The Influence of Reverberation on Externalization

Master Thesis

Peter Maximilian Giller, BSc

Matr. No. 01173072

Master's Degree Programme
Electrical Engineering and Audio Engineering (UV 066 413)
at the University of Music and Performing Arts Graz
and the Graz University of Technology

Supervision:
O. Univ. Prof. Dr. techn. Robert Höldrich
Dipl.-Ing. Florian Wendt

Graz, May 26, 2020

Institute of Electronic Music and Acoustics

Familienname, Vorname

Matrikelnummer

Erklärung

Hiermit bestätige ich, dass mir der *Leitfaden für schriftliche Arbeiten an der KUG* bekannt ist und ich die darin enthaltenen Bestimmungen eingehalten habe. Ich erkläre ehrenwörtlich, dass ich die vorliegende Arbeit selbständig und ohne fremde Hilfe verfasst habe, andere als die angegebenen Quellen nicht verwendet habe und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, den

……………………………………………………………………….
Unterschrift der Verfasserin/des Verfassers

Abstract

Inside-head localization is a common problem of headphone-based reproduction. Realistic binaural synthesis, however, requires well-externalized sound images. Externalization is a subjective quantity referring to the sensation of auditory events being located outside the listener's head. It is a fragile experience that, in addition to the sound field at the entrance of the ear canals, also depends on visual cues, training, and expectation. Various studies have examined the perceptual and technical aspects of the phenomenon. It was found that the presence of reverberation can increase the degree of externalization. However, if the original sound of a recording should be preserved, additional reverberation is undesired. This work investigates how reverberation influences externalization and sound quality in two listening experiments. The first experiment compares the effect of successive de-reverberation of binaural room impulse responses using different methods. The second experiment is an exploratory study considering different scenarios. These include the variation between different spatial distributions as well as binaural, dichotic, and diotic reverberation, the influence of additive and convolutive reverberation, and the effect of discrete specular and diffuse reflections.

$*$

Ein Problem der binauralen Wiedergabe über Kopfhörer ist die sogenannte Im-Kopf-Lokalisation, bei der virtuelle Schallquellen nicht als externalisiert wahrgenommen werden. Der Grad der Externalisierung hängt nicht nur vom Schallfeld am Eingang der Gehörgänge ab, sondern wird auch durch visuelle Reize, Training und Erwartungshaltung beeinflusst. Studien konnten zeigen, dass die Externalisierung unter anderem durch Hinzufügen von Nachhall gesteigert werden kann. Im Rahmen dieser Arbeit wird der Einfluss von Nachhall auf Externalisierung und Klangqualität anhand zweier Hörversuche untersucht. Im ersten Versuch wird mittels verschiedener Verfahren der Nachhall einer binauralen Raumimpulsantwort sukzessive reduziert und ein Vergleich zwischen den Verfahren durchgeführt. Der zweite Versuch behandelt verschiedene explorative Szenarien. Es werden unterschiedliche räumliche Verteilungen sowie binauraler, dichotischer, und diotischer Nachhall verglichen. Darüberhinaus wird der Einfluss von zusätzlichem Nachhall im Anregungssignal sowie der von einzelnen direkten und diffusen Reflexionen untersucht.

Contents

1 Introduction

What we perceive as the surrounding world is not actually the world that surrounds us. We constantly form a picture of the outside world within our heads, and our visual, auditory, and tactile senses connect us with it. But it is impossible to observe reality immediately. Not only are our senses imperfect. Any sensory input is converted into electrochemical neural signals that are transmitted, filtered, and finally, mingled with experience and expectations, put together into a conscious or unconscious, in any case, subjective image by processes that, by far, exceed the scope of this thesis. However, that image is not reality itself.

Sometimes our senses play tricks on us, and sometimes we allow them to. We have the ability to differentiate between what we think is reality and the virtual world. And we also have the ability to voluntarily enter the virtual world and accept virtual experiences with ‘real’ ones as equivalent.

Acoustic virtual reality has experienced great progress in the past decades. With spatial audio techniques, such as Vector Base Amplitude Panning, Wave Field Synthesis, and Ambisonics, we can reproduce sound fields and position virtual sound sources in space using loudspeaker arrays. Binaural audio, in contrast, is spatial audio through headphones. With just two loudspeakers on our ears, we can hear virtual sound sources around us. Much more than with visual virtual reality, this requires some understanding of the involved psychophysical processes. Sec. 1.1 gives a short overview about the peculiarities of binaural hearing.

A perfect simulation is indistinguishable from reality – indistinguishable using any of our senses. The technical realization, however, can never be perfect. But it is often sufficient to provide a certain quality of experience.

It may be a trivial observation, but allow me to point out that the majority of sounds that we are confronted with in our everyday life emerge from some point in space outside of our head. And, in most cases, we perceive them just there – outside the head. This is called externalization. However, what is trivial to observe is surprisingly difficult to reproduce in a simulation. This makes externalization a main quality feature of binaural technology.

Externalization is influenced by various factors which will be explained in more detail in Sec. 1.2 of this chapter. At this point, it is sufficient to state that reverberation is one of these factors. As it is the acoustic fingerprint of all indoor spaces, it provides us with a spatial impression, envelops us. But it is not that simple: The different parameters of reverberation play an important role, as well as the listening room in which the reverberation is being heard. Another complication is that the presence of reverberation may be undesired in different scenarios. This is the field of tension this work has as its topic.

The influence of reverberation on externalization is studied on the example of two listening experiments that constitute the two main chapters of this thesis. The first experiment in Ch. 2 studies different manipulation methods of reverberant binaural room impulse responses. Each method reduces the reverberation in a binaural room impulse response. A comparison between the methods is made with regard to externalization and sound quality. Ch. 3 presents an exploratory study on the effect of varying reverberation parameters in different scenarios. The scenarios include, e.g., varying spatial distributions of binaural reverberation, convolutive reverberation in the input signal, or the use of single reflections rather than persistent reverberation. The conclusion in Ch. 4 summarizes this thesis.

1.1 Binaural Hearing

> *Nature has given man one tongue, but two ears, that we may hear twice as much as we speak.*
>
> *Epictetus* (Higginson et al., 1865, p. 428)

While this ancient proverb is certainly not to be understood literally, it is true insofar as the circumstance of having two ears and being able to hear *binaurally* brings various advantages over *monaural* hearing (i.e., the use of only one ear).

The basis of our binaural perception of the surrounding world is the phenomenon of *binaural fusion* where two different sound signals at the two ears can be merged into a single sound image, under the condition of a certain degree of similarity of the two signals (Gelfand, 2018, p. 321). Listening to two signals that differ in some respect between the two ears is referred to as *dichotic* listening, whereas the presentation of two identical signals at both ears is referred to as *diotic*. Binaural listening leads to a decreased threshold for the detection of tonal, noise, and speech signals compared to monaural listening due to *binaural summation*, and an increased *differential sensitivity* for both intensity and frequency (Jesteadt and Wier, 1977).

Sound localization is greatly improved by the use of two ears. The *Duplex Theory* formulated by Rayleigh in 1907 explains localization by the analysis of *interaural time differences* (ITDs) and *interaural level differences* (ILDs). A widely accepted theory modeled ILDs and ITDs based on a sphere representing the human head. If a sound source moves around this sphere in the horizontal plane, the ILD increases towards lateral directions at high frequencies due to the shadowing effect. As the sound source reaches 90°, the ILD is maximal. Since longer wavelengths are diffracted around the head, the ILD can not function as a localization cue for low frequencies. Instead, the ITD gives information about the angle of the sound source. At higher frequencies, in turn, the ITD is not meaningful due to the phase ambiguity for wavelengths shorter than the travel distance from one ear to the other. As pointed out by Gelfand (2018), the Duplex Theory is inadequate for explaining the localization of elevated sound sources, front/back discrimination, and monaural localization. Another issue is the perception of a sound image inside the head (*intracranially*) on the lateral axis between the ears (*lateralization*) if the signals at the two ears merely differ in ITDs and ILDs as it can be the case when using headphones. In contrast, sound sources in the real world are usually perceived outside of the head and at a certain distance. The perception of sound sources is referred to as *externalization* (sometimes also out-of-head localization or *extracranialization*), whereas distance perception is a related but different phenomenon. The topic of externalization will be treated in more detail in Sec. 1.2.

Monaural spectral cues provide further information on the position of a sound source, as the reflections on pinnae, head, and torso lead to characteristic filtering of the incident sound that can be mapped to a certain angular location, resolving ambiguities of sound sources in the median plane with no level or time differences between the ears (Blauert, 1969). The *head-related transfer function* (HRTF) captures frequency-dependent ITDs and ILDs as well as monaural spectral cues. A widely used method for creating artificial binaural sound images (*binaural synthesis*) is the convolution of a sound signal with the *head-related impulse response* (HRIR) corresponding to the desired source direction.

However, the use of anechoic HRIRs is, in many cases, not sufficient for externalization and distance perception. While the presence of reverberation can reduce the accuracy of localization (Giguère and Abel, 1993), it has been shown that reverberation has a positive effect on the correct perception of distance (Nielsen, 1992). One important measure auditory distance perception depends on is, next to sound intensity and the attenuation of high frequencies (at larger distances), the *direct-to-reverberant energy ratio* (DRR). The DRR increases in a reverberant sound field with increasing distance between the source and the receiver. In addition, the correct perception of distance depends on visual cues and familiarity of experience (Zahorik et al., 2005; Wisniewski et al., 2012).

While the impression of distance can be conveyed even for inside-head localization, true distance perception has externalization as a prerequisite (Kolarik et al., 2016).

1.2 Externalization

The term *externalization* describes the sensation of perceiving a sound image as located outside of the head as it is usually the case in our everyday-world experience. However, using headphone-based reproduction methods, the creation of externalized sound images turns out to be a complex matter.

Headphone presentation usually leads to a *lateralized* sound image inside the head, instead of a localized sound image outside the head (Gelfand, 2018, p. 325). Despite similarities between the two phenomena such as the good correspondence of ILDs and ITDs producing just-noticeable differences in azimuthal localization and dichotic lateralization (Mills, 1960), the problem of inside-head localization was a riddle and a wide range of hypotheses have been studied since.

A list of early assumptions, now seeming rather exotic, can be found in the literature reviews provided by Plenge (1972), and Hartmann and Wittenberg (1996), such as a possible over-modulation of the nervous system or a changed impedance at the eardrum compared to free-field conditions due to the presence of headphones (Blauert, 1974), unavoidable differences between the transmission channels to the left and right ear due to technical reasons (Schirmer, 1966), or a reduced proportion of bone-conducted sound (Sone and Tadamoto, 1968).

The influence of electroacoustic equipment was ruled out by Plenge (1972). Furthermore, examples for inside-head localization with non-electroacoustic presentation were given, such as high sound intensity, manipulation of the shape of the pinnae, or extension of the ear canals using pipes. As possible reasons for intracranial percepts with headphone presentation, he identified poor reproduction of the physiology of the head in dummy-head recordings, missing adaptation, implausible sound level, and contradictory visual information. He concluded with the hypothesis that externalization depends on the classification of stimulus patterns within a long-term memory, and a short time memory containing information about the listening room such as size, amount of reverberation, or source positions (externalization dropped if subjects were unable to acquaint themselves with the sound source and room).

His experimental results suggest that externalization is not part of a continuum as the attempt of the transition of a sound event from inside to outside the head failed – a hypothesis that later has been proven wrong by Hartmann and Wittenberg (1996) who took the frequency dependence of ILDs and ITDs into consideration. Externalization was studied on the basis of a synthesized harmonic sound presented via headphones in comparison to a loudspeaker at a lateral position playing back the same vowel. Amplitude and phase of each harmonic were measured inside the ear canals and adjusted to achieve indiscriminability of the synthesized signal from the loudspeaker (baseline synthesis). The authors systematically introduced deviations from the baseline synthesis in different frequency regions. Being able to move an acoustic image from inside to outside the head, they showed that externalization is a matter of degree. They found that, while ILDs contribute to externalization over the entire frequency spectrum, ITDs only contribute to externalization below 1.5 kHz. The presentation of contradictory ILDs and ITDs led to an intracranial sound image.

In the above study, intracranialization is understood as the result of improper reproduction of the frequency cues provided by head and pinna filtering. This suggests that the reproduction of these cues specific to the physiology of the individual listener is superior to the use of non-individual HRTFs. Consistent with this hypothesis, localization experiments that have shown an increase in localization errors for speech (Møller et al., 1996) and wide-band noise bursts (Wenzel et al., 1991) if generic HRTFs were used for the synthesis. However, Begault et al. (2001) did not find a significant effect of HRTF individualization on externalization with speech.

The experiment by Hartmann and Wittenberg was conducted inside an anechoic room and the possible influence of reverberation was not taken into account. As opposed to their findings, various studies emphasize the importance of reverberation for robust externalization. Sakamoto et al. (1976) found the ratio between indirect and direct sound one of the major factors contributing to externalization. They demonstrated that a monophonic source can be perceived outside the head by adding a time-delayed version of the direct sound to the recording. Durlach et al. (1992) list different studies suggesting a positive relationship between externalization and reverberation. The authors furthermore infer that it ‘seems likely that in many circumstances reverberation is as important as the pinna factor’ (Durlach et al., 1992, p. 255). This is supported by the work of Begault et al. (2001) where, in contrast to individualization, the influence of reverberation on the externalization of speech was found to be positive and significant. The use of *binaural room impulse responses* (BRIRs), capturing not only the reflections of pinnae, head, and torso but also the reflections from the enclosing space, seems a logical consequence. However, Werner et al. (2016a) did find a significant positive effect of both BRIR individualization compared to dummy head BRIRs and the congruence of the simulated with the listening room (regarding reverberation), where both effects were statistically independent. One important question is therefore how important the use of individual BRIRs within a reverberant environment is.

Hassager et al. (2016) investigated how externalization is affected by spectral smoothing of the direct part (HRIR) compared to smoothing of the reverberant part of a BRIR. It was found that only smoothing of the direct part had a negative influence on externalization, whereas no influence of the spectral detail of the reverberant sound was found. This result applies to both frontal and lateral sound, though the externalization of frontal sources was less pronounced in general. Jiang et al. (2020) investigated the influence of the pinnae on externalization. Individual BRIRs were recorded with and without pinnae in a reverberant environment. By crosswise combination of the direct and reverberant parts of the with- and without-pinnae BRIRs, the effect of applying pinna filtering either only to the direct or the reverberant part was studied. For frontal and lateral sound, they demonstrated that pinna filtering of the direct part increased externalization, whereas pinna filtering of the reverberant part had actually a negative effect. Brinkmann et al. (2017) showed that, in reverberant conditions, listeners were less able to discriminate between reality and individual binaural synthesis than in anechoic conditions. Wendt et al. (2019) observed that the importance of HRIR individualization for the externalization of speech diminishes within a reverberant environment independent of prior knowledge of the room, and that externalization can be achieved with generic HRIRs.

Various other reverberation parameters have been studied in the literature. Li et al. (2018) demonstrated by varying the DRR and the BRIR length that reverberation at the ipsilateral ear had less influence on the externalization of lateral white noise than reverberation at the contralateral ear. Yuan et al. (2015) found a greater influence of second-order early reflections on externalization than of late reverberation.

The influence of the length of reverberation on externalization was investigated by Catic et al. (2015). For a frontal and a lateral (30°) sound source, the reverberant part of individual BRIRs was truncated using a variable-length time window. The reverberation outside the window was either replaced with silence or monaural reverberation from one ear and the modified BRIRs were convolved with speech. The results show that truncation of the reverberant part leads to a monotonic decrease of the degree of externalization with decreasing window length, until intracranial localization. The effect is independent of the source direction. A BRIR length of $80 - 100$ ms was found to be sufficient for externalization. Appending monaural reverberation increased the externalization of lateral sources (depending on the length of the time window). For frontal sources, however, monaural reverberation was insufficient at shorter cross-over times ($2.5 - 10$ ms). Further analysis of the presented conditions was carried out regarding possible cues for externalization. In agreement with Catic et al. (2013) where compressing the amount of ILD fluctuations (above 1 kHz) led to a decrease in externalization, it was found that, in turn, the amount of fluctuations of ILDs was in good correspondence with the externalization ratings of the modified BRIRs. Moreover, the increase of the interaural coherence, a measure for the similarity of the signals at both ears, corresponded also well to the decrease in externalization, whereas the DRR was not found to be associated with it.

Many of the above-mentioned studies have followed a *content-based* approach to externalization, i.e., by the variation of parameters of the presented signal content. However, it has been demonstrated that externalization also depends on *context-based* parameters such as expectation, adaptation, and training (Werner et al., 2016a; Klein et al., 2017) or visual cues (Durlach et al., 1992; Udesen et al., 2014). For instance, Udesen et al. (2014) found a significant effect of the variation of the visual appearance of the listening room on externalization using individual HRIR measurements.

The effect of scene rotation in correspondence to head movements is controversial. A positive effect was found by Hendrickx et al. (2017) (different non-individual measurements, with reverberation) and Brimijoin et al. (2013) (individual and non-individual measurements, with and without reverberation), whereas Begault et al. (2001) found no effect (individual and non-individual measurements, with and without reverberation).

Hendrickx et al. (2017) note that the contrary results may be attributed to a short stimulus length (too little time for an effect to establish) and averaging over all directions (the effect may be less pronounced for lateral directions which are well externalized already without head tracking).

Werner et al. (2016a) found that the agreement or disagreement between the synthesized room and the listening room had a major impact on externalization. According to the authors, externalization is supported by the congruence of the two rooms, whereas the so-called *room divergence effect* describes decreased externalization due to a disagreement, or divergence, of the virtual and the real room. Furthermore, they infer that the room divergence effect can be increased or decreased by training to divergent or congruent room combinations. The DRR was identified as one acoustic parameter to characterize the convergence or divergence of two rooms. However, the adjustment of the DRR alone in order to establish convergence was only found to have a small impact on externalization (Werner et al., 2016b).

With increasing research interest in externalization over the past years, a clearer picture of the determining factors has emerged. One important insight is that externalization is a complex multi-dimensional percept and that the possible influence of a certain parameter has always to be regarded within the context.

2 Effect of Subtractive BRIR Modification on Externalization and Sound Quality

This section describes a listening experiment that was conducted to find out how dereverberation using different BRIR modification techniques influences externalization and sound quality. [1] The experiment is motivated by the (to some extent, utopian) question for a way to reduce the perceptive amount of reverberation with minimal detriment to externalization.

Various studies have shown that the presence of reverberation is essential for externalization in different situations (see Sakamoto et al. (1976); Durlach et al. (1992); Begault et al. (2001); Crawford-Emery and Lee (2014); Catic et al. (2013, 2015); Li et al. (2018)). However, it is often desirable to reduce the amount of reverberation to preserve the original sound of the audio material at the input of a binaural decoder. This leads to the question if and to what extent reverberation can be reduced without severely affecting externalization (*subtractive approach*). The question can also be formulated inversely for the synthesis case: how to add only as little reverberation as needed to a binaural impulse response with the largest possible effect on externalization (*additive approach*)?

A subtractive approach brings the advantage that, based on an initial well-externalized condition such as a measured BRIR, parameter variations can be introduced to study the (negative) effects on externalization. Examples can be found in the literature. The truncation of the reverberant part of a BRIR was studied by Catic et al. (2015) who found that externalization increases with BRIR length and that a minimum length of $80 - 100\,\text{ms}$ is required to yield externalized sound images. Moreover, truncating the BRIR at both ears or only at the contralateral ear has a greater negative effect on externalization than at the ipsilateral ear (Li et al., 2018). Crawford-Emery and Lee (2014) observed the emergence of sound colorations with increasing BRIR length.

Another method was studied by Werner et al. (2016b) where it was attempted to adjust the direct-to-reverberant energy ratio (DRR) to counteract the room divergence effect, i.e., a mismatch between the synthesized and the real room. Li et al. (2018) modified the DRR either at both ears or separately at each ear and found a similar effect on the externalization of lateral sound sources as with truncation.

[1] The results of this experiment have been published in a conference paper – *The influence of different BRIR modification techniques on externalization and sound quality* (Giller et al., 2019).

The truncation of an impulse response is purely artificial as there is no equivalent physical phenomenon. It is therefore assumed to likely produce a sound that is also perceived as artificial. In contrast, an increased DRR is usually associated with a reduced distance between source and receiver within enclosed spaces. To extend the set of modification tools, the manipulation of the reverberation time is considered, as well. A decrease of the reverberation time corresponds typically to an increase of the absorption area within a room.

It is expected that step-wise de-reverberation by the application of either of the methods leads to increasing divergence between the synthesized and the real room, and to successively decreasing externalization. It is hypothesized that the use of different methods leads to differences in sound quality considering similarly externalized conditions. The methods were compared in a listening experiment. The results of this endeavor will show that there is a tradeoff between externalization and sound quality, and that the choice of method affects this tradeoff.

The applied modification methods will be explained in detail in the next section. Sec. 2.2 describes the measurement and processing of the BRIR, as well as the general experimental setup. Sec. 2.3 explains the experimental design. The results of the experiment are discussed in Sec. 2.4, followed by the summary of this chapter in Sec. 2.5.

2.1 Studied Modification Methods

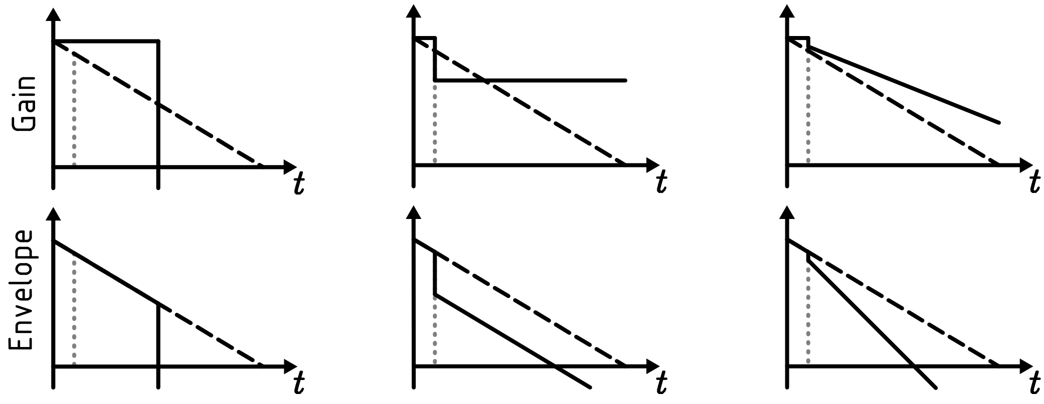
A systematic comparison was made among different modification methods. The reverberant part of a BRIR was manipulated by either

(i) decreasing the impulse response length (truncation),

(ii) increasing the direct-to-reverberant energy ratio (DRR), corresponding to a reduction of the distance between source and receiver inside a room,

(iii) or decreasing the reverberation time, corresponding to the introduction of additional absorption material into a room.

Fig. 1 illustrates the modification methods as the multiplication with a time-varying gain function to shape the distribution of energy. The envelope of the BRIR is represented by a dashed line on the logarithmic scale. The upper row of Fig. 1 shows the simplified gain functions as solid lines, and the lower row illustrates the resulting envelope of the BRIR after modification. It is assumed that the direct sound (the HRIR) and the reverberant tail of the BRIR, including all reflections from the surroundings, are well-separated. Given this premise, the gray dotted line marks the boundary between the direct and the reverberant part.



(a) Truncation (b) DRR (c) Reverberation time

Figure 1: Simplified illustration of the applied modifications. The black and gray dashed lines denote the envelope of the BRIR and the boundary between the direct sound and the reverberant part. The equivalent time-dependent gain function is shown in the upper row, whereas the lower row shows the resulting envelope.

***(i)* Truncation.** Consider a BRIR of length L. As shown in Fig. 1(a), the BRIR truncated to length L'

$$h_{\text{trc},i}(t) = w_{\text{trc}}^{L'}(t) \cdot h_i(t), \tag{1}$$

with $i \in \{\mathsf{L}, \mathsf{R}\}$ denoting the left and right ear channels, is essentially the product of the BRIR and a rectangular window function with unit amplitude

$$w_{\text{trc}}^{L'}(t) = \begin{cases} 1, & t \leq L' \\ 0, & t > L' \end{cases}, \; L' < L, \tag{2}$$

where a cosine-squared window can be applied to the falling edge of the window in order to create a smooth transition. Note that, while this procedure affects the DRR, it does not affect the reverberation time (the slope of the decay) in a diffuse sound field. It is a temporal modification where the energy inside the truncation window stays the same.

***(ii)* Increasing the DRR.** The DRR is defined as the energy ratio between the direct and the reverberant part:

$$DRR = 10 \cdot \lg\left(\frac{\int h_{\text{dir}}^2(t)\text{dt}}{\int h_{\text{rev}}^2(t)\text{dt}}\right). \tag{3}$$

As shown in Fig. 1(b), the DRR can be increased by weighting the reverberant part of the BRIR with a constant factor smaller than one. In order to increase it by a relative difference ΔDRR, the gain factor is $g_{\text{drr}} = 10^{\frac{\Delta DRR}{20}}$ and the modified BRIR is then given by

$$h_{\text{drr},i}(t) = h_{\text{dir},i}(t) + g_{\text{drr}} \cdot h_{\text{rev},i}(t). \tag{4}$$

Obviously, this procedure requires to split up the BRIR into the binaural direct sound $h_{\text{dir},i}(t)$ and the reverberant part $h_{\text{rev},i}(t)$. Consider a BRIR

$$h_i(t) = h_{\text{dir},i}(t) + h_{\text{rev},i}(t) \tag{5}$$

as the superposition of the HRIR $h_{\text{dir},i}$, containing the direct sound and all reflections from pinnae, head and torso, and the reverberant part $h_{\text{rev},i}$ with all reflections from the enclosing space. Given that the BRIR is well separable, the straightforward way to obtain the direct and reverberant parts

$$\begin{aligned} h_{\text{dir},i}(t) &= w_{\text{dir}}(t) \cdot h_i(t), \text{ and} && (6) \\ h_{\text{rev},i}(t) &= w_{\text{rev}}(t) \cdot h_i(t), && (7) \end{aligned}$$

is to apply overlapping windows w_{dir} and w_{rev} with unit amplitude and cosine-squared weighted edges. The window for the reverberant part is given by

$$w_{\text{rev}}(t) = \begin{cases} 0 & 0 \le t < L_{\text{dir}} - \frac{\tau}{2}, \\ \cos^2\left(\frac{(t - L_{\text{dir}} - \tau/2)\pi}{2\tau}\right) & L_{\text{dir}} - \frac{\tau}{2} \le t \le L_{\text{dir}} + \frac{\tau}{2}, \\ 1 & L_{\text{dir}} + \frac{\tau}{2} < t \le L, \end{cases} \tag{8}$$

and the window for the direct part is $w_{\text{dir}} = 1 - w_{\text{rev}}$. The boundary between the direct and the reverberant part is denoted by L_{dir}, and τ is the length of the overlap.

