On Externalization of Virtual Sound Images Presented via Headphones

Von der Fakultät für Elektrotechnik und Informatik
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des akademischen Grades
Doktor-Ingenieur
(abgekürzt: Dr.-Ing.)
genehmigte Dissertation

von

M.Sc. Song Li
geboren am 08. December 1989
in Zhejiang, China

2021

1. Referentin/Referent:	:	Prof. Dr. Jürgen Peissig
2. Referentin/Referent:	:	Prof. Dr. Hyunkook Lee (University of Huddersfield)
Tag der Promotion	:	21 July 2021

M.Sc. Song Li: *On Externalization of Virtual Sound Images Presented via Headphones,* Dissertation, © 2021

ABSTRACT

With the growing market of mobile devices and Virtual, Augmented, and Mixed Reality (VR/AR/MR) applications, headphone-based Three-Dimensional (3D) audio is becoming increasingly important. Head-related Transfer Functions (HRTFs), which represent the acoustic filtering of incoming sounds by the listener's morphology, are essential to create virtual sound images reproduced via headphones.

Various measurement systems have been proposed to fast record personal HRTFs from different directions, but most of them require expensive hardware setups. In addition, these systems are usually limited to estimate directional HRTFs with a fixed source-listener distance. In this thesis, an MR-based mobile system is proposed for fast estimating distance- and direction-dependent HRTFs with only one loudspeaker.

Perceived externalization, i.e., out of the head, is one of the most important features for building up immersive Virtual Acoustic Environments (VAEs). It is well known that reverberation and spectral information of direct sound components are two essential cues related to perceived externalization. This thesis further studies the relative impact of these two cues in contralateral versus ipsilateral ear signals on externalization of lateral sound images. Based on the outcomes of these studies, a series of experiments is designed to build a quantitative model to explain the interplay of important acoustic cues in externalization of lateral sound sources.

Due to the challenge of measuring individual HRTFs for every listener, non-individual HRTFs are commonly applied in binaural rendering systems in combination with simple room models. However, the synthesized sound sources from frontal and rear directions are difficult to be perceived as well externalized. This thesis proposes an advanced binaural rendering system to enhance externalization of frontal and rear sound images based on the localization- and externalization-related auditory cues.

Keywords: Perceived externalization, Head-related transfer function (HRTF), HRTF Measurement, Reverberation, Spectral information, Binaural rendering system.

ZUSAMMENFASSUNG

Mit dem wachsenden Markt mobiler Geräte und Virtual-, Augmented- und Mixed-Reality-Anwendungen (VR/AR/MR) gewinnt kopfhörerbasiertes dreidimensionales (3D) Audio zunehmend an Bedeutung. Kopfbezogene Übertragungsfunktionen (eng: HRTFs), die die akustische Filterung der einfallenden Schallsignale durch die Hörermorphologie darstellen, sind wichtig für die Erzeugung der über Kopfhörer wiedergegebenen virtuellen Klangbilder.

Es wurden verschiedene Messsysteme zur schnellen Aufnahme individueller HRTFs vorgeschlagen, aber die meisten von ihnen erfordern umfangreiche Hardware-Setups. Außerdem sind diese Systeme in der Regel auf die Messung richtungsabhängiger HRTFs mit einem festen Abstand zwischen der Quelle und dem Hörer beschränkt. In dieser Arbeit wird ein MR-basiertes mobiles System zur schnellen Messung von abstands- und richtungsabhängigen HRTFs mit nur einem Lautsprecher vorgeschlagen.

Die Externalisierung ist eine der wichtigsten Eigenschaften für den Aufbau immersiver virtueller akustischer Umgebungen (eng: VAEs). Es ist bekannt, dass die Reflexionen und die spektrale Information der Direktschallkomponenten zwei wesentliche Merkmale in Bezug auf die Externalisierung sind. In dieser Arbeit wird der relative Einfluss dieser beiden Merkmale in kontralateralen gegenüber ipsilateralen Ohrsignalen auf die Externalisierung von lateralen Klangbildern weiter untersucht. Basierend auf den Ergebnissen dieser Studien wird eine Reihe von Experimenten durchgeführt, um ein quantitatives Modell zur Erklärung des Zusammenwirkens der wichtigsten akustischen Merkmale bei der Externalisierung von seitlichen Schallquellen zu erstellen.

Aufgrund der Schwierigkeit zur Messung individueller HRTFs für jeden Hörer werden in binauralen Wiedergabesystemen häufig generische HRTFs in Kombination mit einfachen Raummodellen verwendet. Allerdings sind die synthetisierten Schallquellen aus frontalen und hinteren Richtungen schwierig als perfekt externalisiert wahrzunehmen. In dieser Arbeit wird ein neuartiges Wiedergabesystem vorgeschlagen, um die Externalisierung von frontalen und hinteren Klangbildern basierend auf den lokalisierungs- und externalisierungsbezogenen auditorischen Merkmalen zu verbessern.

Schlagwörter: Wahrgenommene Externalisierung, Kopfbezogene Übertragungsfunktion (HRTF), HRTF Messung, Reflexionen, Spektrale Information, Binaurales Wiedergabesystem.

CONTENTS

GLOSSARY OF ACRONYMS

1D	One-Dimensional
2D	Two-Dimensional
3D	Three-Dimensional
AR	Augmented Reality
BRIR	Binaural Room Impulse Response
BRTF	Binaural Room Transfer Function
CI	Confidence Interval
DRR	Direct-to-Reverberant Energy Ratio
EBU	European Broadcasting Union
ERB	Equivalent Rectangular Bandwidth
FDN	Feedback Delay Network
FFT	Fast Fourier Transform
FFV	Frequency-to-Frequency Variability
FS	Fourier Series
FT	Fourier Transform
HDTV	High-Definition Television
HMD	Head Mounted Display
HpTF	Headphone Transfer Function
HRIR	Head-related Impulse Response
HRTF	Head-related Transfer Function
IACC	Interaural Cross-Correlation Coefficient
IC	Interaural Coherence
IFT	Inverse Fourier Transform
IFFT	Inverse Fast Fourier Transform
ILD	Interaural Level Difference
ILD TSD	ILD Temporal Standard Deviation

IRS	Inverse Repeated Sequence
ITD	Interaural Time Difference
JND	Just Noticeable Difference
LTI	Linear Time Invariant
MESM	Multiple Exponential Sweep Method
MLS	Maximum Length Sequence
MR	Mixed Reality
NLMS	Normalized Least Mean Squares
NMSE	Normalized Mean Square Error
NRMSD	Normalized Root Mean Square Deviation
PWD	Plane Wave Decomposition
RDE	Room Divergence Effect
SD	Spectral Deviation
SG	Spectral Gradient
SH	Spherical Harmonic
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level
SQAM	Sound Quality Assessment Material
UDP	User Datagram Protocol
VAE	Virtual Acoustic Environment
VR	Virtual Reality

Part I

DISSERTATION

1

INTRODUCTION

Humans have the ability to localize sound sources in everyday situations with two ears (binaural hearing). Under normal listening conditions, sounds of the real world are perceived as externalized, i.e., out of the head. In contrast, the sounds presented via headphones are often perceived within the head ("internalized"), i.e., in-head-localization [1]. The main reason for the difference is that the headphone signals directly reach the listener's ears, while the real world sounds are acoustically filtered by the listener's morphology (pinnae, head, and torso) characterized by Head-related Transfer Functions (HRTFs). Head-related Impulse Responses (HRIRs) are the time domain representation of HRTFs.

