Figure 3.15: Medians and corresponding 95% confidence intervals of apparent source width (black) for beampattern designs A^*, B, C using speech as excitation signal and the IKO. Gray lines are mean values of the corresponding auditory distance, normalized to the scale endpoints.

3.3.3 Application of distance models to the IKO

This section applies selected distance models to the experimental results of the IKO. Similar to the lateralization model of Experiment 4 in Section 2.5.1, the distance modeling is based on an approximation of the real sound field using a simple 3rd-order image-source model with uniform frequency independent absorption coefficient, known from the sound field simulation of the previous section, cf. Eq. (2.16), and a weighted energy average of the measured directivity of the IKO. Examined models are the two spatial measures [1-IACC] and LF, and two energy measures DRR and $[1 - ||\boldsymbol{r}_\mathrm{E}||]$. Respective parameters are calculated for each condition and then jointly fitted to the distances determined by listeners using linear regression.

Figure 3.16 compares the distance percepts of Experiment 6 elicit by directivity designs A^*, B, and C with linearly fitted measures of DRR, [1-IACC], LF, and $[1 - ||\boldsymbol{r}_\mathrm{E}||]$. The mapping of model prediction to auditory distance resembles the experimental data. Both DRR and [1-IACC] yield high coefficients of determination similar to the sound field simulation with $R^2 \geq 0.90$. Especially for sigmoid-shaped curves of designs A^* and B the models mostly overlap with corresponding mean distances, while for the more complex mapping function of design C the models distance crosses 95% confidence intervals of most conditions. The correlation of LF and $[1 - ||\boldsymbol{r}_\mathrm{E}||]$ is lower. For design C these models resemble the mapping function found with loudspeaker-based auralization and, thus, overestimate conditions $C_{3...5}$. This results in an underestimation of more

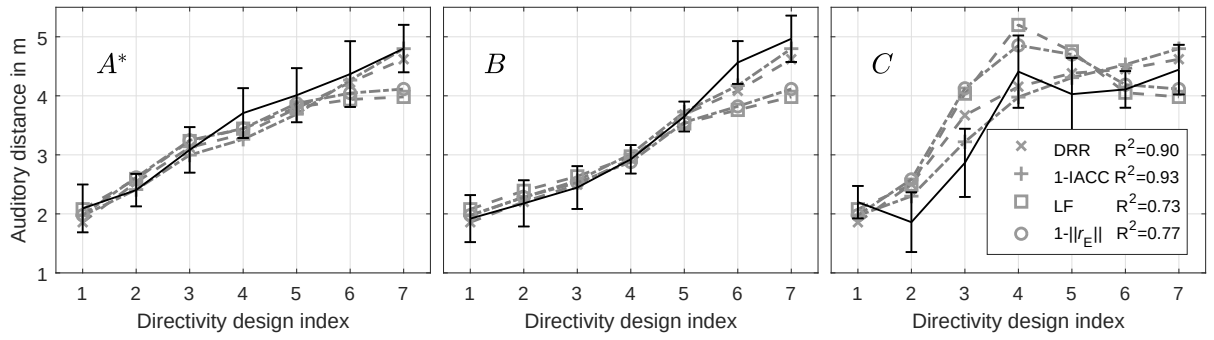


Figure 3.16: Linear fit of distance models to means of distance ratings (thin lines) with corresponding coefficients of determination.

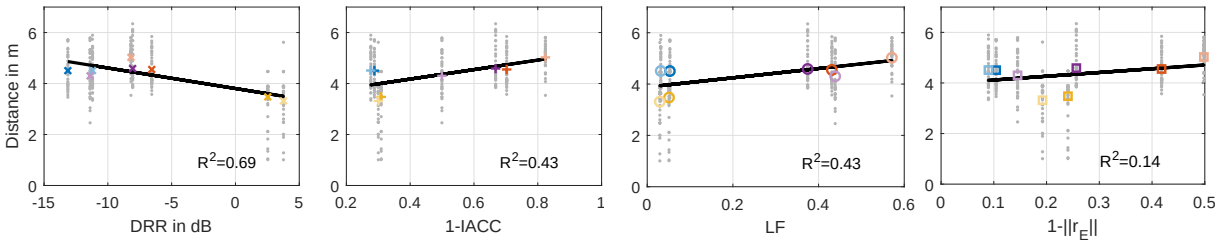


Figure 3.17: Linear regression of distance models to distance ratings (gray markers). Coefficients of determination R^2 are calculated for mean values (colored markers). The color coding of conditions is similar to Figure 2.24.

distant events of designs A^* and B due to the combined fit of the regression function.

To complete the modeling of auditory events in space created by static sound beams of the IKO, the previously examined models are also applied on distance percepts of Experiment 4 from the previous chapter with the image-source model for sound field approximation. The parameters of DRR, [1-ACC], $[1 - ||\boldsymbol{r}_\mathrm{E}||]$, and LF are jointly fitted to all distance ratings of the 4 beam directions and 2 signal types using linear regression. Figure 3.17 compares the linear regression functions with mean distances. Note that slight differences of modeled distances between the two signal types of a beam direction $\phi_\bigcirc$ are due to the signal spectra yielding different averaged directivities of the IKO. The overall variation of distance is relatively low and means fall within the range of 3.3 m to 5.0 m. This yields a poor performance of the models. However, signs of regression function's slopes agree with previous model approaches.

3.3.4 Discussion

This section evaluated directivity designs, known from the loudspeaker-based auralization, with the IKO in a room. A listening experiment could show that in real environments the distance perception is biased due to visual cues. Distant auditory events are warped to fit the reasonable space, whereas events near the IKO are pulled in the direction of the plausible visual event. Nevertheless, significant directivities were found to be the same as those found with the simulated sound field. In addition to distance, the experiment studied the spatial impression caused by the directivity of the sound source. The apparent source width was found to highly correlate with distance impressions. This finding is in contrast to the relationship of physical source-receiver distances and the ASW. For a sound source without pronounced directivity, the ASW is negatively correlated to its distance [Lee13], which explains the performance of spatial measures and enhances the robustness of this effect in real environments.

Modeling the distance task is achieved with the use of spatial and energy measures known from the previous section. For the modeling an approximation of the real sound field is sufficient, and similar to the lateralization an image-source model including the measured source directivity yields successful results as long as the range of modeled distances is wide. However, similar to the auditory system, the models are not able to discriminate small differences in distance [ZBB05].

3.4 Summary

In this chapter, an investigation was carried out into the influence of various beampatterns on the auditory distance. Two-dimensional simulation of a variable-directivity sound source at a single point in the room was shown to provide control of the auditory distance. Different beampattern designs/directivity constellations were proposed that cause pronounced and graduated distance impressions. Additionally, the influence of the auralized room, source-listener distance, signal, and single-channel reverberation was studied. The mapping of beampatterns $A_{1...7}$ and $B_{1...7}$ to auditory distance curves is sigmoid-shaped. It resembles the compressive power functions described in [Zah02a], characterizing the relation between physical and auditory distance. Moreover, agreeing with [Col68, LMC92], signals with an increased relative amount of high-frequency energy appeared to be closer in the study. Both decreasing the auralized room in size and increasing the source-listener distance yield a more compressed curve, which is slightly offset in case of the increased source-listener distance. Despite this, the range of discriminability is persistent. A mild single-channel reverberation is effective at increasing the auditory distance, maintaining distinct distance impressions.

Successful modeling of the experimental results was presented and all models yield curves that are highly correlated with the experimental data. Interestingly, spatial measures used to quantify the ASW provide very accurate predictions.

Based on findings from the loudspeaker-based auralization in the anechoic chamber, an evaluation of the designs synthesized by a variable-directivity sound source in a room was presented. A listening experiment verifies the directivity-controllable range of distinct distance impressions, although the results obtained in real environments were found to be biased due to visual cues. In addition to the auditory distance, the apparent source width was evaluated and it can be concluded that in contrast to the natural environments [Lee13], the width highly correlates with distance impressions caused by the directivity of the sound source. This finding explains the performance of spatial measures in the first experiment and enhances the robustness of this new effect in real environments.

The last section applied the models for predicting the auditory distance in real environments. For beams facing towards or away from the listener, the spatial measure IACC and the energy measure DRR are found to be suitable models. The modeling of distance percepts of arbitrary static beam directions from the previous chapter, confirmed knowledge from literature and the DRR was the only model that worked reasonably well.

4

Asynchrony Effects in the Perception of Height

The azimuthal localization of a sound source is primarily determined by interaural time and level differences. These cues cannot uniquely identify the direction to a sound source in the three dimensions of the frontal hemisphere as sounds on a cone of confusion have no or only weak interaural differences. Nevertheless it is possible to localize a sound source within the median plane or along a cone of confusion and research identified spectral properties to be the most prominent cue for the perception of height.

Outline This chapter is structured as follows. Section 4.1 introduces psychoacoustic phenomena related to height perception and gives an overview of how different rendering systems achieve elevation. Section 4.2 introduces a possible method to control the perceived height. It outlines a listening experiment to evaluate a vertical asynchrony effect, elicited by direct sound and floor reflection. The findings are input to a subsequent experiment, presented in Section 4.3, that examines whether for a frontal sound, the presence of a delayed sound at a different height in the median plane can yield the perception of auditory motion when its delay is varied over time. The experiment presented in Section 4.4 studies the influence of delay alterations more closely to determine the influence of the overall delay on the effect. Finally, Section 4.5 summarizes the chapter and discusses effects observed.

4.1 Relevant cues for height and how they are incorporated in rendering systems

Along a cone of confusion, ITDs and ILDs are essentially constant. To disambiguate directions, access to additional information is required. An early work on sound localization by Wallach [Wal38] assumes that multiple sampled directions achieved by head motion characterize the direction of sound geometrically in azimuth and elevation. More recent research focuses on monaural spectral cues resulting from filtering of sound by the pinna, head, and torso, or echoes and reverberation [HV98, BWLA00]. The resulting conceptual model for auditory localization consists of a pattern recognition system that evaluates the spectral-shape information of the ear input signal.

It is believed that HRTF spectra contain all information about sound localization and convolving the sound with the individual directional transfer function (DTF) mimics the acoustic characteristics of free-field listening [WK89, KW92]. Great effort is made to individualize DTFs as to maintain the localization in binaural reproduction. However, in terms of vertical localization more often than not, the effort fails.

In the horizontal plane, listeners can discriminate between the azimuths of identical sources of about $\Delta\phi = 1°$ with sounds shorter than a millisecond. By contrast, typical minimum audible angles (MAA) for elevation vary inter-subjectively ranging from 2° to 5° [PS90, SDSP91]. The pattern-recognition processes of spectral-shape information from the filtering is limited to broadband signals and requires several tens of milliseconds of acoustic input to be completed. Moreover, for narrow-band 1/3-octave noise pulses Blauert [Bla97] found the perceived elevation to be independent of its presentation direction. He referred to the frequency bands by which the direction of sound image are determined as *directional bands*. In the same contribution, Blauert reports a similar effect for narrow-band signals or pure tones, the so-called *pitch-height effect*. It changes the apparent elevation as a function of stimulus frequency, with higher-frequency tones appearing higher in elevation.

For synchronous presentation of two vertically stacked sound instances with equal intensities Bremen et al. [BVV10] found that the perception of a single fused auditory event is described by weighted averaging, as long as the spatial separation is small. If it exceeds $\Delta\theta = 45°$, response distributions become bimodal. Subsequent studies confirmed this finding for multiple sounds [Pul01, WFZ14] and could show that weights depending on relative sound intensities can shift the auditory event within the array spanned by the sound instances. Accordingly, three-dimensional sound systems with elevated loudspeakers such as VBAP, Dolby Atmos[6], or Auro 3D[7] use inter-channel level differences to vertically move virtual sound sources between loudspeakers with $\Delta\theta < 45°$.

[6] https://www.dolby.com/us/en/brands/dolby-atmos.html
[7] https://www.auro-3d.com/system/

In contrast to the horizontal plane, there is little knowledge about asynchrony effects in the median plane. Wallis and Lee [WL15] studied if inter-channel time differences of a vertically stacked loudspeaker pair in the median plane can be used for panning. They found neither a stable localization curve nor any evidence of a vertical precedence effect for octave-band pink noise. Instead, their results suggest that localization is influenced by the natural comb filtering which alters the spectral content of the stimuli and yields varying pitch-height effects. Conversely for speech, Somerville et al. [SGSN66] found that fused images created by two vertically separated loudspeakers shift towards the leading sound if a delay of $\Delta T = 20\,\text{ms}$ is added. Nevertheless, they state that vertical delay effects are different from those in the horizontal plane. However, other studies did attribute the localization of fused auditory images to a vertical precedence effect [Bla71, LRYH97, RHH00, Aga11, TM15], although none of them could show a clear localization dominance. A recent study by Ege et al. [EOBW18] assumes the perception of direct sound and delayed reflection with equal intensities yields backward masking, resulting in a fused percept at a weighted-average position. According to the study, the auditory system is unable to spatially dissociate sounds that co-occur within a time window of at least 160 ms.

According to this last finding, projecting a broadband sound beam to floor or ceiling for vertical spatialization results in a fused percept between the direct sound and the emphasized reflection. No asynchrony effects are expected to be involved as reflection's delays are typically not more than several milliseconds. This chapter addresses asynchrony effects that might occur and studies elevation cues induced by lagging sound instances from the median plane using a series of listening experiment that build on one another. The results of the experiment presented in Section 4.2 are published in [WHF17], whereas the two subsequent experiments appear in [WFH19].

4.2 The influence of delay and level of a median-plane lag on the apparent height

The direct sound of an emitting sound source is typically followed by attenuated reflections. Lateral reflections contribute mainly to qualitative attributes of the perceived sound [Bar71] without affecting the localization [Oka00]. In contrast, there is evidence that a delayed and attenuated sound instance in the median plane may yield additional localization cues. Guski [Gus90] found that in an anechoic environment the addition of a floor reflection increases localization accuracy of speech, whereas a reflection from the ceiling decreases it.

Considering a sound source and a receiver at a certain distance that are initially at equal height above the floor, a vertical movement of the source along the median plane affect the delay and level alongside with direction. Studies have shown an influence of the reflection pattern on timbre and thus on the spectral-shape of the perceived sound [RJ03] providing additional information on the height of the source. To determine if and what relevant information of the floor reflection is evaluated by the auditory system, Experiment 7 isolates both time delay and level of a floor reflection with fixed direction at $-45°$.

The experimental setup consists of six vertically arranged Genelec 8020A loudspeakers set up in the anechoic laboratory. In the left panel of Figure 4.1 a sketch of the setup is shown. The leading sounds are played back with loudspeakers $\mathrm{LS}_{(-20\ldots20)}$, which are equally distributed between $\theta = -20° \ldots 20°$ in the median plane and covered by an acoustically transparent screen, cf. right panel of Figure 4.1. Corresponding lagging sounds are provided by LS_{-45}, which corresponds to the specular direction for LS_0. For all other leads, there is a small directional mismatch between the lag and the actual specular direction, which is below the minimum audible angle. All loudspeakers are level- and delay-compensated to the central listening position.

Conditions. Conditions are tested in three different sets, each defined by fixed directions for lead and lag $\theta_{\mathrm{lead}}/\theta_{\mathrm{lag}}$. The conditions within each set are created by variations of delay and level of the lag compared to its lead. These variations, indicated with the index j, are derived

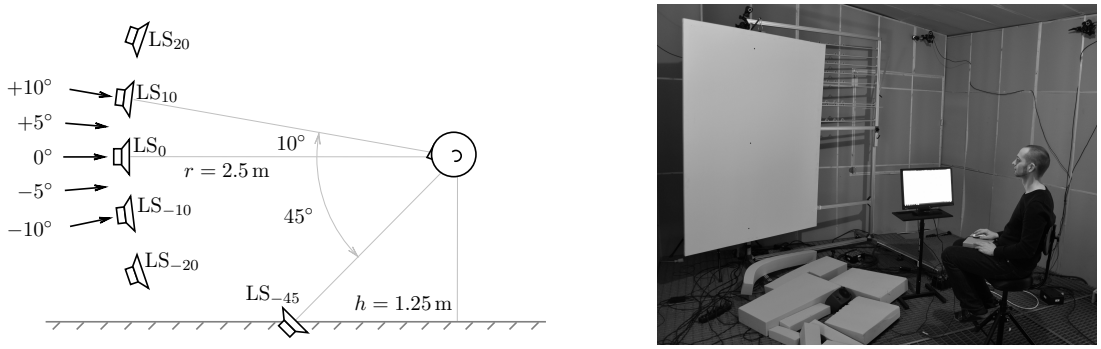


Figure 4.1: Conducting Experiment 7: (left) Sketch of the loudspeaker setup with black arrows exemplary indicating shifts $\Delta\theta_{\mathrm{lead}}$ which are used for creating conditions $j = (-10, -5, 0, 5, 10)$; (right) Setup at the IEM's anechoic laboratory.

Table 4.1: Delay ΔT and level ΔL of the lag compared to its lead for conditions tested in the experiment.

set		low anch.	Condition index j					high anch.
		$\theta_{\text{lead}} - 10°$	-10	-5	0	5	10	$\theta_{\text{lead}} + 10°$
-10/-45	ΔT/ms	1.1	1.1	1.6	2.1	2.6	3.0	3.0
	ΔL/dB	−2.6	−2.6	−3.1	−3.6	−4.0	−4.4	−4.4
0/-45	ΔT/ms	2.1	2.1	2.6	3.0	3.5	3.9	3.9
	ΔL/dB	−3.6	−3.6	−4.0	−4.4	−4.8	−5.1	−5.1
10/-45	ΔT/ms	3.0	3.0	3.5	3.9	4.3	4.7	4.7
	ΔL/dB	−4.4	−4.4	−4.8	−5.1	−5.4	−5.7	−5.7

from specular floor reflections of a source shifted in elevation by $\Delta\theta_{\text{lead}} = (-10°, -5°, 0°, 5°, 10°)$ compared to the loudspeaker providing the lead, with a frequency-independent reflection coefficient of the floor of $R = 0.92$. For example, condition $10/\text{-}45_0$ is composed of a lead played back with LS_{10}. The index $j = 0$ indicates no shift, and delay and level of the lag are those of the corresponding specular reflection but presented from LS_{-45}. For condition $10/\text{-}45_{10}$ the directions remain the same but delay and level of the lag are derived from the specular reflection of LS_{20}. Each set -10/-45, 0/-45, 10/-45 is complemented by two anchor condition yielding 7 conditions per set. The leading directions of the anchors are varied by ±10° (upper/lower anchor) compared to the nominal lead direction of the set and delay and level of the lags are derived from the corresponding specular reflections. In this way the lag of the lower and upper anchor are identical with the lags of conditions $j = -10$ and $j = 10$, respectively (cf. Table 4.1).

The sounds tested are an anechoic sample of female speech[8] S_1 and a one-second-long pink noise burst (onset and release time of 10 ms) followed by a pause of the same length S_2. Sounds are played back in loop at comfortable level of 70 dB(A). In a preliminary experiment, listeners were asked to rate the perceived elevation of each condition using a pointing device [FMSZ10]. Apart from the anchors, there were no tendencies in the response sets. Therefore, Experiment 7 is conducted as multi-stimulus test and each of the three sets -10/-45, 0/-45, 10/-45 allows a direct comparison of the five test conditions $j = -10 \ldots 10$ and both anchors for the excitation signal under test. The listeners' task is to indicate the perceived height of each randomly ordered sample of the set with a continuous slider on a relative scale. They are asked to rate using the entire scale and when listening to a condition they are requested to face the 0° direction while minimizing body movements. However, small head movements were tolerated. Eight listeners participated in this experiment. All of them were experienced listeners and, except for one listener, all performed two runs resulting in 15 answers per condition.

4.2.1 Experimental results

Both anchors with $\Delta\theta_{\text{lead}} = \pm 10°$ were always identified as least and most elevated conditions within each set, except for one out of 90 response sets (3 test sets × 2 signals × 15 runs). This

[8] *Music for Archimedes* CD Bang and Olufsen 101 (1992)

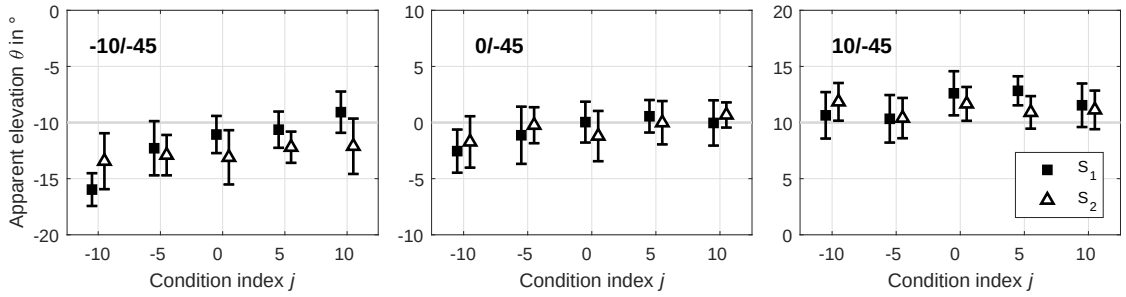


Figure 4.2: Mean and corresponding 95% confidence intervals of responses for the three sets -10/-45, 0/-45, 10/-45 with speech S_1 and noise S_2. Bold horizontal lines indicate the direction of the lead θ_{lead}.

allows to map the relative-scale responses to absolute elevation angles θ by scaling them with the angles of the anchors $\theta_{\text{lead}} \pm 10°$ for each response set. Figure 4.2 shows means and corresponding 95% confidence intervals for responses (normal distribution according to Lilliefors test, $p > 0.05$) for the test conditions $j = (-10, -5, 0, 5, 10)$ of test sets -10/-45, 0/-45, 10/-45.

Speech means of set -10/-45 and 0/-45 increase monotonically along the condition index implying an effect of delay and level on the perceived elevation. Largest shifts were achieved with set -10/-45, which agrees with findings from Guski [Gus90], who showed that a loudspeaker at $-12°$ benefits most (compared to $0°$ and $12°$) when a floor reflector is added. Nevertheless, a statistical analysis (ANOVA corrected using Tukey's HSD) of set -10/-45 and speech S_1 revealed only condition $j = -10$ to be significantly different from $j = 0$ ($p \leq 0.05$; very large effect with Cohen's $d = 1.7$) with a mean shift in the range of the MAA. Contrastingly, for the noise S_2 no significance and no clear tendency is observable.