By applying the gain factor to the reverberant part, the envelope of the reverberant part is shifted downwards as shown in Fig. 1(b). The length and the reverberation time remain unchanged. Therefore, it is a purely energetic modification.

(iii) **Decreasing the reverberation time.** The reverberation time T_{30} can be altered by weighting the reverberant tail with an exponential decay curve. This is shown in Fig. 1(c). As the simplified illustration shows, the direct part is unaffected. However, note that the decay curve begins at $t = 0$ with a value of 1. It is again required to split up the BRIR into its direct and reverberant parts as explained above. The BRIR with altered reverberation time T'_{30} can be computed with

$$h_{\text{dec},i}(t) = h_{\text{dir},i}(t) + w_{\text{dec}}^{T'_{30}}(t) \cdot h_{\text{rev},i}(t) \tag{9}$$

where the exponential decay curve is given by

$$w_{\text{dec}}^{T'\,30}(t) = 10^{-\frac{60}{20}\left(T'^{-1}_{30} - T^{-1}_{30}\right)t}. \tag{10}$$

As can be seen in Fig. 1(c), the envelope of the modified BRIR differs from the original envelope by a steeper slope. This procedure necessarily increases the DRR. While the length of the BRIR technically remains unchanged, the effective (perceived) length can decrease. Altering the reverberation time can thus be regarded as a mixed temporal-energetic method.

2.2 Experimental Setup

The modification methods described above are applied to a BRIR to generate conditions for the listening experiment. The conditions were created based on a BRIR measured for a single frontal direction in a lecture room of the IEM. The BRIR was modified using the above explained methods, and stimuli were created by convolving the modified impulse responses with speech. This section describes every step of the processing chain from the measurement of the BRIR to the playback of the conditions.

2.2.1 Measurement and Processing of the BRIR

Externalization was studied on the example of a single sound source directly in front of the listener. Experience has shown that externalization is more difficult to achieve for frontal than for lateral directions. While externalization is usually less problematic at lateral positions, the phenomenon is also more complex due to differences between the ipsilateral and the contralateral ear (Li et al., 2018).

The BRIR was measured in lecture room PT116 of the IEM. The room has the dimensions 7 m × 8.3 m × 3 m, and a reverberation time of $T_{30} = 0.7\,\mathrm{s}$. Note that the reverberation time was not determined with an omni-directional microphone, but was, instead, derived based on the mean decay curve of the microphones of the dummy head used for the measurement.

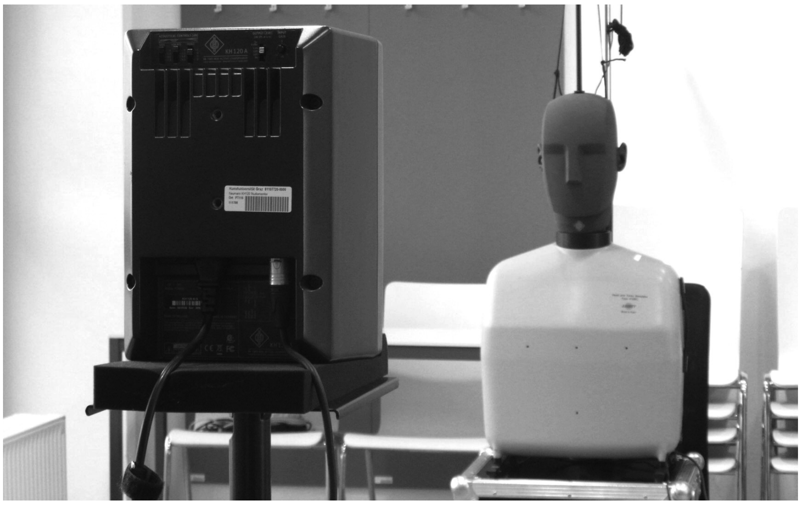


Figure 2: Measurement setup. The loudspeaker was placed directly in front of the Neumann KU 100 dummy head on top of a Brüel & Kjær torso.

A Neumann KU 100 dummy head was used as a receiver, and a Neumann KH 120 loudspeaker was used as a sound source. The sound source was positioned at a distance of 2.5 m to the receiver at a height of 1.25 m directly in front of the receiver with an angular displacement of 0° in azimuth and elevation. The dummy head was equipped with the torso of a Brüel & Kjær HATS in order to provide reflections from breast and torso that may support externalization (Fig. 2). Impulse responses were measured using the swept-sine method (cf. Farina (2000)) with a single exponential sine sweep of length 10 s. To take account of any disturbance due to background noise, multiple runs were carried out and the impulse response with the highest signal-to-noise ratio was selected.

For the measured BRIRs to be separable into the binaural direct sound (HRIR) and the reverberant part, the travel path of the first reflection from the surroundings must be sufficiently long. All items of furniture in the immediate vicinity of the measurement were thus removed so that the first reflection from the surroundings came from the floor between source and receiver. The length of the travel path of the first reflection was approx. 3.5 m, leading to a difference in travel time of 3 ms to the direct sound.

The direct and the reverberant part were separated using overlapping cosine-squared windows (cf. Eq. 8). The magnitude of the measured BRIR as well as the time windows are shown in Fig. 3. For splitting up the BRIR immediately before the arrival of the first reflection, the window length was set to $L_{\text{dir}} = 3\,\text{ms}$ and the overlap to $\tau = 0.5\,\text{ms}$; with the peak of the direct sound at $t_0 = 0.6\,\text{ms}$ this yields a resulting time interval of $L_{\text{dir}} + \frac{\tau}{2} - t_0 = 2.65\,\text{ms}$ for the HRIR to decay.

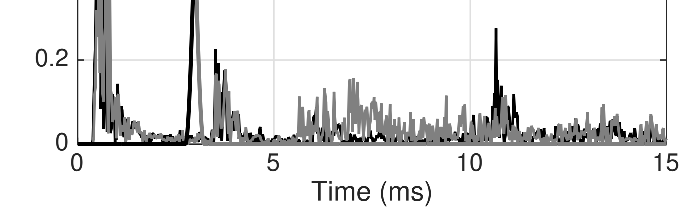


Figure 3: Magnitude of the left and right channels h_{L} and h_{R} of the measured BRIR, and the time windows w_{dir} and w_{rev} used to separate direct and reverberant sound.

2.2.2 Conditions

The three modification methods were applied to the measured BRIR. Conditions were created by varying each parameter (L, DRR, and T_{30}) separately such that the whole range of each parameter from reverberant to anechoic conditions was sampled. In other words, the length of the impulse response and the reverberation time were decreased, and the DRR increased until the same anechoic condition was reached. As mentioned earlier, not every parameter can be accessed independently from the others. For example, truncation leads to increased DRR. Reducing the reverberation time will also lead to an increased DRR and, depending on the strength of the modification, to a reduction of the effective (audible) length of the impulse response. Such effects were not compensated for in this experiment.

Tab. 1 lists the values of the respective varied parameter of each condition. All conditions have the measured BRIR as a starting point. The parameters were then varied successively within each method towards anechoic conditions. The left and right ear signals underwent the same processing. Finally, the HRIR is the last condition in which all modification methods meet. Ergo, the first and the last condition are identical between the methods.

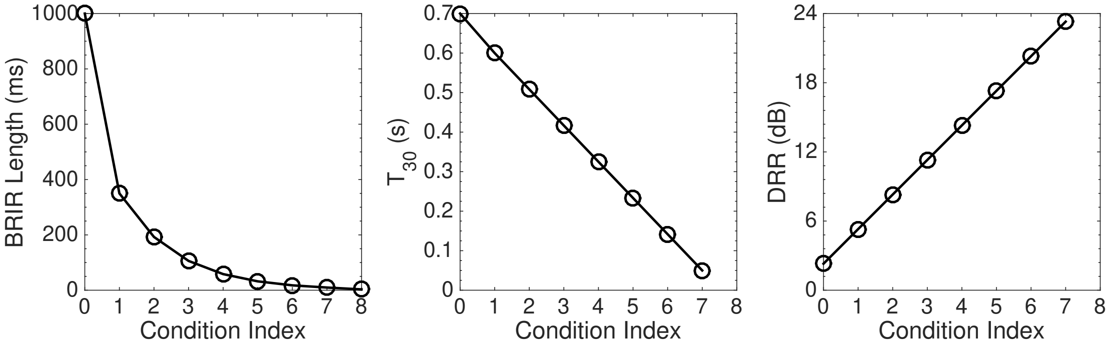
Method		**Truncation**	**Decay**	**DRR**
Parameter		L (s)	T_{30} (s)	DRR (dB)
Condition i	0 (BRIR)	1.0	0.70	2.3
	1	0.350	0.60	5.3
	2	0.193	0.51	8.3
	3	0.106	0.42	11.3
	4	0.059	0.33	14.3
	5	0.032	0.23	17.3
	6	0.018	0.14	20.3
	7	0.009	0.05	23.3
	8 (HRIR)	0.003	-	∞

Table 1: List of parameter values used to create the i-th condition.

The levels of each parameter were selected based on previous informal listening to yield uniform sampling of the range between convergence and divergence of the synthesized and the real room. As Fig. 4 shows, the length of the truncation window follows the heuristically determined function

$$L'_n = 0.55 \cdot L'_{n-1} \tag{11}$$

with $n = 1, ..., 8$ and $L'_1 = 0.35\,\mathrm{s}$ for conditions 1 to 8. The reverberation time T_{30} and the *DRR* were varied linearly, beginning with the parameters of the unmodified BRIR, as shown in Fig. 4(b-c).



(a) Truncation (b) Reverberation time (c) DRR

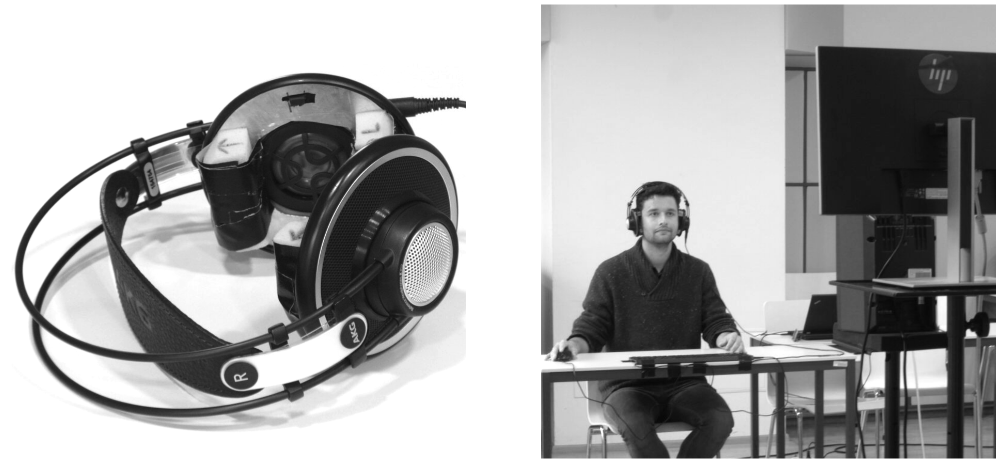
Figure 4: Visualization of the parameter variations.

Stimuli were created by the convolution of an 8 s long sequence of anechoic male speech with the impulse responses corresponding to each of the conditions. Speech has been used in different experiments studying externalization, such as in the studies by Begault et al. (2001); Catic et al. (2015); Hendrickx et al. (2017); Werner et al. (2016a); Wendt et al. (2019). Furthermore, Wisniewski et al. (2012) showed that subjects could judge the distance to a sound source more accurately if they were familiar with the sound, as it is the case with speech. It is therefore assumed that the familiarity of speech makes it also suitable for assessing externalization and that it ensures comparability with the literature.

2.2.3 Playback and Equalization

The stimuli were played back through headphones in comparison to the loudspeaker. The loudspeaker, in turn, played back the original speech sequence as a reference. The loudspeaker remained in the same position as in the measurement and the listeners took the place of the dummy head.

To the comparison, the subjects wore headphones throughout the whole experiment. Unfortunately, the presence of the headphones alters the transfer function from the loudspeaker to the ears. Especially high frequencies are attenuated by the presence of headphones as the sound has to propagate through the ear cups. As a countermeasure,



(a) (b)

Figure 5: Open headphones with modified ear cushions (a) and a participant wearing the modified headphones during the experiment (b).

a widely used open over-ear headphone (AKG K702) was reconfigured with improvised self-made ear cushions. The ear cushions had cutouts at the front ad back, as shown in Fig. 5, in order to reduce the attenuation of high frequencies for frontal incident sound.

Equalization of the loudspeaker transfer path. To investigate the effect of the headphones, far-field responses of the path from the loudspeaker to the microphones of the dummy head were measured inside an anechoic room using exponential sweeps. One measurement was carried out under free-field conditions, and another with the dummy head wearing headphones. Responses were measured for both the original and the modified headphones. The obtained magnitude frequency responses were smoothed within critical bands, and the average of the left and right channels was taken. The resulting frequency responses are shown in Fig. 6. The frequency response with the original headphones (*light black curve*) exhibits a severe attenuation at high frequencies in the range of 20 dB relative to the free-field response (*bold black curve*). The attenuation is greatly reduced after modification of the ear cups (*gray curve*).

To reduce remaining timbral differences of the loudspeaker signal with and without headphones, the headphone signal was equalized to the sound of the loudspeaker. Dividing the with-headphones magnitude frequency response by the without-headphones magnitude frequency response yields only the contribution of the headphones to the transfer function from the loudspeaker to the dummy head (*dashed black curve*). The equalization filter for

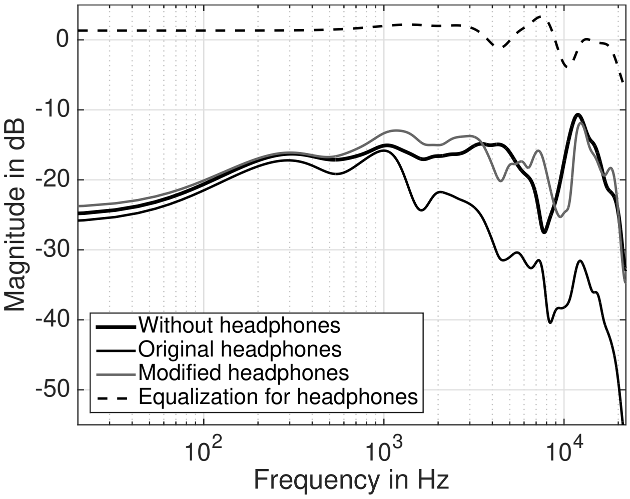


Figure 6: Frequency responses of the headphones, and equalization for frontal sound. Compared to the response without headphones (*bold black line*), the original headphones attenuate high frequencies severely (*light black line*). The attenuation is reduced after modifying the headphones (*gray line*). The equalization filter (*dashed black line*) for the headphone signal further reduces differences in sound.

the headphone signals in the time domain was finally obtained from this by constructing a minimum-phase FIR filter from the magnitude response. Note that, however precise the equalization may be, differences in sound can not be completely removed in practice since the measurement of the BRIR and the equalization were carried out on the dummy head and not individually for each subject.

Equalization of the headphone drivers. To investigate the effect of the modification of the ear cushions, the frequency response from the headphone drivers to the microphones inside the ear canals was measured. As before, the magnitude frequency responses were smoothed and the left and right responses were averaged. The modification of the ear cushions distorts the frequency response of the headphone itself significantly. As shown in Fig. 7, the frequency response of the headphone drivers (*light black curve*) exhibits a drop by about 15 dB/decade towards low frequencies below 1 kHz, and several resonances at mid and high frequencies. Thus, a second equalization filter was needed to linearize the transfer function of the headphone drivers.

A time-domain minimum-phase equalization filter for the headphone signals was derived from the regularized inverse of the magnitude transfer function of the headphone drivers. The *bold black curve* in Fig. 7 is the frequency response of the headphone drivers after equalization. It is reasonably linear below 10 kHz.

A positive side-effect of the equalization is that the undesired contribution of the pinnae, ear canals, and microphones of the dummy head is eliminated. However, in the practice case where the headphones were worn by real humans, the equalization filter can only function imperfectly due to variations in the anatomy of the listeners.

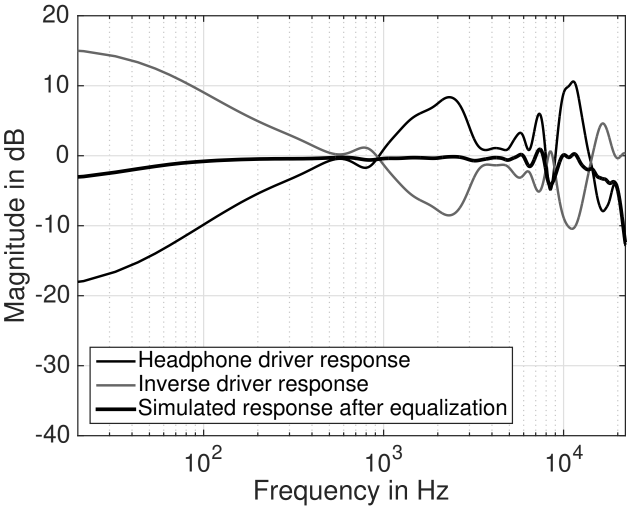


Figure 7: The modification of the headphones leads to a distorted frequency response of the headphone drivers (*light black line*). Applying the inverse response of the drivers (*gray line*) to the headphone signals yields the equalized response (*bold black line*).

The use of acoustically reflective materials for the modified ear cushions (aluminum, foam wrapped with tape) due to the simplicity of construction likely contributed to the observed resonances in Fig. 7. Because the equalization filter was derived for the dummy head and not individually for each subject, the location and strength of the resonances may change when the headphones are worn by humans, compromising the effectiveness of the equalization. But, because of the concentration of the energy of speech in the mid-frequency range, this is expected to be negligible since the resonance at 1 kHz is rather wide, and the narrow resonances lie above 6 kHz.

2.3 Experimental Design

This section discusses the design of the listening experiment. The conditions created in Sec. 2.2.2 were evaluated regarding externalization and sound quality in different parts of the experiment. Firstly, the externalization of the conditions was rated in comparison to the loudspeaker playing back the original speech sequence. Secondly, the sound quality was evaluated by assessing the similarity to a reference signal. The reference was defined as the anechoic binaural signal. This shifts the definition of sound quality towards sound neutrality. The reference signal was generated by convolving the anechoic speech sequence with the HRIR, i.e., the divergent final condition of each modification. To achieve a more fine-grained assessment of sound quality, similarity had to be rated not only in a general sense but also with regard to sound color and the amount of reverberation. Furthermore, the subjects were asked to rate the perceived naturalness of the conditions.

2.3.1 Paradigm

The conditions under test sample the range from convergent to divergent conditions depending on the varied parameter of each modification method. The conditions were rated regarding externalization, and the above explained approaches were used to assess sound quality. The comparison was carried out both within each method and between conditions of all modification methods. A reference stimulus was presented in each experiment, except for the evaluation of naturalness. The conditions were rated in blinded and randomized trials, each of which presenting a set of stimuli to be rated in comparison to the reference (if given). The test method is based on a multiple stimulus comparison with hidden reference and anchor (MUSHRA[2]). Minor deviations from the MUSHRA paradigm have been made that will be discussed further below.

2.3.2 Interface

Each set of stimuli was presented on a graphical user interface. The different trials were presented as consecutive pages, each containing a set of stimuli in randomized order with alphabetical labels. The subjects could freely start and stop playback in an infinite loop. During playback, they could switch between the conditions and the reference. The ratings were to be entered on vertical sliders. A labeled rating scale was located next to the sliders. To provide a better overview, the conditions could be sorted by rating. The participants were able to take as much time as needed for completing the experiment, and they were encouraged to take breaks in between. The interface was displayed on a screen above the

[2] see ITU Recommendation BS 1534-1 (2003)

loudspeaker (see Fig. 5(b)) so that the subjects did not need to turn their heads during the experiment, and the path of the direct sound and the first reflection remained free.

2.3.3 Procedure

As explained above, the listening experiment is divided into the following three parts:

N naturalness according to the subjects' inner reference,

E externalization compared to speech played back through the loudspeaker, and

S sound quality by means of similarity to speech convolved with the HRIR.

The outline of the experiment shown in Tab. 2 in detail. For each part, the table shows which conditions of which modification method are presented together on one page, as well as the reference, hidden reference, and the anchor stimulus. The individual conditions are identified by the condition index i referring to the list of conditions in Tab. 1. An increasing condition index corresponds to a greater modification depth and, thus, decreasing reverberation, where $i = 0$ is always the unmodified BRIR and $i = 8$ is always the HRIR. The table also provides information about the room in which each part was conducted. The tasks for each part are listed in Tab. 3 in the same wording as shown on the graphical interface.

Due to the large number of conditions and time constraints, it was not possible to compare all conditions with each other. Therefore, only a subset of three conditions with index $i = \{2,4,6\}$ of each method was presented in part N. For parts E and S, it was chosen to follow a two-stage procedure: an *indirect* comparison in stage E.I/S.I followed by a *direct* cross-comparison in stage E.II/S.II. First, all conditions were rated within each modification method in the indirect comparison. Three sets of stimuli were presented to the participants in random order, each containing all conditions corresponding to one column of Tab. 1. This allows to draw conclusions on the rating of conditions within one modification technique. But a comparison between the methods can only be made indirectly via the common reference and anchor which is not considered to be very reliable. It was therefore complemented with another page comprising a subset of three conditions from each method for cross-comparison. The conditions with indices $i = \{2,4,6\}$ were selected based on prior informal listening in order to include conditions with low, moderate, and high levels of reverberation. The results of the direct comparison will also be used to scale the results of the indirect comparison in order to obtain corrected ratings for the complete set of conditions.

Part		**Method** (Abbreviation)	**Reference**	**Hidden Reference**	**Conditions** *Cond. index i*	**Anchor**	**Listening Room**
N		all	-	0	2,4,6	8	original
E	I	Truncation (trc)	LS	0	1-7	8	original
		Decay (dec)	LS	0	1-7	8	
		DRR (drr)	LS	0	1-7	8	
	II	all	LS	0	2,4,6	8	
S	I	Truncation (trc)	8	8	1-7	0	anechoic
		Decay (dec)	8	8	1-7	0	
		DRR (drr)	8	8	1-7	0	
	II	all	8	8	2,4,6	0	
	IIa-b	all	8	8	2,4,6	0	

Table 2: Parts of the experiment along with the presented conditions. Condition indices are given corresponding to Tab. 1. The BRIR and the HRIR have the indices 0 and 8; *LS* stands for the loudspeaker.

Part		**Task**
N		'Rate the naturalness!'
E	I-II	'Rate the externalization, compared to the reference!'
S	I-II	'Rate the similarity to the reference *in general*!'
	IIa	'Rate the similarity regarding *sound color*!'
	IIb	'Rate the similarity regarding the *amount of reverberation*!'

Table 3: Task definitions of each part of the experiment.

Part *N*: Perceived naturalness. This first, introductory part was about finding out how natural the participants perceived the conditions. It also served the purpose of familiarizing the subjects with the stimuli. The presented set of conditions was the same subset that was also rated regarding externalization in part *E*.II. The perceived naturalness had to be rated with no reference available for comparison. Instead, participants should frankly rate the stimuli according to their (undefined) inner reference. This part was conducted in the original room, but the participants were instructed that naturalness not necessarily bound to the present listening situation. It was furthermore left open if an anechoic signal can be considered natural. The rating had to be entered on an integer scale between 0 and 100, where 0 corresponded to 'very unnatural', and 100 to 'entirely natural'.

Part *E*: Externalization. Part *N* was immediately followed by the second part in which the subjects were asked to rate the degree of externalization in comparison to the reference. The rating scale was similar to the previous part, whereas now 0 was defined as 'inside the head', and 100 as 'at the position of the loudspeaker'. An additional label 'close to the head' was displayed at one third of the scale range for orientation. As this part was divided into the indirect comparison E.I and the direct cross-comparison E.II, it comprised four pages in total that were presented in random order.

Just like naturalness, externalization was evaluated in the original room at the position of the dummy head in the measurement. The loudspeaker as visible and positioned at the same position as in the measurement. It played back the dry speech signal as the reference. The subjects were instructed to directly face the loudspeaker and avoid any head movements during playback. They were aware that the reference is played back through the loudspeaker.

The loudspeaker was used as a reference to connect the rating criterion to the externalization of a well-externalized compact physical sound source as it would be undefined otherwise. Speech convolved unmodified BRIR (condition 0) was used as a hidden quasi-reference. The anchor was speech convolved with the HRIR (condition 8).

Part *S*: Similarity. After the first two parts were completed, each participant was asked to proceed to the anechoic chamber. Their next task was to rate the same conditions regarding their similarity to the reference. The speech signal convolved with the anechoic HRIR (condition 8), presented through headphones, was used therefor and as the hidden reference. The BRIR (condition 0) was used as the anchor since it contains the largest amount of reverberation of all conditions. This is an indirect approach to assessing sound quality. Here, the definition to sound quality is close to neutrality since it is desired to contribute as little reverberation and sound colorations as possible to the input signal.

The rating scale was defined similarly to the previous two parts. In the present case, 0 corresponds to ‘very different’ from the reference, and 100 to ‘identical’ to reference. As before, similarity was rated within-methods in the indirect comparison S.I, and between-methods in the direct comparison S.II. Similarity was to be rated in a general sense and was not closer specified.

The direct comparison was repeated twice with a variation of the task in order to achieve a more nuanced analysis. The subjects were asked to rate the similarity regarding *sound color* in part S.IIa, and regarding the *amount of reverberation* in part S.IIb. Therefore, part S comprised six pages in total which were presented in random order.

This part was, in contrast to the previous parts, conducted inside an anechoic room to minimize the influence of the spatial properties of the listening room. It is, furthermore, consistent with part E to conduct the experiment in an acoustic environment that is close to the reference being used. The reference as well as all stimuli were played back through the modified headphones. But, in order to prevent that any spatial attributes influence the ratings, all conditions including the reference were presented diotically by playing back the left channel to both ears.

2.4 Results

This section discusses the experimental results. The first part of the experiment was carried out in the lecture room of the institute. First, the subjects had to rate how natural they perceived a reduced set of conditions from all methods subject to their inner reference. Then they were asked to rate externalization in comparison to the loudspeaker. The conditions were compared both within-methods with the complete set of conditions and between-methods with a reduced set. These trials were presented in random order. In the second part, they had to rate the similarity to an anechoic reference (the HRIR convolved with speech). This second part was conducted within an anechoic chamber with the stimuli presented through headphones. The presented sets of conditions were identical to the externalization part, but all stimuli were presented diotically, and the reference was played back through headphones.

The experiment was carried out with twenty-one participants, most of them trained listeners with previous experience in listening experiments. All participants were able to complete the test in not much more than one hour.