To describe the direction of a sound source relative to the listener's head, a special coordinate system with the head at the origin (Head-related Coordinate System) is introduced according to Blauert [2].

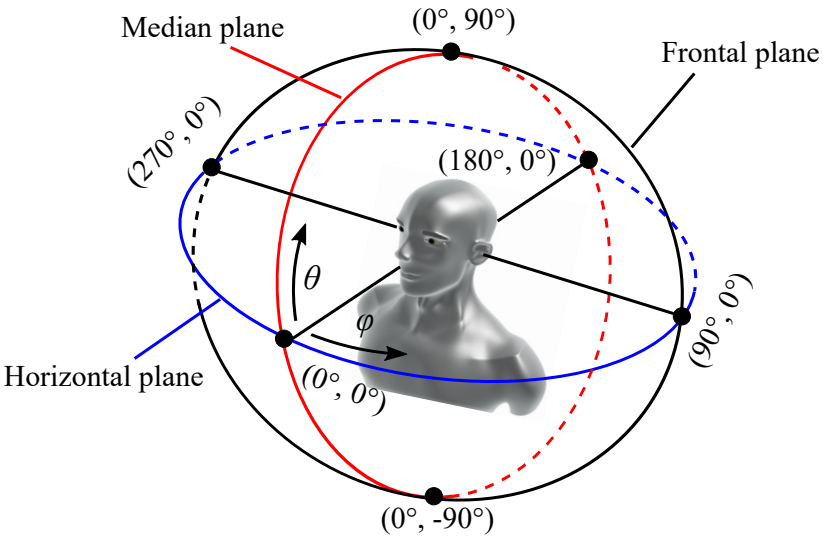


Figure 1.1: Head-related coordinate system according to Blauert [2].

As shown in Figure 1.1, the direction of sound incidence is described by two spherical angles, namely the azimuth angle (φ) and the elevation angle (θ). Further, three planes are defined in this coordinate system, i.e., the horizontal, the

median and the frontal plane. The horizontal plane is at the listener's ear level and divides the space into the upper and lower hemispheres; the frontal plane separates the space into the front and back hemispheres; the median plane is vertical to the horizontal and frontal planes, and divides the space into the left and right hemispheres. The azimuth angle φ describes the source direction in the horizontal plane, starting at 0° (frontal direction), and increasing counter-clockwise up to 360°, while the elevation angle θ indicates the elevated direction and is limited between -90° (bottom) and 90° (top). In order to better illustrate the experimental results in this thesis, we redefine φ on the right-hand side, which decreases clockwise from 0° (frontal direction) to -180° (rear direction). Hence, φ of -180° and 180° indicate the same direction.

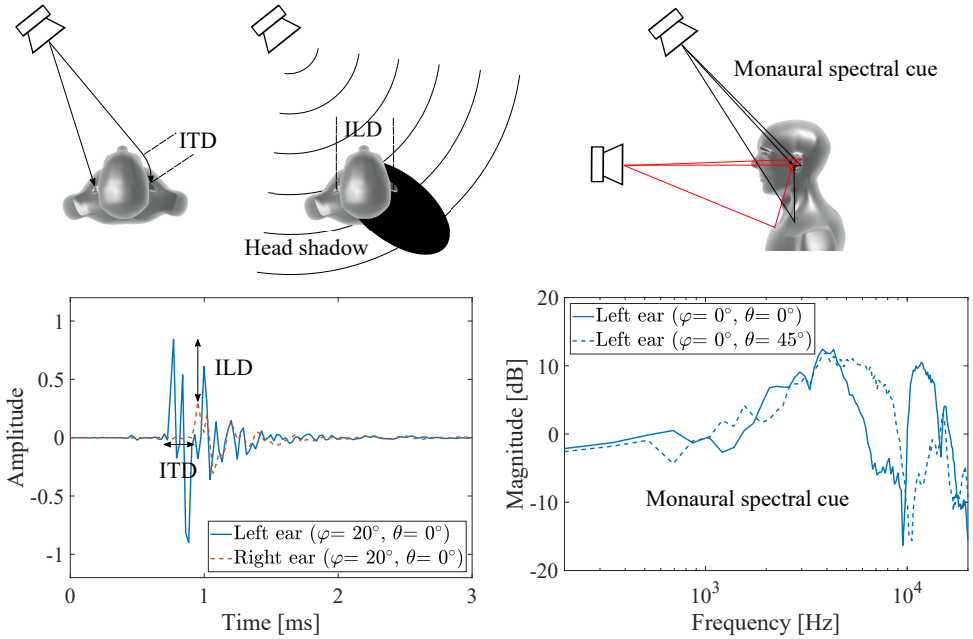


Figure 1.2: Binaural and monaural cues for sound source localization. The upper three figures illustrate the mechanisms of ITD (left), ILD (middle) and monaural spectral cue (right). The bottom left figure shows the binaural cues contained in a pair of HRIRs measured at an azimuth angle (φ) of 20° in the horizontal plane, and the bottom right figure illustrates the difference in the magnitude spectrum of the left ear HRTFs measured at θ of 0° and 45° (φ = 0°). The HRTFs used are taken from the CIPIC database (subject #3) [3].