An alternative approach to evaluate the responses considers the vertical auditory movement direction, perceived when switching between conditions $j = 0$ and one of $j = (-10, 5, 5, 10)$. By interpreting individual relative displacements, the perception can be categorized as *upwards* or *downwards* movement. For the most interesting set -10/-45 and S_1, the intended movement direction is achieved in 73% of responses for conditions $j = (-10, -5)$ (downwards movement) and 67% for conditions $j = (5, 10)$ (upwards movement). For noise S_2 the values are just slightly above chance with 53% and 56%, respectively.

4.2.2 Discussion

This section evaluated the influence of level and delay of a lag derived from a floor reflection and fixed to $-45°$ on the apparent elevation. Two different signals were tested to determine if a floor reflection carries relevant elevation cues for the auditory system. While broadband noise did not result in any significance, decreasing the delay ΔT and at the same time increasing the level ΔL of the lag yields a significant vertical displacement of the auditory event for speech. However, this effect on the apparent elevation was only observed for the differential comparison of two conditions. Moreover, it remains unclear what reflection parameters influence the effect strength, but as reflection levels vary by not more than 1.8 dB it is assumed that primarily the delay is significant.

According to [HV98] the extraction of the veridical elevation angle from the sensory spectrum is an ill-posed problem. The input at the listeners ears results from a convolution of the unknown source signal and an unknown direction-dependent pinna filter. To cope with this, the auditory system has to rely on additional assumptions regarding potential sound locations, pinna filters, and source spectra. The temporal processing of spectral sensory information into a dynamic estimate of sound elevation is assumed to evaluate free-field spectral-shape information. However, given that natural speech spectra are familiar, reflection properties might also contribute to the spectral-shape information evaluated by pattern-recognition processes. This implication is supported as reflection properties of the relatively unfamiliar noise signal did not affect the apparent height.

4.3 The influence of delay changes on the auditory movement

The experiment presented in this section focuses on the delay between two sound instances in the median plane and how its manipulation affects the apparent height. Listeners of the previous experiment reported the comparative rating to be demanding. Therefore, Experiment 8 studies the influence of a continuous delay alteration on the movement perception, which involves similar processes as the discrimination of nearby locations [SG12].

Starting from a fixed delay ΔT of the lag, the listeners task is to adjust both an upper and lower limit, denoted as $\Delta T_+ = \Delta T + t_+$ and $\Delta T_- = \Delta T - t_-$, yielding the most shifted auditory event evoked by a continuous vertical movement. Both limits of the varied time $t_\pm$ are determined separately, using the same task. Using a fader, the time initiated with $t_\pm = 0\,\text{ms}$ is increased as long as a monotone vertical source movement is perceivable. The left panel of Figure 4.3 shows the modification in terms of the impulse response. Once the listener determines the fader position yielding the most vertically deflected event, she or he responds as to the direction in which the auditory event has moved, i.e. *up* or *down*, and the adjusted delay $t_\pm$ is stored by pressing a button. The maximum allowable delay alteration is fixed with times $t_\pm^{\max} = 2\,\text{ms}$, determined by informal listening. If no or only unstable movements are perceived, the listener can also answer with *no*. After logging in the answer, the next sample is loaded and the motorized fader jumps back to the zero position $t_\pm = 0\,\text{ms}$ and the next condition starts automatically.

Conditions. The experimental setup is similar to Experiment 7 with $\text{LS}_{(10,0,-10)}$ for supplying the leading sound, cf. Figure 4.1. Lagging directions are provided by the floor loudspeaker LS_{-45} and an additional loudspeaker at the ceiling LS_{45} with $\theta_{\text{lag}} = 45°$. Conditions are defined by the angles of lead and lag $\theta_{\text{lead}}/\theta_{\text{lag}}$. The delays ΔT are calculated from the distance between the corresponding image source and listener, whereas the level of the lag is fixed with $\Delta L = -4.4\,\text{dB}$ compared to the lead. Hence, a θ_{lead}-dependent reflection coefficient is assumed. Floor conditions are -10/-45, 0/-45, 10/-45 with delays $\Delta T = (2.1, 3.0, 3.9)\,\text{ms}$, respectively. A ceiling condition is studied with 0/45 and $\Delta T = 3.0\,\text{ms}$, and, additionally, a control condition with $\theta_{\text{lead}} = \theta_{\text{lag}} = 0°$

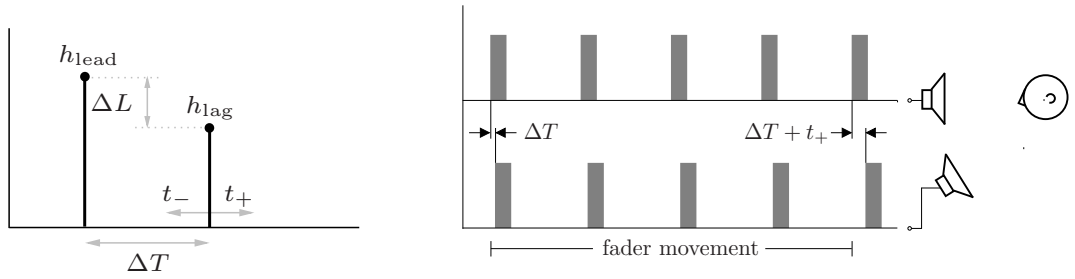


Figure 4.3: Left: Simplified impulse response of direct sound h_{lead} and reflection h_{lag}. The reflection's delay is continuously increased (t_+) or decreased (t_-) by the listeners. Right: Exemplary illustration of the sequence repetition for a condition with increasing delay ΔT_+ and noise bursts. Starting from the specular delay ΔT, moving the fader increases the delay of the lagging reflection until reaches $\Delta T_+ = \Delta T + t_+$ at the beginning of the 5th bursts pair.

and $\Delta T = 3\,\text{ms}$, denoted as 0/0, is included in the test in order to evaluate the influence of comb filter effects. Signals were chosen to investigate the influence of familiarity and envelope to the effect: anechoic female speech[9] S_1, known from the previous experiment, and two 100-ms long pink noise bursts of which S_{21} has onset and release times of 10 ms and S_{22} with 2 ms and 98 ms, respectively. The noise samples are followed by a pause time of 60 ms, whereas to the 7 s-long speech sample no pause is appended. For each condition, the samples are played back in a loop at 70 dB(A) until the listeners confirm their response. The right panel of Figure 4.3 shows an exemplary illustration of the sequence for a condition with increasing delay $\Delta T \rightarrow \Delta T + t_+$. The test sequence for every listener is an individual random permutation of the entire set yielding 3 sounds × 4 conditions × 2 repetitions × 2 delay alterations = 48 adjusted delays and 48 movement directions per listener.

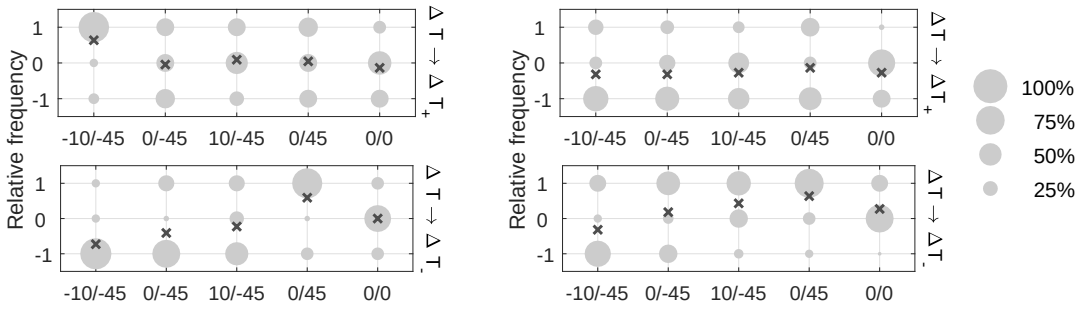
Twelve experienced listeners participated in the experiment. Before conducting the experiment, listeners familiarized with the signals in a short training. For this purpose, single loudspeakers were used for playback and listeners knew which loudspeaker is active. Listeners were requested to face the 0° direction while adjusting the fader and to minimize body movement.

4.3.1 Experimental results

The listeners' consistency is determined by comparing their answers concerning the direction of perceived movement for both repetitions. Overall the consistency is relatively poor and after excluding one listener (consistency 40%), it is 56% on average (min. 50%, max. 67%). The distance between each response category and the neutral location is equidistant on the ordinal scale, what permits a transformation of the data by assigning a value of +1 for *up* responses, −1 for *down* responses, and a 0 for *no* responses yielding a dataset that follows normal distribution (Lilliefors, $p > 0.05$).

A statistical analysis reveals the signal to be a significant parameter ($p \leq 0.05$, ANOVA, Tukey HSD); both noise signals $S_{(21,22)}$ are significantly different from speech S_1, whereas there is no

[9] *Music for Archimedes*, CD Bang and Olufsen 101 (1992)



(a) Speech S_1 (b) Noise S_2

Figure 4.4: Histogram of answers concerning the direction of perceived movement when increasing (upper panel) and decreasing (lower panel) the delay ΔT for speech S_1 and pooled noise signals S_2. The value $+1$ corresponds to *up* responses, -1 to *down* responses, and 0 to *no* responses. Corresponding means are indicated as $\times$.

Table 4.2: Effect size expressed as Cohen's d for speech S_1 and pooled noise samples S_2. Values are color coded according to the classification of [Saw09]: $|d| = 1.2$ very large, $|d| = 0.8$ large effect, and $|d| = 0.5$ medium effect.

condition		-10/-45	0/-45	10/-45	0/45	0/0
ΔT		2.1 ms	3.0 ms	3.9 ms	3.0 ms	3.0 ms
S_1	$\rightarrow \Delta T_+$	0.9	−0.1	0.1	−0.1	−0.2
	$\rightarrow \Delta T_-$	−1.2	−0.5	−0.3	0.8	0.0
S_2	$\rightarrow \Delta T_+$	−0.4	−0.4	-0.4	−0.2	−0.5
	$\rightarrow \Delta T_-$	−0.4	0.2	0.6	1.0	0.6

difference between the envelopes of the noise conditions. Therefore further analysis is done by pooling the responses from the noise signals. Figure 4.4 shows histograms and mean values of answers concerning the direction of perceived movement for speech S_1 and noise S_2.

A one-sample t-test is performed, to determine if condition means are different from 0 (*no* auditory movement). Scaling calculated t-values with the square root of the number of observations yields Cohen's d. Table 4.2 lists effect sizes for incrementing and decrementing conditions with speech S_1 and noise S_2. All conditions with Cohen's $|d| \geq 0.5$ are significant (medium effect; $p \leq 0.05$, Bonferroni-Holm post hoc analysis). A paired-sample t-test reveals the *signal type* to be a significant parameter for most conditions ($p \leq 0.05$, Bonferroni-Holm post-hoc analysis), except for incrementing conditions 0/45, 0/-45 and decrementing condition 0/0. Agreeing with the effect size d listed in Table 4.2 the factor *alteration of the delay* (incrementing vs. decrementing) is significant for conditions -10/-45 and 0/45 ($p \leq 0.05$) for speech S_1, while for noise S_2 it is significant for all other conditions ($p \leq 0.05$).

Interestingly, the effect direction of significant conditions with a floor reflection ($\theta_{\text{lag}} = -45°$) is opposite for the two signal types. First for noise S_2, increasing the delay of condition 10/-45 yields a significant downwards movement, whereas a decrease of ΔT tends to be perceived as

Table 4.3: Mean values and standard deviations of adjusted delays $t_\pm$ for speech S_1 and noise S_2.

condition	t_+ in ms		t_- in ms	
	S_1	S_2	S_1	S_2
-10/-45	1.2 ± 0.6	1.0 ± 0.5	1.0 ± 0.5	0.8 ± 0.4
0/-45	1.1 ± 0.5	1.2 ± 0.6	1.2 ± 0.5	1.1 ± 0.5
10/-45	1.1 ± 0.6	0.8 ± 0.5	1.1 ± 0.6	1.1 ± 0.4
0/45	0.9 ± 0.6	1.1 ± 0.7	0.6 ± 0.4	0.8 ± 0.6
0/0	1.2 ± 0.7	1.0 ± 0.7	1.5 ± 0.6	1.2 ± 0.5

up-moving event. This result, which is obtained similarly for the control condition 0/0, opposes a corresponding physical change of the delay can however be attributed to the pitch-height effect, as decrementing the delay yields comb-filter notches and peaks that move upwards in frequency. Several listeners mentioned that they heard comb filtering for broadband noise signals, perceived as an upward or downward glissando. For the speech signal S_1 on the other hand the effect direction depends on the direction of the reflection. In compliance with physical sound sources, auditory events tend to move upwards if the delay of the floor reflection is increased and downwards if it is decreased. Accordingly, for a reflection from the ceiling the relation is reversed, whereas for the control condition no effect is observed.

Table 4.3 lists adjusted delays as mean values and standard deviations for both alterations. Delays are highly subjective yielding high standard deviations. Nevertheless, mean values of conditions are similar with overall means for incrementing and decrementing conditions of $\bar{t}_\pm = 1\,\text{ms}$. Such a delay alteration of a physical floor reflection is achieved by shifting the height of an emitting source at $\theta = 0°$ and $r = 2.5\,\text{m}$ by approximately $\Delta\theta_\pm = 10°$.

4.3.2 Discussion

The listening experiment presented in this section studied the influence of continuous alterations of the delay between a vertically arranged loudspeaker pair in the median plane on the auditory movement. Depending on the excitation signal, two different effects are deduced. For the relatively unfamiliar broadband noise signal delay alterations $\Delta T \rightarrow \Delta T_\pm$ yield the pitch-height effect as time-variant comb filtering is perceived as a glissando. This relation agrees with [WL15] who found that the comb filter, elicit by two asynchronous noise samples presented the median plane, alters the spectral content and with it the apparent elevation.

For the spectro-temporally sparse and more familiar speech signal findings from the previous experiment in Section 4.2 are confirmed and further evidence is found that median plane reflections contribute to the spectral-shape information for dynamic estimates of sound elevation. In compliance with the physical movement of a sound source in the median plane, increasing the floor reflection's delay yields upwards movements, decreasing it yields downwards movements in the perception. For a refection from the ceiling this relation is reversed, whereas frontal reflections do not elicit any auditory movement.

4.4 Controlling the vertical auditory movement by inter-channel time differences

In the previous experiment two different effects were shown to influence the movement perception, depending on the excitation signal. The pitch-height effect is well-known in localization studies. Results gathered by speech indicate the existence of other localization cues induced by a median plane reflection.

Experiment 9 evaluates the effect that was obtained for speech more closely [WFH19]. Possible influences of the pitch-height effect are minimized by restricting the experiment on floor reflections, for which the movement direction is found to be opposite compared to the physically-motivated effect under investigation. The listening experiment is carried out to assess the influence of the overall delay ΔT and loudspeaker angles θ_{lead} and θ_{lag}.

Conditions. In Experiment 9 listeners are asked to indicate the movement of auditory events for continuous alterations of the delay from ΔT to ΔT_+ and ΔT to ΔT_-. Based on the results from Experiment 8, delay alterations of $t_\pm = 1\,\text{ms}$ are kept constant for all listeners and conditions, and covered a range of $\Delta T = 0...6\,\text{ms}$. To keep the testing time limited, a gap between 4 ms and 5 ms is inserted yielding five delay increments $\Delta T \rightarrow \Delta T_+ = (0 \rightarrow 1, 1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 4, 5 \rightarrow 6)\,\text{ms}$ and corresponding delay decrements $\Delta T \rightarrow \Delta T_- = (1 \rightarrow 0, 2 \rightarrow 1, 3 \rightarrow 2, 4 \rightarrow 3, 6 \rightarrow 5)\,\text{ms}$. The experimental setup is similar to Experiment 7 with $\text{LS}_{(0,-10)}$ supplying the leading sound, cf. Figure 4.1. To study directional dependence, an additional floor loudspeaker LS_{-30} at $\theta_{\text{lag}} = -30°$ is included yielding conditions 0/-45, -10/-45, and 0/-30. Examined 3 s-long signals are non-stationary to prevent the pitch-height effect. S_1 is the female speech sample[10] from Experiment 8 and S_3 is a conga sample[11]. In the experiment, delays of the lags are automatically increased or decreased between 0.5s to 2.5s of the sample length and listeners have to rate the movement of the auditory event with *up*, *down*, or *no* continuous movement. The speech sample S_1 is tested with all conditions, whereas the conga sample S_3 is tested with condition 0/-45 only. Additional control conditions include automatically amplitude-panned speech S_1 using VBAP between LS_0 and LS_{-10} and vice versa. Consistently, the panning starts at 0.5 s and ends at 2.5 s of the sample length.

The test sequence consisted of ((3 speech conditions + 1 conga condition) × 5 delays + 1 VBAP) × 2 repetitions × 2 alterations = 84 samples per listener. Ten experienced listeners participated in the experiment, of which eight already participated in Experiment 8. The training conducted before evaluation was similar to Experiment 8 but included additional amplitude-panned (VBAP) conditions. During the experiment, listeners were asked to minimize body movements.

[10] *Music for Archimedes*, CD Bang and Olufsen 101 (1992)
[11] https://phaidra.kug.ac.at/o:57188

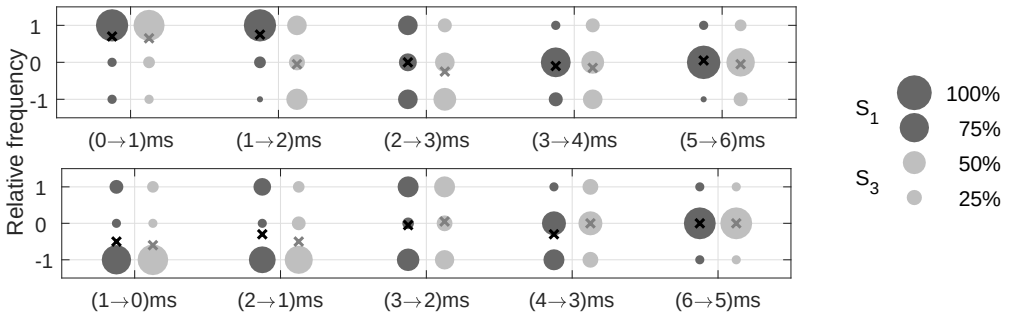


Figure 4.5: Histograms of answers concerning the direction of perceived movement when increasing (upper panels) and decreasing (lower panes) ΔT by $t_{\pm} = 1\,\text{ms}$ for the speech S_1 and conga S_3 (coded in gray) with the setup 0/-45.

Table 4.4: Gray-scale coded effect sizes expressed as Cohen's d for Experiment 9: $|d| > 2$ huge, $|d| = 1.2$ very large, $|d| = 0.8$ large effect, and $|d| = 0.5$ medium effect.

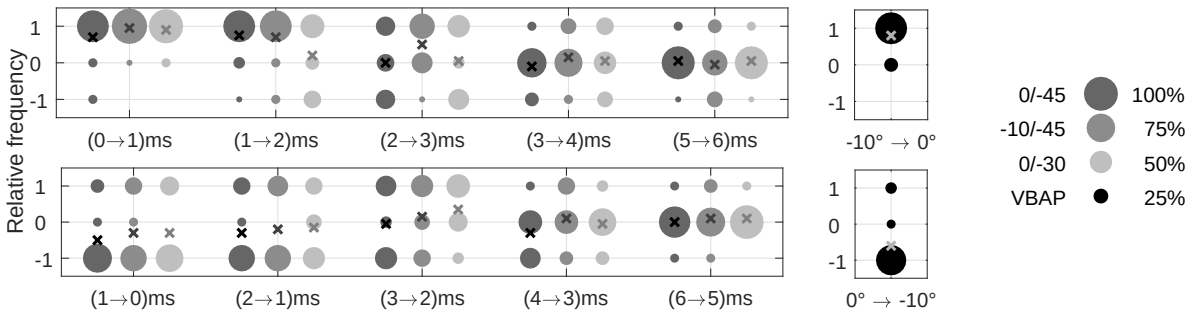
signal	condition		$(0 \rightleftharpoons 1)\,\text{ms}$	$(1 \rightleftharpoons 2)\,\text{ms}$	$(2 \rightleftharpoons 3)\,\text{ms}$	$(3 \rightleftharpoons 4)\,\text{ms}$	$(5 \rightleftharpoons 6)\,\text{ms}$
S_1	0/-45	ΔT_+	1.1	1.4	0.0	−0.2	0.1
		ΔT_-	−0.6	−0.3	−0.1	−0.4	0.0
	-10/-45	ΔT_+	>2.0	1.1	0.8	0.3	−0.1
		ΔT_-	−0.3	0.2	0.2	0.1	0.2
	0/-30	ΔT_+	>2.0	0.2	0.1	0.1	0.1
		ΔT_-	−0.3	−0.2	0.4	−0.1	0.3
S_3	0/-45	ΔT_+	1.0	−0.1	−0.3	−0.2	0.1
		ΔT_-	−0.8	−0.7	0.1	0.0	0.0

4.4.1 Experimental results

The listeners' consistency is monitored by comparing answers of both repetition and the mean consistency of 58% (min. 50%, max. 65%) is just above the value found for Experiment 8. Similarly, responses are transformed by assigning values of 1/0/-1 for movements *up*/*no*/*down*, yielding answers that follow a normal distribution for each conditions (Lilliefors $p > 0.05$).

The ANOVA of conditions with 0/-45 reveals the *signal type* not to be significant ($p > 0.05$) and for both S_1 and S_3 movement directions are physical and resemble those observed by speech in the previous experiment: increasing the delay tends to be perceived as upwards movement, while a decrease is perceived as a downward-moving source. Figure 4.5 delineates histograms and mean values of S_1 and S_3 with condition 0/-45.

One-sample t-tests are performed, to determine if condition means are different from 0 (*no* auditory movement) and the corresponding effect sizes. In the post-hoc analysis, p-values (including VBAP conditions) are corrected using the Bonferroni-Holm method. Table 4.4 lists values of Cohen's d for incrementing and decrementing conditions with speech S_1 and conga S_3. All conditions with Cohen's $|d| \geq 0.8$ are significant ($p \leq 0.05$).