The general data situation and the applicable statistical methods will be discussed in Sec. 2.4.1. In Sec. 2.4.2, an overview of the results of externalization and overall similarity to the reference is given. Sec. 2.4.3 and Sec. 2.4.4 focus on the ratings of the between-methods comparison, considering also naturalness and the similarity regarding sound color and reverberation. In Sec. 2.4.5, the created conditions are related to different physical quantities in order to find a common axis. The findings of Sec. 2.4 are discussed in Sec. 2.4.6.

For the sake of brevity, the manipulation of the BRIR length (truncation), the decay curve, and the DRR will be referred to `trc` , `dec` , and `drr` , respectively, followed by the condition index *i* to indicate a particular condition (cf. Tab. 1). For instance, `drr 3` refers to the third condition of the DRR modifications.

2.4.1 Statistical Methods

The interpretation of the experimental results is based on hypothesis-tests and graphical displays of the average ratings. In either case, the choice of statistical methods is essential for drawing valid conclusions. A choice has to be made whether to use parametric methods for testing and display, which are based on the assumption of a defined distribution of the data, or non-parametric methods making fewer assumptions, or none at all. Beyond that, different methods require different levels of measurement such as ordinal (ranking variable), interval (metric variable with equal differences), or ratio scale (metric variable with equal differences and absolute zero point). Distribution and level of measurement are not only influenced by the studied phenomenon but also depend on the experimental design. In the following, the given data situation is discussed and a decision is made about the methods to be used.

Normal distribution. It is not clear whether the population from which the samples were drawn follows a normal distribution. For a perceptive quantity, it can not be assumed in general that it is normally distributed; this may be particularly true for a percept as complex as externalization. Assuming that it does, it is still likely that the experimental design adopted from the MUSRHA method inherently distorts the distribution. The limitation of the measurement scale at both ends can lead to a skewed or cut-off distribution due to the *ceiling effect*. The presence of a high-quality hidden reference and a low-quality anchor reinforces the ceiling effect (Mendonça and Delikaris-Manias, 2018).

A Lilliefors test for normal distribution (Lilliefors, 1967) was carried out on the ratings of each stimulus. Tab. 4 lists the average percentage of cases in which the null hypothesis *'normally distributed'* (H_0) was rejected in each part of the experiment. The H_0 was rejected in 39% of cases on average; the highest rejection rate was in part S.IIa with 55%. It must be concluded that the given data should not be analyzed based on the assumption of normally distributed data. Non-parametric methods may be used instead.

	Part	**No.**	**Mode of Comparison**	**H_0 rejected** (Mean % of cases)
N	Naturalness	–	Between-methods	36
E	Externalization	I	Within-methods	22
		II	Between-methods	27
S	Similarity	I	Within-methods	48
		II	Between-methods	36
		IIa	Between-methods	55
		IIb	Between-methods	47
Mean				39

Table 4: Percentage of cases in which the the H_0 *'normally distributed'* has been rejected.

Level of measurement. The scale of the present experiment is at least of ordinal level. Rating externalization in comparison to a reference stimulus can be understood as estimating the absolute magnitude of the response variable on a quasi-continuous scale. In theory, all requirements are met to regard the data on an interval scale. But before accepting the hypothesis of interval-scaled data, it has to be put into question if the subjects were able to rate the conditions on an interval scale. More precisely, it is questionable if *(i)* the subjects themselves were able to make an absolute judgment, and if *(ii)* the experimental procedure enables them to do so. For the MUSHRA paradigm, it has been shown by Zielinski et al. (2007) that the distribution of differences in quality among the presented stimuli has a substantial influence on the ratings and may lead to a bias. The impact of this effect can be assumed to grow with increasing number of conditions on each page, and the ratings can be assumed to be more interdependent and less accurate.

Since several stimuli were compared multiple times on different pages, it can be investigated whether the ratings deviate between the different pages. The BRIR, the HRIR, and conditions $i = \{2,4,6\}$ of all modification methods have each been rated two or more times regarding the same task each (cf. Tab. 2). A Friedman test was conducted for each condition that appears multiple times to test for an effect of the page on which the respective stimulus was presented. The Friedman test is a non-parametric alternative to the Analysis of Variance (ANOVA). The results of the test show at a significance level of 5% that the BRIR was rated significantly different between the indirect comparisons of truncation and the modification of the DRR ($p = .02$) regarding externalization. Also, condition `trc 4` received a significantly different rating between the indirect and the direct comparison ($p = .04$). The similarity ratings differ three times significantly between the indirect and the direct comparison for condition `drr 4` ($p = .04$), `trc 6` ($p = .0003$), and `dec 4` ($p = .02$). The presence of deviations between the conditions is an argument against the hypothesis of absolute magnitude estimation and in favor of the interpretation on the ordinal level.

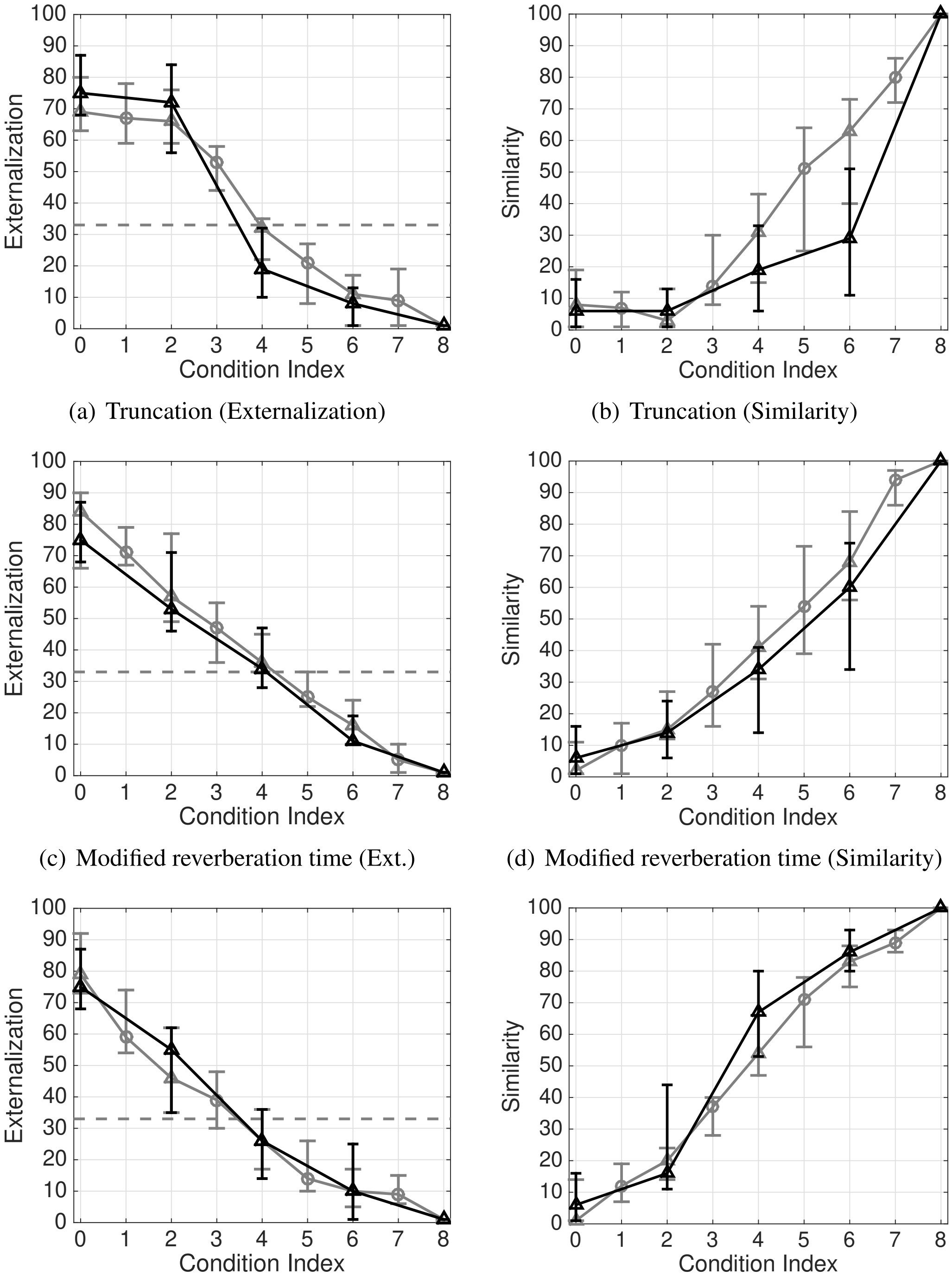
Hypothesis testing and visualization. The above findings show partial non-normality of the data and violations of the hypothesis of direct magnitude estimation. This suggests that the analysis of the experiment should be carried out using non-parametric methods on an ordinal scale. Suitable non-parametric methods for hypothesis testing with the MUSHRA paradigm are the Friedman test as an alternative to the ANOVA, and the Wilcoxon signed-rank test as an alternative to the paired-samples t-test (Mendonça and Delikaris-Manias, 2018). Differences are considered significant if the p-value is below the significance level which will be defined as $\alpha = 0.05$. Multiple paired comparisons lead to an inflation of the α or type I error (false positives). The obtained p-values were therefore corrected to retain a global α level of $\alpha = .05$ using the Bonferroni-Holm method (Holm, 1979). For visualization of the results, the median and the 95 % confidence intervals of the median will be displayed.

Preliminary analysis. The basic prerequisite for the externalization experiment is that the BRIR convolved with speech was perceived as externalized by all subjects. It must be taken into account that the HRIR of the dummy head may not provide sufficient externalization for some of the subjects. Thus, a rating of the BRIR below the 'close-to-the-head' mark was defined as an exclusion criterion. The externalization ratings of every subject were examined and it was found that all subjects perceived the BRIR clearly outside the head. Hence, no subjects had to be excluded.

A Friedman test was carried out to test for the general effect of the 'treatment', i.e., the BRIR modification among stimuli. The effect of the modification is significant with $p < .05$ within each part of the experiment. It remains significant if the anchor (condition 0 or condition 8, respectively) is excluded from the test. Therefore, there is a significant difference between at least one pair of conditions in every trial. As it is unclear which conditions differ significantly, the Wilcoxon signed-rank test was used to perform pair comparisons.

2.4.2 Overview of the Results

Fig. 8 shows the results for externalization (E.I-II) and general similarity to the reference (S.I-II). Each row of Fig. 8 displays the results for one of the three methods, where the externalization is shown in the left, and similarity is shown in the right column. The median ratings of the indirect comparison (E.I and S.I) are plotted with the 95 % confidence intervals of the median in gray in the background. On top, the median ratings of the direct cross-comparison (E.II and S.II) are plotted in black. All conditions that were tested both in the indirect and the direct comparison are highlighted by triangular markers. The condition index i is plotted on the horizontal axis (see Tab. 1 for the list of conditions).



(a) Truncation (Externalization)

(b) Truncation (Similarity)

(c) Modified reverberation time (Ext.)

(d) Modified reverberation time (Similarity)

(e) Modified DRR (Externalization)

(f) Modified DRR (Similarity)

Figure 8: Median ratings and 95 % confidence intervals for externalization (E.I-II) on the *left* and similarity to the reference (S.I-II) on the *right*, linearly interpolated. The ratings of the indirect comparison within-methods are drawn in *gray*, whereas the ratings of the direct comparison between-methods are drawn in *black*. Triangular markers denote common conditions. The condition indices refer to Tab. 1. The position of the ‘close to the head’ label is indicated by a horizontal dashed line.

Increasing i corresponds to increasing modification depth which, in turn, corresponds to decreasing amount of reverberant energy.

By and large, the relation between the modification depth and both externalization and similarity is monotonic for all modification methods. Externalization decreases with increasing modification depth while the similarity to the anechoic reference increases.

Externalization. As expected, the BRIR (condition 0) received the highest externalization rating within each method, whereas the HRIR (condition 8) is not externalized. The differences between all conditions of the direct comparison – including the BRIR and the HRIR – are significant within each modification method. The median ratings of the BRIR lie between 70 and 80.

In the case of the `dec` and `drr` methods, the rating curves of the direct and indirect comparisons essentially follow a straight line. The curves of the direct comparison show only minor deviations from the indirect comparison. In the case of `trc`, however, the ratings of the indirect comparison rather follow an inverse sigmoid shape that appears even more pronounced in the direct comparison. However, this is not a result *per se* as it merely reflects the choice of parameters.

The horizontal dashed line in the externalization plots marks the position of the label 'close to the head' at one third of the rating scale. Only conditions exceeding this threshold will be considered externalized. According to this definition, the minimum BRIR length required for externalization is $L = 59\,\mathrm{ms}$ (condition 4) based on the indirect comparison. In the direct comparison, the transition towards perception outside the head takes place between $L = 59\,\mathrm{ms}$ and $L = 106\,\mathrm{ms}$ (condition 3). If the reverberation time is modified, the threshold is exceeded at condition 4 corresponding to a reduction of the reverberation time by 0.4 s. If the DRR is modified, condition 3 is just above the threshold corresponding to an increase of the DRR by 9 dB.

Similarity to the reference. The hidden reference (condition 8) was correctly recognized in all cases and received the highest rating. The rating decreases within all three methods with increasing reverberation. In the case of the modification of the reverberation time and DRR, the BRIR received the lowest rating. With the truncation method, the first two conditions `trc 1` and `2` received a very low rating as well. Regarding `dec` and `drr`, the direct comparison hardly leads to a change of the curve of the indirect comparison. With trc , however, the ratings differ more distinctly. While the curve of the indirect comparison essentially describes a straight line between trc 2 and 8, condition 4 and 6 are shifted downwards in the direct comparison by 10 and 30 points leading to a rather exponentially shaped curve.

2.4.3 Cross-Comparison Between Methods

This section discusses the ratings obtained in the direct comparison. In contrast to the ratings of the indirect comparison presented in the previous section, they were obtained in a cross-comparison between the methods.

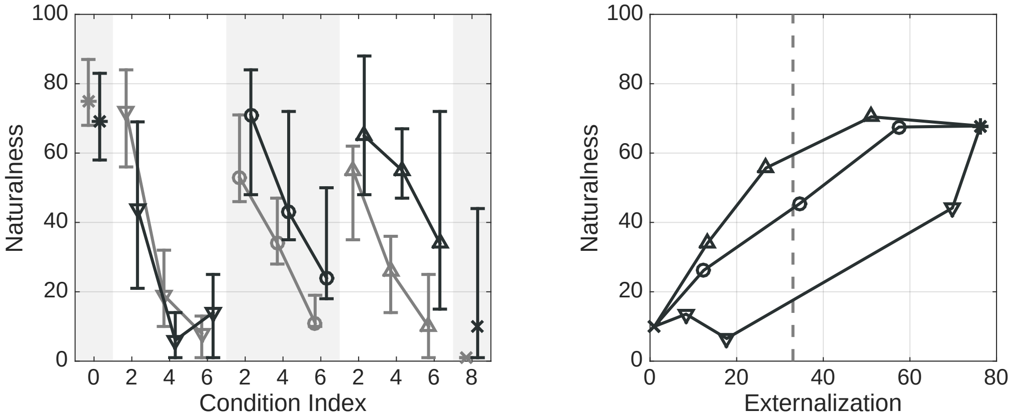
Based on the direct comparison, the ratings can be interrelated between the methods. However, note that a comparison based on the condition index is not meaningful due unrelatedness of the varied parameter between the methods. Instead, naturalness (N), general similarity (S.II), similarity regarding sound color, and regarding the amount of reverberation (S.IIa-b) are regarded relative to the corresponding externalization ratings (E.II). The results of each part are discussed on the basis of Fig. 9-12, each of which showing the same data from two different perspectives.

On the left, naturalness or similarity to the reference are plotted in the foreground with the externalization ratings in the background. The figure shows the median ratings along with the 95 % confidence intervals of the median.

On the right, naturalness or similarity are plotted on the vertical axis with externalization on the horizontal axis to offer a more intuitive view. The figure shows the geometric median[3] of the two-dimensional data set which is the point that minimizes the Euclidean sum of distances to all other points. The data points are linearly interpolated in order to make trends visible.

Perceived naturalness. Fig. 9 shows the naturalness ratings. The conditions were rated in the original room without a reference stimulus available. As can be seen in Fig. 9(a), naturalness shows a tendency to increase with increasing reverberation, similarly to the tendency of the externalization ratings. This decrease is monotonic within each method except for trc 2 and trc 4 where monotony seems to be reversed, albeit these conditions do not differ significantly.

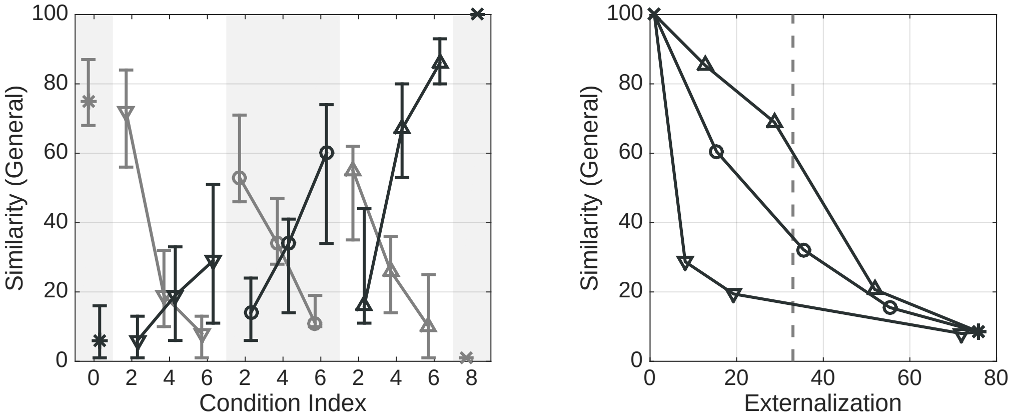
[3] Note that in (Giller et al., 2019, Fig. 5) each data point corresponds to the coordinate given by the individual medians of the two datasets, rather than the geometric median. This leads to minor differences in the location of the data points.



(a) Median naturalness ratings (black) compared to externalization (gray).

(b) Geometric median of naturalness and externalization.

Figure 9: Comparison of perceived naturalness (N) and externalization (E.II). Median ratings and 95 % confidence intervals, linearly interpolated. The dashed vertical line in (b) denotes the 'close-to-the-head' threshold. *Legend:* $\triangledown$ *Truncation,* $\bigcirc$ *Decay,* $\triangle$ *DRR,* $*$ *BRIR,* $\times$ *HRIR.*



(a) Median similarity ratings (black) compared to externalization (gray).

(b) Geometric median of similarity and externalization.

Figure 10: Comparison of similarity to the reference (S.II) and externalization (E.II). Median ratings and 95 % confidence intervals, linearly interpolated. The dashed vertical line in (b) denotes the 'close-to-the-head' threshold. *Legend:* $\triangledown$ *Truncation,* $\bigcirc$ *Decay,* $\triangle$ *DRR,* $*$ *BRIR,* $\times$ *HRIR.*

The ratings of the truncated conditions occupy a lower range than the ratings of the other methods: The highest rating is significantly lower than the highest rating of the DRR and decay modification methods. Plotting naturalness over externalization in Fig. 9(b), it is noticeable that the curves for the decay and the DRR method are rather similar, while the curve for truncation deviates downwards. The truncated conditions are all not significantly different from the HRIR (condition 8), however, all conditions of the other methods are.

Moreover, all truncated conditions were significantly rated less natural than the BRIR (condition 0). With the other methods, only the 6th condition received a significantly lower rating than the BRIR. The confidence interval of the HRIR indicates a high variation in the ratings which, in turn, indicating a disagreement among the subjects whether the anechoic conditions sound natural or not.

Similarity to the reference. Fig. 10 shows the ratings of similarity to the HRIR convolved with speech in comparison to the externalization ratings. In this part, similarity had to be rated in a general sense. Fig. 10(a) shows that the similarity increases monotonically within the ratings of each method with decreasing amount of reverberant energy (increasing condition index). The differences between the three parameter levels of each method are significant.

Except for `trc 2`, all conditions differ significantly from the BRIR. The first condition (condition 2) received the highest externalization ratings (Fig. 10(a), background) with each method. Furthermore, `trc 2` was rated as better externalized than all other conditions, except for the BRIR. Accordingly, the similarity rating of `trc 2` is also lower than of all other conditions. A general decrease of similarity to the HRIR was expected with increasing externalization. However, there are differences between the three methods. Regarding similarity, `dec 4` was rated higher than `trc 4`, although it was also rated higher regarding externalization. The same can be observed for `drr 4` and `trc 6`.

The two-dimensional plot of similarity and externalization in Fig. 10(b) shows that the trajectory of the DRR modification indicates a rather linear tradeoff between similarity and externalization. The trajectories for the decay and the truncation method deviate towards the lower-left corner (low externalization and low similarity). In the truncation curve, this deviation is most pronounced.

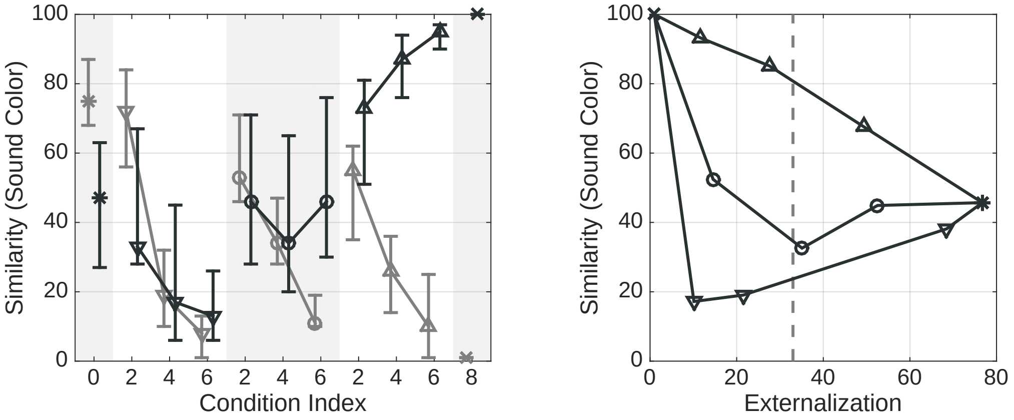
Similarity regarding sound color. The ratings of similarity to the reference regarding sound color are shown in Fig. 11. Here, the differences between the methods are even more pronounced. The similarity ratings of the DRR method in Fig. 11(a) increase monotonically towards less externalized conditions and the differences between all three levels are significant. However, the ratings of the truncated conditions are instead dropping significantly with increasing externalization. While `trc 4` and `trc 6` are not significantly different from each other, both were rated significantly lower than `trc 2`.

Moreover, all conditions with manipulated DRR were rated significantly higher than all truncated conditions. In the case of the decay method, the pair comparison showed that condition `dec 6` received a significantly higher rating than `dec 4`, but no distinct tendency is visible.

While no significant differences between the conditions of the decay method and the BRIR were found, conditions `trc 4` and `6` received a significantly lower rating. Furthermore, all conditions with modified DRR were rated higher than the BRIR with significance. The trajectories through the similarity-externalization space in Fig. 11(b) exhibit pronounced differences. In comparison to the other methods, the convex curve of the DRR method is preferable in terms of the desired maximization of both similarity regarding sound color and externalization. The trajectories of the other methods are characterized by lower similarity scores. The curve of the truncation methods is distinctly concave due to lower similarity ratings of similarly externalized conditions.

Similarity regarding reverberation. Fig. 12 shows the ratings of the similarity regarding the perceived amount of reverberation. It can be seen in Fig. 12(a) that, within each method, similarity increases monotonically with increasing modification depth and towards less externalized conditions. The differences between the three conditions of each method are significant.

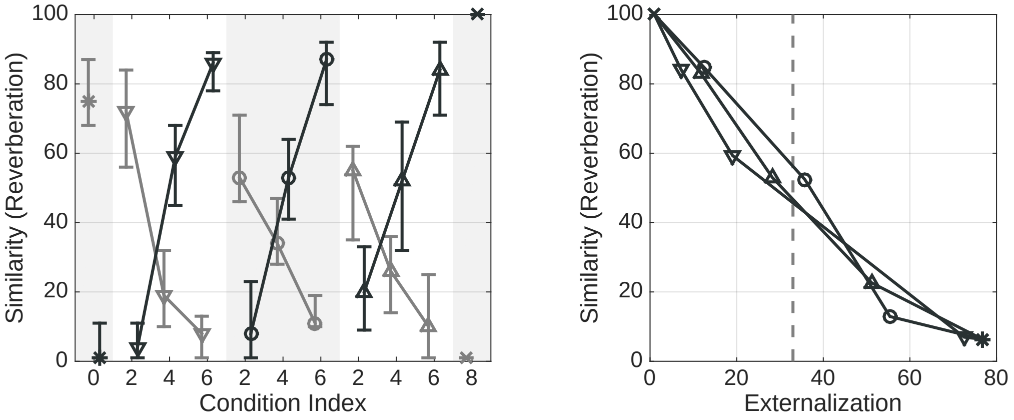
Condition 4 and 6 were rated similarly between the methods. However, `trc 2` was significantly rated lower than `dec 2`, and `drr 2` was, in turn, rated significantly lower than `drr 2`. All conditions, except for `trc 2` and `dec 2`, received a significantly higher rating than the BRIR. However, the differences in similarity regarding reverberation between the conditions with index $i = 2$ are far less pronounced than regarding sound color (cf. Fig. 11(a)). Moreover, while the ratings of the truncated conditions increase regarding reverberation with decreasing reverberation, they decrease with regard to sound color.



(a) Median similarity ratings (black) compared to externalization (gray).

(b) Geometric median of similarity and externalization.

Figure 11: Comparison of similarity to the reference regarding sound color (S.IIa) and externalization (E.II). Median ratings and 95 % confidence intervals, linearly interpolated. The dashed vertical line in (b) denotes the ‘close-to-the-head’ threshold. *Legend:* ▽ *Truncation,* ○ *Decay,* △ *DRR,* $*$ *BRIR,* × *HRIR.*



(a) Median similarity ratings (black) compared to externalization (gray).

(b) Geometric median of similarity and externalization.

Figure 12: Comparison of similarity to the reference regarding the perceived amount of reverberation (S.IIb) and externalization (E.II). Median ratings and 95 % confidence intervals, linearly interpolated. The dashed vertical line in (b) denotes the ‘close-to-the-head’ threshold. *Legend:* ▽ *Truncation,* ○ *Decay,* △ *DRR,* $*$ *BRIR,* × *HRIR.*

Fig. 12(b) shows that the trajectories of each method are very similar. Approximately, they describe straight lines connecting the HRIR with the BRIR. This shows that the perceived amount of reverberation may be a reliable predictor of externalization in the present case.