All directional features of sound sources are encoded in HRTFs, including Interaural Time Differences (ITDs), Interaural Level Differences (ILDs) and the monaural spectral information (see the bottom panel of Figure 1.2). The lower left figure shows a pair of HRIRs measured at φ of 20° in the horizontal plane, and the binaural cues (ITD and ILD) can easily be identified by the difference in onset delays and amplitudes between the left and right HRIRs. The lower right figure shows the magnitude spectra of the left ear HRTFs for a frontal and an elevated sound source, and illustrates a noticeable spectral difference

in HRTFs (at high frequencies) between these two source directions. The upper panel of Figure 1.2 illustrates the mechanisms of these three cues. ITDs and ILDs are caused by the difference in the propagation time from the sound source to two ears, and by the head shadow effect, respectively. The monaural spectral cues describe the direction-dependent spectral energy distribution of sound sources resulting from multiple reflections and diffraction of the pinnae, head and torso. The perception of sound sources in the horizontal plane (lateral-angle perception) is cued by the interaural/binaural cues, where ITDs and ILDs are perceptually dominant at low (below 1.5 kHz) and high (above 1.5 kHz) frequencies, respectively [4]. The monaural spectral cue is relevant to the perception of elevated sound sources (polar-angle perception) [2].

Headphone-based Three-Dimensional (3D) audio technology is becoming increasingly important thanks to the ever-growing market of mobile devices, Virtual Reality (VR)/Augmented Reality (AR)/Mixed Reality (MR) applications, teleconferencing, and High-Definition Television (HDTV), etc. [5]. Filtering a dry audio signal by a pair of HRTFs prior to playback via headphones can create a virtual sound image in the free-field, and the simulated sound is perceived as coming from the direction related to the pair of HRTFs used [6]. However, in our everyday situations, the typical listening environments are rooms (echoic spaces) rather than the free-field (e.g., snowy field). When listening in rooms, not only the direct sound but also multiple reflections from different directions are acoustically filtered by the listener's morphology. The filtering can be considered as an interaction between HRTFs and room transfer functions, and is described by Binaural Room Transfer Functions (BRTFs). Binaural Room Impulse Responses (BRIRs) are the time domain representation of BRTFs. To generate headphone-based reverberant sound sources, BRIRs can be used in the same way as HRTFs in binaural synthesis.

In general, a BRIR can be divided temporally into two parts: a direct and a reverberant part. The direct part is the impulse response from the sound source to the listeners' ears without any room information (free-field condition) as described by HRIRs. The reverberant part contains the information of room acoustics, and can further be separated into an early reflection and a late reverberation part.

Early reflections, consisting of a series of discrete reflections from walls, floors, ceilings, etc., can be observed within a few milliseconds after the direct sound. Reflections arriving within about 1 ms after the direct sound influence the perceived source location ("summing localization") [2]. The perceived source position is dominated by the position of the direct (leading) sound source, if the delay to reflections is larger than 1 ms and below the echo threshold (typically between 5 and 50 ms), above which separate auditory events are

perceived. This effect is called the precedence effect [7–9]. Additionally, early reflections play an important role in source coloration, source width and speech intelligibility [10].

The late reverberation (about 80 ms after the direct sound), consisting of high-density reflections, contributes to the listener envelopment. The energy of the late reverberation is uniformly distributed in the room (diffuse), and individual reflections are not audible. It should be noted that the amount of reverberation is highly dependent on the acoustics of the room. In the case of an anechoic environment (free-field), the amount of reverberation is close to zero, and only the direct sound component remains.

The rest of this Chapter is organized as follows. Section 1.1 provides a brief overview of the acoustic cues contained in binaural signals that are relevant to externalization perception. After that, two important frameworks for modeling perceived externalization are introduced in Section 1.2. Section 1.3 shows a widely used binaural rendering system to create headphone-based virtual sounds. Lastly, the contribution and outline of the dissertation are described in Section 1.4.

1.1 ACOUSTIC CUES FOR PERCEIVED EXTERNALIZATION

Perceived externalization plays an essential role in the construction of immersive Virtual Acoustic Environments (VAEs), and it can be easily affected if there is a mismatch between spatial properties of the synthesized virtual sounds and those of the listener's natural acoustic exposure. Over the years, relevant acoustic cues related to perceived externalization have been investigated [11].

1.1.1 *HRTF-related acoustic cues*

Hartmann and Wittenberg [12] studied the relevance of binaural and monaural acoustic cues contained in HRTFs on externalization, and revealed that ITDs below about 1 kHz, ILDs at all audible frequencies, and the correct spectral information in ear signals were relevant to externalization perception. Some studies investigated the importance of HRTF spectra on externalization. Kulkarni and Colburn [13] expressed the magnitude spectra of HRTFs as Fourier Series (FS), and truncated the number of FS to smooth the magnitude spectra of HRTFs. Listeners were asked to evaluate the localization and externalization performance of virtual sound sources rendered with spectrally smoothed HRTFs. The results showed that the reduction of spectral information affected the elevation perception but not the degree of externalization. In contrast, Baumgartner et al. [14] flattened the magnitude spectra of HRTFs and observed that the sound images

gradually approached listeners' heads. Hassager et al. [15] spectrally smoothed the direct parts of BRIRs (corresponding to HRIRs) by using gammatone filters with different bandwidths, while keeping the reverberant parts unchanged. Listeners were asked to rate externalization of the sound sources generated with these modified BRIRs. The results illustrated that reducing the spectral details in direct parts of BRIRs affected externalization noticeably, in agreement with the observations reported in [14].

1.1.2 *Reverberation-related acoustic cues*

In reverberant environments, reverberation plays an important role in perceived externalization of virtual sound images [16]. Several studies have explored the relationship between the degree of externalization and the length of BRIRs, and have shown that reverberation between 20 ms and 80 ms did affect externalization, but extending reverberation to a longer duration did not further affect externalization [16–18].

Catic et al. [18] demonstrated that diotic reverberation was not sufficient to generate a well externalized sound image, especially when the direct sound provided small interaural differences (e.g., frontal sound sources). Similarly, Leclère et al. [19] concluded that reverberation did improve externalization, but only if it provided interaural differences. Hassager et al. [15] observed unchanged externalization results from virtual sound sources generated with modified BRIRs that had spectrally smoothed reverberation. These studies indicated that the binaural information from reverberation is important for externalization [18, 19], while the spectral information in reverberation is less important for externalization compared to that in direct sound components [15, 20].

In general, two types of psychophysical tasks can be used to quantify sound externalization: binary judgment ("inside" vs. "outside" the head) [21] and continuous scale (from "at the center of the head" to "at the position of the reference sound source") [12]. The binary judgment can be effectively used to detect whether the sound is externalized or not. A continuous rating scale is used to evaluate sound externalization in depth (degree of externalization), since sounds may be perceived inside the head or close to the skull. The latter is commonly applied in externalization studies assuming that externalization is a matter of degree and is thought to mediate the distance perception [1, 12]. Externalization and perceived distance are closely related, and the distance-related acoustic cues may have the potential to indicate perceived externalization of sounds. The main difference between them is that externalization shows a strong dependence on binaural cues, whereas distance perception is dominated by monaural cues [11].