(a) Delay alteration (b) VBAP

Figure 4.6: Histograms of answers concerning the direction of perceived movement with speech S_1 using: (a) delay alterations of ΔT by $t_{\pm} = 1\,\text{ms}$ with loudspeaker angles coded in gray; (b) horizontal VBAP with panning directions $\theta_{\text{pan}} = (-10°, 0°)$.

Figure 4.6 compares the perception of different loudspeaker combinations $\theta_{\text{lead}}/\theta_{\text{lag}}$ tested with S_1. For almost all conditions, most distinct movements of auditory events are achieved at short delays and the effect reduces with increasing the delay ΔT. An ANOVA reveals the *delay* to be a significant parameter for both incrementing and decrementing delays and all three loudspeaker configurations. Effect sizes tend to be higher if the delay is increased with strongest effects found for -10/-45 and 0/-30 at the shortest delay. However, if the effect size considers both incrementing and decrementing conditions by calculating the standardized difference *between* mean values ($= d_{\text{inc}} - d_{\text{dec}}$, cf. Table 4.4), similar values are found independent of the combination $\theta_{\text{lead}}/\theta_{\text{lag}}$. To compare the observed effect with VBAP conditions ($\theta_{\text{pan}} = -10° \rightleftharpoons 0°$, $p \leq 0.05$), speech conditions -10/-45 and 0/-45 with shortest delays are considered (-10/-45: $0 \rightarrow 1\,\text{ms}$; 0/-45: $1 \rightarrow 0\,\text{ms}$). The distance of respective means to 0 is similar to the distance of means of the VBAP conditions, cf. Figure 4.6. Accordingly, effect sizes are comparable with $d_{\text{VBAP}} = 1.9$ for upwards panning and $d_{\text{VBAP}} = -0.8$ for downwards panning (in comparison to $d_{\Delta T_+} = 4.2$ and $d_{\Delta T_-} = -0.6$, cf. Table 4.4).

4.4.2 Discussion

The section examined how continuous increases and decreases of the delay of a lagging sound presented from floor influence the movement perception. Results from the previous experiment were confirmed for a speech and a conga sample; auditory events move upwards if the delay is increased and downwards if the delay is decreased. The overall delay was found to be the significant parameter. If the overall delay exceeds 3 ms, the effect vanishes and no movement was perceived. Strongest effects were observed for delay variations $0\,\text{ms} \rightleftharpoons 1\,\text{ms}$ with effect sizes comparable to vertical amplitude panning using VBAP between $\theta_{\text{pan}} = -10° \rightleftharpoons 0°$.

4.5 Summary

In this chapter, investigations were carried out focusing on the effect of asynchrony on perception of height. It has been shown that a continuous alteration of the delay in the range of ± 1 ms between the leading direct sound and lagging median plane reflection influences the apparent height, which is perceived as vertical auditory movement. Depending on the signal type (speech/noise), the direction of the vertical auditory motion can be directionally opposite.

For stationary broadband noise, incrementing the delay of a lagging sound instance yields auditory events moving downwards. The underlying perceptual phenomena is the pitch-height effect as time-variant comb filtering is perceived as downward glissando. Thus, it is independent of the direction of lag and was observed whenever the overall delay between leading and lagging sound exceeded 3 ms. With instant alterations in the delay caused by switching between conditions in the multi-stimulus test, this effect could not been shown. Obviously, an instant alteration is not perceived as glissando and the resulting tonal change was not pronounced enough to elicit the pitch-height effect for the tested range of the delay.

For spectro-temporally sparse signals such as speech the opposite, more physical relation is observed. A continuous increase of the delay ΔT between direct sound and a floor reflection yields an up-moving auditory event, whereas for a reflection from the ceiling the perceived movement is opposed. In both cases, the delay change corresponds with a thinkable physical change of acoustic floor/ceiling reflection path, and therefore the perception is in compliance with the movement of a physical sound source. One might assume that most distinct movements are perceived whenever the delay matches an acoustic situation, but this is not the case: the strongest effect is obtained for very brief overall delays up to $\Delta T = 1$ ms, which is mostly exceeded by physical floor reflections. The delays of reflections from the chest or shoulders meet this demand. However, they are relatively stable with regard to height of a physical source.

5

Virtual Acoustic Environments for Binaural Reproduction

The previous chapters studied how directional sound sources affect directions and distances in three dimensions, and in greater detail, how various types of acoustic reflections and reverberance in a room contribute. This chapter examines how the perceptual properties of the physical sound field can be plausibly relayed from the original space to a virtual acoustic space for binaural reproduction.

Outline The chapter is structured as follows. Section 5.1 gives an overview of methods and challenges in binaural reproduction. Section 5.2 studies plausibility of a real-time approach for a virtual acoustic environment. It evaluates the perception of auditory events in a virtual room regarding both distance and lateralization. Subsequently, Section 5.3 studies the externalization of the reproduced sound to examine how the people's expectation regarding a listening situation and the individualization of HRIRs affects the plausibility of binaural reproductions. Section 5.4 compares different manipulation strategies of the BRIR reverberation tail to study which physical parameter defining the reverberation of a room is essential for externalization. Finally, Section 5.5 summarizes the chapter.

5.1 Plausibility and authenticity of the virtual auditory space

The headphone-based presentation of sound sources in a virtual room requires the reconstruction of the signals at the listener's ears. The traditional way to emulate room acoustics is to generate binaural signals based on measurements of binaural room impulse responses (BRIRs), capturing not only the effects of pinnae, head, and torso (i.e., the HRIR), but also the reverberation of the room (i.e., the RIR). The drawbacks of this approach are that measurements are time consuming and lack flexibility, as each measurement is restricted to a static scenario with fixed setup of source and receiver (e.g., a dummy head microphone) with fixed orientations within room. To approach this issue, both HRIR and RIR are simulated using mathematical models. For the simulation of RIRs, state of the art algorithms use ray tracing, e.g., [WEJV10], which are extended with image-source models to treat the simulation of early reflections and late reverberant reflections differently in a hybrid approach, e.g., [Nay93, Dal10]. Although these algorithms are able to create complex room acoustical simulations at a computational high cost and require pre-rendering to allow real-time updates of the room acoustical simulation, they are not able to model the physical properties of complex rooms. This is because a comprehensive specification of absorption and scattering for all boundaries of a room is practically impossible. As a consequence, measurement and simulation are perceptually different especially with regard to tone color and the apparent source position [BAA+19].

However, for applications involving perception of human listeners, *authenticity*, defined as the exact perceptual identity [Bla97], is not always necessary. In most cases an auralization has to be perceptually *plausible* and in agreement with the listener's expectation towards an equivalent real acoustic event [LW12]. More efficient methods use feedback delay networks for acoustical modeling, e.g., [Moo79, RS97], or as hybrid approach in combination with an image-source model, where the feedback delay network simulates the late reverberation, e.g. [WvdPE14]. The second section presented in this chapter examines a real-time capable hybrid approach consisting of image-source model and feedback delay network for simulating a virtual directional source and studies its perception in a listening experiment.

For binaural synthesis the synthesized sound field is convolved with measured or modeled HRIRs. This implicitly applies the cues that are evaluated by the auditory system to perceive sound from a certain direction and distance, with a certain source width, and spaciousness. Binaural rendering is of great interest and there are manifold approaches for interpolating a finite subset of HRIRs in order to obtain desired directions, e.g., [BHBS+17, ZSH18, PAB19]. While most of them allow a plausible experience with regard to localization and distance impressions of synthesized sound sources, they can fail when it comes to externalization. This feature describes the fact that humans usually perceive the sound emitted by a sound source to be outside their heads. However, under some listening conditions, this externalized percept breaks down and sounds are perceived inside the head, i.e., internalized. Internalized auditory events are most commonly experienced in headphone listening. One reason therefore is the consideration of

non-individual HRIRs by the binaural renderer yielding a distortion of magnitude spectra of the sound at the ear canal. Studies have shown that especially spectral details in direct sounds are essential cues for perceived externalization of virtual sound sources in reverberant environments [HGD16, LSP19, JSZL20].

Studies about externalized auditory events typically ask for the source distance. Although externalization and distance are closely related when it comes to rating scales, the respective cues are different. While there is no doubt that DRR-related cues play a critical role in distance perception, having access to these cues alone does not necessarily result in externalized sound images. Especially natural ILDs were found to be important for auditory images perceived as convincingly externalized and located at the correct distance in anechoic and reverberant environments [HW96, OLBD10]. The correct ITD and ILD changes with head movements. It is commonly known that when movements are not considered in a static binaural simulations, the externalization decreases, e.g., [BBA13, ORS+20]. Another reason for the breakdown of externalization is the listener's expectation. The divergence of the auralized room and the listening room yields the auditory image to collapse [Ple72, WKH13, NK15]. Although the problem dominates in headphone playback, internalization is not limited to headphones and sometimes also reported when sounds are presented through loudspeakers, e.g., [Too70, BBA13].

Section 5.3 focuses on the interrelation of individualization and expectation on the plausibility of a virtual acoustic environment by measuring the externalization. Subsequently, different strategies to modify binaural room impulse responses maintaining plausible reproduction are examined in Section 5.4.

5.2 Localization of a virtual directional sound source

Experiment 10 studies the perception of static sound beams up to the order $i = 5$ using binaural synthesis. To allow real-time performance, the auralization of the IEM lecture room is based on an image-source model (ISM) to simulate early reflections and a feedback delay network (FDN) to simulate late reverberation. Although this method represents a high degree of simplification of the physical sound field, in the perceptual domain it turned out to be well comparable to state of the art room acoustical modeling [BAA+19]. In contrast to localization experiments presented previously, listeners are not asked to indicate directly the Cartesian coordinates of the auditory event, but to aurally match a location of an omnidirectional virtual source and the location of the given directional virtual source, when both are auralized in the same virtual acoustic environment. To facilitate this task and in order to avoid room divergence the experiment is carried out in the IEM lecture room with the listener's physical position corresponding to the position within the simulation.

The experiment presented in this section can be seen as an extension of Experiment 4 presented in Section 2.5. The content of this section has already been published in [WF18] and it is presented here again to allow for putting it in context with the results of the previous sections.

5.2.1 A simple sound field simulation

The simulation is done using the freely-available tools from IEM plug-in suite[12] and the binaural rendering is done in 7th-order Ambisonics. As depicted in Figure 5.1, the virtual acoustic environment is processed in three parallel paths A – C. Direct sound and early reflections are generated separately for the directional (A) and omnidirectional sound source (B) using the *RoomEncoder* plug-in. Room parameters of the IEM lecture room are identical for both directivity conditions and are derived from Sabine's formula, cf. Eq. (2.16). For the directional-source condition in the path A, a frequency-independent directivity with order i and orientation $\phi_{\bigcirc}$ of the fixed sound source is auralized. Both beam parameters i and $\phi_{\bigcirc}$ are controlled by the *DirectivityShaper* plug-in, which encodes the sound beam into the Ambisonics domain. For the omnidirectional-source condition in the path B, the order $i = 0$ remains constant, but the sound source coordinates (x, y) are varied by the listener throughout the experiment. The plug-ins run in Reaper[13] as DAW on a mobile computer and are controlled according to open sound control (OSC) messages triggered via Pure Data[14] in terms of source position, beam order and orientation, and audio playback. The image-source model in the RoomEncoder plug-ins simulates 126 early reflections, all arriving within a time interval of 50 ms after the direct sound, and they are considered to determine the azimuthal localization. The late reverberation is independent of the source directivity and its position, and does not contribute to the perceived source direction. The signal path C consists of the FDN to simulate a high-density and directivity-independent

[12] `https://plugins.iem.at/`
[13] `https://www.reaper.fm/`
[14] `https://puredata.info/`

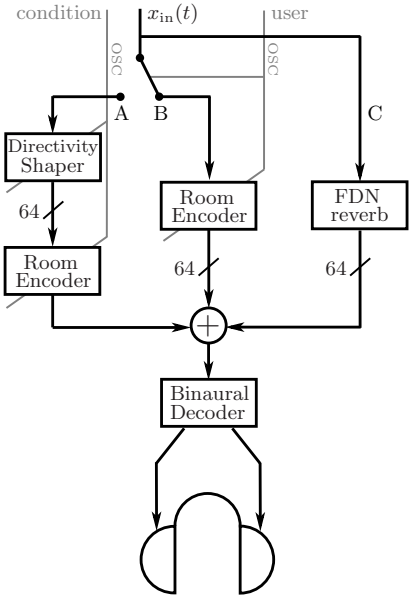


Figure 5.1: Block diagram of binaural rendering on Reaper and the IEM plug-in suite. Gray lines indicate the adjustments by the condition and listener communicated by PureData to the VST plugins in OSC.

diffuse reverberation which is used for both the directional and omnidirectional sound source. Because of the diffuse-field normalization of any directivity involved, it is identically loud for every condition. A very basic FDN consists of a set of N parallel delay lines whose outputs are fed back to their inputs, redistributed according to a $N \times N$ feedback matrix [RS97]. The usage of $N = 64$ channels as inputs of the FDN in the *FDNreverb* plug-in gives a very natural sounding reverberation [Gri17]. The superposition of the impulse response $h_{\text{ISM}}(t)$ (i.e., direct path and early reflections) and $h_{\text{FDN}}(t)$ (i.e., late reverberation) is one of the main challenges in this hybrid approach. By adjusting parameters of the FDN, a continuous temporal energy decay in the transition between both parts of the impulse response is achieved. This can be monitored by backward integration of $h(t) = h_{\text{ISM}}(t) + h_{\text{FDN}}(t)$ over the interval $[0, T]$

$$L(t) = 10 \log_{10} \frac{\int_t^T h^2(t) \mathrm{d}t}{\int_0^T h^2(t) \mathrm{d}t}, \tag{5.1}$$

yielding the energy decay L in dB. Figure 5.2 compares measured and simulated energy decays calculated using Eq. (5.1), which is commonly known as the *Schoeder integration* [Sch65]. Respective room impulse responses of the IEM lecture room are given in Figure 5.3 separately for the ISM and the FDN.

After combining direct sound and early specular reflections with the diffuse reverberation, the signals are filtered by the magnitude-least-squares binaural renderer [ZSH18] based on measurements of the Neumann KU100 [Ber13] using the *BinauralDecoder* plug-in to get high-quality Ambisonics rendering. Playback employed equalized AKG K702 open headphones.

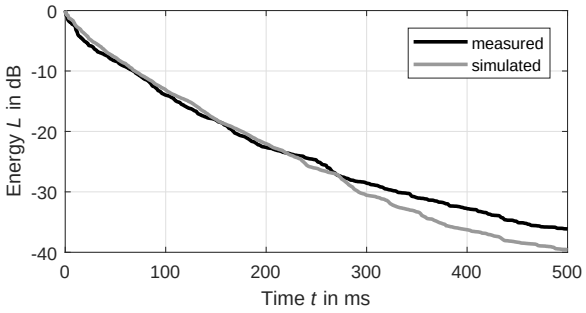


Figure 5.2: Energy decays based on Schroeder integration of the measured and simulated room impulse response of the IEM lecture room at P_1.

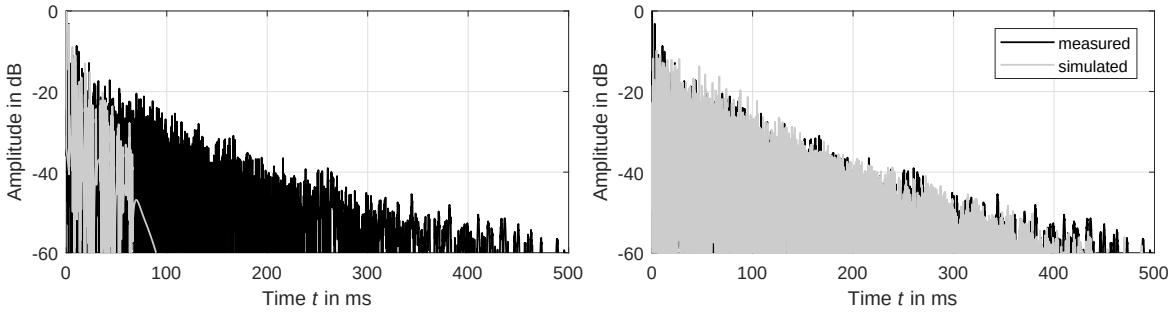
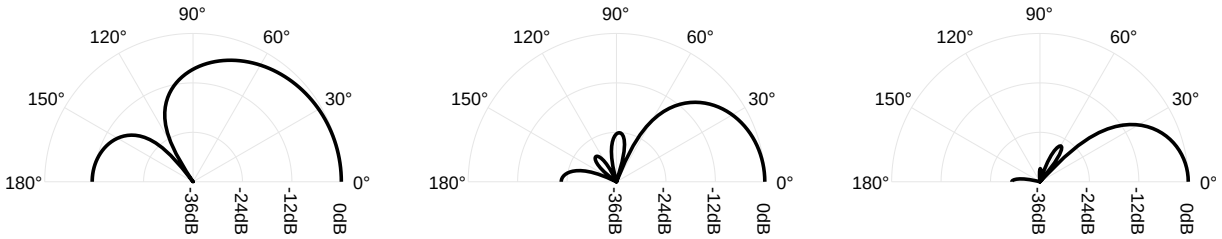


Figure 5.3: Comparison of the measured room impulse response $h(t)$ (black) at position P_1 with simulated direct sound and specular reflections $h_{\mathrm{ISM}}(t)$ using the image-source model (gray, left) and simulated diffuse reverberation $h_{\mathrm{FDN}}(t)$ using the feedback delay network (gray, right).



(a) $i = 1$ (b) $i = 3$ (c) $i = 5$

Figure 5.4: Examined spherical harmonic max-$\boldsymbol{r}_{\mathrm{E}}$ beam patterns of orders i considered by the image source model here in their free-field normalized version to better inspect the difference in shape.

Conditions. Two source positions P_1 and P_2 of the directional source are tested, both with a virtual distance of 4.2 m and azimuth angles of $-10°$ and $40°$ with respect to the listener looking towards the frontal wall. Four beam orientations are tested 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 $\phi_{\bigcirc,1} = (0°, 90°, 180°, 235°)$ for source position P_1 and $\phi_{\bigcirc,2} = (0°, 115°, 180°, 270°)$ for P_2. Directivity orders that are tested for position P_1 are $i = (1, 3, 5)$ and $i = (1, 3)$ for P_2, cf. Figure 5.4. Additionally, for both positions a reference condition with order $i = 0$ is included in the test.

During the experiment, the listener is sitting in the IEM lecture room and her/his physical position corresponds to the position within the simulation. The schematic layout of the room

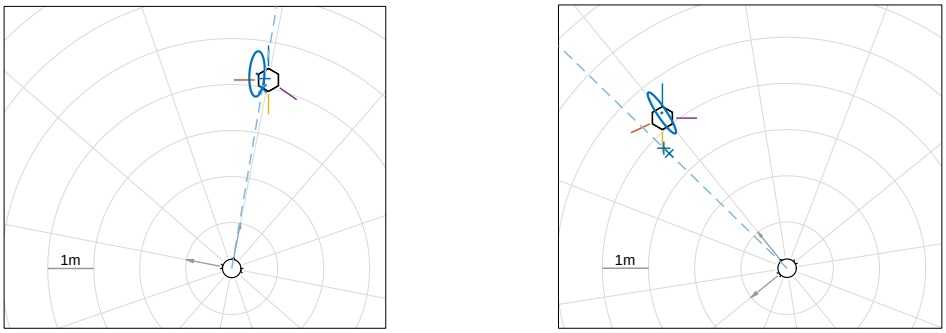
is shown on the computer screen and the listener can only assess the auditory position of the directional sound source through listening (no visual cue given). However, for the omnidirectional source a visual marker within the schematic room layout could be positioned by the listener to adjust its (x, y)-position in the virtual acoustic environment. The listener's task is to match the position of this omnidirectional source in the virtual IEM lecture room with the position of auditory events created by the fixed directional sound source in the same room. While there is the option of just dropping the marker at the expected auditory object location in the floor plan GUI, the listeners are advised to focus on matching the auditory locations, when switching between the given directional condition and the position-parametric omnidirectional auralization. The binaural playback does not include dynamic rendering incorporating the head rotation of the listener. Instead, the rendering is done in a way that the listener is always facing the virtual directional source. Physical markers with numbers are attached to the walls of the room to indicate the viewing direction. In order to match the visual situation with the virtual acoustics, listeners are asked to look at the corresponding number while roughly adjusting the position of the marker (drag and drop) to match the fixed auditory event of the directional-source condition. For fine adjustments, an A/B switch allows to directly compare the auditory event with the omnidirectional source, which can also be adjusted by two sliders controlling its azimuth and distance. Excitation signals are played back in loop until listeners are satisfied with the match. Signal S_1 is a 750-ms-long pink noise burst, linearly faded in and out by $t_{\text{in/out}} = 250\,\text{ms}$, followed by a 500-ms-long pause. Additionally, source position P_1 with directivity order $i = 3$ is tested with male speech[15] S_2. This yields 26 conditions (S_1: 4 beam directions $\times$ [3 orders at P_1 + 2 oders at P_2] + 2 references; S_2: 4 beam directions $\times$ 1 order at P_1), which are 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 of spatial audio. None of the listeners reported that they perceived the auralization as implausible or confusing; some emphasized the naturalness of the binaural synthesis.

5.2.2 Experimental results

The results for the omnidirectional reference conditions with pink noise S_1 are given in Figure 5.5 as 2D mean value and the corresponding 95% confidence ellipse. Ideally, for the reference conditions answers should coincide with the position of the directional sound source. For both view directions answers spread more along the distance compared to the lateralization. This is in agreement with literature according to which we are better in estimating the lateralization compared to the distance [ZBB05]. However, despite these uncertainties mean values and sound source positions largely coincide.