2.4.4 Modeling Similarity

It is unclear whether and to what extent the similarity ratings regarding sound color and reverberation (S.IIa-b) can explain the general similarity score (S.II). This question was investigated using multiple linear regression. Denoting similarity regarding sound color and similarity regarding reverberation with x_{sc} and x_{rev}, the overall similarity ratings

$$y \sim c_1 \cdot x_{sc} + c_2 \cdot x_{rev} + c_3 \cdot \bar{x} + c_4, \quad c_i \in \mathbb{R} \tag{12}$$

can be expressed as a weighted sum of x_{sc}, x_{rev}, an interaction term $\bar{x} = \sqrt{x_{sc} x_{rev}}$ (the geometric mean), and an additive constant c_4. Linear regression models were fitted for every possible combination of the mentioned terms. The Bayesian information criterion (BIC, Raftery (1995)) and the coefficient of determination R^2, defined as the proportion of the variation in the response variable y that can be explained by the model (cf. Moore and McCabe (2009)), serve as criteria for model selection. It was found that the model $y \sim c_3 \cdot \bar{x}$ simultaneously minimizes the BIC and maximizes R^2 with $c_3 = 0.9$ and $R^2 = 0.69$. In other words, a large proportion of the variation in the general similarity can be explained by the geometric mean of the ratings regarding sound color and regarding the amount of reverberation.

2.4.5 Representation on a Common Axis

Due to the different nature of the respective varied parameter, the ratings are not directly comparable between the methods. One possible approach for making a comparison is to relate one response variable with another, as done in the previous section. Otherwise, a common axis has to be found by relating all conditions to the same physical quantity. The DRR, the temporal centroid, and different energy percentiles were reviewed to find a suitable measure.

Linear scaling. The complete set of conditions was only indirectly compared in parts E.I and S.I. These ratings are only comparable between methods via the common reference and anchor. Due to the large number of conditions, there is a risk that the subjects were not able to estimate the response variable precisely for each condition. This could manifest itself in such a way that subjects evaluate the conditions merely as a ranking.

For externalization and similarity, additional redundant ratings were obtained in the direct comparison between the modification methods in part II of each experiment. The analysis in Sec. 2.4.1 has revealed significant differences between the indirect and the direct comparison for certain conditions. It was thus decided that the ratings of parts I and II should not be pooled to gain a more precise estimation of the central tendency. But, since it makes sense to put more trust into the ratings from the direct comparison and to yield a more reliable comparison between methods with the complete set, the externalization and similarity ratings x_i^{I} from the indirect comparison were corrected based on the ratings x_i^{II} from the direct comparison.

First, the redundant ratings from the indirect comparison were replaced for each method on a per-listener basis. This applies to conditions with indices $i = \{0,2,4,6,8\}$. The remaining ratings, x_i^{I} with $i = \{1,3,5,7\}$, were linearly scaled. The complete set of corrected ratings is given by

$$x_i = \begin{cases} x_i^{\mathrm{II}} & \text{even } i, \\ x_{i-1}^{\mathrm{II}} + \frac{x_{i+1}^{\mathrm{II}} - x_{i-1}^{\mathrm{II}}}{x_{i+1}^{\mathrm{I}} - x_{i-1}^{\mathrm{I}}} \left(x_i^{\mathrm{I}} - x_{i-1}^{\mathrm{I}} \right) & \text{odd } i \end{cases}, \tag{13}$$

where $i = 0 \ldots 8$.

Temporal centroid and DRR. A common axis can be found by analyzing the distribution of energy in the BRIRs of each condition. The DRR and the temporal centroid are possible candidates. While the DRR compares the amount of energy in the direct and reverberant part of the BRIR, the temporal centroid T_C is the time instance corresponding to the ‘center of mass’ of the BRIR.

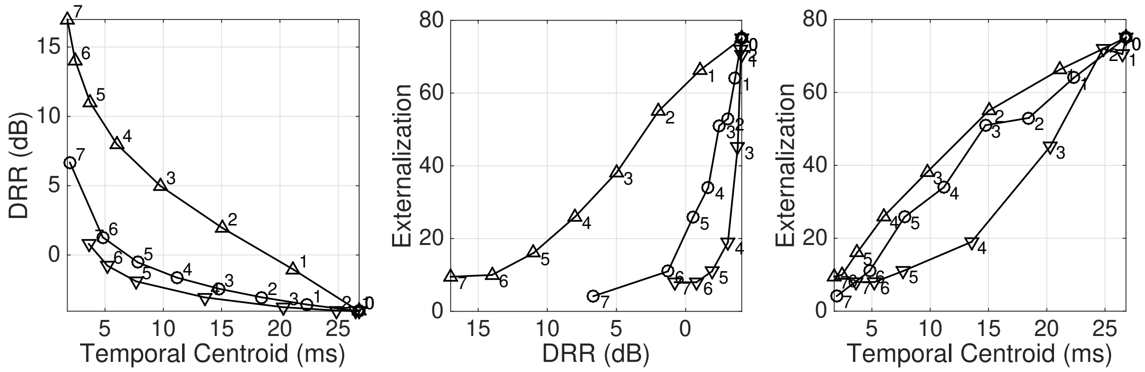
The BRIRs were first weighted in the frequency domain with the spectral envelope of the speech signal used for creating the conditions. The spectral envelope was obtained by taking the square root of the power spectral density estimate using Welch’s method (Welch, 1967). The weighted BRIRs were then used to compute the temporal centroid

$$T_C = \frac{\int_{t=0}^{N} t \cdot \bar{h}_i(t)\,\mathrm{d}t}{\int_{t=0}^{N} \bar{h}_i(t)\,\mathrm{d}t}, \tag{14}$$

where $\bar{h}_i(t) = \frac{1}{2}(h_{i,\mathrm{L}}^2(t) + h_{i,\mathrm{R}}^2(t))$ is the mean of the squared left and right channels of the BRIR corresponding to the i-th condition. The computation was repeated for the conditions of each method.

The DRR was computed for all conditions according to Eq. 3, based on the average of the squared left and right channels, as well.

Fig. 13(a) shows the computed values for the DRR and the temporal centroid corresponding to the weighted BRIRs of all conditions. The DRR is shown on the vertical, and the temporal centroid on the horizontal axis. The conditions can be identified based on the condition index i, where the BRIR has the index $i = 0$, and the indices of the individual conditions range from $i = 1, \ldots, 7$. The HRIR with index $i = 8$ is not shown in the plot because of its infinite DRR.



(a) (b) (c)

Figure 13: Comparison of the conditions in respect of the DRR and the temporal centroid of the weighted BRIRs: (a) shows the trajectories of the methods over temporal centroid and DRR, (b-c) show the median externalization ratings plotted over the DRR or the temporal centroid, respectively. The condition index i is shown beside the conditions. The BRIR has the index $i = 0$, the HRIR is not displayed. *Legend:* ▽ *Truncation,* ○ *Decay,* △ *DRR.*

While the conditions of the different methods cover a similar range of values of the temporal centroid, the range covered by the DRR values varies strongly between the methods with about 5 dB for truncation, 11 dB for the decay, and 21 dB in case of the DRR method. This shows differences in the way the modification methods affect the distribution of energy. Most of the reverberant energy is concentrated in the early reflections. Therefore, the truncation method has only little influence on the DRR at shorter times. The decay method affects also early energy, but less than the late energy. Naturally, the DRR method directly influences the DRR.

Fig. 13(b) shows the scaled externalization ratings as a function of the DRR of the conditions. Although all methods cover different value ranges of the DRR, the externalization ranges from inside the head to the common maximum value for the unmodified BRIR. A comparatively small increase in DRR in the case of truncation has the same negative impact on externalization as the larger increase in DRR caused by the other methods. The DRR is thus unsuitable for modeling externalization independent of the modification method.

In Fig. 13(c), the externalization ratings are plotted with the temporal centroid on the horizontal axis. The curves for the DRR and the decay methods have a very similar shape. Although the curve for truncation deviates slightly downwards between $i = 3$ and 5, the temporal centroid proves to be a better predictor of externalization than the DRR.

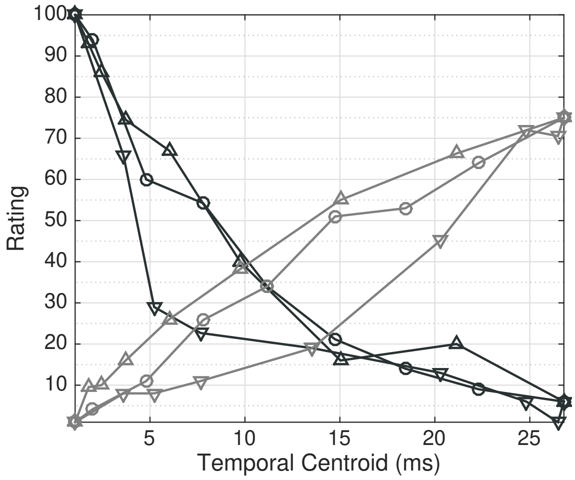
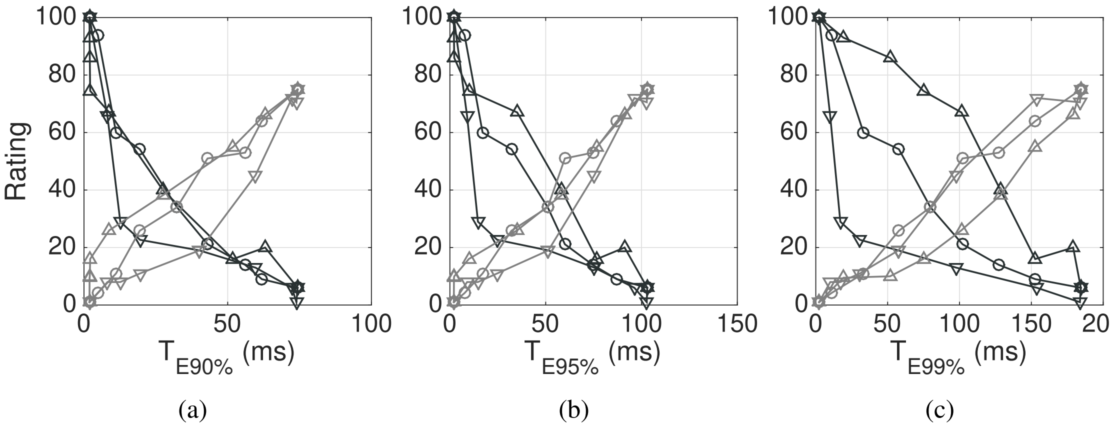


Figure 14: Externalization (gray) and similarity to the reference (black) over the temporal centroid of the weighted BRIRs. *Legend:* ▽ *Truncation,* ○ *Decay,* △ *DRR.*

In order to include sound quality into the analysis, Fig. 14 shows externalization in comparison to the similarity ratings as a function of the temporal centroid. The black curves show the scaled general similarity ratings. The externalization ratings are plotted in gray. From left to right, the condition index decreases, and the amount of reverberant energy increases. The HRIR is the first data point on the left side, and the BRIR is represented by the last data point on the right side.

As already observed for externalization, the truncation curve deviates significantly downwards. The deviation between the curves is most pronounced around the 5th and 6th condition. The similarity curves of the decay and DRR methods are very similar when plotted over the temporal centroid. The downward deviation of the externalization curve for truncation indicates that a similar degree of externalization is associated with more late energy in the BRIR, compared to the other methods. The downward deviation of the similarity curve, on the other hand, shows that, in the same range of the temporal centroid, truncated conditions are perceived as less similar to the HRIR than conditions of the other methods.

Energy percentiles. Another energetic measure closely related to the temporal centroid is the N-th energy percentile $T_{E,N\%}$ which is the time instance at which $N\,\%$ of the total energy of the BRIR has arrived. Fig. 15 shows the scaled ratings of externalization and similarity as a function of the percentiles for $N = \{90, 95, 99\}\,\%$. The 90th percentile in



(a) (b) (c)

Figure 15: Externalization (gray) and similarity to the reference (black) over the N-th energy percentiles of the weighted BRIRs, where (a) $N = 90\,\%$, (b) $N = 95\,\%$, and (c) $N = 99\,\%$. *Legend:* $\triangledown$ *Truncation,* ○ *Decay,* $\triangle$ *DRR.*

Fig. 15(a) leads to curves similar to the temporal centroid. Increasing N to 95 % (Fig. 15(b)), the externalization curves become more similar, but the difference between the curves becomes more pronounced. For $N = 99\,\%$ (Fig. 15(c)), the externalization curves have switched their order, while the similarity curves appear now far apart. This shows that the choice of N affects the result considerably. Because of this dependence, the temporal centroid is regarded as superior to the energy percentiles as a common axis. However, the plot over the energy percentiles nevertheless reveals that there are differences between the methods in terms of similarity to the reference at similarly externalized conditions.

2.4.6 Discussion

As has been shown by Hartmann and Wittenberg (1996), the phenomenon of externalization is a continuum between the extreme cases localization inside or outside the head. In agreement with this finding, the externalization ratings of the indirect comparison in, E.I, show that the perceived location of a sound source can be steered monotonically between these extreme cases by using any of the three methods to successively de-reverberate the BRIR. The externalization ratings agree with the results of both Crawford-Emery and Lee (2014) and Catic et al. (2015) who found that the truncation of the reverberant part causes the externalization to decrease. The main effects that can be assumed responsible for this decrease are that dynamic binaural cues in the reverberant sound, such as ILD fluctuations, are either reduced after cutting off the reverberation (truncation) or masked by the direct sound (decay, DRR). Furthermore, de-reverberation leads to a mismatch between the real and the virtual room (room divergence effect, cf. Werner et al. (2016a)).

The BRIR was perceived as less externalized than the loudspeaker throughout the experiment by a majority of the participants. This can most likely be attributed to individual differences between the HRIRs of the listeners and the dummy head, as well as remaining timbral differences which were unavoidable as the equalization was carried out for the dummy head. Moreover, the experimental design with the loudspeaker available for comparison is rather sensitive to differences in timbre.

By and large, the similarity to the anechoic signal increases with decreasing amount of reverberant energy. But rather than *if* similarity (and sound quality) increases, the question is *to what extent*. Since the interpretation of the results of the indirect comparison in part I is limited to a within-methods comparison of the conditions, the cross-comparison between the methods in part II has to be taken into consideration.

As can be seen in Fig. 8, the ratings of particular conditions in the direct comparison deviate partially from the ratings in the indirect comparison. The strongest deviation between the indirect and the direct comparison can be observed regarding similarity for the truncation method where condition 2 and 6 were rated lower in the direct comparison. On the one hand, this may indicate that the indirect comparison was biased. On the other hand, it could also mean that sound attributes such as sound colorations, which may be inconspicuous in a within-methods comparison, stand out in a comparison between methods and may interfere with externalization. However, both may just two sides of the same coin.

When the BRIR length is modified, the minimum length for externalization was found to be roughly between 59 and 106 ms. The steepest point of the curve of the indirect comparison is at $L = 59$ ms and lies on the threshold 'close to the head'. This agrees reasonably with the results of Catic et al. (2015) where externalization was found to increase substantially between lengths of 20 and 80 ms, whereas a BRIR length greater than 100 ms did not lead to a further increase. Another finding of Catic et al. (2015) was that ILD fluctuations and interaural coherence were well suited as predictors for externalization using truncated BRIRs. However, further experiments are needed to determine how well these measures correspond to externalization for other methods than truncation.

In the present experiment, it was observed that several conditions that are similarly externalized differ between the methods in their ratings regarding attributes of sound quality. The greatest differences between the methods are found between conditions with mediocre externalization. The differences become most apparent in the trajectories of the different methods in the two-dimensional plot of naturalness and similarity over externalization. The trajectories for general similarity over externalization show, considering points of similar externalization, that the DRR received the highest, and truncation the lowest ratings.

No difference was found between the methods regarding the perceived amount of reverberation (Fig. 12(b)). Regarding sound color, however, the curves differ substantially and the DRR method was consistently rated more similar to the reference than the other methods. The convex curve of the DRR method (high similarity) in Fig. 11(b) strongly differs from the concave curve of the truncation method (low similarity). The trajectory of the decay method lies again in between the other two methods.

The overall tendency of the naturalness ratings is that reverberant conditions were perceived as more natural. The differences between the DDR and the truncation method are less pronounced regarding naturalness. Truncated conditions were perceived as less natural at similar externalization than the other methods.

Although no major jumps between the data points are expected due to the single-parametric monotonic BRIR modification, it should nevertheless be noted that the present data is insufficient to provide a precise estimate of the trajectories. However, the hypothesis of differences between the methods is supported by statistical pair comparisons. Especially for the truncation method – the method for which there is no equivalent physical process – impairments of sound quality have been found repeatedly. This fits well with the experience of the author from informal listening experiments, where it was observed that the truncated BRIRs of short to medium lengths stand out from the other conditions with timbral colorations.

A likely cause for these colorations may be the formation of comb filters due to the early reflections interfering with the direct sound. The early reflections in the BRIR are completely unaffected by truncation until reaching a very short length. In contrast, with the DRR and decay methods, the early reflections are attenuated from the beginning. Cutting off reverberant energy at the end of the BRIR may lead to a de-masking of such comb filters that may otherwise be inaudible.

Parts of the present study overlap with an experiment by Crawford-Emery and Lee (2014) who assessed externalization and sound colorations for truncated BRIRs. The design was similar to the present study: Here, the subjects had to rate the similarity to the reference regarding sound color, whereas (Crawford-Emery and Lee, 2014, p. 3) asked 'how much [the stimuli] differed tonally from the reference sample.' Tonal colorations were found to increase with increasing BRIR length. Here, however, an increase of sound colorations with increasing BRIR length was only observed with modified DRR. Instead, differences in sound color using truncation were found to decrease with increasing length within the observed range (Fig. 11). This discrepancy may be due to the different experimental design. Here, the truncated conditions were comparatively rated along with conditions with modified DRR and reverberation time. Furthermore, the participants were instructed to rate the similarity regarding sound color and regarding the amount of reverberation individually, whereas in the experiment of Crawford-Emery and Lee (2014) the perceived amount of reverberation may have been regarded as an aspect of sound color. Furthermore, as Crawford-Emery and Lee (2014) note, the results are specific to the investigated room, and so are the results of the present study.

2.5 Summary

In this chapter, a listening experiment on the influence of different BRIR modifications on externalization and sound quality was presented. The amount of reverberant energy in a BRIR was successively reduced with different methods: the manipulation of the BRIR length by truncation of the reverberant part, the manipulation of the reverberation time by multiplication with an exponential decay curve, and the manipulation of the DRR by weighting the reverberant part with a constant smaller than one. The latter two methods correspond to the introduction of absorbing material into the room and increasing the distance between source and receiver. Truncation, however, is a fully artificial process.

A frontal BRIR was measured using a dummy head. This is the baseline condition from which all other conditions were derived by step-wise de-reverberation of the BRIR. Using either of the three methods, eventually, the modification leads to the same anechoic condition, i.e., the HRIR. Stimuli were generated by convolving the resulting impulse responses with a sequence of male speech. All conditions were played back through open headphones. The ear cushions were modified in such a way that the attenuation of frontal sound was minimized. The headphone signal was furthermore equalized to emulate the sound of the loudspeaker.

The listening experiment consisted of the evaluation of externalization in comparison to the loudspeaker and the evaluation of sound quality. The strategy to assess sound quality was chosen based on the interest in sound neutrality. The conditions had to be rated regarding the perceived naturalness, and the similarity to an anechoic reference stimulus. The reference was created by convolving the speech sequence with the HRIR. To make a more nuanced comparison, the task was complemented by assessing also the similarity regarding sound color and the amount of reverberation. Externalization and overall similarity to the reference were both rated within each method (this involves the complete set of conditions) and in a direct cross-comparison between the methods (with a subset of conditions) Naturalness, similarity regarding sound color, and similarity regarding reverberation were only evaluated between-methods.

Experimental results were collected for twenty-one participants. The results show a monotonic relationship between increasing modification depth and decreasing degree of externalization within each method. The similarity to the reference, in contrast, increases with increasing modification depth. However, due to the incomparability of the varied parameter of each modification method and the subjectiveness of choice of the parameter values, a comparison between the methods based on the condition index is not meaningful. Instead, each attribute of sound quality was regarded relative to the degree of externalization of the same condition. Naturalness and similarity were plotted against externalization. Linear interpolation yields trajectories in the space spanned by each pair of response variables. Statistical pair comparisons have been made to support the interpretation of the results. To gain an alternative perspective, the ratings were also compared on a common axis by relating the conditions to a derived physical quantity. The DRR, the temporal centroid, and the energy percentiles of the conditions were considered. The DRR turns out to be unsuitable for the comparison since a similar decrease in externalization is associated with different DRR ranges for each method. The temporal centroid and the different energy percentiles, however, are more suitable. While these measures do not allow an accurate prediction of the degree of externalization regardless of the modification method, the differences of the externalization curves are reduced between the methods so that differences of the same conditions in similarity to the reference become more apparent.

Using these different methods to evaluate the results, it has been found that differences in the relationship between externalization and sound quality exist between the methods. The greatest differences were found in conditions with mediocre externalization, whereas the ratings of well-externalized conditions differ less. The DRR method yields the highest overall similarity ratings relative to externalization, followed by the decay and, with the lowest ratings, the truncation method. Furthermore, the DRR and decay methods appear to be perceived as more natural than truncation in the mid externalization range. While there is no indication of differences between the methods regarding the perceived amount of reverberation, they differ distinctly in their similarity to the reference regarding sound color. The results suggest that truncation is more likely to produce sound colorations than increasing the DRR. It could be shown using multiple linear regression that the decomposition of general similarity into similarity regarding sound color and regarding reverberation is plausible.

In summary, it can be said that each method has its advantages depending on the application. For instance, in a scenario where reverberation should be added to an HRIR to support externalization, it makes sense to give the reverberation natural decay time and DRR. But, in order to enhance the efficiency of computations, truncation may be required in some cases, with the possible consequence of timbral artifacts. Since both the modification of the DRR and the reverberation time can reduce the effective BRIR length (especially the decay method), these methods may be combined with truncation to remedy sound colorations. The combination of methods is one of various scenarios that were tested in the follow-up experiment in Ch. 3.

Finally, the limitations of this study must be pointed out. Different choices regarding the experimental design do not allow a broad generalization. Non-individually measured BRIRs have been used for the binaural synthesis. Due to this circumstance, it can be expected that the externalization ratings are generally lower.

Furthermore, the results are confined to the studied scenario in which speech was played back by a single source in a frontal direction. While externalization of speech is a realistic and familiar scenario, and the use of only a single speech sequence brings comparability between trials, it is not representative of other scenarios.

Rating externalization in comparison to the real loudspeaker has the great advantage of ‘grounding’ the ratings by a physically distant sound source. However, it also brings the necessity for the subjects to listen to the loudspeaker while wearing headphones. The awareness of the headphones, as well as differences between the sound of the headphones and the loudspeaker, may hinder the establishment of a state of immersion and, hence, have a negative effect on externalization.

The chosen method to assess sound quality by narrowing it down to naturalness and the similarity to the HRIR does not evaluate sound quality in a general sense, but rather from the standpoint of sound neutrality.

As discussed in Sec. 2.4.1, the use of a MUSHRA-like design may have introduced a bias into the results due to the large number of conditions.

All in all, the experiment regarded the topic rather on a macroscopic level. It contributes to the big picture, but further research is needed to provide more insight about the role of fine structure in the reverberation, e.g., by the analysis of dynamic binaural cues of each method with the methodology of Catic et al. (2015).

3 Exploratory Study on the Influence of Reverberation on Externalization in Different Scenarios

The previous chapter investigated how the de-reverberation of BRIRs affects externalization and different attributes of sound quality. A systematic comparison between three different modification methods was made: the manipulation of either the BRIR length, the DRR, or the decay time. This chapter investigates a number of open questions in a series of exploratory experiments that build on the preceding work.

Sec. 3.1 explains the different experiments. In the first experiment, the combination of different modification techniques is evaluated based on conditions from the previous experiment. The second experiment uses purely diffuse artificial reverberation. Different spatial distributions as well as binaural, dichotic, and diotic reverberation are compared. In the third experiment, it is investigated how additive stereo reverberation and convolutive mono reverberation influence externalization. Finally, in the fourth experiment, discrete diffuse and direct reflections are added to the HRIR and compared. The results are discussed in Sec. 3.2, followed by the summary of this chapter in Sec. 3.3.

3.1 Experiments

This section presents the design of the experiments. In contrast to Ch. 2, the experiments are rather diverse. However, certain parameters have been adopted from Ch. 2 and stayed the same. In each part, externalization was assessed in comparison to the loudspeaker. The loudspeaker stood at the same position in the same room. Conditions were created by the convolution of binaural impulse responses with speech. The same speech sequence was used as before. The stimuli were presented through the modified open headphones. Head movements were disallowed. All conditions have the measured BRIR from Sec. 2.2.1 as a basis. Some conditions tested already in Ch. 2 appear again in this experiment. Accordingly, the same nomenclature is used, indicating the modification methods with their respective abbreviation (`trc`, `dec`, and `drr`, cf. Tab. 2), followed by the condition index as defined in Tab. 1. In some cases of interest, sound quality was evaluated via the similarity to a reference stimulus as explained on pp. 29f. As a reference, the speech sequence was convolved with the HRIR. The conditions were then presented in diotically and via the modified headphones.

Tab. 5 gives an overview of the experiments. This section is divided into four experiments, each of which consists of one or two associated trials. Here, *trial* refers to one set of stimuli that are presented to the subject at the same time. All experiments and trials were presented in random order, and so were the stimuli within each trial. The experimental design is based on the MUSHRA paradigm and implemented using a graphical interface (cf. Sec. 2.3.1-2.3.2).

Part		**Trial**		**Task**
1	Two-fold BRIR modification	**a**	DRR and reverberation time	E/S
		b	DRR and BRIR length	E/S
2	Diffuse binaural reverberation	**a**	Varying spatial distribution	E
		b	Dichotic and diotic reverberation	E
3	Additive and convolutive reverberation	**a**	Convolutive mono reverberation	E/S
		b	Additive stereo reverberation	E
4	Discrete reflections		Diffuse/specular reflections	E/S

Table 5: List of experiments and trials. Externalization (E) had to be rated in every trial. Some trials also included rating the similarity to the reference (S).

3.1.1 Part 1: Two-fold BRIR Modification

As a follow-up to the preceding experiment, the effect of combining the previously tested methods was investigated. This experiment consists of two parts. The first part is about the joint modification of the reverberation time and the DRR, and the second part treats the joint modification of the DRR and the BRIR length. In both parts, participants had to rate externalization in comparison to the loudspeaker, as well as the similarity to the HRIR convolved with speech. It was evaluated how externalization and similarity are affected by the two-fold modification, in comparison to single-method modifications. Because of limited time, only a small set of combinations was tested. These comparisons were presented on two separate pages to yield a manageable number of stimuli on each page.