Direct-to-Reverberant Energy Ratio (DRR) has long been regarded as an important cue for distance perception of sound sources [22]. Shinn-Cunningham et al. [23] observed that the Frequency-to-Frequency Variability (FFV), which describes the frequency variability in the magnitude spectra of BRIRs, increasing with source-listener distance. Hence, DRR and FFV can be considered as potential indicators of the degree of externalization. Catic et al. [24] analyzed the distributions of short-term ILDs collected from reverberant binaural speech signals, and pointed out that the ILD temporal fluctuations, represented by standard deviation of ILDs over time, contributed to externalization of sound sources that contain high-frequency components (above 1 kHz). In their second study, Catic et al. [18] found that reverberation-related short-term binaural cues, i.e., Interaural Coherence (IC), ILD and IC temporal fluctuations, were responsible for externalization perception, while the externalization results were not well reflected in DRR values. Similar results can be found in [19, 20], suggesting that the reverberation-related binaural cues can be used as indicators for predicting externalization.

1.1.3 *Head and source movements*

In natural listening situations, listeners are free to rotate their heads, resulting in sound sources that move relative to the subjects' heads. To realize the behavior of head movements in binaural synthesis, head-tracking devices (placed on the listeners' heads) are commonly used to detect the head movements so that the virtual listening environments can be rotated accordingly to fix the absolute positions of the virtual sound sources.

Compared to static scenarios (binaural reproduction without head movement), dynamic cues introduced by head movements can effectively improve the localization performance of virtual sound images [25, 26]. Several recent studies revealed that large head movements can improve externalization of virtual sound images especially for frontal and rear sound sources, while small head movements have little effect on externalization [21, 27–29]. The improvement in externalization can persist even when the head movement stops. In contrast, moving the head without head tracking, i.e., the relative direction between the virtual sound source and the listener remains unchanged during head movements, deteriorates the degree of perceived externalization [21, 27, 28]. Li et al. [29] showed that large head movements in the horizontal and median planes have almost the same influence on externalization of a frontal sound source, indicating that the influence of head movements on externalization does not depend on the movement pattern. Additionally, Li et al. [30] found that the rel-

evance of head movements on externalization is reduced when reverberation is present.

Large source movements can also improve externalization but to a smaller extent than the improvement caused by head movements [28]. Moreover, large source movements in the horizontal and median planes have nearly the same impact on externalization, while source movements in the front/back direction have no effect on externalization [29].

1.1.4 *Visual information*

The presentation of congruent visual information that supports the existence of well externalized sound images improves perceived externalization of virtual sounds presented via headphones.

Werner et al. [31] divided listeners into two groups with and without the presence of visual cues. Test stimuli rendered with BRIRs were played back through headphones, with one group of listeners tested in darkness, while the listeners in the second group could see the room and the loudspeakers. The listening test was performed in the same room where the BRIRs were measured. In the experimental results, higher externalization ratings were observed for listeners who were presented with visual cues. Udesen et al. [32] reported reductions in externalization of virtual sounds presented with incongruent vision of the listening room, i.e., the test stimuli were recorded in one room while the subjects listened to them in a different room. A similar result can be found in [31], which is referred to as the Room Divergence Effect (RDE).

Gil-Carvajal et al. [33] further studied the influence of incongruent auditory and visual information in perceived externalization. Listeners were divided into two groups to rate externalization of test stimuli presented over headphones, where one group was provided only with the auditory information, and the second group was given only the visual information of the playback room. The listeners in the first group could not see the listening environment (blindfolded) but could hear sounds from a speaker positioned in the listening room ("auditory" group), while the listeners in the second group could see the listening environment but no sound was reproduced except for test stimuli presented via headphones ("visual" group). The experimental results showed that the externalization ratings were mainly influenced by the incongruent auditory information, but not by the visual impression of the playback room, suggesting that the congruent auditory cues were more relevant to externalization perception than the room-related visual information.

All these studies indicate that virtual sounds can be perceived as well externalized when the played back signals match the listeners' expectations.

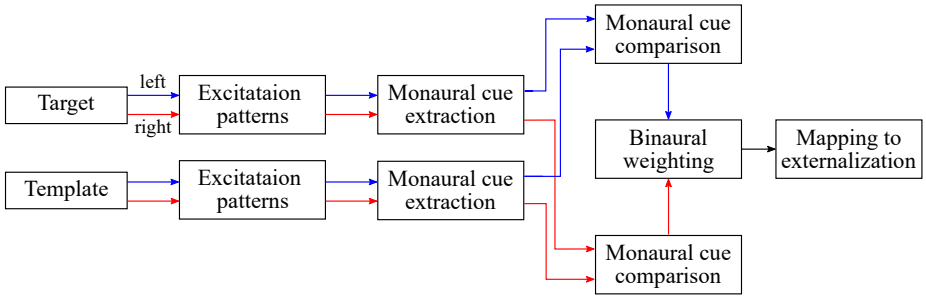
1.2 EXTERNALIZATION MODEL

Externalization is commonly quantified through psychoacoustic listening experiments, i.e., subjects listen to test signals and rate them based on a given rating scale.

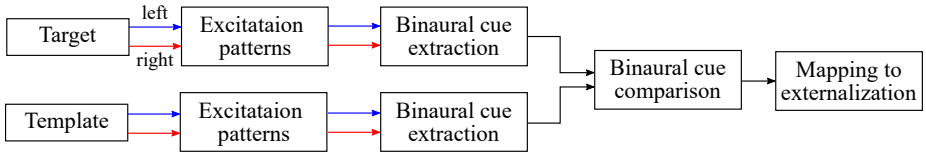
Plenge [34] introduced a conceptual model to explain the localization (direction and distance) perception by humans, consisting of a long-term and a short-term memory. The long-term memory stores the auditory localization cues (ILDs, ITDs and monaural spectral cues) presented in HRTFs, while the short-term memory is mainly used to store the properties of sound sources and the room-related information (e.g. reverberation and visual information). The HRIRs are initially learned in childhood and then continuously re-learned until the individual anatomies (e.g., head size, pinnae structure) reach their final sizes [34]. Hence, an adaptation to altered head-related localization cues takes a long time (several days) [35]. In contrast, the adaptation of information in short-term memory is relatively fast and is needed every time the listening environment changes.

Perceived Externalization of virtual sound images tends to be disrupted, when the "information" provided by binaural signals (target) differs from that stored in both memories (reference templates). Based on Plenge's conceptual framework, two externalization models have been proposed by calculating the deviations in monaural (Figure 1.3a) or binaural ((Figure 1.3b) cues between target and template signals [15, 36]. All acoustic cues are extracted after the auditory peripheral filtering of ear signals ("externalization patterns"). Note that the deviation of monaural cues between the target and template signals is calculated for each ear, and the binaurally weighted deviation is applied to map to externalization results.