Influence of directivity order. Figure 5.6 shows means and corresponding 95% confidence ellipses for all conditions tested with noise S_1 at source positions P_1 and P_2. Beam orientations

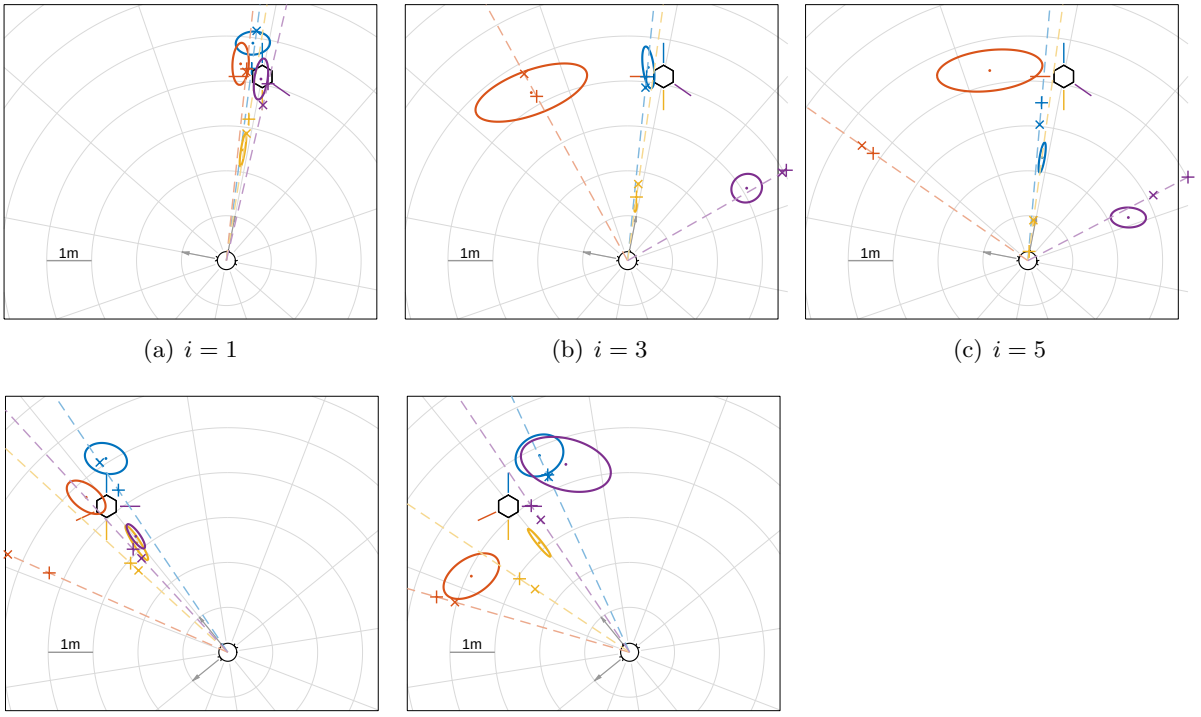
[15] *Music for Archimedes*, CD Bang and Olufsen 101 (1992)



(a) P_1 (b) P_2

Figure 5.5: 2D-means and corresponding 95% confidence ellipses for the omnidirectional reference condition ($i = 0$) at P_1 and P_2 with pink noise S_1. The dashed line indicates the modeled direction $\phi_{\boldsymbol{r}_E}$ and distance models are shown as markers with $\times \ldots f(\text{DRR})$ and $+ \ldots f(1 - \text{IACC})$.

$\phi_{\bigcirc\,(1,2)}$ are color coded and schematically depicted as solid lines around the corresponding sound source. For both source positions, the distance of confidence ellipses to the directional sound source tends to increase with the increasing directivity order i. Like in Experiment 4 of Section 2.5, at P_1 this shift away from the directional sound source agrees with the beam orientation and from the listeners perspective lateral orientations $\phi_{\bigcirc} = (90°, 235°)$ mainly yield lateral shifts, whereas orientations along the median plane $\phi_{\bigcirc} = (0°, 180°)$ shift the smaller ellipses along the distance.



(a) $i = 1$ (b) $i = 3$ (c) $i = 5$

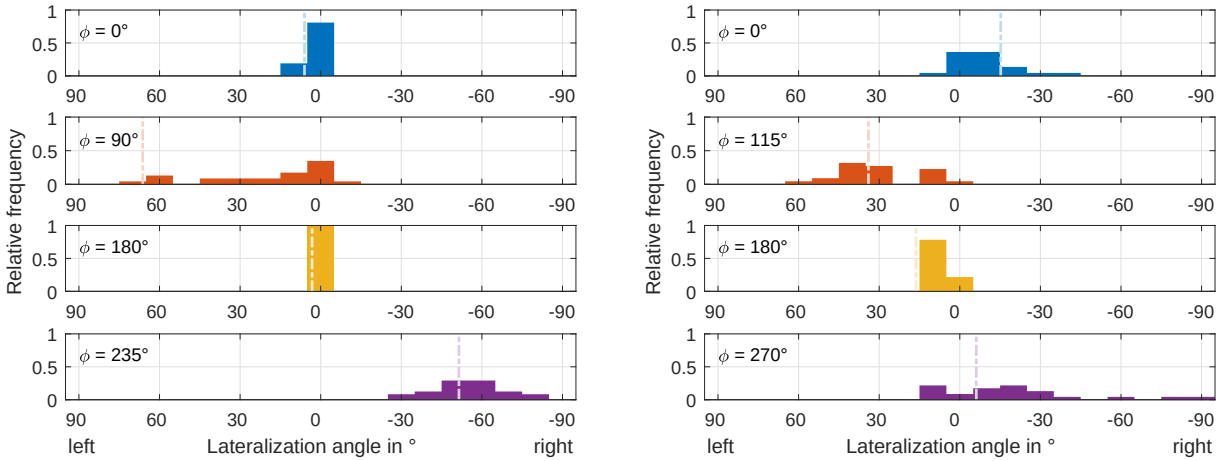
(d) $i = 1$ (e) $i = 3$

Figure 5.6: Order dependent 2D-means and corresponding 95% confidence ellipses for tested beam directions color coded for P_1 (a-c) and P_2 (d-e) with pink noise S_1. Modeled lateralizations $\phi_{\boldsymbol{r}_E}$ (dashed lines) are base on the signal-dependent precedence weight $\beta = -0.36\,\text{dB/ms}$. Distance models are shown as markers with $\times \ldots f(\text{DRR})$ and $+ \ldots f(1 - \text{IACC})$.

Interestingly, increasing the order $i = 1 \rightarrow (3, 5)$ decreases the distances for the median-plane orientation $\phi_{\bigcirc} = 0°$ and the specular lateral orientation $\phi_{\bigcirc} = 235°$. For these conditions, it is assumed that the reflected energy from the wall with $\Delta T \approx 10\,\text{ms}$ serves as increased direct energy in the distance cue, e.g., DRR. Similarly, for $\phi_{\bigcirc} = 90°$ increasing $i = 3 \rightarrow 5$ yields a decrease in the lateralization of auditory events. One reason might be that with a more narrow beam, the angular misalignment of its direction (compared to the specular direction) has a greater influence on the reflected energy. However, in our case the reflected energy for $i = 5$ is only 1 dB below the reflected energy for $i = 3$. Examining the individual responses for $i = 5$ reveals a bimodal distribution of individual angles, and source splitting could serve as alternative explanation, cf. left panel of Figure 5.7. The delay between direct sound and the strong lateral reflection $\Delta T = 21\,\text{ms}$ exceeds the individual echo threshold and it is therefore likely that the perception of a single auditory event breaks apart. Some listeners perceive the auditory event near the direction of the prominent reflection, whereas others report it near the direction of the directional sound source.

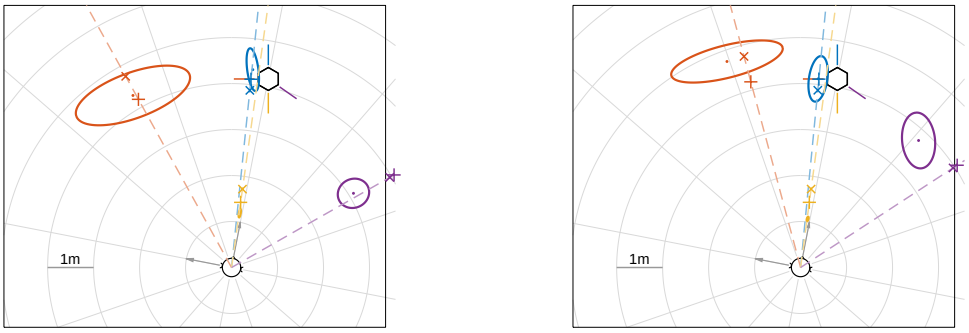
At position P_2 no clear classification of beam directions is possible, except for the specular orientation $\phi_{\bigcirc} = 115°$. Interestingly, confidence ellipses for $\phi_{\bigcirc} = (115°, 270°)$ with $i = 1$ are shifted along the distance, whereas increasing the order to $i = 3$ yields lateral shifts. An explanation therefore can be found in the histogram of individual responses, cf. right panel Figure 5.7. With $i = 1$ mainly the direction of the direct sound is heard. For the increased beam order $i = 3$ the answer distribution becomes bimodal and some listeners report the auditory image in the direction of the prominent reflection.

Influence of signal. The influence of the signal $S_{1/2}$ is examined in Figure 5.8 showing means and corresponding 95% confidence ellipses for both signals with order $i = 3$ at source position P_1. For orientations aligned to the median plane $\phi_{\bigcirc} = (0°, 180°)$ there is almost no influence, neither on the distance nor on the lateralization. Conversely, for lateral orientations $\phi_{\bigcirc} = (90°, 235°)$ the precedence effect becomes active, which is seen by the signal-dependent perception: auditory events of the more transient speech signal S_2 are less lateralized than those of the smooth-onset pink noise S_1.



(a) $i = 5$ at P_1 (b) $i = 3$ at P_2

Figure 5.7: Histogram of collected directions for signal S_1 from the listeners perspective (virtual source at $\phi = 0°$); for directivity order $i = 5$ at source position P_1 (left panel) and for directivity order $i = 3$ at source position P_2 (right panel). Dashed vertical lines indicate modeled directions $\phi_{\boldsymbol{r}_E}$.



(a) S_1 (b) S_2

Figure 5.8: 2D-means and corresponding 95% confidence ellipses for pink noise S_1 and speech S_2. Modeled lateralizations $\phi_{\boldsymbol{r}_E}$ are shown as dashed lines are based on the signal-dependent precedence weight $\beta = -0.35\,\mathrm{dB/ms}$ for S_1 and $\beta = -0.43\,\mathrm{dB/ms}$ for S_2; distance predictions are shown as markers with $\times \ldots f(\mathrm{DRR})$ and $+ \ldots f(1 - \mathrm{IACC})$.

5.2.3 Modeling the perception

Lateralization. The lateralization angle of auditory events can be predicted using the extended energy vector, cf. Eq. (2.14). It takes into account amplitude, delay, and direction information of the image-source model and incorporates the strength of the precedence effect by a weight β that considers the delays of the reflection paths with regard to the direct sound. Similar to Section 2.5, the signal-dependent weight β is found by minimizing the cost function J_{RMS} of Eq. (2.17). Input of the cost function are all collected directions of tested condition of a signal S_j except for conditions that exhibit source splitting (i.e., S_1 at P_1 with $i = 5$ and $\phi_{\bigcirc} = 90°$; S_1 at P_2 with $i = 3$ and $\phi_{\bigcirc} = (115°, 270°)$). Orientations of the calculated energy vector $\phi_{\boldsymbol{r}_E}$ are represented as dashed lines in Figures 5.5 – 5.8 with $\beta = -0.36\,\mathrm{dB/ms}$ for noise S_1 and $\beta = -0.43\,\mathrm{dB/ms}$ for speech S_2.

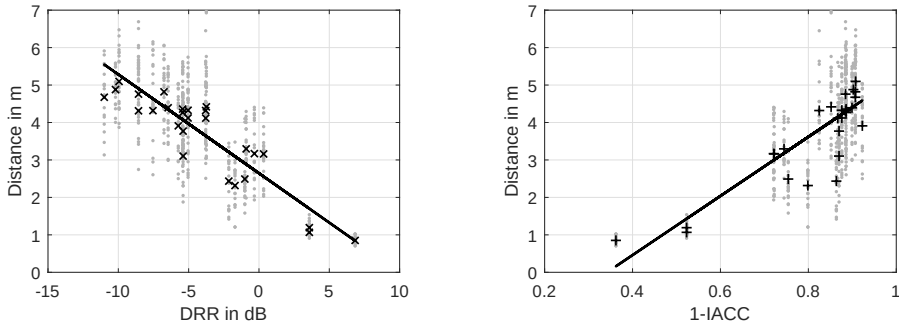


Figure 5.9: Distance models fitted to the experimental data (gray markers). The correlation to mean distances (black markers) is $R^2 = 0.82$ for the DRR (left panel) and $R^2 = 0.76$ for the 1−IACC.

For listening position P_1, the model achieves matching results for both signals and almost all vectors intersect the corresponding confidence ellipse. A major deviation is only seen for the disregarded source splitting-condition with $i = 5$. Nevertheless, the model predicts one of the two cluster points, cf. right panel of Figure 5.7. In contrast, the modeling of source position P_2 is not as accurate as it is for P_1. This is explained by competing direct sound and prominent reflection yielding bimodal answer distributions for some conditions, i.e., $i = 3$ with $\phi_{\bigcirc} = (115°, 270°)$. For others, however, the model fails, i.e., $i = 1$ with $\phi_{\bigcirc} = 115$ and $i = 3$ with $\phi_{\bigcirc} = 180°$.

Distance. The auditory distance is modeled with the metrics of the spatial sound field introduced in Section 3.2.3 and includes the direct-to-reverberant energy ratio (DRR) and the interaural cross correlation coefficient (IACC). To account for the increased reflected energy, the upper bound of integration is shifted for the DRR. Thus, in contrast to the traditional DRR, the time constant T of Eq. (3.2) does not only regard the direct sound but all sound instances up to the most prominent sound, which for some conditions is the prominent reflection of the targeted wall, i.e., $\phi_{\bigcirc,1} = (0°, 90°, 235°)$ with $i = (3, 5)$ and $\phi_{\bigcirc,2} = (0°, 115°)$ with $i = 3$. For the early IACC, no adaptation is necessary and the time window up to 80 ms, cf. Eq. (3.3), considers all important reflections.

Input to the linear regression analysis are all distances collected for speech and pink noise conditions. The left panel of Figure 5.9 compares the linear regression function $f(\mathrm{DRR}) = -0.35\,\mathrm{DRR} - 2.61$ with distances ratings. The correlation of the DRR to mean distances is high with $R^2 = 0.82$ and markers are within or (if the direction of $\boldsymbol{r}_\mathrm{E}$ does not hit) at the same distance as the corresponding confidence ellipse for most conditions. Note that if the integration bound of Eq. (3.2) is restricted to the direct sound only, the correlation of the DRR to the mean results drops to $R^2 = 0.34$.

The linear regression of the IACC yields $f(1 - \mathrm{IACC}) = 7.61(1 - \mathrm{IACC}) - 2.55$ with a lower correlation to mean distances $R^2 = 0.76$, cf. right panel of Figure 5.9.

Validation Generally, the listeners reported a plausibly sounding virtual acoustic scenery. Although head movements were not taken into account in the rendering, the auditory feedback

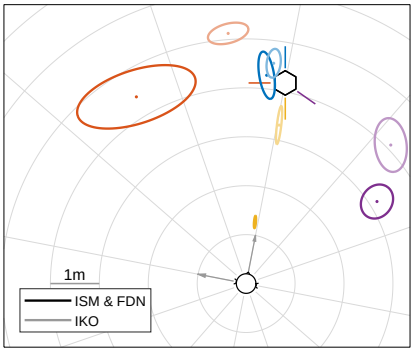


Figure 5.10: Comparison of the perceptions elicit by the sound field simulation (dark) with those elicit by the IKO in the IEM lecture room (light). The data is given as 2D means and 95% confidence ellipses for beam order $i = 3$ and pink noise signals.

of the listener-controlled omnidirectional sound source was convincing. Most listeners reported that they did not have to rely on the visual feedback of the surrounding room for the marker placement.

The setup location of listener and directional sound source for P_1 corresponds to the setup studied in Experiment 4 of Section 2.5 with a similar excitation signal (1.5-s-long pink noise burst linearly faded in and out by 0.5 s), what justifies a direct comparison of the results. Figure 5.10 compares the 95% confidence ellipses of the real sound source (IKO) with the results from the virtual acoustic environment with $i = 3$ and S_1. Although ellipses do not coincide for all beam orientations, the same tendencies can be observed. Different locations of confidence ellipses are due to the obvious acoustic differences, e.g., different excitation signals, directivity patterns, and walls' absorption coefficients, but can also arise from differences in the experimental procedure, i.e., drag and drop vs. aurally matching the location of an omnidirectional source. Moreover, in the simulation visual cues about the directional-source position were absent. It is assumed that the absence of a fixed physical position as a visual hint has successfully avoided the ventriloquism effect in the present response locations.

5.2.4 Discussion

Experiment 10 studied how directivity and orientation of a virtual directional sound source influences the localization of auditory events. The auralization of frequency-independent directivity patterns in a reverberant room utilizes a simplistic approach consisting of an image-source model for direct sound and early reflections and a feedback delay network for the diffuse reverberation. To evaluate the localization of the auditory events caused by the fixed directional source, the listeners had to aurally match a movable omnidirectional source in terms of auditory location in the same virtual acoustic environment.

For source position P_1 results from previous experiments with a real sound source were confirmed, and the source directivity shifts auditory events along the direction of the main energy. For lateral beam orientations distinct lateral shifts are observed. While lateralizations

tend to increase with the beam order, level and delay of reflections excited by higher-order directivities can exceed the individual echo threshold yielding auditory events to split. Moreover, the critical effect of onset is observed comparing lateralizations of more transient speech S_2 with smooth-onset pink noise S_1.

Orientations aligned to the median plane mainly affect the auditory distance, and for a listener-facing orientation $\phi_\bigcirc = 180°$, the increase of the directivity order yields a decrease of the auditory distance, whereas for directivities facing away from the listener $\phi_\bigcirc = 0°$, the auditory distance increases. For higher-order directivities the frontal wall reflection is strong enough to prevail the direct sound and dominates the auditory distance. This is not only seen for the median-plane orientation $\phi_\bigcirc = 0°$ but also for the lateral beam $\phi_\bigcirc = 235°$. At source position P_2 a similar tendency is observed. However, as tested orientations $\phi_\bigcirc$ can no longer be classified as median-plane or lateral beam for every beam, the dominant perceptual cue is either distance or lateralization, depending on the order of the beam-pattern.

Successful modeling of the experimental results was presented. For modeling the lateralization, the extended energy vector is used yielding the signal-dependent precedence weight $\beta = -0.36\,\mathrm{dB/ms}$ for the noise burst train S_1 with smooth onset time $t_\mathrm{on} = 250\,\mathrm{ms}$. This parametrization agrees with values found in previous experiments as it is above the weight of the transient noise bursts $\beta = -0.57\,\mathrm{dB/ms}$ but below the weight $\beta = -0.12\,\mathrm{dB/ms}$ for noise with $t_\mathrm{on} = 500\,\mathrm{ms}$, cf. Section 2.5.3. The weight for speech $\beta = -0.43\,\mathrm{dB/ms}$, however, is below the slopes determined from echo threshold curves of literature $\beta_\mathrm{ET} = -0.20\,\mathrm{dB/ms}$ to $-0.35\,\mathrm{dB/ms}$, cf. Section 2.5.2. The auditory distance is modeled by physical measures of the sound field introduced in Chapter 2, including the IACC and an adapted version of the DRR. Interestingly, in comparison to Experiment 4 of Section 2.5, for the DRR, the time constant that is typically chosen to isolate the direct sound had to be extended to the include most prominent reflections. The level difference between direct sound and prominent reflection exceeding the threshold of backward masking is considered as possible cause.

Finally, the results for the 3rd-order beam patterns are compared to the perceptions of the IKO studied in Section 2.5. Although the virtualization is simplistic, the same tendencies are observed by the experimental data indicating a plausible virtual acoustic environment.

5.3 The influence of individualization and training of BRIRs on the externalization

It is commonly known that the presence of reverberation increases the plausibility of binaural reproduction as it helps with externalization, e.g., [SGK76, DRP+92, BLW17]. Binaural spatialization systems like Oculus Spatializer[16], Resonance Audio[17], or Steam Audio[18] reverberate HRIRs by applying measured or simulated room impulse responses (RIR) in order to move the auditory image out of the head. However, if the synthesized room does not acoustically match the room the listener is sitting in, the externalization may remain poor. Research has shown that room impulse responses of a congruent amount of reverberation and direct-to-reverberant energy ratio that fit the auditory expectations of a listener with respect to the surrounding environment better is often decisive to get an externalized auditory image, e.g., [BWLA00, WKS16]. Similar to generic HRIRs in the anechoic case, the so-called *room divergence effect* yields auditory images that are perceived either close to the head or internalized.

An early study presented by Plenge [Ple72] introduces a conceptual model of the externalization process. It assumes that externalization includes top-down processing, where the resulting auditory impression depends on prior knowledge about the auditory event. It consists of two memory stages: The instrumental means of human localization ability, i.e., the HRIRs, are stored in the long-term memory and the short-term memory is filled with information on the sound source and the room characteristics, i.e., the RIR.

In contrast to the learning and adaptation process to new HRIRs, which is rather slow [KW13], the content in the short-term memory is volatile and adaptation to new RIRs happens each time one enters a new listening situation. According to Plenge's model, sound images are externalized if the information provided by ear signals complies with the information in both memories. If there is contradiction between the memory stages and what is received by the ears, the sound image is internalized and localized in the head.

Experiment 11 focuses on the relative contribution of individual HRIRs and the congruency of RIRs on externalization. Given that externalization is particularly fragile for the frontal direction, e.g., [CSD15, LSP19], and many important, everyday events involve, e.g., face-to-face conversations, the experiment considers a relatively near sound source without pronounced directivity, directly in front of the listener. Simulated acoustics provide a maximum of control of the condition. The listening experiment was conducted at DTU, and in comparison to what has been published in [WHM19], this section presents a revised statistical analysis and introduces a model to predict the experimental results. The experiment took place in the audiovisual immersion lab (AVIL), a 6 m × 7 m × 8 m large anechoic chamber equipped with a 64-channel spherical loudspeaker array at a distance of 2.4 m from the central listening position, cf. right panel of Figure 5.11. For a plausible experience, the room simulation employs the LoRa toolbox [FB10]

[16] https://developer.oculus.com/downloads/package/oculus-spatializer-unity/
[17] https://resonance-audio.github.io/resonance-audio/
[18] https://valvesoftware.github.io/steam-audio/

and ODEON room acoustic software [Nay93]. The latter is a commercial tool for room acoustical simulation based on a hybrid ray tracing approach for detecting early specular reflections and calculating late reverberation.