Trial 1a: Joint modification of DRR and reverberation time. To implement the joint modification of the DRR and the reverberation time, the reverberation time was first decreased by multiplying the BRIR with an exponential decay curve as described in

Sec. 2.1. The DRR was then increased by applying a relative difference in the level of the reverberant part.

Tab. 6 lists the conditions that were compared in this experiment. Because of the connection to the preceding experiment in Ch. 2, the same nomenclature was used for the conditions (cf. Tab. 1). The subjects were asked to rate externalization in comparison to the loudspeaker. The presented set of stimuli can be found in the left half of Tab. 6. The BRIR (condition 0) is considered the hidden quasi-reference, and the HRIR (condition 5) functions as an anchor. Conditions 1 to 4 were created by manipulating the reverberant part of the measured BRIR.

Response variable:	**Externalization**				**Similarity**	
Function	**ID**	**Impulse response**	T_{30} (s)	ΔDRR (dB)	**ID**	**Impulse response**
Reference	-	*(real loudspeaker)*			5	HRIR
Hidden ref.	0	BRIR	0.7	0	5	HRIR
	1	drr 1+dec 1	0.6	3	1	drr 1+dec 1
Stimuli	2	drr 2+dec 2	0.5	6	2	drr 2+dec 2
	3	drr 1	0.7	3	3	drr 1
	4	drr 2	0.7	6	4	drr 2
Anchor	5	HRIR	-	∞	0	BRIR

Table 6: Two-fold BRIR modification: Joint modification of DRR and reverberation time. Presented conditions and corresponding parameters.

To create the first two conditions, both the DRR and the reverberation time were varied. The DRR was increased by 3 dB and 6 dB corresponding to conditions drr 1 and drr 2 in Ch. 2. The resulting impulse responses were tested with and without alteration of the reverberation time. The reverberation time was decreased, corresponding to dec 1 and dec 2, by 0.1 and 0.2 s. In order to reduce the number of stimuli, not all possible combinations were tested: drr 1 was only combined with dec 1, and drr 2 only with dec 2. The other two conditions (3 and 4) were created by modifying just the DRR by the same relative differences. Thus, condition 3 and 4 are identical with drr 1 and drr 2 in Ch. 2.

The subjects were also asked to rate the similarity to the anechoic binaural speech signal (condition 5, HRIR). The conditions presented in this part can be found in the right half of Tab. 6. In contrast to the externalization part, all conditions were presented via headphones inside the anechoic room. The conditions were presented in mono by playing back the left channel to both ears. The HRIR (condition 5) was used as the hidden reference and the BRIR (condition 0) as an anchor. Besides these changes, all conditions remained the same.

Trial 1b: Joint modification of DRR and BRIR length. The conditions presented in the second part were created similarly. Tab. 7 lists the conditions and parameters. The loudspeaker was used as a reference in the externalization part. The BRIR (condition 0) and the HRIR (condition 6) were used as the hidden reference and anchor. The conditions for the externalization part are shown in the left part of the table. Again, two combined conditions were tested (conditions 1 and 2), as well as two conditions where only the DRR was modified (conditions 3 and 4). For creating these, the BRIR underwent the same relative DRR modification as before (`drr 1` and `drr 2`). Conditions 1 and 2 were created by truncating both resulting impulse responses to a length of $L = 106\,\mathrm{ms}$. This particular length corresponds condition `trc 3` of the first experiment (cf. Tab. 1). Condition `trc 3` - with unaltered DRR - was presented as an additional 5th condition for comparison. As shown in the right half of Tab. 7, the same conditions were also evaluated regarding the similarity to the reference.

Response variable:		**Externalization**				**Similarity**
Function	**ID**	**Impulse response**	Length (s)	ΔDRR (dB)	**ID**	**Impulse response**
Reference	-	*(real loudspeaker)*			6	HRIR
Hidden ref.	0	BRIR	1.0	0	6	HRIR
	1	`trc 3+drr 1`	0.1	3	1	`trc 3+drr 1`
	2	`trc 3+drr 2`	0.1	6	2	`trc 3+drr 2`
Stimuli	3	`drr 1`	1.0	3	3	`drr 1`
	4	`drr 2`	1.0	6	4	`drr 2`
	5	`trc 3`	0.1	0	5	`trc 3`
Anchor	6	HRIR	0.003	∞	0	BRIR

Table 7: Two-fold BRIR modification: Joint modification of DRR and BRIR length. Presented conditions and corresponding parameters.

3.1.2 Part 2: Diffuse Binaural Reverberation

This experiment investigates the variation of different reverberation parameters using artificial BRIRs. The BRIRs consist of the measured binaural direct sound and purely diffuse synthetic reverberation. In the first part, the directional distribution of the reverberation was varied, whereas, in the second part, omni-directional reverberation was compared with dichotic and diotic reverberation.

The use of artificial BRIRs has the advantage of a high degree of flexibility regarding the independent variation of different sound field parameters. However, this requires a convincing synthesis. As the experiment is carried out inside the original room and in comparison to the loudspeaker, the baseline synthesis should agree with the measured BRIR in reverberation time, DRR, and sound color.

The first part of this section describes the impulse response synthesis. The two studied scenarios are described thereafter.

Impulse response synthesis. A synthetic BRIR was generated as the basis to create further conditions for the experiments. The BRIR consisted of two elements: The frontal HRIR of the dummy head was concatenated with artificial binaural reverberation. The reverberation was generated based on so-called *velvet noise*, a sparse pseudo-random pulse sequence proposed for modeling late reverberation (Järveläinen and Karjalainen, 2007; Välimäki et al., 2017). Two seconds of velvet noise with unit amplitude and an average inter-pulse interval of 4 ms were generated.[4] A random direction was assigned to each of the discrete pulses. The directions were drawn from all possible directions on the sphere with uniform probability. Corresponding to each direction, an HRIR was assigned to the respective pulse by selecting the closest match (least distance) from a dense spherical HRIR set. The HRIR set was made available by Bernschütz (2013) for the same dummy head model as used in the measurement of the binaural direct sound (cf. Sec. 2.2.1).

Convolving each pulse with the assigned HRIR and superimposing the resulting signals would yield diffuse binaural noise with uniform envelope modeling an isotropic sound field. To transform the binaural noise signal into artificial reverberation, it must be temporally weighted in order to establish a decay over time. Simply weighting the sequence with an exponential decay curve as in Sec. 2.1 would result in a rather unnatural sound since,

[4] The noise sequences were obtained using a noise generator [available online]: `https://github.com/mmorise/NoiseGenerators` (last accessed: 2020/05/14)

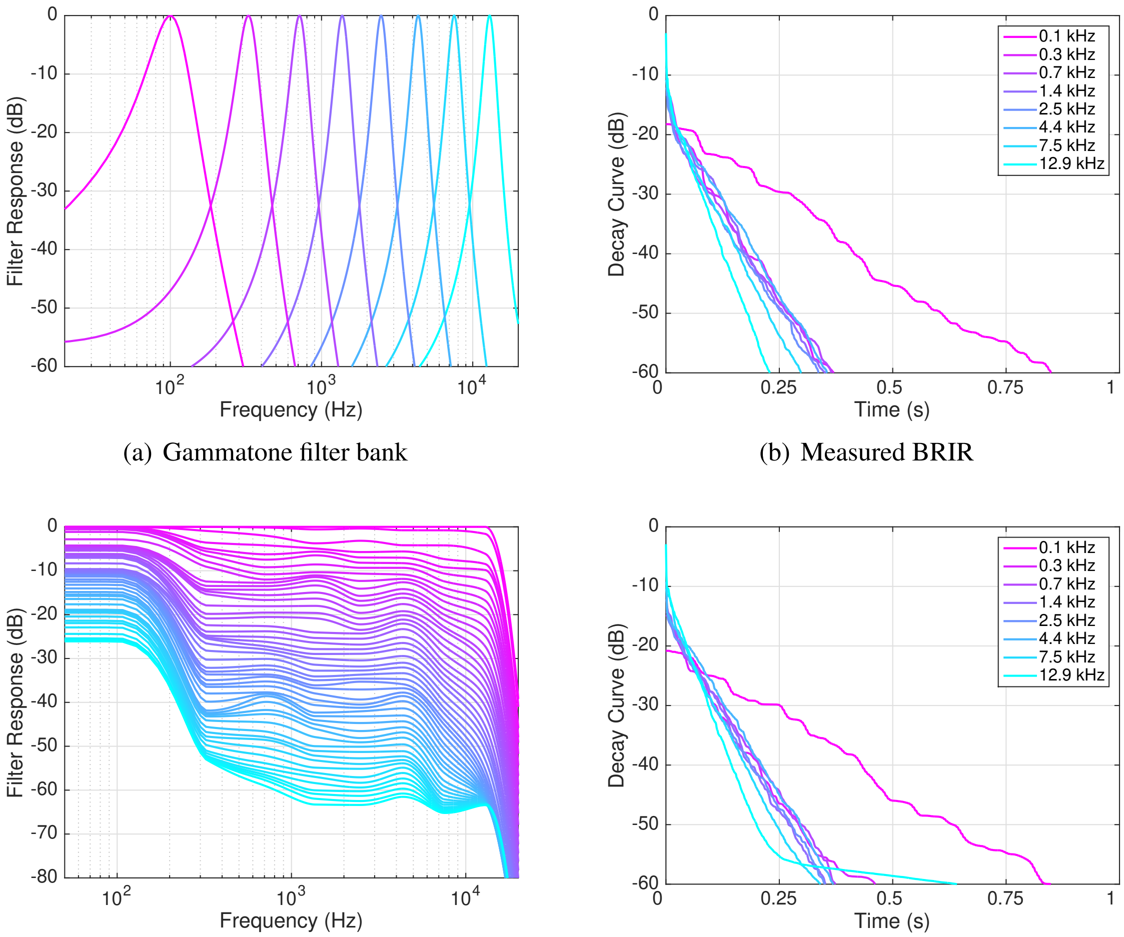
typically, the sound decays faster at high than at low frequencies within a room. The decay should be frequency-dependent and resemble the reverberation of the measured BRIR. It is thus necessary to filter the reverberant tail, taking into account the instantaneous spectral envelope of the BRIR.

To analyze the frequency-dependent decay, the BRIR was split into eight frequency bands using a Gammatone filter bank. The frequency response of the analysis filters is shown in Fig. 16(a). Energy decay curves were computed via backward integration of the squared reverberant part within each frequency band (cf. Schroeder (1965)). Fig. 16(b) shows the resulting decay curves. The normalized energy decay curves were equally split into 0.75 ms long time segments. The curves were then evaluated within each band by taking the root mean square of the curve within each segment. For every segment, this yields a magnitude frequency response sampled at the eight center frequencies of the filter bank. The instantaneous frequency responses were interpolated using piecewise cubic interpolation. In order to obtain a smooth frequency response, it was defined to be constant below the lowest band (100 Hz), and to decay by 60 dB between the highest band (12.9 kHz) and the Nyquist frequency. Linear-phase impulse responses were derived from the interpolated frequency response. Every binaural noise pulse was convolved separately with the impulse response corresponding to the segment in which the pulse lies. The convolved pulses were compensated for the group delay of the filter and then superimposed, yielding the diffuse decaying binaural reverberation signal.

The synthetic reverberation was equalized to the reverberant part of the BRIR to reduce further differences in sound color. It was then concatenated with the binaural direct sound from the measured BRIR in such a way that the diffuse reverberation begins at the time instant of the first reflection in the room.

Using the measured binaural direct sound, and matching frequency-dependent reverberation time, DRR, and sound color, are measures to increase the similarity to the measured BRIR. However, no early reflections were added, and the pulse density was kept constant during the decay. While this is a strong simplification, it is considered unproblematic, provided that the synthetic BRIR is sufficiently externalized so that the experiment is sensitive enough to detect a decrease in externalization possibly produced by the introduced parameter variations.

To verify the synthesis, the energy decay curves of the synthetic BRIR were computed (Fig. 16(d)) and compared to those of the measured BRIR (Fig. 16(b)). The comparison



(a) Gammatone filter bank

(b) Measured BRIR

(c) Interpolated filters over time

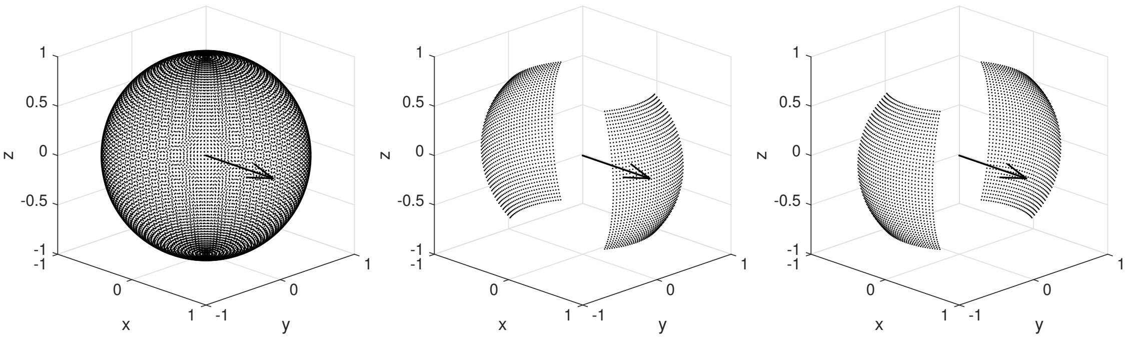
(d) Synthesized BRIR

Figure 16: Analysis and synthesis framework. The Gammatone filters in (a) were used to obtain frequency-dependent energy decay curves (b). The energy decay relief (c) is obtained by temporal segmentation of the interpolated decay curves (shown for $t \in [0, 0.5]$ s). The energy decay curves of the synthesized reverberation are shown in (d).

shows that the curves agree well in all frequency bands between 0 dB and −55 dB. In addition, the spectrograms of both conditions were analyzed. A slightly increased decay time was observed at high mid frequencies, but the difference is smaller than 50 ms. It is not surprising to find frequency-dependent differences between the measured and the synthesized impulse responses because the instantaneous frequency responses are produced by two different signal types - discrete early reflections in the case of the BRIR, and diffuse filtered noise in the case of the simplifying synthesis. Considering this fundamental difference between synthesis and physical reality, the resulting differences in sound color can be regarded as acceptable.

Trial 2a: Varying spatial distribution. Binaural reverberation is important for the externalization of frontal sound (Catic et al., 2015). For lateral sound, the reverberation at the contralateral ear was found by Li et al. (2018) to have a greater influence on externalization than at the ipsilateral ear. For the externalization of frontal sound, both ears can be expected to be equally important, but the effect of reverberation may be direction-dependent. This experiment investigates how varying the spatial distribution of the reverberation affects externalization.

An artificial BRIR was synthesized as explained above. The spatial distribution of the diffuse reverberation was changed by restricting the set of contributing source directions on the sphere. Fig. 17 shows that all distributions are symmetrical to the median plane. Either all directions contributed to the reverberation (Fig. 17(a)), or only two sphere segments located in the front and back, or on the left and right (Fig. 17(b)-17(c)). Each segment covered an azimuth angle of 60° and an elevation angle of 90°. The frontal segments where located at $\varphi = \{0^\circ, 180^\circ\}$, and the lateral segments were at $\varphi = \{-90^\circ, 90^\circ\}$.



(a) Omni-directional (b) Front/Back (c) Left/Right

Figure 17: Source positions of the HRIR set contributing to each of the three investigated scenarios. The black arrow denotes an azimuth of 0° and an elevation of 90°.

As explained above, directions from a dense grid were randomly drawn and assigned to the noise pulses. To achieve a certain distribution, the source directions were restricted by discarding a drawn direction if it is outside the defined regions. By doing so, the pulse density is reduced in proportion to the reduced number of active source positions.

Tab. 8 shows the conditions. The conditions were rated regarding externalization in comparison to the loudspeaker playing back the anechoic speech sequence. The BRIR (condition 0) was used as a hidden quasi-reference. Condition 1 is the omni-directional synthetic BRIR convolved with speech. Conditions 2 and 3 contain the frontal and lateral reverberation. The HRIR was used as the anchor (condition 4).

Function	**ID**	**Impulse response**	**Spatial distribution of reverberation**
Reference	-	*(real loudspeaker)*	*(real room)*
Hidden ref.	0	BRIR	measured
	1	Synthetic BRIR	omni-directional
Stimuli	2	Synthetic BRIR	front/back segments
	3	Synthetic BRIR	left/right segments
Anchor	4	HRIR	*none*

Table 8: Diffuse binaural reverberation: Varying spatial distribution. Presented conditions and corresponding parameters.

Trial 2b: Dichotic and diotic reverberation. Studies have shown decreasing importance of the accurate reproduction monaural cues in the reverberant sound for externalization (Hassager et al., 2016; Jiang et al., 2020). In this experiment, it is compared if externalization can also be achieved with different types of dichotic and diotic reverberation. Here, diotic means that the reverberation at both ears was identical, whereas dichotic refers to two different reverberation signals at both ears.

Tab. 9 shows the conditions of this trial. The conditions were rated regarding externalization in comparison to the loudspeaker. Condition 0, the measured BRIR convolved with speech, is the hidden quasi-reference. Conditions 1 to 4 are created with artificial dichotic or diotic reverberation that was appended to the measured HRIR.

In the dichotic case, the same omni-directional binaural reverberation as in the previous trial (condition 1) was compared with ‘stereo’ reverberation (condition 2). For this purpose, synthetic reverberation was generated as before, but without providing meaningful binaural cues. Instead of convolving a single pulse sequence with HRIRs of random directions, two independent realizations of velvet noise were taken for the left and the right ear. Otherwise, the sequences underwent the same processing. The reverberant signal was then appended to the HRIR to obtain a ‘stereo’ equivalent of the synthetic binaural impulse response. The term ‘stereo’ is somewhat misleading since ‘stereophony’ implies the presence of meaningful time or level differences which is not the case here. It is, however, used to distinguish it from the binaural reverberation.

Function	ID	Impulse response	Type of reverberation
Reference	-	*(real loudspeaker)*	*(real room)*
Hidden ref.	0	BRIR	measured
	1	Synthetic BRIR	binaural, omni-directional
Stimuli	2	Synthetic BRIR	two independent noise sequences
	3	Synthetic BRIR	one channel of condition 1, diotic
	4	Synthetic BRIR	one channel of condition 2, diotic
Anchor	5	HRIR	*none*

Table 9: Diffuse binaural reverberation: Dichotic and diotic reverberation. Presented conditions and corresponding parameters.

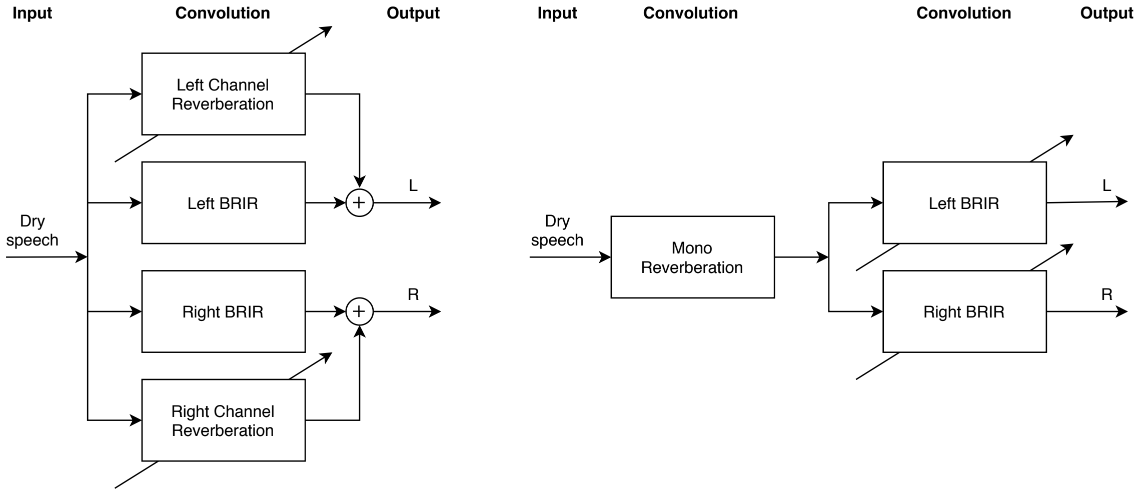
In the diotic case, the reverberation of the left-ear channel from condition 1 and condition 2 was played back to both ears. Appending the reverberation to the HRIR yields condition 3, the ‘monaural’ condition derived from the left channel of the binaural reverberation, and condition 4, the ‘mono’ condition derived from the left channel of the ‘stereo’ reverberation. Again, these names were assigned only for memorability.

As in the previous experiments, the HRIR convolved with speech served as the anchor (condition 5).

3.1.3 Part 3: Additive and Convolutive Reverberation

Two different scenarios are studied in this experiment. The first scenario investigates the influence of additive stereophonic reverberation. In Sec. 3.2.3, the reverberation of the BRIR was replaced by a dichotic surrogate with decorrelated left and right channels meant to resemble the reverberation time, DRR, and sound color – but not the binaural cues – of the measured original reverberation of the room. Here, reverberation was added to the BRIR. In order to find out if and how additional reverberation affects externalization, the amount of additive reverberation was successively increased. This method is equivalent to convolving the excitation signal with the BRIR and a reverberant impulse response in parallel and superimposing the results, as shown in Fig. 18(a).

The second scenario takes account for a fact that was disregarded so far in this thesis (and various other studies): in many practical cases, the signal at the input of a binaural decoder is already reverberant. It is therefore investigated if externalization is less sensitive to the deteriorative effect of de-reverberation by increasing the DRR if the excitation signal



(a) Additive stereo reverberation

(b) Convolutive mono reverberation

Figure 18: Schematic showing how additive and convolutive reverberation were applied. Diagonal arrows denote which block is being modified to create the different stimuli.

contains reverberation. Furthermore, it is tested whether the reverberation in the BRIR has a decreased influence on sound quality by evaluating the similarity to the reverberant speech signal convolved with the anechoic HRIR. Conditions were created by adding monophonic reverberation to the speech signal at the input prior to convolution with the modified BRIR (Fig. 18(b)).

In both of the scenarios under test, the same reverberation was used. It was obtained by simulating the stereophonic room impulse response (RIR) of a room with a volume of $V = 2484\,\mathrm{m}^3$ and a reverberation time of $T_{60} = 1.2\,\mathrm{s}$ using a common audio plug-in. In contrast to the experiments in Sec. 3.1.2, the reverberation includes both early reflections and late diffuse reverberation. The reverberation time was set to be longer than the reverberation time of the listening room, and the reverberation was not equalized to the sound color of the BRIR.

Trial 3a: Additive stereo reverberation. In this part of the experiment, binaural signals with different amounts of stereo reverberation were rated regarding externalization. A binaural signal was created by convolving the speech sequence with the unmodified BRIR. Furthermore, a reverberant signal was generated by convolving the same speech signal with each of the two channels of the stereo RIR. The direct sound was removed from the RIR before convolution, and the reverberant part was aligned to the BRIR by matching the time instants of the arrival of the first reflection. The reverberation signals were then added channel-wise to the binaural signal as shown in Fig. 18(a). Different conditions were generated by varying reverberant energy ratio

$$RER := 10 \cdot \lg \left(\frac{\int_0^\infty h_{\mathrm{SR}}^2(t)\mathrm{d}t}{\int_0^\infty h^2(t)\mathrm{d}t} \right) \tag{15}$$

which is here defined as the ratio between the energy of the stereo reverberation h_{SR} and the energy of the BRIR h.

Function	ID	Impulse response	Additive reverberation
Reference	-	*(real loudspeaker)*	
Hidden ref.	0	BRIR	-
	1	BRIR	$RER = -10\,\mathrm{dB}$
	2	BRIR	$RER = -5\,\mathrm{dB}$
Stimuli	3	BRIR	$RER = 0\,\mathrm{dB}$
	4	BRIR	$RER = 5\,\mathrm{dB}$
	5	*HRIR*	equiv. $RER = 0\,\mathrm{dB}$
Anchor	6	*HRIR*	-

Table 10: Additive stereo reverberation. Presented conditions and corresponding parameters.

Tab. 10 lists the presented conditions. The loudspeaker, playing back the input speech signal without reverberation, was used as a reference. Condition 0 is the hidden reference, i.e., the BRIR with no additional reverberation. Conditions 1 to 4 contain additive reverberation with increasing RER, varied in steps of 5 dB. One further condition with additive reverberation was included: to generate condition 5, reverberation was added to the HRIR. In this case, the level of the additive reverberation was set to the equivalent level yielding $RER = 0\,\mathrm{dB}$ with the BRIR. The HRIR without reverberation (condition 6) was also included in the experiment as an anchor.

Trial 3b: Convolutive mono reverberation. This experiment investigates how the presence of reverberation at the input of a binaural decoder influences externalization. In this case, the reference loudspeaker plays back a reverberant signal, as well. Furthermore, it is assessed if the input reverberation leads to a higher tolerance of reverberation in the BRIR. The same conditions were assessed regarding their similarity to a likewise reverberant reference created by convolving the reverberant speech sequence with the HRIR.

As shown in Fig. 18(b), the speech signal is first convolved with one channel of the RIR used in the previous trial. But, in contrast, the RIR contained a direct sound. A unit impulse was placed 3 ms before the reverberant part. The DRR was adjusted to the same DRR as at the listening position. The reverberated speech signal was convolved with the BRIR. Not the level of the input reverberation was varied, but the DRR of the BRIR before convolution.

Response variable:	**Externalization**			**Similarity**		
Function	**ID**	**Impulse Response**	**Input**	**ID**	**Impulse Response**	**Input**
Reference	-	*(real loudspeaker)*	Reverb.	4	HRIR	Reverb.
Hidden ref.	1	BRIR	Reverb.	4	HRIR	Reverb.
Stimuli	0	BRIR	Anech.	1	BRIR	Reverb.
	2	drr 2	Reverb.	2	drr 2	Reverb.
	3	drr 5	Reverb.	3	drr 5	Reverb.
	4	HRIR	Reverb.	5	HRIR	Anech.
Anchor	5	HRIR	Anech.	0	BRIR	Anech.

Table 11: Convolutive mono reverberation. Presented conditions and corresponding parameters. Different impulse responses are convolved either with reverberant (‘Reverb.’) or anechoic (‘Anech.’) speech.

Tab. 11 shows how the individual conditions were generated. The impulse response and their modification, as well as the type of input signal, are listed Conditions with reverberant speech at the input are marked with ‘Reverb.’, whereas ‘Anech.’ stands for anechoic speech. The left side of the table shows the conditions for the externalization part. The loudspeaker playing back the reverberant speech signal was used as a reference. Condition 1, the BRIR with reverberant input, is the hidden reference. Condition 0 is the BRIR with no reverberation at the input. The index 0 was kept for this condition for consistency with all other experiments. Condition 2 and 3 were created by convolving the reverberant speech signal with the BRIR with modified DRR, where drr 2 corresponds to $DRR = 8.3\,\mathrm{dB}$, and drr 5 corresponds to $DRR = 17.3\,\mathrm{dB}$. Condition 4 is the HRIR with reverberant speech at the input, and condition 5 is, for comparison, the same without reverberation.