Hassager et al. [15] observed that spectral smoothing of the direct sound component of BRIRs resulted in ILD deviations from the reference signal generated with individually measured BRIRs (see Section 1.1.2). They have therefore developed a model that calculates the deviations of frequency-dependent ILDs between target (generated with spectrally smoothed BRIRs) and reference signals to predict externalization ratings, and the results were consistent with the subjective data. Baumgartner and Majdak [36] compared the performance of different metrics, i.e., ILDs, Spectral Gradients (SGs), ILD temporal fluctuations, IC, inconsistencies between ITDs and ILDs, and the Sound Pressure Level (SPL), in terms of externalization prediction of anechoic sounds. They applied these metrics to four previous studies ([12, 14, 15, 37]) and compared the predicted results with externalization ratings subjectively obtained in those studies. The validation results suggested that the monaural spectral cues represented by SGs



(a) Externalization model based on monaural acoustic cues.



(b) Externalization model based on binaural acoustic cues.

Figure 1.3: Simple structures of externalization models by comparing monaural (Figure 1.3a) and binaural (Figure 1.3b) cues between target and template signals [36]. The blue and right lines represent left and right ear signals, respectively.

were important for predicting externalization of anechoic sounds. In addition, the joint evaluation of deviations in binaural and monaural cues provided the most reliable results.

1.3 BINAURAL RENDERING

BRIRs are highly individual and depend on source and listener positions (directions and distances) in a room. For dynamic binaural rendering applications with moving sound sources and a moving listener, a large set of BRIRs with different source and listener positions in the room is required. Further, storing a complete set of BRIRs needs a lot of memory, and the convolution with long BRIRs in real-time is computationally complex. Instead of convolving with BRIRs, a widely used method for binaural rendering is to synthesize the direct sound and the reverberant part separately.

Figure 1.4 shows the block diagram of a typical binaural rendering system. The direct sound is generated by filtering the input audio signal with a pair of HRTFs ("$\mathrm{HRTF_L}(\varphi_0, \theta_0)$" and "$\mathrm{HRTF_R}(\varphi_0, \theta_0)$") according to the direction of the sound source relative to the listener's head. The reflections can be considered as delayed, attenuated and low-pass filtered input signals because of the propagation paths, wall and air absorption, etc. Since early reflections are sparse and the direction of each reflection is discernible, the reflected sounds ("early reflection buffers") are filtered by corresponding HRTFs. The late reverberation is diffuse and does not contain distinct directional information. The

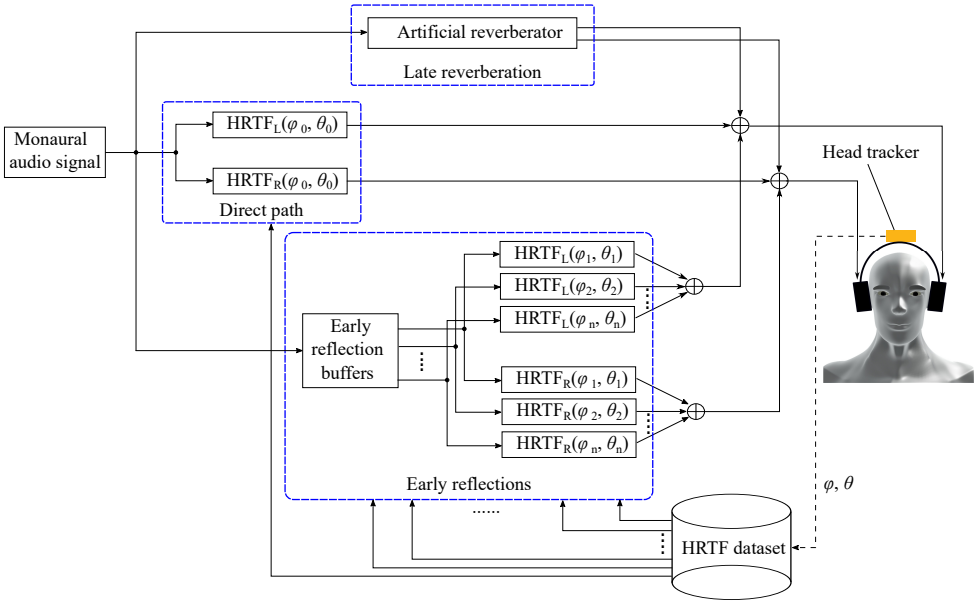


Figure 1.4: Structure of a typical binaural rendering system.

image-source method [38] and the Feedback Delay Network (FDN) method [39] are commonly used to simulate early reflections and the late reverberation, respectively. For dynamic scenarios, a head tracking device is used to record the listener's head orientation, allowing the VAE to be rotated accordingly.

Since HRTFs do not contain room information and only the relative position between the sound source and the listener should be considered for the measurement, measuring HRTFs is relatively easier than measuring BRIRs. Even so, recording high-resolution individual HRTFs is time-consuming and can hardly be performed for every listener. In practice, non-individual HRTFs taken from available databases are often used for binaural synthesis.

1.4 THESIS CONTRIBUTION AND OUTLINE

The main contributions of this thesis related to perceived externalization of headphone-based virtual sounds are listed below and are described in the rest of this thesis.

Chapter 2: HRTF measurement
HRTFs are essential for creating immersive VAEs reproduced over headphones. Different systems/approaches have been proposed to fast measure direction-dependent individual HRTFs. Most systems require comprehensive hardware setups, e.g., loudspeaker arrays, and a large space for the HRTF measurement, and subjects should be kept still during the measurement. He et al. [40] proposed an approach to rapidly record individual HRTFs using only one loud-

speaker. With this system, subjects are asked to actively perform head movements to cover different measurement positions.

Almost all systems are designed to measure Two-Dimensional (2D) HRTFs from different directions with a fixed distance between the loudspeaker and the subject [41]. In this thesis, an overview of the state of the art in different HRTF measurement systems/approaches is provided. Furthermore, an MR-based mobile measurement system is proposed for rapid acquisition of distance- and direction-dependent (3D) individual HRTFs with a single loudspeaker. With this system, subjects are asked to rotate their heads and move their bodies towards or away from the loudspeaker to cover measurement positions visualized by the MR device. In addition, the estimation errors of the HRTFs are calculated and provided to the subjects in real-time. The proposed system shows the potential to quickly measure individual 3D HRTFs with a mobile setup.