5.3.1 Individual HRIRs and congruent RIRs

Conditions studying the long-term memory consist of *individual* or *generic* HRIRs. For individual HRIRs, the loudspeaker sphere is used as playback device, whereas conditions with generic HRIRs are played back over open Sennheiser HD800 headphones. The headphone signals are created by convolving the loudspeaker signals with directionally consistent HRIRs from a Brüel & Kjær HATS. The left panel of Figure 5.11 shows the processing scheme for headphone and loudspeaker playback. As listeners wear the headphones throughout the experiment, the loudspeaker signals are somewhat distorted because the sound has to propagate through the open headphones' ear cups. This attenuation is considered in the headphone conditions by equipping the HATS with the same headphones during the HRIR measurement. Figure 5.12 shows the attenuation due to headphone insulation for different loudspeaker directions. To equalize the transfer function from the headphone transducers to the corresponding ear canal, a minimum-phase filter is applied to the headphone conditions.

Familiarizing with the room is essential for externalization, e.g., [Ple72, WKS16], and conditions with *congruent* or *divergent* RIRs are used to study the short-term memory. Congruent conditions are established by a familiarization phase in the room before evaluation, and for divergent conditions, the training room differs from the evaluation room. The seven-minutes-long training is done on loudspeakers simulating four talkers at different azimuth angles at a distance of 2.4 m around the listener. During training, listeners can freely move their heads. After training,

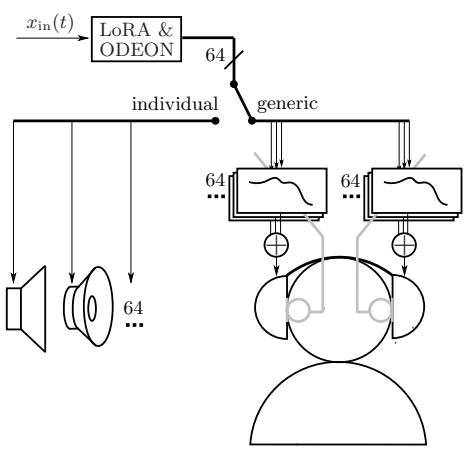


Figure 5.11: Processing scheme (left) and conducting the Experiment 11 in the AVIL (right). The room acoustics are simulated with LoRA and ODEON for the given 64 channel loudspeaker sphere. Individual HRIRs are tested using loudspeaker playback. Generic HRIRs are tested using open headphones. The headphones signal is created by convolving the loudspeaker signals with corresponding HRIR measurements of a dummy head with headphones on (grey).

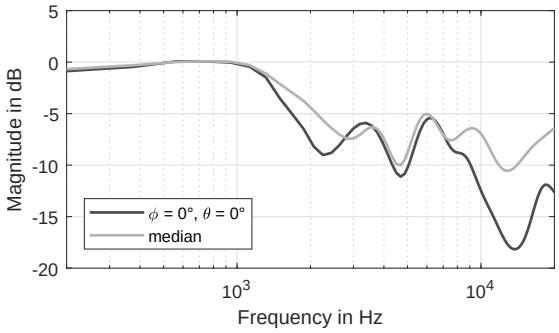


Figure 5.12: Measured attenuation of the external sound vs. frequency of the Sennheiser HD800 headphone for the frontal direction and the median attenuation over all 64 loudspeaker directions. Magnitudes are third-octave smoothed and normalized to 1 kHz.

listeners are instructed to face the loudspeaker directly in front of them and rate the degree of externalization of the condition under test. The stimulus consists of a 2.5 s-long unknown speech sample[19], simulated at the position of the frontal loudspeaker. Ratings are provided using a computer keyboard with a rating scale inspired by [HW96], extended with an additional rating for cases in which listeners are unable to decide on a direction:

0 The source is in my head.
1 The source is not well externalized. It is at my ear, or at my skull.
2 The source is externalized. It is between me and the loudspeaker.
3 The source is well externalized, compact, and located at the loudspeaker.
4 The source is externalized, but with no specific direction.

Conditions can be either with individual or generic HRIRs (loudspeaker or headphone playback) and with congruent or divergent RIRs (trained on congruent or divergent room). For the headphone playback, head-tracking is omitted because it would be another degree of freedom and for the evaluation listeners are requested to not move their heads. The position of the head is monitored using an OptiTrack system and playback of the condition under test starts automatically when the listener is looking at the frontal direction. Visual feedback of the heads position allows compensation of deviations from the frontal direction during the playback of the condition under test. Hiekkanen et al. [HMK09] reported that head rotations of 2.5° already produce audible differences in binaural transfer functions. However, informal listening of the author revealed no influence on externalization for such small rotations and deviations of the listeners head from the frontal direction of 5° are allowed. Whenever this value is exceeded, the playback (loudspeaker and headphone) is stopped and the condition has to be restarted again.

Three different room simulations $R_{0,1,2}$ (anechoic, dry, wet) are studied, cf. Table 5.1. All listeners start with the training on anechoic R_0. Conditions rated after the training are the congruent condition (simulated with R_0 and denoted as $R_{0/0}$) and a divergent condition simulated with R_1 (denoted as $R_{1/0}$). Both conditions are tested four times, with individual/generic HRIRs and with two repetitions. After each rating of a condition, the training on R_0 is continued for

[19] *Music for Archimedes*, CD Bang and Olufsen 101 (1992)

Table 5.1: Description and characteristics of the simulated rooms as used in Experiment 11.

room	reverberation	T_{60}	DRR	room size in m	simulation
R_0	anechoic	0.0 s	∞ dB		free field
R_1	dry	0.5 s	7.5 dB	$9.5 \times 7.6 \times 3.0$	classroom
R_2	wet	1.2 s	2.8 dB	$14.2 \times 9.0 \times 5.5$	auditorium

another 10 s before the next condition is tested. In this way the information in the short-term memory is maintained. Then the dry room R_1 is trained and conditions $R_{0/1}$, $R_{1/1}$, and $R_{2/1}$ are tested similarly. Lastly, the wet room R_2 is trained and conditions $R_{1/2}$ and $R_{2/2}$ are tested. This sequence of training is the same for all listeners, whereas the conditions presented after each training are an individual random permutation yielding $(2 + 3 + 2)$ conditions $\times 2$ playback devices $\times 2$ repetitions $= 28$ rating tasks for each listener.

Seven listeners participated in the experiment and except for the author, who participated in the experiment, none of them was experienced in terms of binaural perceptual experiments. At the beginning of the experiment, each listener was familiarized with the sensation of externalization and internalization by comparing an anechoic speech stimulus presented over the frontal loudspeaker against presenting the mono loudspeaker excitation signal over the headphone without binaural processing but phase-inversion for one of the channels.

5.3.2 Experimental results

The externalization index is used for evaluation. It indicates the ratio of conditions under which the source is perceived as "well externalized, compact, and located at the loudspeaker" (rate 3, counted as one) in relation to all other ratings (counted as zero).

Mean values of the externalization index responses are given in Figure 5.13 as mean values by filled symbols for loudspeaker and open symbols for the headphone playback, grouped according to the training room. The corresponding binomial 95% confidence intervals are constructed with the method after Clopper and Pearson [CP34]. Fisher's *exact test* for count data is used for all significance tests and the odds ratio as a measure for the effect size. The odds ratio (OR) is defined in the interval $[0, \infty]$ with a value OR $= 1$ meaning no effect.

As expected, the externalization is best for congruent rating/training conditions $R_{0/0}$, $R_{1/1}$, and $R_{2/2}$ with loudspeaker playback. The corresponding conditions with headphone playback on the other hand are rated lower. While the room itself is not a significant factor under congruent rating/training conditions using loudspeaker playback ($p > 0.05$, OR $= 1.8$), a pairwise statistical analysis reveals the room to be a significant factor when presented on headphones. In this case, condition $R_{1/1}$ has about 10 and condition $R_{2/2}$ about 7 times higher odds of appearing externalized than $R_{0/0}$ ($p \leq 0.05$). With headphones, increasing the reverberation increases externalization, and the significance of the playback device diminishes. While for $R_{0/0}$ the externalization using loudspeakers is significantly better than with headphones ($p \leq 0.05$, OR $= 13.0$), for $R_{1/1}$ and $R_{2/2}$ no significant difference is observed ($p > 0.05$) and the effect sizes

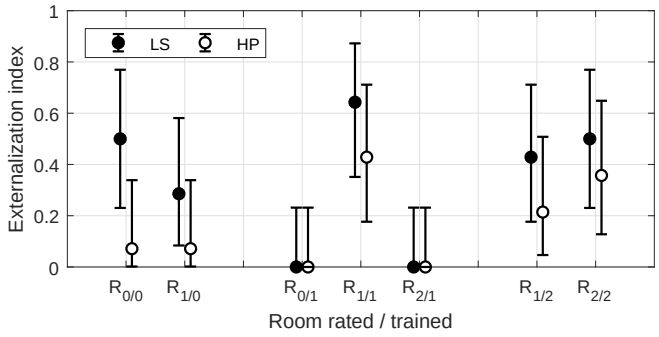


Figure 5.13: Means and binomial 95% confidence intervals of the externalization rating for different room simulations after training on congruent or divergent room. Playback employed loudspeakers (LS) or headphones (HP).

decrease to OR = 2.4 and OR = 1.8, respectively.

The influence of training with individual HRIRs (loudspeaker playback) gets apparent by comparing congruent and divergent conditions of the same room. Considering the ratings of room R_0, the externalization under the congruent rated/trained condition $R_{0/0}$ is rated significantly higher for presentation on loudspeakers (individual HRIRs) than the divergent condition $R_{0/1}$ ($p \leq 0.05$, OR $= \infty$). Conversely on headphones (generic HRIRs), training does not yield any improvement and externalization remains poor ($p > 0.05$) under the convergent $R_{0/0}$ and divergent $R_{0/1}$ condition. For ratings of the room R_2, the externalization of $R_{2/1}$ significantly improves after training ($R_{2/2}$) for both loudspeaker and headphone playback ($p \leq 0.05$, OR $= \infty$), with ratings generally higher compared to R_0.

Surprisingly, for room R_1 the training is not a significant factor ($p > 0.05$) for any playback method. Overall R_1 yields highest externalization indices for congruent $R_{1/1}$ and divergent $R_{1/0}$ and $R_{1/2}$ conditions compared to corresponding conditions of other rooms. However, effect sizes between congruent and divergent R_1 conditions are not negligible ($R_{1/1}$ vs. $R_{1/0}$: $\mathrm{OR_{LS}} = 4.5, \mathrm{OR_{HP}} = 9.8$, $R_{1/1}$ vs. $R_{1/2}$: $\mathrm{OR_{LS}} = 2.4, \mathrm{OR_{HP}} = 2.8$) and it is assumed that with a larger number of listeners training on R_1 would become significant.

5.3.3 Are individual pinna cues negligible under reverberant conditions?

Most of the individual variations in the HRIR result from the shape of the pinnae [ADTA01]. Reproducing these individual spectral details in the direct part of the BRIR is thought to be essential for externalization [HGD16, LSP19], especially in a static reproduction [ORS+20]. Conversely, Jiang et al. [JSZL20] recently found out that pinna filtering in the reverberant parts of a BRIR produces less externalized sound images. Assuming externalization to be a result of the evaluation of HRIR-related and RIR-related cues [Ple72], this finding implies that the consideration of individual pinnae becomes more dispensable within more reverberant environments.

Brinkmann et al. [BLW17] examined the influence of reverberation on authenticity of individual dynamic binaural simulations. The differences between real loudspeakers and individual binaural

synthesis is less detectable in reverberant spaces in comparison to anechoic conditions. Accordingly, the results of Experiment 11 have shown, that in a reverberant environment generic HRIR are sufficient for a plausible simulation.

Shinn-Cunningham et al. [SCKM05] studied the effects of reverberation on spatial cues by analyzing dummy-head BRIRs. Spatial acoustic cues are distorted due to the presence of reverberation. While it systematically reduces the ILD magnitude, the reverberant energy increases temporal fluctuations in both short-term ITD and ILD cues. This effect is thought to reduce accuracy of directional localization in reverberant environments. Contrastingly, externalization involving binaural cues improves with reverberation. Catic et al. [CSB+13] assume that the combined binaural cues that arise from the interaction of the head-related binaural cues with reverberant energy play a role in externalization. Accordingly, they could show that listeners are sensitive to changes in the shape of the ILD distributions, and manipulation of fluctuations of the ILDs resulted in a reduced degree of externalization.

To investigate how reverberation individually affects binaural cues, fluctuations of the ILD measured with a Brüel & Kjær HATS are compared against a Neumann KU100 dummy head. The method is the same as described in [CSB+13]. Measured HRIRs are convolved the simulated RIR of $R_{0...2}$, yielding two sets of BRIRs. Another convolution with the speech material from Experiment 11 results in binaural signals that are analyzed by a gammatone filterbank with a bandwidth of 1 equivalent rectangular bandwidth. The Hilbert envelopes of each filter output in the left and right ear signals are used to calculate the ILD at each time instant. These short-term ILDs are lowpass filtered at 500 Hz and collected from 1 s of speech to obtain ILD distributions. Figure 5.14 shows the measured ILD distributions determined by the two dummy head microphones for tested rooms $R_{0,1,2}$ (left, middle, and right column, respectively). The distributions are shown for a low-frequency channel at 0.3 kHz (top row), a mid-frequency channel at 2.1 kHz (middle row), and a high-frequency channel at 4.1 kHz (bottom row) with a resolution of the histograms of 1 dB.

For anechoic R_0 (left column) the ILD distribution is narrow for all three frequency channels, and fluctuations are mostly in the range of ±2 dB around their means. For the low frequency channel (upper panel) mean ILDs of both HRIRs are 0 dB. For higher frequencies, ILDs are highly individual as mean values of HATS and KU100 deviate differently from 0 dB. However, as reverberation is added, individual pinna cues become negligible and individual differences in the histogram diminish; ILDs are less likely to take on values close to their mean and become distributed over a larger ILD range because the reflected sounds reach the ear from different directions.

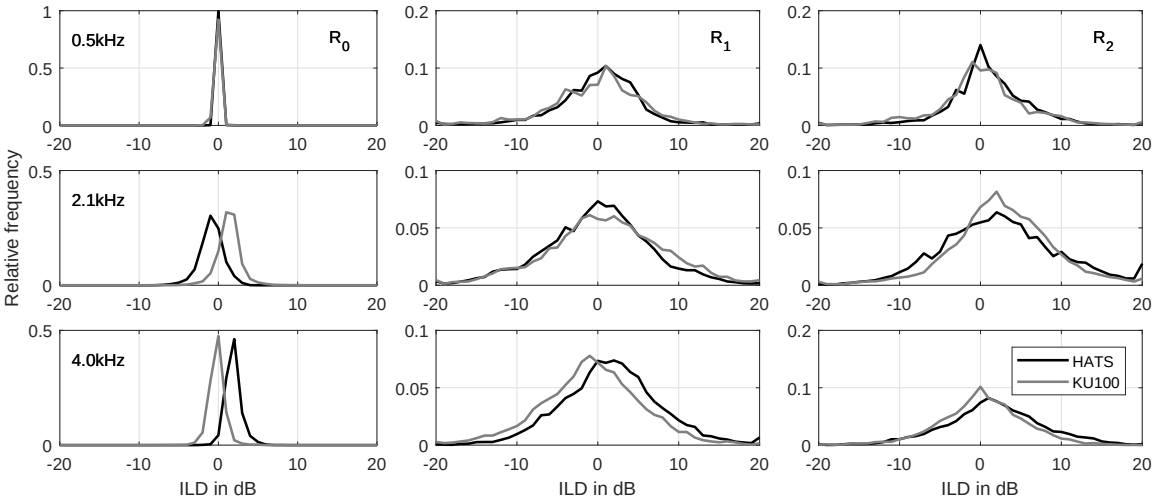


Figure 5.14: ILD distributions for speech, processed with BRIRs acquired on a Brüel & Kjær HATS (black) and a Neumann KU100 (grey). Columns show distributions for different rooms R, rows show different frequency channels with a bandwidth of 1 ERB.

5.3.4 Discussion

This study considered the influence of BRIRs on the externalization of speech. Based on a conceptual model of externalization, a listening experiment could be presented that examines the contributions of individual HRIRs and expectations on RIRs on the externalization. The influence of individual as opposed to generic HRIRs is studied by comparing loudspeaker playback (individual HRIRs) against headphone playback using generic HRIRs. The influence of listeners' expectations is studied by comparing ratings of the externalization that was trained (congruent) or diverged from the room that was trained.

The externalization index revealed that the advantage of individual HRIRs compared to generic HRIRs diminishes with increasing reverberation energy. This is independent of the expectation of the listeners and applies similarly for congruent and divergent conditions. Agreeing with the findings from Werner et al. [WKS16], training on congruent RIRs can improve externalization and corresponding conditions with individual HRIRs are rated highest. It remains unclear if the same applies for generic HRIRs, as the training was done only with individual HRIRs. However, in [WKS16] the influence of training was found to be less distinctive for generic HRIRs.

Interestingly, for the externalization in room R_1 training was not found to be significant and effect sizes calculated between congruent and divergent conditions are lower compared to the other rooms. As vision can affect externalization [UPG14], a possible explanation is that listeners imagined that if the anechoic chamber was more natural than anechoic, it would sound like R_1 due to its similarity in size. The quality of the acoustic simulations could serve as alternative explanation that would imply that the model of R_1 is more plausible than the simulation of R_2.

Brinkmann et al. [BLW17] found that reverberation increases the authenticity of individual binaural synthesis. Accordingly, Experiment 11 extended this finding to generic HRIRs and listening through a dummy head's ears resulted in a plausible perception for the dry room R_1

and the wet room R_2. Catic et al. [CSB+13] suggest that the ILD fluctuations resulting from the combination of head-related and room-related binaural information play an important role in the externalization. Accordingly, individually calculated short-term ILDs were used to demonstrate an effect of reverberation on perceived externalization: For the anechoic condition, short-term ILDs deviate significantly across individuals at mid- to high frequency channels. However, as reverberant energy is added, spatial acoustic cues are distorted yielding a reduced individuality of short-term ILDs.

5.4 BRIRs shortening strategies that maintain externalization

Results of the previous section showed that with reverberation, individual binaural cues are not necessarily needed, and a binaural synthesis of a reverberant room using generic HRIR yields convincing and externalized auditory images. Several studies identified the early reflection part to be essential to support externalization. For a virtual source directly in front of the listener, truncation of the reverberant tail to 20–80 ms has a critical effect on externalization, and increasing the reverberation above this threshold does not improve externalization any further [CEL14, CSD15, LSP18]. Li et el. [LSP19] studied the impact of reverberation on the externalization of lateral sounds and found differences between the two ears, suggesting that reverberation received at the contralateral ear has a greater effect on perceived externalization than reverberation received at the ipsilateral ear.

However, all these studies modified the BRIR by simply truncating its reverberant tail yielding unnatural conditions, because there is no natural equivalent to the cues of a truncated BRIR. To study which parameter related to reverberation is important for externalization, Experiment 12 systematically examines three BRIR shortening strategies and compares their effects. Furthermore, as reverberation is not always desired since the original sound of a recording should usually be preserved, the study examines the sound quality. This allows a rating of proposed strategies that ideally yield both an externalized but at the same time dry-sounding auditory impression. The experiments presented in this section are excerpts of a series of experiments that were conducted within Peter M. Giller's Master thesis and supervised by the author. The comprehensive results are published in [GWH19, Gil20].

5.4.1 The influence of different modifications of BRIRs on externalization

In the time domain, a BRIR can be written as:

$$h(t) = h_{\mathrm{dir}}(t) + h_{\mathrm{rev}}(t). \tag{5.2}$$

The direct part h_{dir} represents the HRIR, and the reverberant part h_{rev} includes early reflections and diffuse reverberation from all directions, convolved with the HRIRs for respective directions. Experiment 12 examines three different modifications applied to h_{rev} to shorten reverberation.

Strategy A truncates the length of the reverberant part by a multiplication with a rectangular window w with unit gain and target length L_i of the ith condition:

$$h_{\mathrm{rev},i} = h_{\mathrm{rev}}(t) \cdot w_{L_i}. \tag{5.3}$$

This method has been topic to many studies on externalization examining the influence of reverberation of congruent BRIRs, e.g., [LSP18].

Strategy B increases the DRR yielding a reduction (and hereby shortening) of the reverberant

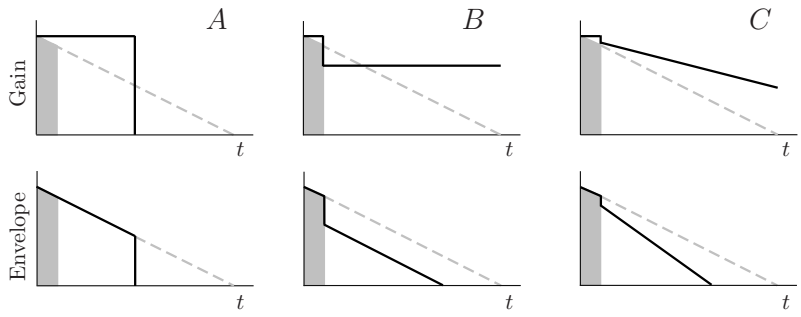


Figure 5.15: Schematic illustration of the modifications A, B, and C applied to the reverberant part h_{rev} of the BRIR manipulating its length, the DRR, or the reverberation time, respectively. Filled polygons indicate the direct part and dashed lines the reverberation envelope of the unmodified BRIR. In the upper row black solid lines show the equivalent time-dependent gain in dB, whereas in the lower row they show the resulting envelope of the modified BRIR in dB. Figure adapted from [GWH19].

part of the BRIR by attenuating it with a gain $g_{\mathrm{DRR}} = 10^{(\mathrm{DRR}-\mathrm{DRR}_i)/20}$:

$$h_{\mathrm{rev},i}(t) = h_{\mathrm{rev}}(t) \cdot g_{\mathrm{DRR}}, \tag{5.4}$$

where DRR is the direct-to-reverberant energy ratio estimated from the original BRIR, cf. Eq. (3.2), and the target DRR_i defined by the condition. Such a modification of the BRIR naturally arises if the source distance is reduced. Werner [WKS16] found that externalization slightly improves if a BRIR is adapted to the actual listening room by adjusting its DRR (i.e., reduction of the room divergence effect).