In the second part, the subjects were asked to rate the similarity to the reference stimulus. The HRIR, convolved with the reverberant speech signal and played back through headphones, was used as the reference. The right side of Tab. 11 shows the conditions and their function. All conditions were identical to the externalization part, except that they were presented diotically by playing back the left channel to both ears.

3.1.4 Part 4: Discrete Reflections

It has been shown that a certain duration of persistent reverberation is required for convincing externalization (Catic et al., 2015). Still, it is an interesting question if the mere presence of single reflections can lead to an increase of externalization whatsoever. The findings of Yuan et al. (2015) suggest a positive influence of early reflections.

Conditions were created by adding isolated binaural reflections to the measured HRIR. Lateral first-order reflections from two parallel walls were simulated. Fig. 19 illustrates the configuration of the simulated space. Specular and diffuse reflections were compared, motivated by the assumption that diffuse reflections may produce less strong timbral artifacts due to comb filter effects.

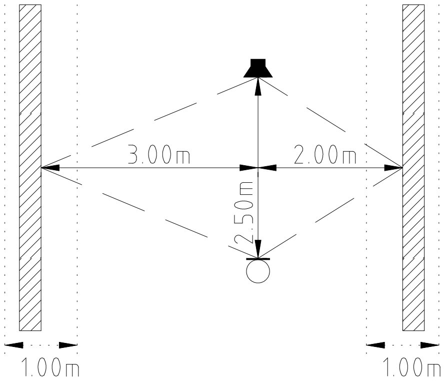
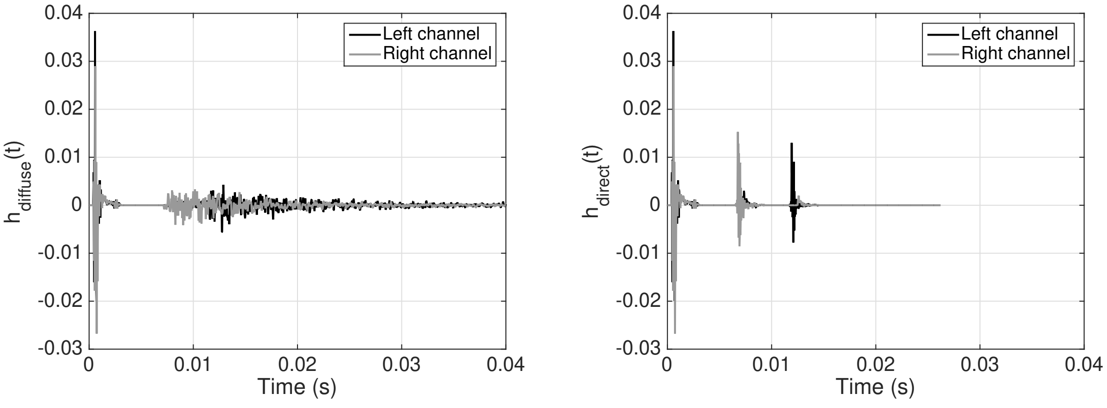


Figure 19: Simulated configuration of a source and a receiver between two parallel walls. Dashed lines denote the acoustic transfer path of first-order specular reflections on each of the walls. Dotted lines illustrate the maximal amplitude of displacement due to the simulated wall roughness for diffuse reflections.

A diffuse reflection can be thought of as the reflection on a rigid wall that scatters the impinging sound, and thus acts as a diffusor. In the case of a pure specular reflection, the incident angle of sound is equal to the emergent angle. In contrast, the sound is scattered in all directions according to Lambert's cosine law if a wall is assumed to produce only diffuse reflections (Wendt and Höldrich, 2018). Specular and diffuse reflections constitute the extreme cases of reflection properties; in reality, one will usually encounter a mixture of both. Diffuse reflections were simulated by modeling two such 'diffuse walls', placed as shown in Fig. 19. The arriving sound power is computed by the numeric integral of the sound intensity that is radiated in the direction of the receiver over all surface elements dA of each wall (cf. Wendt and Höldrich (2018)). In addition, the wall was given a rough surface structure divided into quadratic patches with random displacements orthogonal to the wall. Average displacements of ±0.5 m were used, denoted by dotted lines in Fig. 19. This jitter leads, together with the purely diffuse scattering properties of each patch, to a spatial and temporal spread of the total sound power that arrives at the receiver (cf. Fig. 20(a)). But, most importantly, the jitter functions as a spatial filter leading to a whitening of the frequency response of the reflection. As described by Wegler (2020), it is designed to avoid sound colorations introduced from the interference of partial reflections with different phase due to different travel times. To create binaural reflections, the sound from each contributing direction was convolved with an HRIR corresponding



(a) Diffuse reflections

(b) Specular reflections

Figure 20: Synthesized impulse responses with diffuse (a) and specular first-order reflections (b) of two parallel walls. The reflections are appended to the measured binaural direct sound. Impulse responses with $DRR = 6\,\text{dB}$ are shown.

to the respective incidence angle at the receiver. An HRIR set provided by Bernschütz (2013) was used, obtained with the same dummy head model as used in the measurement of the BRIR in Sec. 2.2.1. Two simulated binaural diffuse reflections were appended to the measured binaural direct sound with varying DRR. The resulting impulse response for $DRR = 6\,\mathrm{dB}$ is shown in Fig. 20(a).

The specular reflections were generated using a first-order image source model. Two Dirac pulses were delayed corresponding to the travel paths shown as dashed lines in Fig. 19, and filtered with the HRIR corresponding to the incidence angle of the respective reflection. The delayed and filtered pulses were then added to the HRIR, resulting in the impulse response shown in Fig. 20(b).

Response variable:	**Externalization**			**Similarity**		
Function	**ID**	**Impulse Response**	**DRR**	**ID**	**Impulse Response**	**DRR**
Reference	-	*(real loudspeaker)*		6	HRIR	
Hidden ref.	0	BRIR		6	HRIR	
Stimuli	1	HRIR + diff. refl.	2 dB	1	HRIR + diff. refl.	2 dB
	2	HRIR + diff. refl.	6 dB	2	HRIR + diff. refl.	6 dB
	3	HRIR + diff. refl.	11 dB	3	HRIR + diff. refl.	11 dB
	4	HRIR + spec. refl.	6 dB	4	HRIR + spec. refl.	6 dB
	5	`trc 3`		5	`trc 3`	
Anchor	6	HRIR		0	BRIR	

Table 12: Discrete reflections. Presented conditions and corresponding parameters.

Tab. 12 shows the conditions of this experiment. For each part (externalization, similarity to the reference), the table lists the impulse responses used for creating the conditions as well as the value the DRR was set to. Participants were asked to rate externalization in comparison to the loudspeaker through which the original speech sequence was played back. The conditions evaluated regarding externalization are shown on the left side of the table. The BRIR (condition 0) is considered the hidden reference. Conditions 1 to 3 were created based on the impulse response with two binaural diffuse reflections. Three different conditions were created by varying the DRR. The diffuse reflections were added with $DRR = \{2, 6, 11\}\,\mathrm{dB}$. Condition 4 contains two specular reflections with $DRR = 6\,\mathrm{dB}$. For comparison, a truncated condition was added.

Condition 5 is identical with condition `trc3` in Ch. 2. This condition led to slightly decreased externalization and increased sound colorations. The HRIR convolved with speech (condition 6) was included as an anchor.

The participants were also asked to rate the similarity to speech convolved with the HRIR within the anechoic room. The conditions are shown in Tab. 12 on the right. Condition 6 is identical to the reference and functions as the hidden reference. In turn, the BRIR (condition 0) is the anchor. The presented set of stimuli stayed the same, except that all conditions were presented through headphones diotically.

3.2 Results

This section presents the results of the above-described listening experiments. The experiments were conducted with seventeen participants, most of them with trained hearing and previous experience in listening experiments. Each trial was carried out by all of the participants. It took between 30 and 60 minutes to complete the experiment. Analogously to the first experiment, the participants first completed all trials in which externalization to the loudspeaker had to be rated. This part took place in the lecture room of the IEM. Then they were led into the anechoic chamber for the trials in which similarity to the reference was to be assessed via headphones. All trials were presented in random order. Before the analysis of the results, it will be discussed what statistical methods to apply.

3.2.1 Statistical Methods

In the first part of this thesis in Sec. 2.4.1, it was found that a substantial part of the collected experimental data did not follow a normal distribution. Moreover, it was argued that, due to significant differences in the rating of repeatedly occurring conditions, the data should not be interpreted on higher than ordinal level. It was therefore decided to rely only on non-parametric methods. In this section, the same analyses will be carried out to determine the applicable methods.

Normal distribution. Since the choice of methods depends on the distribution of the data, a Lilliefors test for normal distribution was carried out on the ratings of every stimulus in each part of the present experiment. Tab. 13 lists the results for each part. The H_0 that a sample stems from a normally distributed population was rejected in 30% of cases on average for externalization and in 43% of cases on average for similarity. The percentage of non-normally distributed data is therefore on a similar scale as in the first experiment.

Level of measurement. As already discussed in Sec. 2.4.1, it is questionable if the subjects were able to perform a direct estimation of magnitude of the response variable, or if the results should rather be understood as a ranking. Or, to put it differently: if the data should be interpreted on the interval or ordinal level. While it is generally advisable to regard data obtained using the MUSHRA paradigm on the ordinal scale and to employ non-parametric statistical tests (Mendonça and Delikaris-Manias, 2018), as it was also

		H_0 **rejected (% of cases)**	
Part	**Trial**	Externalization	Similarity
1 Two-fold BRIR modification	**a**	17	50
	b	29	29
2 Diffuse binaural reverberation	**a**	40	–
	b	33	–
3 Additional reverberation	**a**	33	50
	b	29	–
4 Discrete reflections		29	43
Mean		30	43

Table 13: Percentage of cases in which the the H_0 *'normally distributed'* has been rejected.

done in Ch. 2, this also means to discard potential further information on the interval scale. The comparison of repeatedly occurring conditions may again serve as a benchmark. Of course, only those conditions can be taken into consideration that were tested with the same reference and anchor. That includes all parts except trial 3b where reverberation was added to the reference stimulus. Tab. 14 shows which stimuli have been tested repeatedly in what parts of the experiment. For each redundant stimulus and each type of task (externalization or similarity to the reference), a Friedman test was carried out to test for an effect of the membership of the respective stimulus to a particular trial. Subject to a significance level of 5%, all of the, in total, eleven Friedman tests led to a non-significant result. Therefore, the hypothesis that the membership of any redundant stimulus to a particular trial has no effect on the ratings has not been rejected. Based on this finding it was decided that the data can be viewed on an interval scale. This allows also for the use of the paired-samples t-test where samples do follow a normal distribution.

Hypothesis testing. To find out which stimuli are rated different from each other, statistical tests were carried out for every pair of stimuli within one part of the experiment. It was decided to follow a three-step procedure.

For every pair, it was first tested if the normality assumption is violated for any of the two samples. In the second step, the appropriate statistical test was selected based on the result. If the H_0 'the data follow a normal distribution' had been rejected, the Wilcoxon signed-rank test was carried out. A dependent-samples t-test was carried out otherwise. All obtained p-values were corrected using the Bonferroni-Holm method to compensate for

Part	**Trial**	BRIR	HRIR	drr 1	drr 2	trc 3	*Bin. Omni.*
1 Two-fold modification	**a**	✓	✓	✓	✓		
	b	✓	✓	✓	✓	✓	
2 Diffuse binaural reverb.	**a**	✓	✓				✓
	b	✓	✓				✓
3 Additional reverberation	**a**	✓	✓				
	b						
4 Discrete reflections		✓	✓			✓	

Table 14: Redundant conditions throughout the different trials. Check marks indicate stimuli that were tested multiple times with the same reference and anchor. '*Bin. Omni.*' refers to the omni-directional binaural reverberation in trial 2a-b. Except for the latter, all conditions were rated both regarding externalization and similarity.

the inflation of the α error due to multiple testing. In the following, the term 'significantly different' refers to a corrected p-value below the significance level $p < .05$, found either using the signed-rank or the dependent-samples t-test depending on the distribution of the samples under test. In the third step, the effect size was computed.

Effect size. While the p-value of a hypothesis test merely signals the presence of an effect, however small, the effect size indicates the magnitude of the observed effect. The importance of reporting effect sizes has recently been emphasized in the literature (Fritz et al., 2012; Tomczak and Tomczak, 2014). Furthermore, due to the exploratory character of this study, the computation of effect sizes is considered useful to put the analyses on more solid ground.

Widely used measures of effect size for normally distributed data to accompany the t-test are Cohen's d, the mean difference divided by the (estimated) population standard deviation (Cohen, 1988), and Hedge's g, the mean difference divided by the weighted pooled-samples standard deviation (Hedges and Olkin, 1985). The latter is more appropriate for small sample sizes as no bias is introduced due to the possibly false assumption of equivalence of the population and the sample standard deviations.

Both d and g are defined on the interval of $[0, \infty)$, where the value 0 corresponds to the absence of an effect. Values of 0.2, 0.5, and 0.8 are generally associated with small, medium, and large effects if no typical magnitude of effect size is available for comparison from the field of study (Cohen, 1988).

Non-parametric measures that are compatible with the use of the signed-rank test are less widely adopted in the literature (Kerby, 2014). An appropriate measure is the *matched-pairs rank-biserial correlation r* (King and Minium, 2008). The coefficient r takes on values in the interval of $[-1,1]$, where 0 is considered no effect, and the sign tells about the direction of the difference. For different correlation measures of effect size, the criteria for small, medium, and large effects are usually considered to be 0.1, 0.3, and 0.5 (Cohen, 1988).

Since the ranges and characteristic values of g and r are quite different, it is desirable to find a conversion between both coefficients in order to obtain a comparable and comprehensive measure. A conversion from d to r (and, thus, from g to r) is given by Cohen (1988) with

$$r' = 1.253 \cdot \frac{d}{\sqrt{d^2+4}} \tag{16}$$

for the case of two samples of equal size, where r' is the equivalent *point*-biserial correlation. The conversion factor 1.253 takes account of mapping the characteristic values of d to those of r'. In the following, the absolute value of r will be reported in the case of non-normal samples, and g will be computed and converted to r' elsewhere.

Visualization. While it was decided to accept the hypothesis of interval-scaled data and, therefore, the mean value can be computed, the confidence intervals of the mean are not meaningful for the portion of non-normal data. Therefore, the ratings will be visualized using the median along with its 95 % confidence intervals. Since no significant differences were found between the ratings of repeatedly occurring conditions, the ratings are pooled for each of the redundant stimuli (cf. Tab. 14). This will be done for visualization only; the following plots of the test results will show the pooled ratings next to the original ratings. In cases where also the similarity to the reference was evaluated, additional plots are provided showing externalization and similarity together in two dimensions. The purpose of these supplementary plots is to provide a more intuitive, qualitative view of the data. In this case, the mean and 95 % confidence ellipses will be displayed since, for the median, the computation of the corresponding confidence areas is non-trivial.

Validation. To validate the obtained ratings, it was inspected whether the subjects perceived the original BRIR as externalized. This was carried out for all externalization parts except for the experiment with convolutive reverberation in Sec. 3.1.3 because of the different reference stimulus. The BRIR received ratings below the ‘close-to-the-head’ threshold four times from a single subject. All other subjects rated the BRIR consistently as externalized. The ratings of the subject in question were, as a consequence, excluded from the following analysis. This reduces the effective number of participants to sixteen.

Statistical power. How reliable are the results of the pairwise comparison, given the ratings of sixteen subjects? The reliability of a statistical test is characterized by the type I and type II errors. The type I error is the probability α of a false positive, i.e., rejecting the H_0 while being true. This probability can be controlled with the significance level defined as $\alpha = 0.05$ in the present case. The type II error is the probability β of a false negative, i.e., retaining a false H_0. Power analysis can be used as a means to determine both the type II error and the required sample size to yield a given power. The power of a test is the probability $1-\beta$ of rejecting a false H_0 and can be computed for the paired-samples t-test according to (Cohen, 1988). [5] Given a desired power of $1-\beta = 0.9$ and a significance level of $\alpha = 0.05$, the number of subjects needed for detecting a small effect of $d = 0.2$ would be $N = 265$; for detecting a medium and large effect, $d = \{0.5, 0.8\}$, the required number would be $N = \{44, 19\}$, respectively. Unfortunately, with the ratings of the remaining sixteen subjects, the expected probability of detecting a small, medium, or large effect is only $1-\beta = \{0.12, 0.46, 0.85\}$. This means that, in the case of normally distributed samples, the t-test is considerably underpowered to detect small and medium differences. Large effects can be detected with reasonably high probability. A non-parametric test such as the signed-rank test depends on the underlying distribution, but can be expected to have lower power than a parametric test if its assumptions would be met. With this knowledge, non-significant results have to be regarded with special caution.

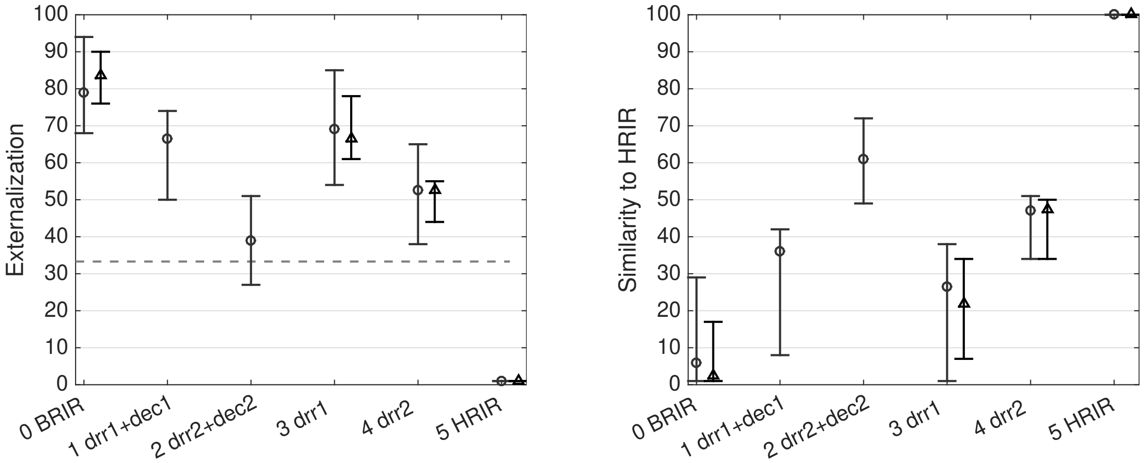
Preliminary analysis. As a first approach to the collected data, a Friedman test was carried out within each part of the experiment to test whether the *treatment* (the variation of the stimulus) has a significant effect. The non-parametric Friedman test is, rather than the single-factor ANOVA, appropriate for the collected data because a not negligible fraction of it is not normally distributed. A significant effect of the treatment was found in every trial of the experiment with $p < 0.05$.

[5] The power analysis was carried out using the R library pwr [available online]: `https://cran.r-project.org/web/packages/pwr/pwr.pdf` (last accessed: 2020/05/14)

3.2.2 Part 1: Two-fold BRIR Modification

This section discusses the results of the joint modification of the measured BRIR using two different methods from the experiment in Ch. 2. The modification of the DRR has been tested in combination with either altering the decay time or the impulse response length. The conditions are denoted by their abbreviations. The abbreviations and the corresponding parameter values can be found in Tab. 6 and 7.

Trial 1a: Joint modification of DRR and reverberation time. The externalization ratings of this trial are shown in Fig. 21(a). The BRIR received tendentially the highest ratings. Only `drr 1` is not significantly different from the BRIR. Compared to condition `drr 1`, the combination of `drr 1` and `dec 1` led only to a small, insignificant decrease in externalization. The same applies for `drr 2` and the combination of `drr 2` and `dec 2`. With or without being combined with the decay modification, `drr 2` is consistently rated significantly lower than `drr 1` with large effects of $r' > .5$. With a score of 0 to 1, the HRIR received the lowest ratings.

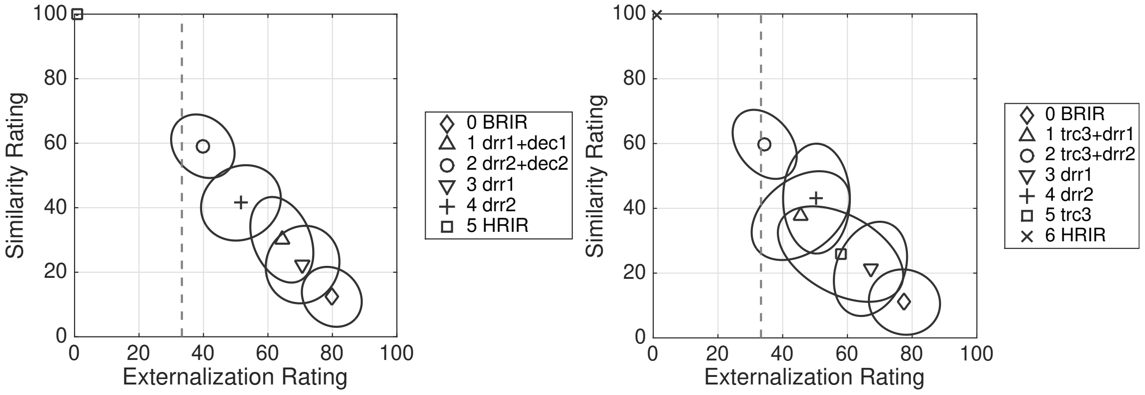


(a) Externalization

(b) Similarity to the reference

Figure 21: Two-fold BRIR modification: Joint modification of DRR and reverberation time. Median and 95 % confidence intervals. Black triangular markers denote pooled data.

Fig. 21(b) shows the ratings of similarity to the HRIR. If one imagines the median ratings to be connected by line segments, the resulting curve is nearly the reflection of the externalization ratings about the centerline of the plot: Each positive difference in externalization appears to be associated with a negative difference in similarity of comparable magnitude. The two-dimensional plot of similarity over externalization in Fig. 22(a) shows that the mean values of each method are almost perfectly aligned on a straight diagonal line with a negative slope, indicating a mere anti-proportional relationship between externalization and similarity.



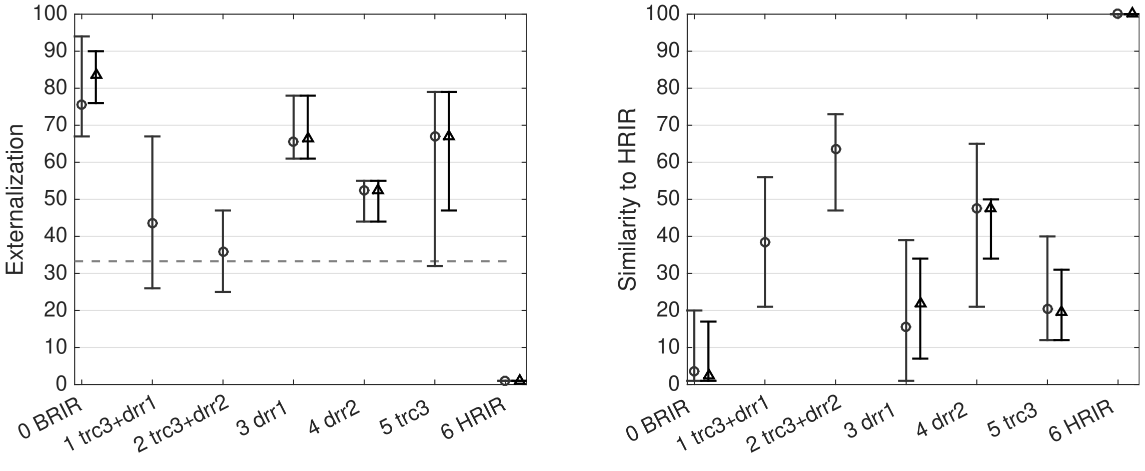
(a) DRR and reverberation time

(b) DRR and BRIR length

Figure 22: Two-fold BRIR modification: Mean value along with the 95 % confidence ellipses. Similarity to the reference on the vertical axis over externalization on the horizontal axis.

Trial 1b: Joint modification of DRR and BRIR length. Fig. 23(a) shows the externalization ratings for the combination of truncation and modified DRR. All participants gave the HRIR a score of either 0 or 1. The BRIR yielded the highest ratings, but is not significantly different from `drr 1` and `trc 3`. Condition `trc 3` exhibits the largest variation of all presented conditions. The confidence interval includes values from inside-head localization up to decent externalization. Except for the HRIR, the ratings of `trc 3` do not differ significantly from any other stimuli. Combining `trc 3` with `drr 2` results in a significant decrease with $r' = .58$ (large effect). The combination of `trc 3` with `drr 1` leads only to slight, insignificant decrease with $r' = .30$ (medium effect). As observed before when combining the DRR and decay methods, `drr 2` was rated significantly lower than `drr 1` with $r = .98$. However, the difference is not significant anymore if both conditions are combined with `trc 3`, and also less pronounced both visually and in effect size ($r' = .36$ in contrast to $r' = .81$).

Fig. 23(b) shows the ratings of similarity to the HRIR. All conditions were rated lower than the HRIR which received a rating of 100 from all subjects. Condition `trc 3` was rated



(a) Externalization

(b) Similarity to the reference

Figure 23: Two-fold BRIR modification: Joint modification of DRR and BRIR length. Median and 95 % confidence intervals. Black triangular markers denote pooled data.

significantly less similar to the HRIR than drr 2 ($r' = .43$) as well as the combination of trc 3 with drr 2 ($r' = .47$). But, as expected, both were rated more similar to the HRIR than the BRIR. The similarity ratings are again roughly a mirror image of the externalization ratings. In the two-dimensional plot in Fig. 22(b), the conditions are not as clearly aligned as in the previous part, but, considering the confidence intervals in Fig. 23, the relationship is similarly anti-proportional.

Discussion. In the two parts of this experiment, the combination of the DRR and the decay method, and the combination of the DRR and the truncation method were evaluated in comparison to applying the modifications separately. The ratings of redundant conditions between the two trials are consistent. In order to check for the consistency between this experiment and the experiment in Ch. 2, the ratings of the conditions that were tested in both experiment (the BRIR, drr 1, drr 2, trc 3, and the HRIR) were pooled within trials and then compared between the experiments using the rank-sum test. Significant differences were found for trc 3 regarding externalization and drr 2 regarding similarity to the HRIR. As a consequence, ratings of the combined conditions can, unfortunately, not directly be compared with the results from Ch. 2.

By and large, the relationship between externalization and similarity appears to be anti-proportional in the first trial, the combination of the DRR and the decay methods. Regardless of the method used, an increase in similarity to the HRIR is associated with a reduced degree of externalization. In the second trial, the combination of the DRR and the truncation methods, the trend is similar but less obvious.