Chapter 3 and Chapter 4: **Acoustic cues on perceived externalization**
It is well known that reverberation is relevant for perceived externalization. In the case of a frontal sound source, the reverberation is equally important for both ears for externalization. In this thesis, the relative influence of reverberation at the ipsilateral versus the contralateral ear on externalization of a lateral sound source is investigated. Different acoustic cues are extracted from binaural signals to explain the influence of lateralized reverberation on externalization. Additionally, this influence is investigated for different source directions (see Chapter 3).

Afterwards, the role of ILDs and monaural spectral information on externalization of a lateral sound source is further studied. Moreover, the relevance of spectral details of the HRTFs on externalization in the presence of reverberation is explored. Based on the findings of these studies, a novel externalization model is proposed that incorporates three important acoustic cues, namely ILDs, SGs, and ILD temporal fluctuations, to predict externalization ratings of anechoic and reverberant lateral sound sources. This developed model can be used to generate hypotheses for externalization experiments in the future (see Chapter 4).

Chapter 5: **Externalization enhancement of virtual frontal and rear sound sources**
Non-individual HRTFs are widely applied in binaural rendering systems in combination with a simple room model to create virtual sounds reproduced via headphones (see Section 1.3). However, many studies have shown that the externalization ratings of frontal and rear sound sources are relatively low compared to that of lateral sound sources [19, 21]. The reason could be that the interaural differences contained in sound sources close to the median plane are smaller than those present in lateral sound sources. In this thesis, an advanced binaural rendering system is proposed to improve externalization of frontal and rear sound sources by using the localization- and externalization-related auditory information.

Chapter 6 concludes the dissertation and suggests a perspective for future work.

2

MEASUREMENT OF INDIVIDUAL HRTFS

2.1 INTRODUCTION

HRTFs are highly individual and direction-dependent. Moreover, in the near-field (proximal region), HRTFs vary as a function of source-listener distance [42]. A high-resolution HRTF dataset is required for each listener to experience immersive VAEs, but its measurement is time-consuming when using traditional methods. Although interpolation and extrapolation approaches [43–47] can dramatically reduce the measurement points, the required measurement number is still high [48, 49]. Over the years, different measurement systems/methods have been proposed for rapid acquisition of individual HRTFs, which are mostly based on multi-loudspeaker setups. In this Chapter, a mobile measurement system based on an MR device is proposed to quickly measure individual 3D HRTFs with only one speaker. Parts of this Chapter have been published in [41, 50, 51].

This Chapter is organized as follows: Section 2.2 introduces the fundamentals of HRTF measurements. After that, a survey of HRTF measurement systems/methods is provided in Section 2.3. Then, a mobile measurement system for fast capturing individual 3D HRTFs is presented in Section 2.4. Concluding remarks are drawn in Section 2.5.

2.2 FUNDAMENTALS OF HRTF MEASUREMENTS

2.2.1 *HRTF as an LTI system*

An HRTF can be approximated as a Linear Time Invariant (LTI) system between a point sound source in the free-field and a defined position in the listener's ear canal. Theoretically, all methods for identifying the transfer functions of LTI systems can be used to measure HRTFs.

Figure 2.1 shows the principle of signal processing through an LTI system in the time and frequency domain. In the time domain, the output signal $y(t)$

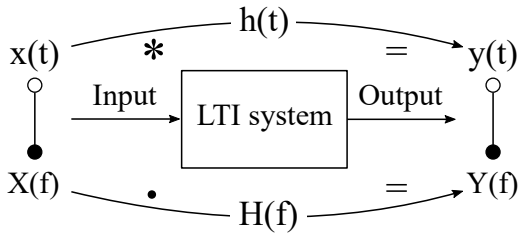


Figure 2.1: Basic principle of signal processing through an LTI system (adapted from Figure 7.7 in [52]).

is calculated by convolving the input signal $x(t)$ with the system impulse response $h(t)$:

$$y(t) = x(t) * h(t) = \int_{-\infty}^{\infty} x(\tau)\, h(t-\tau)\, d\tau, \tag{2.1}$$

where $*$ denotes the convolution operator. In the frequency domain, the output signal $Y(f)$ can be calculated by multiplying the input signal $X(f)$ and the system transfer function $H(f)$:

$$Y(f) = X(f) \cdot H(f). \tag{2.2}$$

The transformation of signals between the time and frequency domain is realized with the Fourier Transform (FT) and its inverse (IFT). In order to determine the impulse response/transfer function of the LTI system, the input and output signals are prerequisites. Since the multiplication is easier than the convolution operation, the calculation of the system impulse response, i.e., deconvolution process, is typically performed in the frequency domain:

$$h(t) = \mathrm{IFT}\{H(f)\} = \mathrm{IFT}\left\{\frac{\mathrm{FT}\{y(t)\}}{\mathrm{FT}\{x(t)\}}\right\}. \tag{2.3}$$

In the digital domain, the signal transformation between the time and frequency domain is usually performed by the Fast Fourier Transform (FFT) and its inverse (IFFT) for efficiency. Note that the signal length should be a power of two to apply FFT and IFFT. In practice, the input and output signals are padded with zeros to double the signal length (or FFT length) before transforming to the frequency domain to avoid aliasing errors (circular shift of the calculated impulse response in the time domain). Further, to avoid division by small values in the frequency domain, the denominator is appropriately regularized [53]. In addition to this typical deconvolution method (Equation 2.3), some other approaches, such as time-reversed filter method, circular cross-correlation method, have been proposed based on the properties of the excitation signals to be used to accelerate the deconvolution process [54, 55].

2.2.2 *Basic HRTF measurement*

HRTF measurements are usually performed in an anechoic chamber to avoid noticeable reflections. A subject is equipped with a pair of miniature microphones, and a loudspeaker (typical sound source) is placed at a defined distance and direction (measurement position) to the subject. An excitation signal is played back from the loudspeaker and is recorded by the miniature microphones. Then, a pair of HRTFs for this measurement position is calculated based on the emitted and recorded signals. This process is then repeated until all measurement positions are covered. The measured HRTFs need to be further post-processed to remove reflections and compensate for the influences of the loudspeaker and microphones.

The position of miniature microphones has an influence on the measurement results, since the sound pressure changes along the subject's ear canals [56–58]. Based on the model of the ear canal with a transmission-line [59], Hammershøi and Møller [60] demonstrated that almost all localization information of sound sources can be well captured with the microphones placed at blocked entrances of the ear canals. An extensive subjective experiment was performed in [61] to compare HRTFs with blocked and open ear conditions for different source directions, and the results confirmed the concepts presented in [59, 60]. The "blocked ear technique" is therefore widely used to measure HRTFs of human subjects because it is more convenient than placing microphones inside the ear canals.