The third strategy C studied in the experiment decreases the reverberation time T_{30} by multiplying the reverberant part with an exponentially decaying function:

$$h_{\mathrm{rev},i}(t) = h_{\mathrm{rev}}(t) \cdot 10^{-\frac{60}{20}(T_{30}^{-1}-T_{30,i}^{-1})t} \tag{5.5}$$

where T_{30} is the reverberation time estimated from the original BRIR using Schroeder backward integration, cf. Eq. (5.1), and $T_{30,i}$ is the target reverberation time defined by the condition. This modification is similar to the ModRIR algorithm proposed by Pörschmann et al. in [PSA17] for a predictive auralization of a room with additional absorbers.

Figure 5.15 schematically illustrates impacts of the three modification strategies on the envelope of the reverberant part of the BRIR (lower row) by applying corresponding gains (upper row). Strategy A corresponds to a pure temporal modification, strategy B to an energetic modification, and strategy C can be interpreted as a mixed temporal/energetic modification.

Setup and conditions. In addition to *externalization*, listeners were asked to rate the *similarity* of the headphone condition compared to the loudspeaker condition in a separate task. The externalization task was carried out in the IEM lecture room, cf. Table 3.2, and each listener was sitting on a chair wearing open headphones AKG K702 for playback of the conditions. Additionally, a Neumann KH120 loudspeaker was placed 2.5 m directly in front of the listeners at ear height and served as externalized reference condition. This distance results in a slightly

Table 5.2: Conditions defined by levels i created by strategies A, B, and C modifying the parameters window length L, direct-to-reverberant energy ratio DRR, and reverberation time T_{30}, respectively

Strategy		A	B	C
Parameter		L	DRR	T_{30}
Level i	0	1.000 s	2.3 dB	0.70 s
	1	0.350 s	5.3 dB	0.60 s
	2	0.193 s	8.3 dB	0.51 s
	3	0.106 s	11.3 dB	0.42 s
	4	0.059 s	14.3 dB	0.33 s
	5	0.032 s	17.3 dB	0.23 s
	6	0.018 s	20.3 dB	0.14 s
	7	0.009 s	23.3 dB	0.05 s
	8	0.003 s	∞ dB	0.00 s

emphasized direct sound field. The similarity task was conducted in the IEM anechoic laboratory with the same setup of listener and reference loudspeaker.

The aim of the experiment is to measure relative differences of externalization with modified BRIRs to an original, full-length BRIR rather than the absolute differences to a physical sound source. Measurements of the studied BRIR were taken in the IEM lecture room with a Neumann KU100 dummy head congruent to the listening position. To account for reflections from breast and shoulders, the dummy head was combined with the torso of a Brüel & Kjær HATS.

Table 5.2 lists the condition under investigation, and their parameters that differently modify the reverberation h_{rev}. The parameter values were defined by the authors of [GWH19] in an informal listening session to fill out the range spanned by the full-length BRIR ($i = 0$) and the HRIR ($i = 8$). While T_{30} and DRR are varied linearly, the window length L follows the function $L_{i+1} = 0.55 \cdot L_i$ starting from $L_1 = 0.35\,\text{s}$.

For creating the binaural signal, the different versions of the BRIR are convolved with anechoic male speech[20]. Open headphones facilitate the comparative rating against the reference loudspeaker condition and are worn throughout the whole task. Obviously, the headphone alters the sound from the loudspeaker reference as the sound has to propagate through the ear cups. Attenuation of frontal sound at $\phi = 0°$ is reduced by replacing the ear cushions of an AKG K702 headphone with two pieces of Basotect on a sheet aluminum plate, cf. Figure 5.16. To account for the modified attenuation of the ear cups, far-field responses from the loudspeaker to the dummy head, both with and without the modified headphone, are measured in the IEM's anechoic chamber. The magnitude spectra are smoothed within critical bands and a minimum-phase equalizer is determined by dividing the with-headphones magnitude frequency response by the without-headphones response. Moreover, like in Experiment 11 an additional minimum-phase filter equalizes the magnitude transfer function from each headphone driver to the corresponding ear canal.

[20] *Music for Archimedes*, CD Bang and Olufsen 101 (1992)

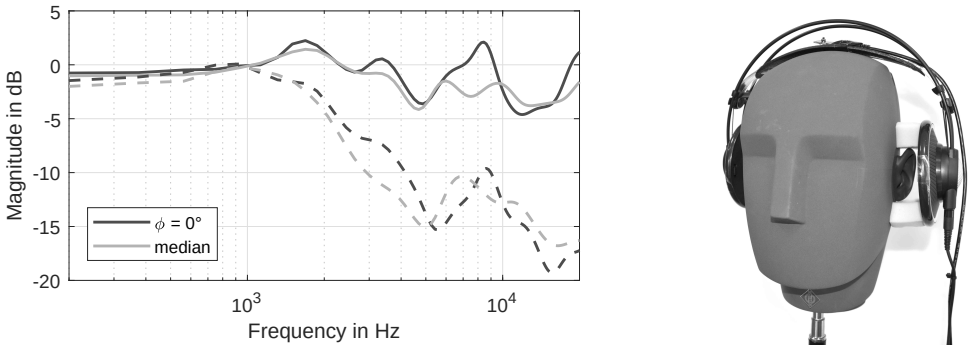


Figure 5.16: Left: Measured attenuation of the external sound over frequency of the modified (solid) and original (dashed) AKG K702 headphone for the frontal direction $\phi = 0°$ and the median over 64 sampled directions on a sphere. Magnitudes are third-octave smoothed and normalized to 1 kHz. Right: AKG K702 with modified cushions on the Neumannn KU100. Instructions to replicate this DIY solution can be found in [MKRB$^+$20].

The listening test is carried out as a MUSHRA-like test [ITU01], where listeners comparatively rate the nine conditions $i = 0 \ldots 8$ of a modification as set against the reference loudspeaker condition using a continuous slider for each condition. Each condition is repeated at will and played back in loop. An LCD screen is used as visual interface and placed directly above the loudspeaker in front of the listener. As for the externalization task and the similarity task loudspeaker playback serves as reference, the (hidden) binaural reference and anchor varies between the tasks. In the externalization task, conducted in the reverberant lecture room, the unmodified BRIR ($i = 0$) represents the binaural reference and the HRIR ($i = 8$) represents the anchor, whereas in the similarity task, conducted in the anechoic laboratory binaural reference and anchor are reversed. Rating scales are in the interval $[0, 1]$ with lowest values labeled "inside the head" or "very different" and highest values labeled with "at the position of the loudspeaker" or "identical" for externalization and similarity, respectively. Additionally, for the externalization task a label "close to the head" at 0.33 is added.

An additional 11-stimulus comparison set is included in both the externalization and the similarity task. It consists of conditions $i = (2, 4, 6)$ of each modification A, B, and C and the respective anchor and binaural reference. By individually supplementing responses $i = (0, 2, 4, 6, 8)$ from the cross-comparison set with corresponding linearly re-mapped responses from the modification set $i = (1, 3, 5, 7)$ using Eq. (3.1), the ranges between levels $(0, 2)$, $(2, 4)$, $(4, 6)$, and $(6, 8)$ are filled out. This enables a fine-grained quantitative comparison of modifications techniques.

Twenty-one experienced listeners participated in the experiment. Ten of it started with rating the similarity, the rest with externalization. As the binaural synthesis did not consider dynamic cues, listeners were asked to keep their heads immobile when performing the externalization task.

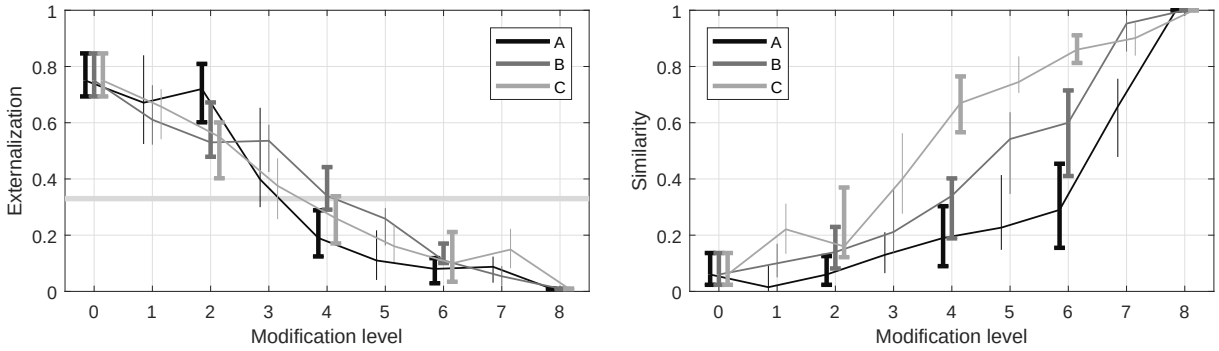


Figure 5.17: Medians and corresponding 95% confidence intervals of externalization ratings (left) and similarity ratings (right) for strategies A, B, and C. Individual responses for modification levels $= 1, 3, 5, 7$ are linearly re-mapped (thin lines) to fill out the ranges between modification levels $0, 2, 4, 6, 8$ (bold lines). Externalized conditions exceed the "close to head"-threshold at 0.33.

5.4.2 Experimental results

A Lilliefors test rejects the null-hypothesis of normal distribution for 6 out of $3 \times 9 + 11 = 38$ externalization conditions ($p \leq 0.05$). For similarity, the percentage of conditions that do not follow normal distribution is tripled. Therefore, Figure 5.17 shows medians with corresponding 95% confidence intervals for the externalization and similarity ratings for each shortening strategies A, B, C. Ratings of conditions with levels $i = 0, 2, 4, 6, 8$ are directly obtained with the comparison set, all others come from the three modification sets and are linearly re-mapped.

The quantitative comparison of ratings of the different modifications is not feasible, due to the different nature of the respective varied physical parameter. Hence, statistical analysis considers only the ratings from the cross comparison set $i = 0, 2, 4, 6, 8$ performing paired comparisons within the modification using the nonparametric Wilcoxon signed-rank test and Bonferroni-Holm post hoc analysis. As expected, the unmodified BRIR yields highest externalization ratings. However, only 3 of the 21 listeners rated the corresponding auditory image to be at the position of the loudspeaker. This might be due to the differences between individual and generic HRIRs. Another reason might be the fact that the physical loudspeaker was available for comparison (authenticity) and therefore the setup was particularly sensitive. Remaining timbral differences were still audible, partly because the filter considering the attenuation of the ear cups regards only the frontal direction. For all strategies, increasing the modification level results in a monotonic decrease of externalization. Modifications with levels $i \leq 3$ are perceived beyond the externalization threshold 0.33 ("close to the head"), whereas all others yield an internalized percept with the HRIR ($i = 8$) at the bottom end of the BRIR length. Paired sample tests of levels $i = 0, 2, 4, 6, 8$ within one shortening strategy reveal that differences of the condition to be significant ($p \leq 0.05$), except for neighboring conditions $i = 0, 2$ of strategy A. Thus, for the truncation strategy, findings from literature are confirmed and externalization remains unaffected until it drops, if the BRIR length is below a certain threshold, i.e., $L = 193\,\text{ms}$, cf. Table 5.2.

For the results from the similarity task in the anechoic chamber the relation is reversed; suppressing the reverberation in the speech by increasing the modification level increases its similarity, and the HRIR condition ($i = 8$) was rated to be identical to the anechoic reference

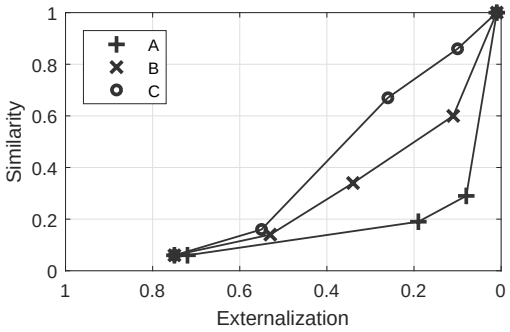


Figure 5.18: Median externalization ratings over corresponding similarity ratings form the cross-comparison set including modifications levels $0, 2, 4, 6, 8$. The horizontal line indicates the externalization threshold.

loudspeaker by the majority of listeners (18/21). While the statistics reveal similar performance of the three strategies regarding externalization, there are major differences when it comes to the similarity task. Still, for strategies B and C, all conditions are significantly different ($p \leq 0.05$), except for neighboring conditions $i = 0, 2$ of strategy B. Conversely, for strategy A, a significantly improved similarity is only reached with the HRIR ($i = 8$) in comparison to its modification levels, and between the pairs $i = 0, 6$ and $i = 2, 6$. However, a direct comparison of ratings of the different modifications is not feasible. Therefore the strategies are compared by plotting median externalization ratings of levels $i = 0, 2, 4, 6, 8$ over corresponding median similarity ratings, cf. Figure 5.18. This analysis allows a general rating of the strategies and reveals that decreasing the reverberation time utilizing strategy C yields the best trade-off of between externalization and similarity.

5.4.3 Modeling the externalization by physical measures of reverberation

In this section the responses to the different conditions are related to physical measures of reverberation. This allows to examine which reverberation parameter affects the externalization by which extent. Figure 5.19 shows median externalization and similarity ratings for modification levels $i = 0, 2, 4, 6, 8$ over the DRR (left), the temporal energy centroid (middle), and the time it takes for the reverberant energy to drop by 30 dB (hereafter referred as 30 dB decay time; right). These measures are calculated by weighting the impulse responses of respective conditions with weights accounting for the signal spectrum of the anechoic speech signal and the relative loudness (A-weighting). The mapping of the externalization curves for the strategies B and C determined by plotting them with regard to the DRR as x-axis yield values below ratings (medians) of the strategy A. Disregarding the HRIR, a linear regression of medians shows a coefficient of determination of only $R^2 = 0.32$. The temporal energy centroid as x-axis increases the goodness of fit of all condition medians to $R^2 = 0.87$. However, relating the conditions to the 30 dB decay time yields the best prediction of externalization. Medians of each strategy describe an almost perfectly straight line yielding $R^2 = 0.95$, and similarity-rating curves resemble the respective curves in Figure 5.17. Thus, the 30 dB decay time appears to be a suitable measure to predict externalization when listening to shortened-BRIR rendering in an otherwise congruent

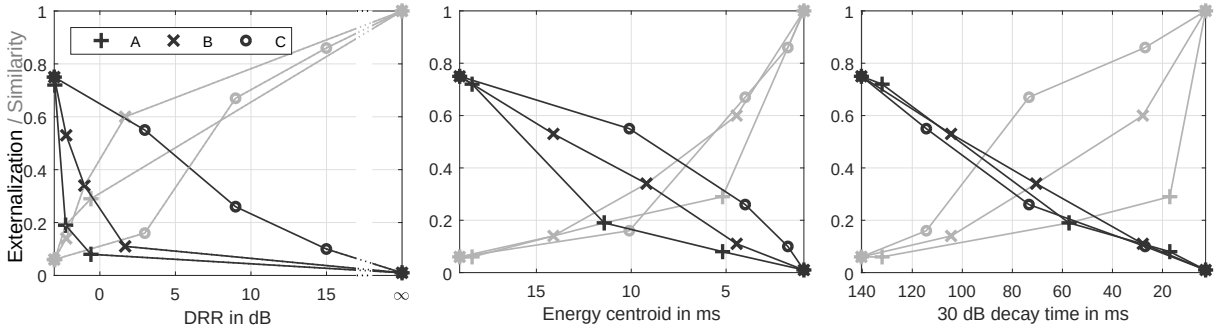


Figure 5.19: Medians of externalization (black) and similarity (gray) for modification levels $0, 2, 4, 6, 8$ over the DRR (left), the temporal energy centroid (middle), and the time it takes for the reverberant energy to drop by 30 dB (right) for the strategies A, B, and C.

listening environment for $i = 0$. It reveals similar differences between the strategies as the direct relation to dry-HRIR similarity, cf. Figure 5.18, and decreasing the reverberation time (strategy C) performs best with regard to similarity to an anechoic environment, followed by increasing the DRR (strategy B) and truncating the reverberant part (strategy A).

5.4.4 Discussion

This section studied the influence of reverberation parameters on the externalization of sound. Three different BRIR shortening strategies, manipulating either A the impulse response length, B the DRR, or C the reverberation time of a BRIR were examined in a listening experiment. Additionally to externalization, listeners were asked to rate the similarity of modifications to the dry excitation signal. In doing so, the dilemma of creating an externalized dry-sounding signals is concerned. As expected, none of the shortening strategies provides well-externalized conditions that, at the same time, yield satisfactory ratings of similarity to the anechoic HRIR.

As each strategy varies another physical parameter, no direct comparison of externalizations of different strategies is feasible, but an indirect comparison by relating the externalization responses to respective ratings for similarity. This reveals that strategy C, decreasing the reverberation time by multiplying the reverberant part with an exponential decay function, yields the best trade-off between externalization and similarity. Contrastingly, the well studied truncation strategy A is rated worst. Informal listening tests lead to the impression that truncated BRIR of medium and short lengths stand out from the other strategies due to timbral colorations, most likely caused by comb filtering of the interference between early reflections and the direct sound. However, for strategy A, findings from literature were confirmed and externalization remains unaffected until a certain threshold length $L \approx 200\,\text{ms}$ of the BRIR is reached, below which the externalization collapses. Based on this value, it can be assumed that with generic HRIRs the overall BRIR length L has to be longer than with individual HRIRs, where literature suggests values in the range of $L = 80 - 100\,\text{ms}$, [CEL14, CSD15, LSP18].

An alternative comparison between the strategies is drawn by relating them to a common physical measure. The relationship between modification level and a decrease in externalization

is modeled, and it complies with the increase to the dry-HRIR similarity. In particular, the energy decay time defined as time it takes for the reverberant energy to drop by 30 dB is found to correlate well with externalization of the experimental results.

5.5 Summary

In this chapter, investigations were carried out into the plausibility and authenticity of binaural reproduction. The first section presented a simplistic method for a virtual acoustic environment. It consists of an image-source model and a feedback delay network which provides a real-time control of the parameters. Experiment 10 examined an acoustic auralization of a directional sound source to describe and compare its auditory properties to known real-world experiments. Findings from previous experiments with the IKO are confirmed and the proposed method is capable to create a plausible binaural reproduction. Strongly focused sound beams shift the auditory event away from the physical loudspeaker position either directly or via a wall reflection. Results from Chapter 3 are confirmed and median-plane beams of a frontal source mainly affect distance perception. Increasing the order of such a listener-facing beam reduces the auditory distance, whereas for beams in the opposite direction the distance is increased. However, if the level of the reflection of the adjacent wall is strong enough, it can prevail the direct sound, and the effect is reverse so that an increase of the beam order decreases the auditory distance when the beam is facing away from the listener. Successful modeling of distance is achieved by an adapted DRR measure. To account for the increased reflected energy, the temporal integration limit of the direct energy is shifted and the adapted DRR describes the ratio of the energy up to the most prominent sound instance of the BRIR to the energy in the remainder of the BRIR.

Lateral beams yield distinct lateral shifts. Knowledge from the precedence effect is applicable and a critical influence of signal onset is observed reducing lateralizations of transient sounds. If delay and level of the targeted wall reflection exceed the echo threshold, auditory events were moreover found to split. The extended energy vector provides a basis to predict directional localization. Precedence weights of noise bursts agreed with weights used in previous modeling approaches, whereas the weight for transient speech is in accordance with values from the available literature on the echo threshold. If directivity directions are not distinctly lateral or aligned to the median plane but something in between, direct sound and prominent reflection compete. For these conditions the energy vector sometimes fails, and distance and lateralization curves of the means are not monotonic with the beam order. This is because source splitting is more likely to occur and the representation of the experimental data with mean values is insufficient.

The subsequent sections focused on externalization of frontal sound in the virtual acoustic environment. Experiment 11 is based on a conceptual model of externalization and evaluates the influence of the listener's expectation on RIR and individualization of HRIR using simulated acoustics of an anechoic, a dry, and a wet room. The contribution of expectation on externalization is tested by comparing training on a congruent or divergent room; the influence of individualization is studied by comparing loudspeaker playback against binaural reproduction using dummy head recordings. The results of the listening test showed that in anechoic environments, both individual HRIRs and training on the RIR are essential for externalization. The addition of reverberation was shwon to reduce individual spatial cues, and externalization involving generic binaural

cues improves in the dry and wet room. For these conditions, no significant difference between externalization measured with loudspeaker and headphones is observed. To investigate how reverberation individually affects binaural cues, fluctuations of the ILD are measured with two different dummy-head microphones. The analysis revealed highly individual ILDs for mid to high frequency components. However, this individuality gets lost as reverberation reduces the ILD magnitude and increases temporal fluctuations in short-term ILDs.

Experiment 12 examined the influence of reverberation parameters on externalization. Three BRIR-shortening strategies were studied that either manipulate the impulse response length, the DRR, or the reverberation time of a generic BRIR. A listening experiment evaluated the degradation of different modification levels on the externalization together with an improvement of the BRIR's similarity to the dry HRIR. While all three manipulation strategies increase the similarity of the BRIR to the dry HRIR, their degradation in externalization differs, and the decrease in reverberation time of the BRIR yields the best trade-off between externalization and preserving the original sound of the HRIR. The time it takes for the reverberant energy to drop by 30 dB was shown to be applicable as a model to commonly describe the deterioration in externalization for all of the three BRIR-shortening techniques tested.

6 Conclusion

This thesis investigated the perception of auditory events spatialized by directional sound sources in reverberant rooms. The first chapters considered how the interaction of direct sound and reflections affect the localization of auditory events in three dimensions, i.e., lateralization (Chapter 2), distance (Chapter 3), and height (Chapter 4). The localization of auditory events was evaluated by a series of listening experiments. The bottom-up approach started with evaluating the perception of a few precisely controlled sound instances played back over loudspeakers or headphones in an anechoic laboratory. Based on the generic results, simple predictive measures were introduced. Building on this, the perception of a real directional sound source in a reverberant environment was studied and insights of the laboratory were used to explain the perception of auditory events yielding more complex perceptual models. Finally, in Chapter 5 knowledge of the previous chapters was brought together for creating a virtual acoustic environment of a directional sound source for plausible binaural reproduction with headphones.