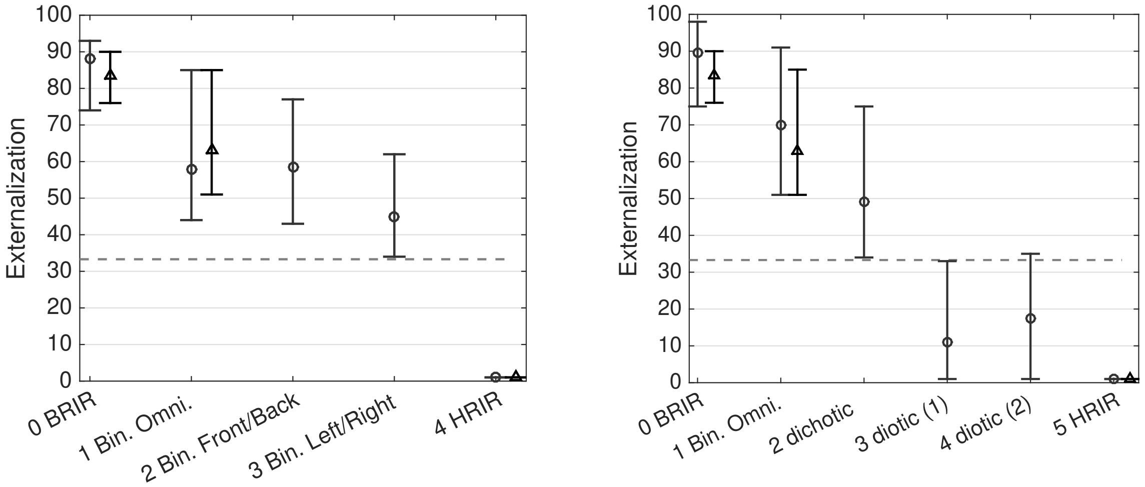
The combined-methods BRIR modifications were tested to find out if a better sound quality compared can be achieved to the single modifications while maintaining externalization. The results show no evidence for this hypothesis. However, since only a few particular conditions were tested, the results can not be considered representative of the whole parameter ranges of the modification methods. Thus, further studies would be needed to investigate this question.

3.2.3 Part 2: Diffuse Binaural Reverberation

In this part, different parameters of diffuse binaural reverberation have been varied. Following a simple algorithm, diffuse reverberation was created to resemble the measured reverberation as best as possible regarding frequency-dependent decay and sound color. The reverberation was appended to the binaural direct sound, i.e., the measured HRIR. Two different scenarios have been implemented. In the first scenario, binaural reverberation with varying spatial distribution was compared. The second scenario investigated differences between different types of dichotic and diotic reverberation.

Trial 2a: Varying spatial distribution. Fig. 24(a) shows the externalization ratings of the conditions with varying spatial distribution of the reverberation. Omni-directional binaural reverberation (condition 1) was compared with a scenario in which only two sphere segments at the front and back (condition 2) or on the left and right (condition 3) contributed to the reverberation.

The BRIR (condition 0) received the highest ratings of all conditions with significance and $r > .5$. All subjects perceived the HRIR to be located inside the head. The conditions with artificial reverberation yielded ratings of externalization in the mid-upper range. Condition 3, the HRIR with reverberation originating only from two lateral segments, was rated significantly lower compared to the HRIR with reverberation from the full sphere, $r' = .40$. In contrast, for reverberation originating from two segments in the front and back (condition 2), no significant difference to omni-directional reverberation was found, and the effect size is small with $r' = .13$. The difference between frontal and lateral reverberation is not significant, $r' = .32$



(a) Varying spatial distribution

(b) Dichotic and diotic reverberation

Figure 24: Diffuse binaural reverberation. Median and 95 % confidence intervals. Black triangular markers denote pooled data.

Trial 2b: Dichotic and diotic reverberation. This trial compared different types of dichotic and diotic reverberation regarding externalization. Omni-directional diffuse binaural reverberation (condition 1, identical to condition 1 in Trial 2a) and decorrelated dichotic reverberation (condition 2) were presented. In contrast to condition 1, condition 2 was generated using two independent noise instances for the left and right ear. Therefore, it provides no meaningful directional binaural cues in the reverberation. Condition 3 and 4 are the diotic conditions – diotic in the sense that the reverberation at both ears is identical, but not the binaural direct sound. One channel of the binaural reverberation in condition 1 was played back both ears to create condition 3, and one channel of the reverberation of condition 2 was used for condition 4. The conditions are therefore named ‘diotic (1)’ and ‘diotic (2)’ in Fig. 24(b) where the results of this trial are shown.

All stimuli were rated significantly lower than the BRIR with $r > .5$. The HRIR was given the lowest ratings. All stimuli received ratings that differ from each other significantly and distinctly ($r > .49$), except for diotic reverberation, conditions 3 and 4, with a non-significant difference and a very small effect size $r' = .08$. The latter two conditions were furthermore not externalized; the confidence intervals do not exceed the ‘close-to-the-head’ threshold. The dichotic ‘stereo’ condition (condition 2) was rated significantly lower than omni-directional binaural reverberation (condition 1). However, it was still perceived clearly outside of the head.

Discussion. Two different experiments have been carried out based on speech convolved with synthetic BRIRs. The BRIRs consisted of the measured direct sound and synthetic diffuse reverberation. The use of these synthetic BRIRs allowed for the variation of different parameters. In the first trial, omni-directional, frontal, and lateral binaural reverberation were compared. Different types of dichotic and diotic reverberation were tested in the second trial. Both trials included the measured BRIR and HRIR in the presented set of stimuli, as well as a synthetic BRIR comprising diffuse binaural reverberation from all directions on the sphere. The latter was expected to yield the best ratings in the externalization of all synthetic conditions. It received an externalization score of 68 on average of both parts. Taking into account the simplicity of synthesis, this can be considered an acceptable baseline synthesis with enough footroom to resolve detriments in externalization within the outside-the-head range.

In the first trial, no significant differences could be observed between the binaural synthesis with omni-directional reverberation and the synthesis featuring only spherical segments in the front and back. Likewise, no differences were significant between the frontal and the lateral synthesis. The lateral synthesis was rated significantly lower than the synthesis on the full sphere. Based on the obtained data, it can not be concluded with certainty if the frontal synthesis differs in externalization from the omni-directional or the lateral synthesis.

However, the confidence intervals of the frontal reverberation overlap completely with those of the omni-directional binaural reverberation. The apparent tendency that the frontal reverberation appears to be superior to the lateral one differs from the expectations.

In order to analyze the conditions regarding the dynamic binaural cues, the short-time interaural coherence (IC) of the presented stimuli was computed as the maximum of the moving normalized interaural cross-correlation function (IACF, cf. Blauert and Lindemann (1986)), considering only the interval $\tau \in [-1,1]$ ms of meaningful interaural time differences (cf. Li et al. (2018)). Different studies have found a negative relationship between the IC and externalization (Catic et al., 2015; Li et al., 2018). Because the mean IC of the stimulus with frontal reverberation is with 0.84 higher than the mean IC of the stimulus with lateral reverberation with 0.60, the expected result would be a reduced externalization for the frontal condition. A possible explanation for the opposite tendency is that a frontal sound source with lateral-only reverberation may lead to an odd room impression. Due to the ‘acoustic gap’ between the reverberant lateral sectors and the frontal speech stimulus, both may be perceived as separate sound events, whereas the frontal reverberation merges nicely with the speaker. Reverberation in the back is unlikely to interfere because of the assumable high chance of front/back confusion. However, this is a hypothesis. Further experiments could investigate if and under what circumstances such an 'acoustic gap' can emerge. Beyond that, it is yet to quantify the influence of the frontal and lateral sector width on externalization.

In the second trial, the synthetic BRIR with diotic 'stereo' reverberation was perceived outside the head in the mid-range of the scale. Its externalization score is comparable to the frontal and lateral synthesis in the preceding trial. This indicates that sufficiently long dichotic reverberation, in contrast to diotic reverberation, may already increase externalization substantially (compared to the HRIR) – at least in the present case where the left and right channel were created based on uncorrelated noise. However, further experiments are required to draw a definite conclusion.

The outside-the-head-locatedness of the dichotic 'stereo' condition overlaps with both the findings of Hassager et al. (2016) and Jiang et al. (2020) who showed that the importance of accurate reproduction of spectral detail and pinna cues in the reverberant part of a BRIR diminishes. However, this condition constitutes a special case as it was generated using uncorrelated noise sequences for both ears that were equalized to the instantaneous spectral envelope of the original BRIR. It should be pointed out that, in mentioned publications, mainly monaural cues were affected by the applied modifications. Here, by equalization to the smoothed spectral envelope of the BRIR, only the average direction-independent monaural cues were retained. All meaningful direction-dependent binaural cues were removed. This is a likely explanation for the significantly lower rating of the 'stereo' condition, compared to binaural reverberation.

In the case of the diotic conditions, 3 and 4, the left and right channels of the reverberation were identical, leading to a high interaural coherence. The conditions were localized inside or close to the head. Since condition 3 was created by taking the left channel of the binaural reverberation, it contains the monaural cues of the left ear. However, due to the equalization to the original reverberation, the average monaural cues of the BRIR are also contained in condition 4. As a result, no difference in externalization was found between condition 3 and 4. The poor externalization of the diotic conditions in comparison with dichotic reverberation agrees with the findings of Catic et al. (2015)

who found that monaural reverberation was insufficient for the externalization of frontal sources.

It can be concluded that the uncorrelatedness of the left and right channels of the dichotic ‘stereo’ reverberation, causing ILD fluctuations and decreased interaural coherence between the left and right channels, likely led to the increased externalization in comparison to the diotic reverberation.

3.2.4 Part 3: Additive and Convolutive Reverberation

This section presents the results of two experimental trials. In both of them, the effect of additional reverberation was investigated. Reverberation was generated using the simulation of a stereophonic impulse response. The simulated room had a longer reverberation time than the listening room. No equalization to the measured BRIR was performed. Two scenarios were investigated: Stereo reverberation added to the BRIR, and mono reverberation added to the speech signal prior to convolution with the BRIR.

Trial 3a: Additive stereo reverberation. In the first part of this experiment, stereophonic reverberation was added to the binaural signal obtained by the convolution of the BRIR with the speech sequence. The amount of additional reverberant energy to the total amount energy of the BRIR was successively increased from $RER = -10\,\mathrm{dB}$ to $RER = +5\,\mathrm{dB}$ (cf. Eq. 15). For comparison, the HRIR was presented with and without stereo reverberation at an equivalent RER of $0\,\mathrm{dB}$. The conditions are listed in Tab. 10. Externalization was evaluated in comparison to the loudspeaker without any reverberation.

Fig. 25 shows the externalization ratings. The unmodified BRIR was rated significantly different from conditions 4-6, i.e., the HRIR, the HRIR with additional reverberation, and the BRIR with $RER = 5\,\mathrm{dB}$. Judged by the range covered by the confidence intervals, a general downwards trend of the ratings can be observed with increasing condition index. While the ratings of the other conditions do not deviate from the BRIR with significance, the downwards trend of the ratings goes along with increasing magnitude effect size from $r = .2$ for condition 1 to $r' = .8$ for condition 5.

The addition of reverberation to the HRIR was, however, sufficient to move the produced sound image from inside the head to outside. Only four of sixteen subjects rated the HRIR with reverberation (condition 5) lower than the ‘close-to-the-head’ mark. The BRIR with the largest portion of reverberation, condition 4, was rated very similar with five out of sixteen subjects perceiving it inside or close to the head. Condition 3 and 4 do not differ significantly from the HRIR with reverberation ($r' = \{.31, .11\}$). All other conditions do differ from the reverberated HRIR with significance and effect sizes of $r, r' > .5$.

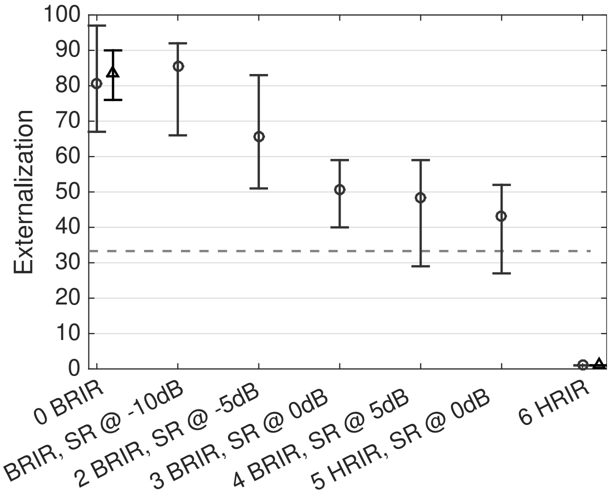
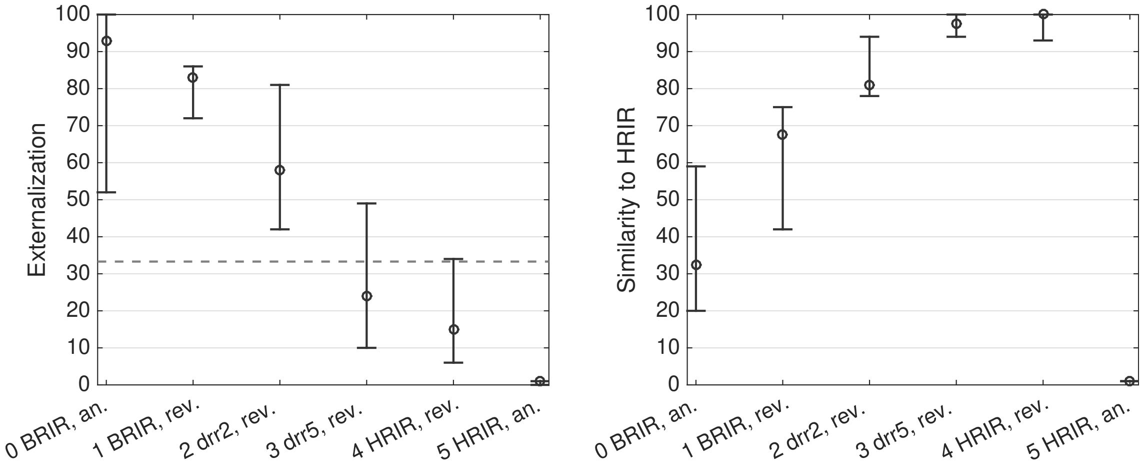


Figure 25: Additive Stereo reverberation. Median and 95 % confidence intervals. Black triangular markers denote pooled data. Conditions with additive reverberation are denoted by *SR* and the corresponding *RER* value.

Trial 3b: Convolutive mono reverberation. In contrast to the previous trial, monophonic reverberation was added to the input speech signal. The reverberant excitation signal was then convolved with the following impulse responses: the measured BRIR (condition 1), the BRIR with modified DRR (condition 2 and 3 corresponding to `drr 2` and `drr 5`), and the HRIR (condition 4). See Tab. 11 for the list of conditions. Externalization was rated relative to the loudspeaker playing back the reverberant speech sequence. The measured BRIR and HRIR were also presented without reverberation for comparison (condition 0 and 5). Furthermore, the similarity to a reverberant reference had to be rated.

The externalization ratings are shown in Fig. 26(a). The HRIR without reverberation was given a score of 0. The original BRIR received the highest externalization ratings and differs significantly from the HRIR with reverberation ($r = .85$), and `drr 5` with reverberation ($r = .94$). Both were perceived as poorly externalized. The difference between the BRIR and condition 2, `drr 2` with reverberation, is not significant but has a magnitude effect size of $r = .53$ with the sign indicating higher externalization of the BRIR.

The difference between the BRIR with and without reverberation is not significant, and the effect size $r = .15$ is small. It can be observed that the ratings of the BRIR with reverberant speech vary a lot less than those of the BRIR with anechoic speech. In contrast to the latter, the BRIR with reverberant input differs significantly from `drr 2` with reverberant input ($r = .57$).



(a) Externalization

(b) Similarity to the reference

Figure 26: Convolutive mono reverberation. Median and 95 % confidence intervals.

Fig. 26(b) shows the similarity to the reference, i.e., the HRIR convolved with monophonic reverberation. The BRIR with reverberation (condition 1) was rated significantly more similar to the reference than the BRIR without reverberation (condition 0) with $r' = .59$. The original, 'dry' HRIR (condition 5) was rated 0 by all subjects, whereas the HRIR with mono reverberation, i.e., the hidden reference (condition 4) received a score close to 100. The ratings of the hidden reference exhibit some variation. A closer look at the individual ratings reveals that condition 3 (`drr 5` with reverberation), receiving ratings close to 100 as well, was confused with the hidden reference by four subjects. It should be noted that the subjects were not explicitly instructed to identify the hidden reference or to rate at one stimulus with 100. However, in the other parts, the similarity rating of the hidden reference was unanimously at 100 without any confusions. Condition 2 (`drr 2` with reverberation) was rated significantly less similar to the reference than condition 3 (`drr 2` and `drr 5` with reverberation) with $r = .89$. Next, the BRIR with reverberation (condition 1) was rated lower than condition 2 with $r' = .67$, and the BRIR without reverberation (condition 0) was rated lower than the BRIR with reverberation with $r' = .57$.

For all stimuli except for the HRIR without reverberation, an increase in externalization is still associated with a decrease in similarity. But, as Fig. 27 shows, the ratings of similarity over externalization do not lie on a diagonal line anymore. Instead, drr 2 with reverberation and the BRIR with reverberation lie in the first quadrant of the graph, indicating simultaneously increased externalization and similarity to the reference.

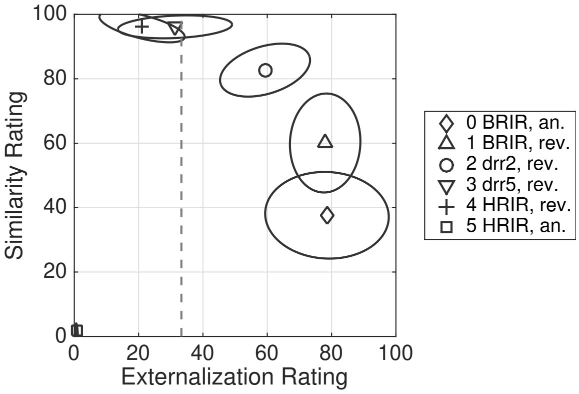


Figure 27: Convolutive mono reverberation: Mean value along with the 95 % confidence ellipses. Similarity to the reference on the vertical axis over externalization on the horizontal axis.

Discussion. This experiment investigated the effect of additional reverberation in two different scenarios. In the first scenario, stereophonic reverberation was added at different levels after convolution with the BRIR. The subjects had to rate externalization in comparison to speech without reverberation played back through the loudspeaker. The second scenario was about the influence of monophonic reverberation present in the excitation signal. Externalization and similarity were rated in comparison to a reference containing reverberation, as well.

In the first part, no evidence was found for the hypothesis that stereo reverberation may have a positive effect on externalization. Instead, the tendency of decreasing externalization could be observed with increasing amount of reverberation added to the BRIR. This may well be an example of the emergence of the room divergence effect, i.e., a mismatch between the real and the synthesized room affecting externalization. It is also plausible that binaural cues present in the BRIR have been masked by the additional reverberation.

It was found that the externalization of speech convolved with the HRIR (condition 5) could be substantially improved by adding stereophonic reverberation. This result is similar to trial 2b in Sec. 3.2.3 where the HRIR with dichotic reverberation (condition 2) was similarly externalized (Fig. 24(b)). However, there are differences between these conditions. In the present experiment, a room simulation with longer reverberation time than in the listening room was used. In contrast, the reverberation time in trial 2b was adjusted to match the listening room. The noise sequences used to generate the reverberation were uncorrelated between the two ears and were equalized to the BRIR in order to reduce differences in sound color. Thereby, the average monaural spectral cues of the dummy head were applied. In contrast, no equalization was carried out here. It can be assumed based on these findings and relying on (Hassager et al., 2016) and (Jiang et al., 2020) that monaural spectral cues are dispensable within the reverberant part of the BRIR.

In the second part, externalization and similarity were assessed with reverberant conditions and reference signals. A reverberant speech signal was convolved with the BRIR. The BRIR, in turn, was manipulated by increasing the DRR. Furthermore, the BRIR and HRIR were also tested with an anechoic speech signal. Externalization was rated in comparison to the loudspeaker playing the reverberant speech sequence. Similarity was rated in comparison to the HRIR convolved with reverberant speech.

Reducing the reverberation in the BRIR had the expected negative effect on externalization and a positive effect on the similarity to the reference. This has been observed similarly in the previous experiments. What is more interesting is the comparison of reverberant with anechoic speech, considering also the preceding experiments. In the present experiment, the difference between the BRIR with and without reverberation in the excitation signal is not significant. Moreover, the effect size is small with $r = .15$. The difference seems to lie rather in the noticeably greater spread of the ratings of the unmodified BRIR. This may indicate that the subjects were confused by the presented scenario in which a stimulus (the BRIR anechoic speech), though externalized, sounded very different from the reference with reverberant speech.

The BRIR, the HRIR, and condition `drr 5` have also been evaluated with the combined-methods BRIR modifications in Sec. 3.2.2, and in the first experiment in Ch. 2. Condition `drr 2` has been tested in Ch. 2, too. The HRIR with reverberant speech was rated more externalized than with anechoic speech, although still perceived close to the head. Moreover, the externalization score of the HRIR with reverberation is higher than the score of the corresponding anechoic condition in Ch. 2. The HRIR with convolutive reverberation might have been rated higher because of a more similar sound to the reverberant loudspeaker reference, compared to the HRIR with anechoic speech. Assessing the externalization of the other reverberant conditions with a reverberant reference has hardly led to different ratings. The BRIR BRIR, drr 2, and drr 5 were rated similarly externalized with anechoic speech in Sec. 3.2.2 and in the previous experiments in Ch. 2.

Considering the perceived similarity to reverberant speech convolved with the HRIR, one difference to the other experiments without reverberation is striking. Both the HRIR with reverberation (the hidden reference) and condition drr 5 received a very high similarity score and there is no indication of a difference between the conditions based on the present data. The comparison with Ch. 2, however, shows that drr 5 was rated 25 points lower than the hidden reference in the case of anechoic speech. A similar trend can be observed for the BRIR and condition drr 2. Ergo, a small amount of reverberation in the impulse response of drr 5 leads no longer to significant similarity differences to the anechoic reference if the excitation signal contains a larger amount of reverberant energy.

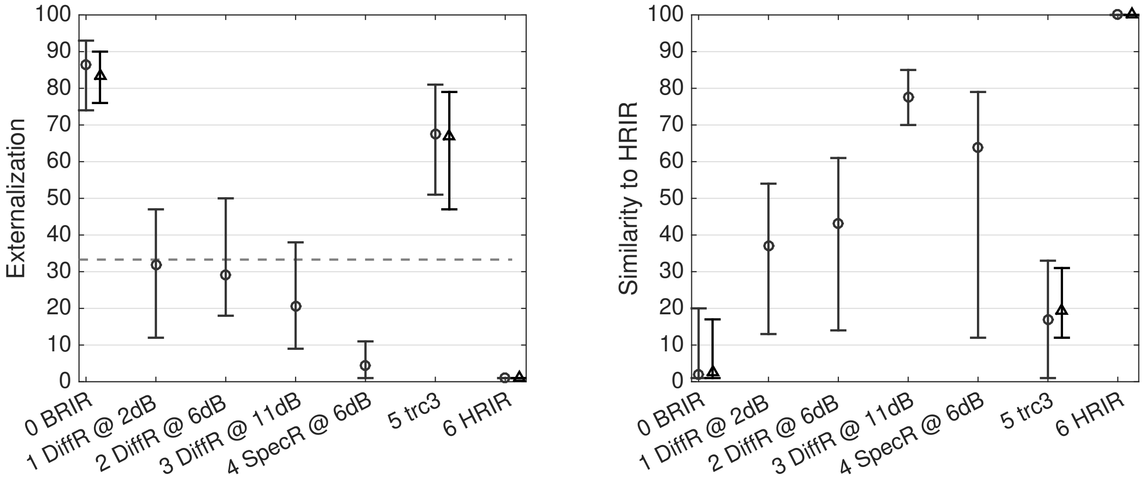
The similarity ratings of the BRIR with reverberation at the input are significantly higher than of the BRIR without reverberation, which in turn received a higher median score than the same BRIR in Ch. 2 and Sec. 3.2.2. Expressed in numbers: while the BRIR was rated 95 points lower than the HRIR in Ch. 2 and 94 points lower than the HRIR in Sec. 3.2.2, here, the reverberant BRIR was only 31 points less similar to the reverberant HRIR than the hidden reference. So even for the larger amount of reverberant energy in the impulse response – compared to condition drr 5 – the similarity to the reference increases if the excitation signal contains reverberation.

In summary, it can be said that there is no strong indication for the original binaural reverberation to be less important for externalization in the presence of monophonic reverberation at the input. However, from the comparison with other experiments with anechoic speech can be inferred that the binaural reverberation in the BRIR may have less influence on sound quality if the input signal is reverberant itself. Furthermore, while adding stereophonic reverberation did not improve externalization if added to the BRIR, it did lead to an increase in externalization if added to the HRIR. Yet, only the reverberation from one particular room simulation was treated here. The 'added room' had a longer reverberation time than the listening room in which the BRIRs were measured. One can imagine that a variation of the present constellation may lead to different results.

3.2.5 Part 4: Discrete Reflections

To study the possible influence of a small number of reflections on externalization, two parallel walls were simulated. From each wall, one reflection was added to the measured HRIR. Diffuse reflections at three different levels were compared with direct (specular) reflections and the measured BRIR. For comparison, also a truncated condition with persistent reverberation was added to the set of stimuli. The subjects had to rate externalization as well as the similarity to the reference.

The results in Fig. 28(a) show that the BRIR received the highest externalization score. The truncated condition (`trc 3`, cf.Tab. 2.2.2) received the second-highest score and was rated significantly less externalized than the BRIR. The other conditions lag far behind, being only poorly externalized or not at all. Condition 4, the impulse response with specular reflections, was rated only slightly higher than the HRIR which, in turn, was rated zero by all subjects but one. All conditions with diffuse reflections were rated significantly higher than the one condition with specular reflections. Yet, the ratings scarcely exceed the ‘close-to-the-head’ mark. Diffuse reflections with $DRR = 6\,\mathrm{dB}$ were rated significantly higher than diffuse reflections with $DRR = 11\,\mathrm{dB}$. No significant differences were found between diffuse reflections with $2\,\mathrm{dB}$ and $\{6,11\}\,\mathrm{dB}$ DRR.



(a) Externalization

(b) Similarity to the reference

Figure 28: Discrete diffuse and specular reflections. Median and 95 % confidence intervals. Conditions with diffuse reflections are denoted by *DiffR*, followed by the respective DRR. The condition with specular reflections is abbreviated *SpecR*. Black triangular markers denote pooled data.