The excitation signals used to obtain impulse responses have a wide range. Linear/exponential sweeps [62, 63], white noises [64], Maximum Length Sequence (MLS) [65] and Inverse Repeated Sequence (IRS) [66] are commonly used for measuring HRTFs. The energy of the excitation signal should be set high enough compared to the ambient noise to provide a sufficient Signal-to-Noise Ratio (SNR) of the measurement result (usually above 60 dB). On the other hand, the signal energy must be limited due to the dynamic range of the measurement equipment [54]. More details on the fundamentals of HRTF measurements can be found in [41, 64, 67].

2.3 STATE OF THE ART IN INDIVIDUAL HRTF MEASUREMENTS

2.3.1 *Multi-loudspeaker setups*

Increasing the number of sound sources (loudspeakers) is a straightforward way to fast measure HRTFs from different directions. In general, two design options of the multi-loudspeaker setup can be found in literature. The first

option is that multiple loudspeakers are placed in a spatial or spherical layout (see [68, 69]). Since the speaker positions may already cover the complete measurement directions, there is no need to reposition the subject or the loudspeaker system. The second case is the combination of a loudspeaker array and a single-axis positioning system (see [70–73]). The speakers are placed on a vertical, horizontal or circular arc, and the positioning system is used to rotate the loudspeaker array or the subject to cover the measurement points. Considering the measurement space and infrastructure costs, the latter option is widely applied in acoustic laboratories, and different fast HRTF measurement approaches have been developed based on this setup.

Majdak et al. [74] introduced a Multiple Exponential Sweep Method (MESM) to reproduce interleaving and overlapping exponential sweeps from multiple loudspeakers. This approach takes advantage of using exponential sweeps as excitation signals, where the harmonic distortions and the linear impulse response of the system under test are separated after deconvolution. The MESM consists of two mechanisms, i.e., interleaving and overlapping. For the interleaving mechanism, a group of exponential sweeps is played back with a short delay, so that the group of linear impulse responses of the measured systems lies between the linear impulse response and the 2^{nd} order harmonic distortion of the system response. For the overlapping strategy, a group of exponential sweeps for the following system (grouped loudspeakers) can be reproduced in overlapping form with the previous sweep signals if the highest order harmonic distortion of the system response does not interfere with the linear impulse response of the previous system. Dietrich [75] optimized the MESM by applying a generalized overlapping strategy instead of overlapping and interleaving mechanisms to further reduce the measurement time.

The MESM and its optimized form can reduce the time to measure HRTFs from different fixed source (loudspeaker) directions. However, for the measurement of other source directions, the subject or the loudspeaker system must be repositioned, which takes a lot of time, i.e., stop & go mechanism. Richter and Fels [70] extended the MESM to measure HRTFs continuously with a rotating subject (continuous measurement mechanism). The measurement setup consists of a vertical loudspeaker array and a turntable, where the turntable can be rotated in the horizontal plane. The exponential sweeps are played back in an overlapped form according to Dietrich [75], and after the last loudspeaker starts to emit sweep signals, the first loudspeaker is re-started with an overlap. The subject is positioned on the turntable and is rotated continuously in the horizontal plane. Compared to the traditional MESM applied to the stop & go mechanism, this approach can save the time required to reposition the subject or loudspeaker system. However, this method may introduce

noticeable frequency-dependent offsets at the measurement positions, since the sweep signal changes its instantaneous frequency with time and the azimuth angles between the loudspeakers and the listener also vary with time. Richter and Fels [70] adjusted the offsets by interpolating HRTFs in the Spherical Harmonic (SH) domain. The rotational speed of the turntable has a large influence on measurement results. Subjective experiments have been performed to evaluate the quality of HRTFs measured with different rotational speeds. The results revealed that the measurement errors were not audible when the rotational speed was less than 3.8 °/s.

Instead of using classic deconvolution methods, adaptive filtering approaches can alternatively be applied for recursive estimation of HRTFs [76]. Enzner [77] first introduced a measurement system for continuously capturing One-Dimensional (1D) HRTFs in the horizontal plane with a single fixed loudspeaker and a rotating subject. Various adaptive algorithms can be applied to identify HRTFs, and among them, the Normalized Least Mean Squares (NLMS) method is often used due to its low computational cost and high performance [76]. Different broadband signals, such as white noises and pseudo random sequences can be used as excitation signals for adaptive filtering approaches. Antweiler et al. [78] derived perfect sweeps from perfect sequences to enable a fast convergence rate of the adaptive system with a robustness against non-linearities of the measurement system, where the perfect sequences are periodic and pseudo-noise sequences with an impulse-like periodic autocorrelation function [79, 80]. The perfect sweep is actually a repeated linear sweep signal and is generated with a constant magnitude spectrum and a linear group delay in the frequency domain [78]. The results showed the use of perfect sweeps as excitation signals outperformed white noise in terms of SNRs of measured HRTFs. Enzner [81] further extended the NLMS method to fast measure 2D HRTFs with multi-loudspeaker setups and a rotating subject. With this approach, multiple loudspeakers can simultaneously reproduce orthogonal excitation signals (independent of each other) to uniquely identify HRTFs from different loudspeaker directions.

Most multi-loudspeaker setups are designed to measure HRTFs with a fixed source-listener distance (2D HRTFs), where multiple loudspeakers are mounted on an arc. Since such setups can not be flexibly changed to measure HRTFs with different distances, they are commonly applied for measuring HRTFs in the far-field (the source-listener distance is typically larger than 1 m), where the HRTFs are distance-independent. Yu and Xie [82] developed a multi-loudspeaker setup to measure 3D HRTFs (distance- and direction-dependent HRTFs) in which the loudspeakers are mounted on an arc with length-adjustable support rods. The HRTF measurement for different source directions and distances is achieved by

rotating the subject with a turntable and by adjusting the length of the support rods, respectively. With this setup, not only the far-field HRTFs but also the near-field (the source-listener distance is typically less than 1 m) HRTFs can rapidly be measured.

2.3.2 *Single-loudspeaker setups*

The HRTF measurement systems mentioned above require comprehensive hardware setups and a large space for the placement of these setups. Further, in the case of individual HRTF measurement, subjects should be kept still during the whole measurement. For the measurement of high-density HRTF datasets with only one loudspeaker, the conventional stop & go mechanism is not suitable due to the long measurement time. Several approaches have been developed based on the continuous mechanism to accelerate the measurement process.