Lateralization

The precedence effect resolves the complex array of stimuli by weighting corresponding localization cues in the lateralization process. Starting from a very basic constellation, first it was determined how strongly delay, level, and direction of a single reflection relative to the direct sound affect the precedence effect strength by measuring the echo threshold. Indications were found that the weak but significant influence of direction is a rather basic result of the directivity of the human ear. The modeling of the precedence effect considered the more prominent parameters level and delay. Their trade off is described by the shape of echo threshold curves and modeling was achieved by a linear approximation yielding a signal-dependent parameter for the echo threshold slope.

Subsequently, the influence of acoustic variations in the reflection type was examined regarding the precedence strength. A model for diffuse reflections was introduced yielding a temporal and directional diffusion of the reflected sound. The listening experiment showed that the temporal diffusion weakens the precedence effect as the perceptive center of the reflection, the temporal energy centroid, emerges later in time. Similar to the directional separation of direct sound and reflection, the directional diffusion of the reflection was found to have a minor influence on precedence.

The lateralization of auditory events of more complex reflection patterns was considered in

another study. Different types of the precedence effect, known from two-source experiments, were examined using multiple reflections with increased reflection levels. It was shown that localization cues of a leading brief transient sound or the cues of the transient onset of an ongoing sound dominate lateralization.

Lastly, lateralizations of auditory events created by a directional sound source in a room were studied. The experiment employed the icosahedral loudspeaker array (IKO), a compact spherical loudspeaker array which allows Ambisonics beamforming, and is able to project strongly focused sound beams in any direction. It was found that with pronounced beams steering to a reflective lateral wall, the auditory event is lateralized away from the physical position of the source. Agreeing with knowledge obtained in the acoustically controlled, anechoic environment, this displacement shows a critical effect of signal onset. Smooth-onset signals elicit no or only weak precedence and were fully lateralized to the reflective wall. For transient signals on the other hand, the direct sound dominated localization to a stronger amount and fused percepts at the vicinity of the physical sound source were heard. Successful modeling of lateralization was achieved by a customized version of the extended energy vector. In addition to the attenuation due to the propagation of sound, the model considers the precedence effect by converting the delays of the reflected sounds into attenuations based on the corresponding signal-dependent echo threshold slope. Further simplification was achieved by approximating the energy vector's input parameters, which are gain, delay, and angle of sounds reaching the listener, using a simple image-source model. Similarly, for the perception of a dynamic directivity, knowledge from laboratory was applicable and the less-lateralized transient sounds implied a stronger precedence effect than sounds with a more smooth onset. Additionally, the experiments showed that the perception of a time-variant sound field involves additional higher order perceptual effects that are known from the perception of physically moving sound sources.

Distance

The perception of a dynamic beams of the IKO showed not only variations in lateralization but also in distance. This is because the directivity of a sound source influences the direct-to-reverberant energy ratio, which is known to be the most relevant distance cue in reverberant environments. Chapter 3 focused on how the directivity can be used to physically control this cue from a single beamforming source in the room. Starting from an acoustically controlled environment, the perception of different higher-order beampattern designs/beam constellations was studied using a 2D loudspeaker-based simulation. The experiment confirmed the option of a pronounced and graduated distance impression through beampattern control. Like the perception of physical source-listener distances, distance perceptions elicited by the directivity of a sound source resemble a compressive power function. Successful modeling of the distance perception was achieved by both well-known distance models, such the obvious monaural energy measure DRR, but also by models initially developed to predict the apparent source width, such as the spatial measure IACC.

To confirm the findings from the laboratory, the IKO was employed in a reverberant room and the auditory distance and apparent source width was studied. Although the study contained the physical IKO as visual bias, knowledge from the acoustically controlled environment was confirmed and the same directivities yielded distinct distance impressions. Moreover, it was proven that the apparent source width and the auditory distance highly correlate for the directional sources, which explains the successful modeling of both DRR and IACC.

Height

In contrast to lateralization, only a few studies in the available literature examine asynchrony effects in the perception of height. None of them could prove a prominent perceptual effect on the apparent height for sound instances presented from different elevations. Chapter 4 focused on potential asynchrony effects that might be triggered with beamformed sounds. The hypothesis of the approach was that median plane reflections carry additional cues for vertical localization.

The first experiment studied if delay and level of a simulated floor reflection with a fixed direction can affect the apparent elevation of the direct sound in the anechoic laboratory. Results showed a significant but rather weak effect for speech signals; increasing the delay and at the same time decreasing the level of the reflection yielded a more elevated auditory event, whereas altering the parameters in the other direction yielded less elevated events. This in accordance with the physical movement of a sound source.

Subsequent experiments focused on the delay of median plane reflections and studied the effect of continuous alterations. In accordance with the previous experiment, it was found that continuous alterations of the delay induce a movement of the auditory event: Increasing the delay of a floor reflection yielded an upwards movement, while a decrease of the delay was perceived as downwards movement. The effect direction was reversed if instead of a floor reflection, the attenuated and delayed sound was presented from the ceiling. Movements were most distinct for very brief overall delays up to 1 ms with alterations in the range of ± 1 ms. This finding, however, is not in agreement with the hypothesis of a physically motivated effect, as neither floor/ceiling reflections nor reflection from the body meet both demands at the same time. Moreover, the effect was observed only for spectro-temporally sparse signals such a speech. Altering the reflection's delay of broadband signals yielded the pitch-height effect, as the pronounced time-variant comb filters varied the spectral content of the signals at the ears. Overall, none of the asynchrony effects was as prominent as the effects examined in previous chapters, which is why studies on the apparent height did not go beyond the acoustically controlled, anechoic environment.

Virtual Acoustic Space

The last chapter approached the headphone-based presentation of directional sound sources in a virtual room. The first study focused on RIR simulations and introduced a real-time capable

method consisting of an image-source model for simulating direct sound and early reflections, and a feedback delay network for providing diffuse reverberation. The plausibility of the method was examined in a listening experiment that studied lateralization and distance of auditory events created by a simulated directional sound source. The localization within the simulation was in accordance with the findings obtained in the physical sound field, and lateralization and distance were modeled by the measures introduced in the previous chapters. However, the displacements of auditory events were more pronounced in the simulation and it is assumed that the absence of the physical loudspeaker array as a visual hint avoided the ventriloquism effect.

Subsequently, it was examined how the listener's expectation on the RIR and individual HRIR affect the plausibility of a virtual environment. For this purpose, the externalization of sound sources simulated in known or unknown environments with individual or generic HRIRs was measured. The results showed that in an anechoic environment both training on the room and individual HRIRs are essential for plausible reproduction. However, the benefit of individual HRIRs diminishes if the room provides reverberation. An analysis of the short-term interaural level differences showed that in an anechoic room the distribution of fluctuations is individual. This individuality in the spatial cues diminishes if reverberation is added.

Lastly, a study examined how physical parameters related to the room reverberation influence the externalization. Three different BRIR-shortening strategies were studied that truncate the BRIR, increase its DRR, or decreasing its T_{60}. Although all modifications yield a degradation of externalization together with an improvement of the BRIR's similarity to the dry HRIR, the trade-off relations of the methods are different. The decrease in reverberation time of the BRIR yields both an acceptable externalized and at the same time a relatively dry-sounding auditory impression.

Bibliography

[AB79] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-room Acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[ADTA01] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 99–102.

[Aga11] M. Y. Agaeva, "The precedence effect in the horizontal and vertical planes in experiments with a moving lagging signal," *Human Physiology*, vol. 37, no. 5, pp. 545–549, 2011.

[AZ14] P. W. Anderson and P. Zahorik, "Auditory/visual distance estimation: Accuracy and variability," *Frontiers in Psychology*, vol. 5, no. SEP, pp. 1–11, 2014.

[BAA+19] F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, and S. Weinzierl, "A round robin on room acoustical simulation and auralization," *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. 2746–2760, 2019.

[Bar71] M. Barron, "The subjective effects of first reflections in concert halls-The need for lateral reflections," *Journal of Sound and Vibration*, vol. 15, no. 4, pp. 475–494, 1971.

[BBA13] W. O. Brimijoin, A. W. Boyd, and M. A. Akeroyd, "The contribution of head movement to the externalization and internalization of sounds," *PLoS ONE*, vol. 8, no. 12, 2013.

[BC78] J. Blauert and W. Cobben, "Some consideration of binaural cross correlation analysis," *Acustica*, vol. 39, no. 2, pp. 96–104, 1978.

[BD88] J. Blauert and P. Divenyi, "Spectral Selectivity in Binaural Contralateral Inhibition," *Acta Acustica united with Acustica*, vol. 66, no. 5, pp. 267–274, 1988.

[Bec98] S. Bech, "Spatial aspects of reproduced sound in small rooms," *Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 434–445, 1998.

[Ber88] A. J. Berkhout, "Holographic approach to acoustic control," *AES: Journal of the Audio Engineering Society*, vol. 36, no. 12, pp. 977–995, 1988.

[Ber13] B. Bernschütz, "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100," *Fortschritte der Akustik – AIA-DAGA 2013*, pp. 592—-595, 2013.

[BH99] A. W. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," *Nature*, vol. 397, no. 6719, pp. 517–520, 1999.

[BHBS+17] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, and B. Rafaely, "Spectral equalization in binaural signals represented by order-truncated spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4087–4096, 2017.

[Bio68] M. A. Biot, "Generalized boundary conditions for multiple scatter in acoustic reflection," *Journal of the Acoustical Society of America*, vol. 44, pp. 1616–1622, 1968.

[Bla71] J. Blauert, "Localization and the Law of the First Wavefront in the Median Plane," *The Journal of the Acoustical Society of America*, vol. 50, no. 2B, pp. 466–470, 1971.

[Bla97] ——, *Spatial Hearing - The Psychophysics of Human Sound Localization.* The MIT Press, 1997.

[Blu31] A. Blumlein, "Improvements in and relating to Sound-transmission, Sound-recording and Sound-reproducing Systems," 1931.

[BLW17] F. Brinkmann, A. Lindau, and S. Weinzierl, "On the authenticity of individual dynamic binaural synthesis," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 1784–1795, 2017.

[BM81] M. Barron and A. H. Marshall, "Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure," *Journal of Sound and Vibration*, vol. 77, no. 2, pp. 211–232, 1981.

[BPWJ77] C. D. Bremer, J. B. Pittenger, R. Warren, and J. J. Jenkins, "An Illusion of Auditory Saltation Similar to the Cutaneous "Rabbit"," *The American Journal of Psychology*, vol. 90, no. 4, pp. 645–654, 1977.

[BST15] A. D. Brown, G. C. Stecker, and D. J. Tollin, "The Precedence Effect in Sound Localization," *Journal of the Association for Research in Otolaryngology*, vol. 16, no. 1, pp. 1–28, 2015.

[BVV10] P. Bremen, M. M. Van Wanrooij, and A. J. Van Opstal, "Pinna cues determine orienting response modes to synchronous sounds in elevation," *Journal of Neuroscience*, vol. 30, no. 1, pp. 194–204, jan 2010.

[BWLA00] D. R. Begault, E. M. Wenzel, A. S. Lee, and M. R. Anderson, "Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source," in *Proceedings of the 108th Convention of the Audio Engineering Society*, vol. 5133, Paris, 2000.

[CDD+06] T. J. Cox, B.-I. I. L. Dalenback, P. D'Antonio, J. J. Embrechts, J. Y. Jeon, E. Mommertz, M. Vorländer, P. D 'antonio, J. J. Embrechts, J. Y. Jeon, E. Mommertz, and M. Vorländer, "A Tutorial on Scattering and Diffusion Coefficients for Room Acoustic Surfaces," *Acta Acustica united with Acustica*, vol. 92, pp. 1–15, 2006.

[CEL14] R. Crawford-Emery and H. Lee, "The subjective effect of BRIR length perceived headphone sound externalisation and tonal colouration," in *136th AES Convention Berlin, April 2014*, Berlin, 2014.

[CF98] Y.-c. 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.

[CG92] D. W. Chandler and D. W. Grantham, "Minimum audible movement angle in the horizontal plane as a function of stimulus frequency and bandwidth, source azimuth, and velocity," *Journal of the Acoustical Society of America*, vol. 91, pp. 1624–1636, 1992.

[Cli87] R. K. Clifton, "Breakdown of echo suppression in the precedence effect," *Journal of the Acoustical Society of America*, vol. 82, no. 5, pp. 1834–1835, 1987.

[Col68] P. Coleman, "Dual Role of Frequency Spectrum in Determination of Auditory Distance," *Journal of the Acoustical Society of America*, vol. 44, no. 2, pp. 631–632, 1968.

[CP34] C. J. Clopper and Pearson E. S., "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934.

[Cre48] L. Cremer, *Die wissenschaftlichen Grundlagen der Raumakustik: Geometrische Raumakustik*. S. Hirzel Verlag, 1948.

[CSB+13] J. Catic, S. Santurette, J. M. Buchholz, F. Gran, and T. Dau, "The effect of interaural-level-difference fluctuations on the externalization of sound," *The Journal of the Acoustical Society of America*, vol. 134, no. 2, pp. 1232–1241, 2013.

[CSD15] J. Catic, S. Santurette, and T. Dau, "The role of reverberation-related binaural cues in the externalization of speech," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 1154–1167, 2015.

[Dal10] B.-I. Dalenbäck, "Engineering principles and techniques in room acoustics prediction," *Baltic-Nordic Acoustics Meeting*, 2010.

[Dam71] P. Damaske, "Die psychologische Auswertung akustischer Phänomene," in *7th International Congress on Acoustics*, 1971.

[Dan01] J. Daniel, "Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia," Ph.D. dissertation, Université Paris 6, 2001.

[DB01] T. Djelani and J. Blauert, "Investigations into the Build-up and Breakdown of the Precedence Effect," *Acta Acustica united with Acustica*, vol. 87, no. November 1999, pp. 253–261, 2001.

[DC06] R. M. Dizon and H. S. Colburn, "The influence of spectral, temporal, and interaural stimulus variations on the precedence effect," *The Journal of the Acoustical Society of America*, vol. 119, no. 5, pp. 2947–2964, may 2006.

[Div92] P. L. Divenyi, "Binaural suppression of nonechoes," *The Journal of the Acoustical Society of America*, vol. 91, no. 2, pp. 1078–1084, feb 1992.

[DK86] L. Dietsch and W. Kraak, "Ein objektives Kriterium zur Erfassung von Echostörungen bei Musik- und Sprachdarbietungen," *Acta Acustica united with Acustica*, vol. 60, no. 3, pp. 205–216, 1986.

[DRP+92] N. I. Durlach, A. Rigopulos, X. D. Pang, W. S. Woods, A. Kulkarni, H. S. Colburn, and E. M. Wenzel, "On the Externalization of Auditory Images," *Presence: Teleoperators and Virtual Environments*, vol. 1, no. 2, pp. 251–257, 1992.

[DRP98] J. Daniel, J.-B. Rault, and J.-D. Polack, "Ambisonics encoding of other audio formats for multiple listening conditions," in *Proc. of the 105th AES*, 1998.

[DT98] P. D'Antonio and C. Trevor, "Two Decades of Sound Diffusor Design and Development, Part 2: Prediction, Measurement, and Characterization," *J. Acoust. Soc. Am.*, vol. 46, no. 12, pp. 1075–1091, 1998.

[DT00] ——, "Diffusor application in rooms," *Applied Acoustics*, vol. 60, no. 2, pp. 113–142, 2000.

[EB02] M. O. Ernst and M. S. Banks, "Humans integrate visual and haptic information in a statistically optimal fashion," *Nature*, vol. 415, no. 6870, pp. 429–433, jan 2002.

[EOBW18] R. Ege, A. J. V. Opstal, P. Bremen, and M. M. V. Wanrooij, "Testing the Precedence Effect in the Median Plane Reveals Backward Spatial Masking of Sound," *Scientific Reports*, vol. 8, no. 1, 2018.

[ESN68] M. Ebata, T. Sone, and T. Nimura, "On the perception of direction of echo," *J. Acoust. Soc. Am.*, vol. 44, no. 2, pp. 542–547, 1968.

[FB10] S. Favrot and J. M. Buchholz, "LoRA: A Loudspeaker-Based Room Auralization System," *Acta Acustica united with Acustica*, vol. 96, no. 2, pp. 364–375, 2010.

[FCL91] R. L. Freyman, R. K. Clifton, and R. Y. Litovsky, "Dynamic processes in the precedence effect," *The Journal of the Acoustical Society of America*, vol. 90, no. 2, pp. 874–884, 1991.

[FGZ14] R. L. Freyman, A. M. Griffin, and P. M. Zurek, "Threshold of the precedence effect in noise," *The Journal of the Acoustical Society of America*, vol. 135, no. 5, pp. 2923–2930, 2014.

[FMS11] M. Frank, G. Marentakis, and A. Sontacchi, "A simple technical measure for the perceived source width," *Fortschritte der Akusitk*, vol. 37, pp. 691–692, 2011.

[FMSZ10] M. Frank, L. Mohr, A. Sontacchi, and F. Zotter, "Flexible and intuitive pointing method for 3D auditory localization experiments," in *Proceedings of the 38th AES International Conference, Piteå, Sweden*, 2010.

[Fra13] M. Frank, "Phantom Sources using Multiple Loudspeakers in the Horizontal Plane," Ph.D. dissertation, University of Music and Performing Arts Graz, 2013.

[FZ06] H. Fastl and E. Zwicker, *Psychoacoustics - Facts and Models*, 3rd ed. Hugo Fastl, Eberhard Zwicker, 2006.

[Ger73] M. A. Gerzon, "Periphony: With-Height Sound Reproduction," *Journal of the Acoustical Society of America*, vol. 21, no. 1, pp. 2–10, 1973.

[Ger92] ——, "General Metatheory of Auditory Localization," in *92nd Convention of Audio Engineering Society*, 1992.

[Gil20] P. M. Giller, "The Influence of Reverberation on Externalization," Master's Thesis, University of Music and Performing Arts, 2020.

[GL07] S. Getzmann and J. Lewald, "Localization of moving sound," *Perception & psychophysics*, vol. 69, no. 6, pp. 1022–1034, 2007.

[GMVM13] E. Georganti, T. May, S. Van De Par, and J. Mourjopoulos, "Sound source distance estimation in rooms based on statistical properties of binaural signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 8, pp. 1727–1741, aug 2013.

[Gri17] S. Grill, "VST-Implementierung Ambisonischer Nachhalleffekte," Institute of Electronic Music, University of Music and Performing Arts, Graz, Tech. Rep. Toningenieur-Projekt, 2017.

[Gus90] R. Guski, "Auditory localization: effects of reflecting surfaces." *Perception*, vol. 19, no. 6, pp. 819–830, 1990.

[GvdPT17] J. Grosse, S. van de Par, and C. Trahiotis, "Stimulus coherence influences sound-field localization and fusion/segregation of leading and lagging sounds," *J. Acoust. Soc. Am.*, vol. 141, no. 4, pp. 2673–2680, 2017.

[GWH19] P. M. Giller, F. Wendt, and R. Höldrich, "The influence of different BRIR modification techniques on externalization and sound quality," in *Proceedings on the Spatial Audio Signal Processing Symposium*, Paris, 2019.

[GYL12] M. J. Goupell, G. Yu, and R. Y. Litovsky, "The effect of an additional reflection in a precedence effect experiment," *The Journal of the Acoustical Society of America*, vol. 131, no. February, p. 2958, 2012.

[Haa72] H. Haas, "The Influence of a Single Echo on the Audibility of Speech," *J. Audio Eng. Soc*, vol. 20, no. 2, pp. 146–159, 1972.

[HBO95] T. Hidaka, L. L. Beranek, and T. Okano, "Interaural cross-correlation, lateral fraction, and low- and high-frequency sound levels as measures of acoustical quality in concert halls," *The Journal of the Acoustical Society of America*, vol. 98, no. 2, pp. 988–1007, 1995.

[HGD16] H. G. Hassager, F. Gran, and T. Dau, "The role of spectral detail in the binaural transfer function on perceived externalization in a reverberant environment," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2992–3000, 2016.

[HHO05] K. Hamasaki, K. Hiyama, and R. Okumura, "The 22.2 Multichannel Sound System and Its Application," in *Proceedings of the 108th Convention of the Audio Engineering Society*, 2005.

[HMK09] T. Hiekkanen, A. Mäkivirta, and M. Karjalainen, "Virtualized listening tests for loudspeakers," *Journal of the Acoustical Society of America*, vol. 57, no. 4, pp. 237–251, 2009.

[HR89] W. M. Hartmann and B. Rakerd, "Localization of sound in rooms, IV: The Franssen effect." *The Journal of the Acoustical Society of America*, vol. 86, pp. 1366–1373, 1989.

[HV98] P. M. Hofman and A. J. Van Opstal, "Spectro-temporal factors in two-dimensional human sound localization," *The Journal of the Acoustical Society of America*, vol. 103, no. 5, pp. 2634–2648, may 1998.

[HW96] W. M. Hartmann and A. Wittenberg, "On the externalization of sound images," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3678–3688, 1996.

[ITU01] ITU-R Recommendation, "Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA)," *International Telecommunication Union*, vol. BS. 1534-1, 2001.

[JSZL20] Z. Jiang, J. Sang, C. Zheng, and X. Li, "The effect of pinna filtering in binaural transfer functions on externalization in a reverberant environment," *Applied Acoustics*, vol. 164, p. 107257, 2020.

[Kee89] D. B. D. Keele, Jr., "Effective Performance of Bessel Arrays," *Audio Engineering Society Convention 87*, vol. 5, 1989.

[KF17] E. Kurz and M. Frank, "Prediction of the listening area based on the energy vector," in *Proceedings of the 4th ICSA, Graz*, Graz, 2017, pp. 52–58.

[KKvdH+97] A. Kohlrausch, R. Kortekaas, M. van der Heijden, S. van de Par, A. J. Oxenham, and D. Püschel, "Detection of Tones in Low-noise Noise: Further Evidence for the Role of Envelope Fluctuations," *Acta Acustica united with Acustica*, vol. 83, pp. 659–669, 1997.

[Kle01] S. A. Klein, "Measuring, estimating, and understanding the psychometric function: A commentary," *Perception & Psychophysics*, vol. 63, no. 8, pp. 1421–1455, 2001.

[Kro14] M. Kronlachner, "Plug-in Suite for Mastering the Production and Playback in Surround Sound and Ambisonics," *136th AES Convention Berlin, April 2014*, pp. 3–7, 2014.