Fig. 28(b) shows the perceived similarity to the reference. The reference was created by convolving anechoic speech with the HRIR. All participants recognized the hidden reference correctly. Both the BRIR and conditions `trc 3` received a very low rating. All conditions with added reflections were rated significantly higher than the BRIR with large effects of $r' \geq .77$. Diffuse reflections with $DRR = \{2, 6\}$ dB were rated significantly less similar to the HRIR than specular reflections with $DRR = 6$ dB. Diffuse reflections with $DRR = 11$ dB tend to be rated higher than the specular reflections. The difference is not significant, but the much smaller variation in the ratings and the large effect size of $r' = .78$ in favor of the diffuse reflections indicate that a difference may nonetheless be present, in spite of the insignificant result. Furthermore, diffuse reflections with $DRR = 11$ dB were rated significantly higher than diffuse reflections with $DRR = 6$ dB, both with regard to externalization and similarity to the reference.

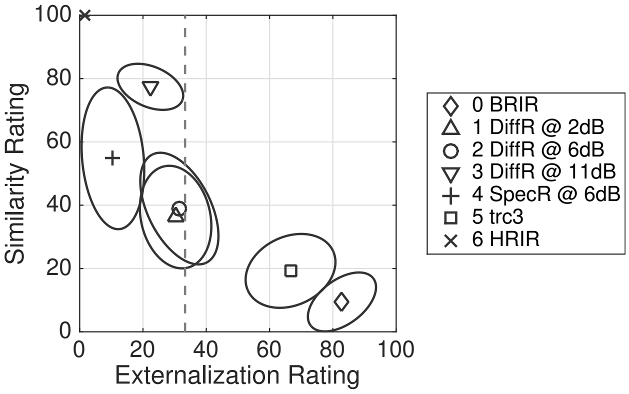


Figure 29: Discrete diffuse and specular reflections: Mean value along with the 95 % confidence ellipses. Similarity to the reference on the vertical axis over externalization on the horizontal axis.

Discussion. Fig. 29 shows the externalization and similarity ratings in a two-dimensional scatter plot. From this perspective, it becomes very even more clear that none of the compared methods achieve moderate externalization and similarity ratings at the same time. Specular reflections must be discarded as a means to enhance externalization in the studied configuration. Diffuse reflections led to significantly higher externalization ratings, yet in a range of the scale where it is debatable to speak of actual externalization. Based on the present data, it is impossible to state if the higher rating in this range corresponds to an actual increase in externalization (or maybe rather: tendency to be externalized), or if the ratings were influenced by attributes such as perceived source width and diffuseness, strength of reverberation, or sound color. However, considering the individual ratings, six subjects perceived condition 2 (diffuse reflections, $DRR = 6\,\mathrm{dB}$) outside of their head, while condition 4 (specular reflections, $DRR = 6\,\mathrm{dB}$) was only rated externalized by one to two subjects (one rating is on the threshold). This may indicate that, in a scenario where only a few reflections should be added, it makes more sense to use diffuse reflections. While the outcome does not justify the time-consuming numeric simulation of diffuse reflections, it is plausible that this degree of physical accuracy may not be necessary to achieve a similar effect.

The differences between diffuse reflections with different DRRs are more pronounced in similarity to the HRIR than in externalization. Eventually, it is unknown which attributes the subjects took into consideration when rating the similarity to the HRIR. But sound color likely played a major role since sound colorations, while most pronounced for specular reflections, were clearly audible for all stimuli with added reflections according to informal listening.

The very large variation in the similarity ratings of the specular reflections is striking. According to the impression of the author, the specular reflections lead to more pronounced sound colorations than the truncated condition which, in turn, has longer audible reverberation. This may have caused a disagreement among the subjects due to their individual choice of the rating criterion.

In the context of the influence of single reflections on externalization, the early reflections of the listening room may play a special role. The importance of accurately modeled early reflections for externalization does not seem fully clear yet. The findings of Yuan et al. (2015) suggest that moderate externalization can be reached by adding second-order (specular) early reflections. No similar effect could be achieved here with only two methods. However, neither here nor in the study of Yuan et al. (2015) it was investigated to what extent the agreement of the modeled early reflections with the listening room matters. Wendt et al. (2014) showed that BRIRs can be simulated with a reasonable degree of plausibility using strongly simplified low-order image source models of different rooms. It should be noted that the subjective evaluation by Wendt et al. (2014) was carried out inside a sound-attenuating room, whereas Yuan et al. (2015) give no information about the listening room or the mode of comparison. In contrast, in the present study, the evaluation took place in the original room and in comparison to the real loudspeaker. Furthermore, the reflections were not simulated to resemble the early reflections in the listening room. This may likely have led to differences in sound color and spatial impression, possible reasons for the poor performance of the diffuse and specular reflections.

Neither do the present data allow conclusions about a scenario with an increased number of diffuse or specular reflections nor about the importance of the accurate modeling of early reflections. But, in order to avoid the room divergence effect, it can be assumed that any differences between the synthesized and the real room should be minimized to maximize externalization. It may be worthwhile for future research to further investigate the effect of early reflections on externalization, e.g., by varying the level, delay, and order of early reflections in a more complex (measured or simulated) BRIR and in comparison to a real room.

3.3 Summary

This chapter treated a series of different experiments on the influence of reverberation on externalization. The experiments were all carried out in the same room and under the same conditions as the first experiment in Ch. 2. To generate conditions, impulse responses were prepared based on a measured BRIR and convolved with a speech sequence. The experimental design was based on the MUSHRA paradigm. Different experimental trials were carried out in random order. In each trial, a randomized and blinded set of stimuli was presented on a graphical interface. The stimuli were played back through the modified headphones and the degree of externalization was rated in the original room and in comparison to the loudspeaker. Some parts also included the evaluation of differences in sound quality by assessing the similarity to a reference signal. The HRIR convolved with speech was used therefor. In order to exclude any spatial attributes from the comparison, similarity was rated in diotic headphone listening inside an anechoic room.

Seventeen subjects participated in the experiment. The ratings of one particular subject were excluded from the analysis because the BRIR had not been perceived as externalized multiple times.

Part 1 consisted of two trials concerning the modification of a BRIR, similar to the experiment in Ch. 2. In each trial, two different modification methods were combined. The conditions were rated regarding externalization and regarding similarity to the anechoic reference signal. Sec. 3.1.1 explains the experimental design, and the conditions are listed in Tab. 6 and 7. In the first trial (trial 1a), the DRR was increased while decreasing the reverberation time. In the second trial (trial 1b), the DRR was increased while truncating the BRIR. First, the reverberation time or the BRIR length were manipulated. The DRR of the modified BRIRs was then increased by a relative difference of either 3 or 6 dB.

The relationship between similarity to the HRIR and externalization was roughly anti-proportional in both parts (see pp. 75ff.) In the tested configuration, no evidence was found that the combination of methods had a positive or negative effect on externalization or similarity to the reference, compared to the single-modification methods. Due to time constraints, the number of individual conditions that were compared was rather small. The results can therefore not be generalized to the modification techniques as such.

In part 2, the reverberation of the measured BRIR was replaced with artificial diffuse reverberation. The participants were asked to rate externalization in comparison to the speech sequence played back through the loudspeaker. Two different experiments were carried out. In the first part (trial 2a), the spatial distribution of binaural diffuse reverberation was varied. Either all directions on the sphere contributed to the reverberation or only two sphere segments located either at the front and back or on the left and right of the listener. In the second part (trial 2b), omni-directional diffuse binaural reverberation was compared with different types of dichotic and diotic reverberation. All conditions are listed in Tab. 8 and Tab. 9.

As explained in Sec. 3.1.2, the diffuse reverberation was generated based on sparse pseudo-random sequences (‘velvet noise’). To generate binaural reverberation, each noise pulse was convolved with the HRIR corresponding to a randomly drawn direction on a sphere. Different spatial distributions were created for trial 2a by restricting the set of directions. To generate dichotic ‘stereo’ reverberation in trial 2b, two independent realizations of noise were used for the left and right ear channels. Further diotic conditions were created by presenting only one channel of the reverberation of either the binaural or the ‘stereo’ condition to both ears. In either case, the noise sequence underwent a time-variant filtering operation in order to equalize each pulse based on the instantaneous spectral envelope of the measured BRIR at the corresponding time instance. The generated decaying noise signals were then appended to the binaural direct sound so that the DRR of the original BRIR was established.

The results of trial 2a concerning the spatial distribution of reverberation are discussed on pp. 78f. The synthesis with omni-directional binaural reverberation was perceived outside the head. However, it was rated less externalized than the BRIR. Reducing the distribution of the reverberation to two lateral segments, in turn, led to a further significant decrease of externalization. Reverberation from two segments in the front and back was not significantly different from any of the two. Nevertheless, the frontal reverberation tends to be more externalized than the lateral reverberation. This is surprising because of the higher interaural coherence of the front/back reverberation. A possible explanation may be the emergence of an unnatural ‘acoustic gap’ between the lateral reverberation and the frontal speech signal. At any rate, there is no evidence that narrowing down the contributing directions to the reverberation from the full sphere two merely two segments in the front and back had an effect on externalization.

In trial 2b, it was found that the concatenation of the HRIR with uncorrelated dichotic reverberation was still perceived as externalized, however, less externalized than the omni-directional binaural reverberation (see pp. 79f.). Diotic reverberation was perceived close to or inside the head but rated higher than the HRIR. Due to the instantaneous equalization to the original reverberation, the synthesized reverberation signals contain the smoothed direction-independent monaural cues of the dummy head. However, since the left and right channels of the dichotic condition were uncorrelated and contain no meaningful binaural cues of the room, it can be assumed that increased ILD fluctuations and lower interaural coherence are responsible for the main difference in externalization compared to the diotic reverberation.

Part 3 was about the influence of additional simulated reverberation. In contrast to part 2, the reverberation was not adjusted to resemble the measured BRIR, but the simulation of a different room with a longer reverberation time was used.

In trial 3a, stereo reverberation was added to a binaural speech signal. This signal was generated by convolving the anechoic speech sequence with the measured BRIR. The amount of additional reverberation was successively increased. Furthermore, the HRIR with stereo reverberation was added to the set of conditions which is listed in Tab. 10. The conditions were rated regarding the degree of externalization in comparison to the loudspeaker playing back the anechoic speech signal. The results on pp. 82f. show that externalization decreases with increasing additive reverberation. Whether a small amount of additional reverberation ($RER = -10\ldots0\,\mathrm{dB}$) has a positive or negative effect on externalization can, however, not be answered with certainty. But a continuous decrease makes sense insofar as increasing additive reverberation leads to a divergence between the synthesized and the listening room. It was furthermore found that the HRIR with additive reverberation, while unrelated to the listening room and not equalized to the BRIR (in contrast to trial 2b), still produced a sound image outside of the head.

In trial 3b, monophonic reverberation was added to the speech sequence before convolution. The reverberated speech signal was then convolved with the BRIR, of which the DRR was successively varied. Furthermore, the BRIR and the HRIR were convolved with anechoic speech and added to the set of conditions (see Tab. 11). The subjects had to rate externalization in comparison to the loudspeaker, now playing back the reverberant speech signal. This design had two different ambitions. On the one hand, to find out whether or not increasing the DRR leads to a smaller decrease in externalization compared to the experiment in Ch. 2 if the excitation signal is reverberant. On the other hand, if the BRIR with or without convolutive reverberation will be perceived as more externalized.

In addition, the influence of the DRR modification on sound quality was evaluated by assessing the similarity to the reference. The reference was the HRIR convolved with the reverberant speech signal.

The results are discussed on pp. 83ff. The externalization ratings show that, also in the presence of monophonic reverberation, increasing the DRR had a negative effect on externalization and a positive effect on the perceived similarity to the reference. The BRIR with anechoic speech received the highest ratings but is not significantly different from the BRIR with reverberant speech. Furthermore, the variation in the ratings of the BRIR with reverberation is much higher than without. Another observation is that the HRIR with reverberant speech at the input was rated higher than the HRIR convolved with anechoic speech (nonetheless, still inside or close to the head). All conditions with convolutive reverberation were rated more similar to the reference than in the other experiments where an anechoic speech signal was used. This supports the hypothesis that the original reverberation of the BRIR may be perceived as less disturbing in the case of a reverberant input signal.

Part 4 dealt with the effect of single reflections on externalization and sound quality. Using a model of two parallel lateral walls, one either diffuse or direct binaural reflection from each wall was simulated. Diffuse reflections are characterized by a wider spread in time and space compared to direct (specular) reflections. Impulse responses were created by appending the reflections to the HRIR at the time instants corresponding to the travel times of the reflections. The amplitude of the diffuse reflections was varied in three steps in order to yield $DRR = \{2, 6, 11\}$ dB. Furthermore, specular reflections were added at a DRR of 6 dB. For comparison, the BRIR, the HRIR, and a truncated condition with $L = 106$ ms were included in the set of conditions. Tab. 12 lists all conditions. The impulse responses were convolved with anechoic speech. The participants were asked to rate the conditions regarding externalization in comparison to the loudspeaker, and regarding similarity to anechoic speech convolved with the HRIR.

The results in Sec. 3.2.5 show good externalization for the BRIR and mediocre externalization for the truncated condition. The ratings are on a similar scale as in the other experiments. The conditions with diffuse reflections, on the other hand, only received ratings around the 'close-to-the-head' threshold. The specular reflections were rated significantly lower than all diffuse reflections, and not externalized. But they were also rated higher than the HRIR without added reflections. The diffuse reflections with a DRR of 2 dB and 11 dB as well as the direct reflections were rated more similar to the HRIR than the truncated condition. However, the large variation in the ratings of the specular reflections indicates different preferences among the participants. In summary, it can be said that the addition of single reflections in this scenario was not sufficient to increase externalization substantially, although there are differences in the lower-externalization range. It is conceivable that differences between the simulated reflections and the early reflections in the listening influenced the results.

Note that the limitations of the previous experiment listed in Sec. 2.5 also apply to this study. This includes in particular the remarks regarding HRIR individualization, the choice of audio material, headphone vs. loudspeaker presentation, the approach to assess sound quality, and the test paradigm.

4 Conclusion

In this thesis, the influence of reverberation on externalization and sound quality was investigated based on a literature review and two listening experiments. Different studies have shown that reverberation has a substantial positive effect on externalization. But externalization is a phenomenon that depends on context as much as content: Beyond the presented audio material, other factors play an important role, such as the acoustic and visual impression of the listening room, learning and prior experiences, and the ability and willingness to get involved in the experience. This entanglement of externalization with context makes it a demanding research object.

While the literature agrees on the importance of reverberation for externalization in the main, additional reverberation can be perceived as a disturbance. The use of head-related impulse responses (HRIRs), capturing the necessary cues for angular localization, is often the preferred minimal solution. However, the use of anechoic HRIRs leads only to poor externalization. To study the apparent tradeoff between externalization and sound quality in scenarios where additional reverberation is undesired, two experiments were carried out.

The first experiment in Ch. 2 investigated how this tradeoff manifests itself when modifying a binaural room impulse response (BRIR) in ways that aim for greater similarity to an anechoic stimulus. The reverberation was successively reduced by manipulating either the length of the BRIR, the direct-to-reverberant energy ratio (DRR), or the reverberation time. The modified BRIRs were then convolved with speech and compared between the methods. The participants assessed externalization as well as sound quality via the similarity to an anechoic reference signal. While this approach to sound quality is by no means universal, it is strongly associated with sound neutrality. To get a more nuanced picture, the perceived naturalness, as well as the similarity to the reference regarding sound color and the amount of reverberation were also evaluated.

Using either of the methods, the relationship between reverberation and externalization is positive and monotonic, in agreement with the literature. Increasing reverberation goes along with increasing perceived naturalness, but also with decreasing similarity to the reference. Considering similarly externalized conditions, the modification methods differ in naturalness and similarity to the reference. The differences are most pronounced regarding sound color. Compared to the other methods, truncation received the lowest ratings. This is likely a consequence of comb filters arising from the interference of the direct sound with the early reflections. In contrast, no differences between the methods were found regarding the perceived amount of reverberation of conditions with similar externalization.

In Ch. 3, multiple hypotheses were investigated. The studied scenarios include different spatial distributions of binaural diffuse reverberation, a comparison of binaural, dichotic, and diotic diffuse reverberation, the addition of stereophonic reverberation, the influence of monophonic reverberation in the input signal, and the effect of discrete diffuse and specular reflections.

It was found that externalization was improved by diffuse dichotic reverberation with parameters matching the listening room, but also with stereophonic reverberation with a longer reverberation time. Narrowing down the spatial distribution of the reverberation to two lateral segments had a negative effect on externalization, possibly due to an ‘acoustic gap’ between the frontal speech signal and the lateral reverberation. Monophonic reverberation in the input signal could not compensate for a negative effect of the DRR increment on externalization. However, the increased similarity to the reference indicates that the influence on sound quality diminishes. The addition of merely two reflections was, in contrast to persistent reverberation, found to be insufficient to achieve convincing externalization. However, diffuse reflections led to slightly increased externalization, compared to specular reflections.

This thesis gave an overview of the influence of reverberation on externalization. The conducted experiments, as diverse as they are, integrate into the current research dialog and help to improve the understanding of externalization. In a follow-up experiment, it would be interesting to study how the size and location of reverberant segments influence the externalization of frontal sound, and under what circumstances an ‘acoustic gap’ emerges. Moreover, the effect of reverberation in the input signal on externalization and sound quality should be quantified by varying different reverberation parameters. The importance of an agreement between synthesized early reflections and the real room is not fully clear yet. Further studies may investigate this question by varying the level and delay of early reflections, considering the resulting spatial impression and sound color.

References

Begault, D. R., Wenzel, E. M., and Anderson, M. R. (2001). Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *J. Audio Eng. Soc.*, 49(10):904–916.

Bernschütz, B. (2013). A spherical far field HRIR/HRTF compilation of the Neumann KU 100. In *Proc. 39th German Annu. Conf. Acoust. (DAGA)*, page 29.

Blauert, J. (1969). Sound localization in the median plane. *Acta Acustica united with Acustica*, 22.

Blauert, J. (1974). *Räumliches Hören*. Hirzel Stuttgart.

Blauert, J. and Lindemann, W. (1986). Spatial mapping of intracranial auditory events for various degrees of interaural coherence. *J. Acoust. Soc. Am.*, 79(3):806–813.

Brimijoin, W. O., Boyd, A. W., and Akeroyd, M. A. (2013). The contribution of head movement to the externalization and internalization of sounds. *PloS one*, 8(12).

Brinkmann, F., Lindau, A., and Weinzierl, S. (2017). On the authenticity of individual dynamic binaural synthesis. *J. Acoust. Soc. Am.*, 142(4):1784–1795.

Catic, J., Santurette, S., Buchholz, J. M., Gran, F., and Dau, T. (2013). The effect of interaural-level-difference fluctuations on the externalization of sound. *J. Acoust. Soc. Am.*, 134(2):1232–1241.

Catic, J., Santurette, S., and Dau, T. (2015). The role of reverberation-related binaural cues in the externalization of speech. *J. Acoust. Soc. Am.*, 138(2):1154–1167.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum, 2 edition.

Crawford-Emery, R. and Lee, H. (2014). The subjective effect of BRIR length on perceived headphone sound externalization and tonal coloration. In *Audio Eng. Soc. Conv. 136*.

Durlach, N. I., Rigopulos, A., Pang, X. D., et al. (1992). On the externalization of auditory images. *Presence: Teleoperators and Virtual Environments*, 1(2):251–257.

Farina, A. (2000). Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Audio Eng. Soc. Conv. 108*.

Fritz, C. O., Morris, P. E., and Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *J. Exp. Psychol.*, 141(1):2.

Gelfand, S. A. (2018). *Hearing: An introduction to psychological and physiological acoustics*, volume 6. CRC Press.

Giguère, C. and Abel, S. M. (1993). Sound localization: Effects of reverberation time, speaker array, stimulus frequency, and stimulus rise/decay. *J. Acoust. Soc. Am.*, 94(2):769–776.

Giller, P. M., Wendt, F., and Höldrich, R. (2019). The influence of different BRIR modification techniques on externalization and sound quality. In *EAA Spatial Audio Signal Process. Symp.*, pages 61–66, Paris, France.

Hartmann, W. M. and Wittenberg, A. (1996). On the externalization of sound images. *J. Acoust. Soc. Am.*, 99(6):3678–3688.

Hassager, H. G., Gran, F., and Dau, T. (2016). The role of spectral detail in the binaural transfer function on perceived externalization in a reverberant environment. *J. Acoust. Soc. Am.*, 139(5):2992–3000.

Hedges, L. V. and Olkin, I. (1985). *Statistical methods for meta-analysis*, chapter Estimation of a single effect size: Parametric and non-parametric methods. Academic press.

Hendrickx, E., Stitt, P., Messonnier, J.-C., et al. (2017). Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis. *J. Acoust. Soc. Am.*, 141(3):2011–2023.

Higginson, T. W. et al. (1865). *The Works of Epictetus: Consisting of His Discourses, in Four Books, the Enchiridion, and Fragments*, volume 5. Little, Brown & Co.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pages 65–70.

ITU Recommendation BS 1534-1 (2003). *Method for the subjective assessment of intermediate quality level of coding systems*, volume 14. International Telecommunications Union, Geneva, Switzerland.

Järveläinen, H. and Karjalainen, M. (2007). Reverberation modeling using velvet noise. In *30th Int. Conf.: Intelligent Audio Environments*.

Jesteadt, W. and Wier, C. C. (1977). Comparison of monaural and binaural discrimination of intensity and frequency. *J. Acoust. Soc. Am.*, 61(6):1599–1603.

Jiang, Z., Sang, J., Zheng, C., and Li, X. (2020). The effect of pinna filtering in binaural transfer functions on externalization in a reverberant environment. *Applied Acoustics*, 164:107257.

Kerby, D. S. (2014). The simple difference formula: An approach to teaching nonparametric correlation. *Comprehensive Psychology*, 3:11–IT.

King, B. M. and Minium, E. W. (2008). *Statistical reasoning in the behavioral sciences*. John Wiley & Sons Inc.

Klein, F., Werner, S., and Mayenfels, T. (2017). Influences of training on externalization of binaural synthesis in situations of room divergence. *J. Audio Eng. Soc.*, 65(3).

Kolarik, A. J., Moore, B. C., Zahorik, P., et al. (2016). Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. *Attention, Perception, & Psychophysics*, 78(2):373–395.

Li, S., Schlieper, R., and Peissig, J. (2018). 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. *J. Acoust. Soc. Am.*, 144(2):966–980.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *J. Am. Stat. Assoc.*, 62(318):399–402.

Mendonça, C. and Delikaris-Manias, S. (2018). Statistical tests with MUSHRA data. In *Audio Eng. Soc. Conv. 144*.

Mills, A. W. (1960). Lateralization of high-frequency tones. *J. Acoust. Soc. Am.*, 32(1):132–134.

Moore, D. S. and McCabe, G. P. (2009). *Introduction to the Practice of Statistics*, chapter Linear Regression. W. H. Freeman and Company, 6 edition.

Møller, H., Sørensen, M. F., Jensen, C. B., and Hammershøi, D. (1996). Binaural technique: Do we need individual recordings? *J. Audio Eng. Soc.*, 44(6):451–469.

Nielsen, S. H. (1992). Auditory distance perception in different rooms. In *Audio Eng. Soc. Conv. 92*.

Plenge, G. (1972). Über das Problem der Im-Kopf-Lokalisation. *Acustica*, 26:213–221.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, 25.

Sakamoto, N., Gotoh, T., and Kimura, Y. (1976). On-out-of-head localization-in headphone listening. *J. Audio Eng. Soc.*, 24(9):710–716.

Schirmer, W. (1966). On the explanation of errors in head-related stereophonic and monophonic reproduction. *Acustica*, 17:228–233.

Schroeder, M. R. (1965). New method of measuring reverberation time. *J. Acoust. Soc. Am.*, 37(3):409–412.

Sone, T. and Tadamoto, N. (1968). On the difference between localization and lateralization. In *Proc. 6th Int. Congr. Acoust., Tokyo*.

Tomczak, M. and Tomczak, E. (2014). The need to report effect size estimates revisited. an overview of some recommended measures of effect size. *Trends in Sport Sciences*, 1(21):19–25.

Udesen, J., Piechowiak, T., and Gran, F. (2014). Vision affects sound externalization. In *55th Int. Conf.: Spatial Audio*.

Välimäki, V., Holm-Rasmussen, B., Alary, B., and Lehtonen, H.-M. (2017). Late reverberation synthesis using filtered velvet noise. *Applied Sciences*, 7(5):483.

Wegler, K. (2020). *Über den Einfluss unterschiedlicher Reflexionseigenschaften auf den Präzedenzeffekt*. Master Thesis. University of Music and Performing Arts, Graz.

Welch, P. (1967). The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73.

Wendt, F. and Höldrich, R. (2018). Reflection properties influencing the precedence effect. *44. Jahrestagung für Akustik - DAGA*.

Wendt, F., Höldrich, R., and Marschall, M. (2019). How binaural room impulse responses influence the externalization of speech. *45. Jahrestagung für Akustik - DAGA*.

Wendt, T., van de Par, S., and Ewert, S. D. (2014). A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation. *J. Audio Eng. Soc.*, 62(11):748–766.

Wenzel, E. M., Wightman, F. L., and Kistler, D. J. (1991). Localization with non-individualized virtual acoustic display cues. In *Proc. SIGCHI Conf. Human Factors in Computing Systems*, pages 351–359.

Werner, S., Klein, F., Mayenfels, T., and Brandenburg, K. (2016a). A summary on acoustic room divergence and its effect on externalization of auditory events. In *8th Int. Conf. Quality of Multimedia Experience (QoMEX)*.

Werner, S., Klein, F., and Sporer, T. (2016b). Adjustment of the direct-to-reverberant-energy-ratio to reach externalization within a binaural synthesis system. In *2016 AES Int. Conf. Audio for Virtual and Augmented Reality*.

Wisniewski, M. G., Mercado III, E., Gramann, K., and Makeig, S. (2012). Familiarity with speech affects cortical processing of auditory distance cues and increases acuity. *PLoS One*, 7(7).

Yuan, Y., Fu, Z., Xu, M., Xie, L., and Cong, Q. (2015). Externalization improvement in a real-time binaural sound image rendering system. In *2015 Int. Conf. Orange Technologies (ICOT)*, pages 165–168. IEEE.

Zahorik, P., Brungart, D. S., and Bronkhorst, A. W. (2005). Auditory distance perception in humans: A summary of past and present research. *Acta Acustica united with Acustica*, 91(3):409–420.

Zielinski, S., Hardisty, P., Hummersone, C., and Rumsey, F. (2007). Potential biases in MUSHRA listening tests. In *Audio Eng. Soc. Conv. 123*.