[Kur18] E. Kurz, "Efficient prediction of the listening area for plausible reproduction," Ph.D. dissertation, University of Music and Performing Arts Graz, 2018.

[Kut09] H. Kuttruff, *Room Acoustics*, 5th ed., H. Kuttruff, Ed. Taylor & Francis, 2009.

[KW92] D. J. Kistler and F. L. Wightman, "A Model Of Head-Related Transfer Functions Based On Principal Components Analysis And Minimum-Phase Reconstruction," *Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1637–1647, 1992.

[KW13] F. Klein and S. Werner, "HRTF adaptation and pattern learning," in *Proceedings of ISAAR 2013: Auditory Plasticity - Listening with the Brain*, T. Dau, Ed., Nyborg, DK, 2013.

[LB58] J. P. A. Lochner and J. F. Burger, "The subjective masking of short time delayed echoes their primary sounds and their contribution to the intellegibility of speech," *Acustica*, vol. 8, no. 1, pp. 1–10, 1958.

[Lee13] H. Lee, "Apparent Source Width and Listener Envelopment in Relation to Source-Listener Distance," in *Audio Engineering Society Conference*. Audio Engineering Society, 2013.

[Lev71] H. Levitt, "Transformed Up-Down Methods in Psychoacoustics," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 467–477, 1971.

[LMC92] A. D. Little, D. H. Mershon, and P. H. Cox, "Spectral content as a cue to perceived auditory distance." *Perception*, vol. 21, no. 3, pp. 405–416, 1992.

[Lös14] S. Lösler, "MIMO-Rekursivfilter ur Kugelarrays," Ph.D. dissertation, Univeristy of Music and Performing Arts, 2014.

[LPHP15] M.-V. Laitinen, A. Politis, I. Huhtakallio, and V. Pulkki, "Controlling the perceived distance of an auditory object by manipulation of loudspeaker directivity," *The Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. EL462–EL468, 2015.

[LPT+11] T. Lokki, J. Pätynen, S. Tervo, S. Siltanen, and L. Savioja, "Engaging concert hall acoustics is made up of temporal envelope preserving reflections," *The Journal of the Acoustical Society of America*, vol. 129, no. 6, pp. EL223–EL228, jun 2011.

[LR05] H.-K. Lee and F. Rumsey, "Investigation Into the Effect of Interchannel Crosstalk in Multichannel Microphone Technique," in *AES 118th Convention*, Barcelona, 2005.

[LRYH97] R. Y. Litovsky, B. Rakerd, T. C. T. Yin, and W. M. Hartmann, "Psychophysical and physiological evidence for a precedence effect in the median sagittal plane," *Journal of Neurophysiology*, vol. 77, no. 4, pp. 2223–2226, 1997.

[LSC01] R. Litovsky and B. Shinn-Cunningham, "Investigation of the relationship among three common measures of precedence: Fusion, localization dominance, and discrimination suppression," *The Journal of the Acoustical Society of America*, vol. 109, no. 1, pp. 346–358, 2001.

[LSP18] S. Li, R. Schlieper, and J. Peissig, "The effect of variation of reverberation parameters in contralateral versus ipsilateral ear signals on perceived externalization of a lateral sound source in a listening room," *Journal of the Acoustical Society of America*, vol. 144, no. 2, pp. 966–980, 2018.

[LSP19] ——, "The Role of Reverberation and Magnitude Spectra of Direct Parts in Contralateral and Ipsilateral Ear Signals on Perceived Externalization," *Applied Sciences*, 2019.

[LW12] A. Lindau and S. Weinzierl, "Assessing the plausibility of virtual acoustic environments," *Acta Acustica united with Acustica*, vol. 98, no. 5, pp. 804–810, 2012.

[LWFZ18] J. Linke, F. Wendt, M. Frank, and F. Zotter, "How the perception of moving sound beams is influenced by masking and reflector setup," in *Proceedings of the 30th Tonmeistertagung*. Köln: Verband Deutscher Tonmeister, 2018.

[LWZF18] J. Linke, F. Wendt, F. Zotter, and M. Frank, "How Masking affects Auditory Objects of Beamformed Sounds," in *Fortschritte der Akusitk*, 2018, pp. 355–357.

[Mak62] Y. Makita, "On the directional localisation of sound in the stereophonic sound field," *EBU Review*, vol. 73, no. A, pp. 1536–1539, 1962.

[Mar67] A. H. Marshall, "A note on the importance of room cross-section in concert halls," *Journal of Sound and Vibration*, vol. 5(1), pp. 100–112, 1967.

[Mas17] R. Mason, "How Important Is Accurate Localization in Reproduced Sound?" in *Proceedings of the 142th Convention of the Audio Engineering Society*, Berlin, 2017.

[MBL+89] D. H. Mershon, W. L. Ballenger, A. D. Little, P. L. McMurtry, and J. L. Buchanan, "Effects of room reflectance and background noise on perceived auditory distance," *Perception*, vol. 18, no. 3, 1989.

[MK75] D. H. Mershon and L. E. King, "Intensity and reverberation as factors in the auditory perception of egocentric distance," *Perception & Psychophysics*, vol. 18, no. 6, pp. 409–415, nov 1975.

[MKRB+20] N. Meyer-Kahlen, D. Rudrich, M. Brandner, S. Wirler, S. Windtner, and F. Matthias, "DIY Modifications for Acoustically Transparent Headphones," in *Proc. of the 148th AES*, 2020.

[MLK09] S. D. Miller, R. Y. Litovsky, and K. R. Kluender, "Predicting echo thresholds from speech onset characteristics," *The Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. EL134–EL140, 2009.

[MM88] M. Morimoto and Z. Maekawa, "Effects of Low Frequency Components on Auditory Spaciousness," *Acta Acustica united with Acustica*, vol. 66, no. 4, pp. 190–196, 1988.

[MMP16] C. Mendonça, P. Mandelli, and V. Pulkki, "Modeling the Perception of Audiovisual Distance: Bayesian Causal Inference and Other Models," *PLOS ONE*, vol. 11, no. 12, p. e0165391, 2016.

[Moo79] J. A. Moorer, "About This Reverberation Business," *Computer Music Journal*, vol. 3, no. 2, p. 13, 1979.

[Nay93] G. Naylor, "ODEON - Another hybrid room acoustical model," *Applied Acoustics*, vol. 38, no. 2-4, pp. 131–143, 1993.

[NK15] A. Neidhardt and N. Knoop, "Investigating the room divergence effect in binaural playback," in *Proceedings of the 28th Tonmeistertagung*, 2015.

[Oka00] T. Okano, "Image shift caused by strong lateral reflections, and its relation to inter-aural cross correlation," *The Journal of the Acoustical Society of America*, vol. 108, no. 5, pp. 2219–2230, 2000.

[OLBD10] B. Ohl, S. Laugesen, J. Buchholz, and T. Dau, "Externalization versus Internalization of Sound in Normal-hearing and Hearing-impaired Listeners," in *Fortschritte der Akusitk*, Berlin, 2010, pp. 633–634.

[OOL+20] F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, and B. V. Saunders, "NIST Digital Library of Mathematical Functions," 2020.

[ORS+20] J. Oberem, J.-G. Richter, D. Setzer, J. Seibold, I. Koch, and J. Fels, "Experiments on localization accuracy with non-individual and individual HRTFs comparing static and dynamic reproduction methods," *bioRxiv*, p. 2020.03.31.011650, 2020.

[PAB19] C. Porschmann, J. M. Arend, and F. Brinkmann, "Directional equalization of sparse head-related transfer function sets for spatial upsampling," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 27, no. 6, pp. 1060–1071, 2019.

[PB15] M. T. Pastore and J. Braasch, "The precedence effect with increased lag level," *The Journal of the Acoustical Society of America*, vol. 138, no. 4, pp. 2079–2089, 2015.

[Ple72] G. Plenge, "Über das Problem der Im-Kopf-Lokalisation.pdf," *Acustica*, vol. 26, no. 5, pp. 241–252, 1972.

[PS90] D. R. Perrott and K. Saberi, "Minimum audible angle thresholds for sources varying in both elevation and azimuth." *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1728–1731, 1990.

[PSA17] C. Pörschmann, P. Stade, and J. M. Arend, "Binaural auralization of proposed room modifications based on measured omnidirectional room impulse responses," in *Proceedings of meetings on acoustics Acoustical Society of America*, vol. 30, 2017.

[PSM87] D. R. Perrott, T. Strybel, and C. Manligas, "Conditions under which the Haas precedence effect may or may not occur," *J. Aud. Res.*, vol. 27, pp. 59–72, 1987.

[Pul97] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, 1997.

[Pul01] ——, "Localization of Amplitude-Panned Virtual Sources II: Two- and Three-Dimensional Panning," *Journal of the Audio Engineering Society*, vol. 49, no. 9, pp. 753–767, 2001.

[RH85] 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.

[RH86] ——, "Localization of sound in rooms, III: Onset and duration effects." *The Journal of the Acoustical Society of America*, vol. 80, no. 6, pp. 1695–1706, 1986.

[RHH00] B. Rakerd, W. Hartmann, and J. Hsu, "Echo suppression in the horizontal and median saggittal planes," *Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 1061–1064, 2000.

[RJ03] P. Rubak and L. G. Johansen, "Coloration in Natural and Artificial Room Impulse Responses," *23rd International AES Conference*, pp. 1–19, 2003.

[RS97] D. Rocchesso and J. O. Smith, "Circulant and elliptic feedback delay networks for artificial reverberation," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 1, pp. 51–63, 1997.

[RWFB13] P. W. Robinson, A. Walther, C. Faller, and J. Braasch, "Echo thresholds for reflections from acoustically diffusive architectural surfaces," *The Journal of the Acoustical Society of America*, vol. 134, no. 4, p. 2755, 2013.

[San76] F. Santon, "Numerical prediction of echograms and of the intelligibility of speech in rooms," *The Journal of the Acoustical Society of America*, vol. 59, no. 6, pp. 1399–1405, 1976.

[Saw09] S. S. Sawilowsky, "New Effect Size Rules of Thumb," *The Journal of Modern Applied Statistical Methods*, vol. 8, no. 2, pp. 597–599, 2009.

[Sch65] M. R. Schroeder, "New Method of Measuring Reverberation Time," *The Journal of the Acoustical Society of America*, vol. 37, no. 6, pp. 1187–1188, 1965.

[Sch75] ——, "Diffuse sound reflection by maximum-length sequences," *Journal of the Acoustical Society of America*, vol. 57, no. 1, pp. 149–150, 1975.

[Sch79] ——, "Binaural dissimilarity and optimum ceilings for concert halls : More lateral sound diffusion," *J. Acoust. Soc. Am.*, vol. 65, pp. 958–963, 1979.

[SCKM05] B. G. Shinn-Cunningham, N. Kopco, and T. J. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *The Journal of the Acoustical Society of America*, vol. 117, no. 5, pp. 3100–3115, 2005.

[SCZD93] B. G. Shinn-Cunningham, P. M. Zurek, and N. I. Durlach, "Adjustment and discrimination measurements of the precedence effect," *J. Acoust. Soc. Am.*, vol. 93, pp. 2923–2932, 1993.

[SDSP91] K. Saberi, L. Dostal, T. Sadralodabai, and D. R. Perrott, "Minimum Audible Angles for Horizontal, Vertical, and Oblique Orientations: Lateral and Dorsal Planes," *Acta Acustica united with Acustica*, vol. 75, no. 1, pp. 57–61, 1991.

[SE06] V. P. Sivonen and W. Ellermeier, "Directional loudness in an anechoic sound field, head-related transfer functions, and binaural summation," *The Journal of the Acoustical Society of America*, vol. 119, no. 5, pp. 2965–2980, 2006.

[Sep61] H. P. Sepharim, "Über die Wahrnehmbarkeit mehrerer Rückwurfe von Sprachschall," *Acustica*, vol. 11, no. 2, pp. 80–91, 1961.

[SG12] G. C. Stecker and F. Gallun, "Binaural Hearing, Sound Localization, and Spatial Hearing," in *Translational Perspectives in Auditory Neuroscience: Normal Aspects of Hearing*, 2012, ch. 14, pp. 383–433.

[SGK76] N. Sakamoto, T. Gotoh, and Y. Kimura, "On -Out-of-Head Localization- in Headphone Listening," *Journal of the Audio Engineering Society*, vol. 24, no. 9, pp. 710–716, 1976.

[SGSN66] T. Somerville, C. L. S. Gilford, N. F. Spring, and R. D. M. Negus, "Recent work on the effects of reflectors in concert halls and music studios," *Journal of Sound and Vibration*, vol. 3, no. 2, pp. 127–134, 1966.

[Sha16] G. K. Sharma, "Composing with Sculptural Sound Phenomena in Computer Music," Ph.D. dissertation, University of Music and Performing Arts Graz, 2016.

[SLTS12] S. Siltanen, T. Lokki, S. Tervo, and L. Savioja, "Modeling incoherent reflections from rough room surfaces with image sources," *The Journal of the Acoustical Society of America*, vol. 131, no. 6, pp. 4606–4614, 2012.

[Sti15] P. Stitt, "Ambisonics and Higher-Order Ambisonics for Off-Centre Listeners : Evaluation of Perceived and Predicted Image Direction," Ph.D. dissertation, Queen's University Belfast, 2015.

[Tak05] S. Takumai, "Loudspeaker array device and method for setting sound beam of loudspeaker array device," 2005.

[TH99a] D. J. Tollin and G. B. Henning, "Some aspects of the lateralization of echoed sound in man. II. The role of the stimulus spectrum," *The Journal of the Acoustical Society of America*, vol. 105, no. 2, p. 838, 1999.

[TH99b] ——, "Some aspects of the lateralization of echoed sound in man. II. The role of the stimulus spectrum," *The Journal of the Acoustical Society of America*, 1999.

[TM15] A. Tregonning and B. Martin, "The vertical precedence effect: Utilizing delay panning for height channel mixing in 3D audio," *Audio Engineering Society Convention*, pp. 1–9, 2015.

[Too70] F. E. Toole, "In Head Localization of Acoustic Images," *The Journal of the Acoustical Society of America*, vol. 48, pp. 943–949, 1970.

[TPKL13] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial decomposition method for room impulse responses," *Journal of the Audio Engineering Society*, vol. 61, no. 1/2, pp. 17–28, 2013.

[TTY00] B. T. G. Tan, S. H. Tang, and G. Yu, "Perception of a secondary auditory image with three sound sources," in *Acustica*, vol. 86, 2000, pp. 1034–1037.

[UPG14] J. Udesen, T. Piechowiak, and F. Gran, "Vision Affects Sound Externalization," in *AES Conference: 55th International Conference: Spatial Audio*, 2014, pp. 27–30.

[WAKW93] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 111–123, 1993.

[Wal38] H. Wallach, "Über die Wahrnehmung der Schallrichtung," *Psychologische Forschung*, vol. 22, no. 3-4, pp. 238–266, dec 1938.

[WDC97] O. Warusfel, P. Derogis, and R. Causse, "Radiation Synthesis with Digitally Controlled Loudspeakers," in *Audio Engineering Society Convention 103*, sep 1997.

[Weg20] K. Wegler, "Über den Einfluss unterschiedlicher Reflexionseigenschaften auf den Präzedenzeffekt," Master's Thesis, University of Music and Performing Arts, 2020.

[WEJV10] A. Wabnitz, N. Epain, C. T. Jin, and A. Van Schaik, "Room acoustics simulation for multichannel microphone arrays," *International Symposium on Room Acoustics*, no. August, pp. CDROM: 1—-6, 2010.

[Wen63] K. Wendt, "Das Richtungshören ber der Überlagerung zweier Schallfelder bei Intensitäts- und Laufzeitstereophonie," Ph.D. dissertation, RWTH Aachen, 1963.

[Wen17] F. Wendt, "Investigations on Perceptual Phenomena of the Precedence Effect using a Bessel Sequence," in *Proceedings of the 142th Convention of the Audio Engineering Society*, 2017.

[WF01] B. A. Wright and M. B. Fitzgerald, "Different Patterns of Human Discrimination Learning for Two Interaural Cues to Sound-Source Location," pp. 12 307–1231, 2001.

[WF18] F. Wendt and M. Frank, "On the localization of auditory objects created by directional sound sources in a virtual room," in *Proceedings of the 30th Tonmeistertagung*. Köln: Verband Deutscher Tonmeister, 2018.

[WFH19] F. Wendt, M. Frank, and R. Höldrich, "The role of median plane reflections in the perception of vertical auditory movement," in *Proceedings of the 23th International Congress on Acoustics*, 2019.

[WFZ14] F. Wendt, M. Frank, and F. Zotter, "Panning with height on 2, 3, and 4 loudspeakers," in *2nd International Conference on Spatial Audio (ICSA)*, Erlangen, 2014.

[WFZH16a] F. Wendt, M. Frank, F. Zotter, and R. Höldrich, "Directivity patterns controlling the auditory source distance," in *Proceedings of the 19th International Conference on Digital Audio Effects (DAFx-16)*, Brno, 2016, pp. 295–300.

[WFZH16b] ——, "Influence of directivity pattern order on perceived distance," in *Fortschritte der Akustik*, 2016.

[WH20] F. Wendt and R. Höldrich, "Precedence effect for specular and diffuse reflections," *Acta Acustica, Manuscript in revision*, 2020.

[WHF17] F. Wendt, R. Höldrich, and M. Frank, "The Influence of the Floor Reflection on the Perception of Sound Elevation," in *Fortschritte der Akustik*, 2017, pp. 767–770.

[WHM19] F. Wendt, R. Höldrich, and M. Marschall, "How binaural room impulse responses influence the externalization of speech," in *Fortschritte der Akusitk*, 2019.

[WK89] F. L. Wightman and D. J. Kistler, "Headphone Simulation Of Free-Field Listening. I: Stimulus Synthesis," *Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 858–867, feb 1989.

[WKH13] S. Werner, F. Klein, and T. Harczos, "Context-dependent quality parameters and perception of auditory illusions," *4th International Symposium on Auditory and Audiological Research*, no. August, pp. 1–8, 2013.

[WKS16] S. Werner, F. Klein, and T. Sporer, "Adjustment of the Direct-to-Reverberant-Energy-Ratio to Reach Externalization within a Binaural Synthesis System," in *Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality*, 2016.

[WL15] R. Wallis and H. Lee, "The effect of interchannel time difference on localization in vertical stereophony," *Journal of the Audio Engineering Society*, vol. 63, no. 10, pp. 767–776, 2015.

[WMA19] T. Wühle, S. Merchel, and M. E. Altinsoy, "Perception of auditory events in scenarios with projected and direct sound from various directions," in *146th Audio Engineering Society Convention*, 2019.

[WNR49] H. Wallach, E. B. Newman, and M. R. Rosenzweig, "The precedence effect in sound localization," *The American journal of psychology*, vol. 62, no. 3, pp. 315–336, 1949.

[WRS13] A. Walther, P. Robinson, and O. Santala, "Effect of spectral overlap on the echo suppression threshold for single reflection conditions." *The Journal of the Acoustical Society of America*, vol. 134, no. 2, pp. EL158–64, 2013.

[WSF+17] 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, no. 1, pp. 76–88, mar 2017.

[WT02] H. Wittek and G. Theile, "The Recording Angle - Based on Localisation Curves," in *Proceedings of the 112th Convention of the Audio Engineering Society*, 2002.

[WvdPE14] T. Wendt, S. van de Par, and S. D. Ewert, "A Computationally-Efficient and Perceptually-Plausible Algorithm for Binaural Room Impulse Response Simulation," *Journal of the Acoustical Society of America*, vol. 62, no. 11, 2014.

[WWH19] K. Wegler, F. Wendt, and R. Höldrich, "How level, delay, and spatial separation influence the echo threshold," in *Fortschritte der Akusitk*, Rostock, 2019.

[WZFH17] 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.

[XSC11] J. Xia and B. Shinn-Cunningham, "Isolating mechanisms that influence measures of the precedence effect: Theoretical predictions and behavioral tests," *The Journal of the Acoustical Society of America*, vol. 130, no. 2, pp. 866–882, aug 2011.

[YG97] X. Yang and D. W. Grantham, "Echo suppression and discrimination suppression aspects of the precedence effect," *Perception and Psychophysics*, vol. 59, no. 7, pp. 1108–1117, 1997.

[Yos07] W. A. Yost, "Lead-Lag Precedence Paradigm as a Function of Relative Level and Number of Lag Stimuli," in *Proceedings of the 19th International Congress on Acoustics*, Madrid, 2007.

[Zah01] P. Zahorik, "Estimating sound source distance with and without vision," *Optometry and Vision Science*, vol. 78, no. 5, pp. 270–275, 2001.

[Zah02a] ——, "Assessing auditory distance perception using virtual acoustics," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1832–1846, 2002.

[Zah02b] ——, "Direct-to-reverberant energy ratio sensitivity," *The Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2110–2117, 2002.

[ZBB05] 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.

[ZF12] 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.

[ZF15] ——, "Investigation of auditory objects caused by directional sound sources in rooms," *Acta Physica Polonica A*, vol. 128, no. 1, pp. 1–12, 2015.

[ZF19] ——, *Ambisonics - A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, 2019.

[ZFFR14] F. Zotter, M. Frank, A. Fuchs, and D. Rudrich, "Preliminary study on the perception of orientation-changing directional sound sources in rooms," *Proc. of Forum Acusticum*, no. c, pp. 1–6, 2014.

[ZFKC14] F. Zotter, M. Frank, M. Kronlacher, and J.-W. Choi, "Efficient phantom source widening and diffuseness in ambisonics," *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics*, vol. 2, no. April, pp. 69–74, 2014.

[Zot09] F. Zotter, "Analysis and synthesis of sound-radiation with spherical arrays," Ph.D. dissertation, University of Music and Performing Arts Graz, 2009.

[ZPS08] F. Zotter, H. Pomberger, and A. Schmeder, "Efficient directivity pattern control for spherical loudspeaker arrays," *Proceedings of the European Conference on Noise Control*, pp. 4231–4236, 2008.

[ZSH18] M. Zaunschirm, C. Schörkhuber, and R. Höldrich, "Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3616–3627, 2018.

[Zur80] P. M. Zurek, "The precedence effect and its possible role in the avoidance of interaural ambiguities." *The Journal of the Acoustical Society of America*, vol. 67, no. February 1979, pp. 953–964, 1980.

[ZZFK17] 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, 2017.