Ranjan et al. [83] proposed a continuous 2D HRTF measurement approach using a head tracking device. A loudspeaker is placed in front of a subject and reproduces an excitation signal. The subject is equipped with a pair of miniature microphones and is asked to freely perform head movements to reach different measurement directions. The head tracking device placed on the subject's head is used to record the head's orientation, which is further synchronized with the excitation signal and the recorded ear signals. The pair of HRIRs for each source direction is then adaptively calculated using the NLMS method.

Under the assumption that the time-varying HRTF/HRIR is a time-varying linear system, the recorded binaural signals $y(k)$ (neglecting the subscripts designating the left and right ear) at discrete time k can be formulated as [77]:

$$y(k) = \sum_{n=0}^{N-1} x(k-n)\, h(n, \varphi_k, \theta_k) + v(k)\,, \tag{2.4}$$

where $h(n, \varphi_k, \theta_k)$ represents one sample of the HRIR at an azimuth angle of φ_k and an elevation angle of θ_k . $v(k)$ describes the measurement noise, and N denotes the HRIR length. Note that this model is valid when the time of HRIR changes is larger than the HRIR length [77]. The orientation data (φ_k, θ_k) is recorded by the head tracking device during the measurement. Equation 2.4 can be rewritten with the vector representation of the HRIR and the input signal [77]:

$$y(k) = \mathbf{h}^T(\varphi_k, \theta_k)\, \mathbf{x}(k) + v(k)\,, \tag{2.5}$$

where $\mathbf{h}(\varphi_k, \theta_k)$ and $\mathbf{x}(k)$ denote the HRIR and the input signal in vector form, respectively. $\mathbf{x}(k)$ consists of the most recent N samples of the excitation signal

at time k. By applying the NLMS algorithms, the HRIR at the next time point $\left(\widehat{\mathbf{h}}(\varphi_{k+1}, \theta_{k+1})\right)$ is expressed as [83]:

$$\begin{aligned} e(k) &= y(k) - \widehat{\mathbf{h}}^{T}(\varphi_k, \theta_k)\, \mathbf{x}(k), \\ \widehat{\mathbf{h}}(\varphi_{k+1}, \theta_{k+1}) &= \widehat{\mathbf{h}}(\varphi_k, \theta_k) + \mu\, \frac{\mathbf{x}(k)}{||\mathbf{x}(k)||^2 + \epsilon}\, e(k), \end{aligned} \tag{2.6}$$

where $e(k)$ denotes the residual error between the recorded and predicted ear signals, and ϵ is a regularization factor to ensure that the denominator in the equation is non-zero. μ is the step size of the NLMS algorithms which is limited between 0 and 2 [76]. It should be chosen as a compromise between the convergence behavior of the algorithms and the noise rejection performance. Instead of choosing a fixed value for μ, some studies have proposed to recursively adjust μ based on estimation errors [40, 84, 85].

It is possible that several measurement points are visited many times because of arbitrary head movements. Ranjan et al. [83] suggested to adapt HRIRs at new (unvisited) and old (already visited) measurement positions separately. If the measurement direction is not yet visited ($\widehat{\mathbf{h}}(\theta_{k+1}, \varphi_{k+1}) = 0$), the adaption equation is the same as Equation 2.6, namely progressive-based NLMS (P-NLMS). In contrast, if the measurement direction is already visited ($\widehat{\mathbf{h}}(\theta_{k+1}, \varphi_{k+1}) \neq 0$), the HRIR for this direction is updated based on its old weights for the same direction ($\widehat{\mathbf{h}}_{\mathbf{old}}(\varphi_{k+1}, \theta_{k+1})$), namely activation-based NLMS (A-NLMS):

$$\widehat{\mathbf{h}}(\varphi_{k+1}, \theta_{k+1}) = \widehat{\mathbf{h}}_{\mathbf{old}}(\varphi_{k+1}, \theta_{k+1}) + \mu\, \frac{\mathbf{x}(k)}{||\mathbf{x}(k)||^2 + \epsilon}\, e(k). \tag{2.7}$$

$\widehat{\mathbf{h}}_{\mathbf{old}}(\varphi_{k+1}, \theta_{k+1})$ is then updated with the new filter coefficients estimated at time $k+1$. Their studies have shown a substantial improvement in the accuracy of the measured HRIRs by using the P- and A-NLMS (PA- NLMS) algorithms compared with the conventional NLMS approach. In addition to the adaptive filtering methods, several studies have used periodic excitation signals combined with classic deconvolution methods to continuously estimate time-varying HRIRs [86–88].

During the measurement, subjects should be able to see the target measurement directions and their head movement pattern in real-time so that they can be prompted to cover all measurement points. We have developed an HRTF measurement system based on a video monitor and a head tracking device [50]. As shown in the left panel of Figure 2.2, a head tracker is placed on the subject's head to record the orientation data, and a video monitor is placed in front of the subject to display the measurement positions and the head movement pattern. The right panel of Figure 2.2 shows an example of the visualized information of

the subject's head orientation (blue lines) and the target measurement positions (red circles). However, the subject can only see the information when looking at the video monitor, i.e., the video monitor provides only intermittent visual feedback to the subject [89].

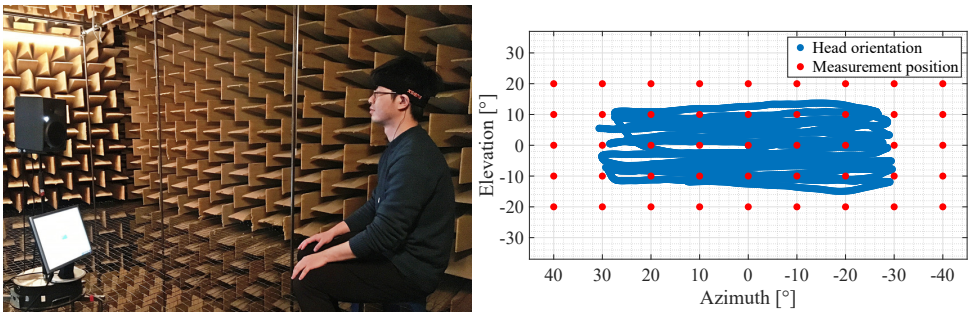


Figure 2.2: Fast 2D HRTF measurement system based on a video monitor and a head tracking device (left panel). Visualization of the head orientation and the measurement points (right panel).

To overcome this problem, we have further developed a mobile HRTF measurement system based on a VR Head Mounted Display (HMD). The measurement setup can be seen in the left panel of Figure 2.3, where a subject wears a VR HMD and stands in front of a loudspeaker. As shown in the right panel of Figure 2.3, the VR HMD introduces the subject in a virtual measurement environment (virtual anechoic chamber). The green balls presented through the VR HMD indicate desired measurement positions and the small white ball (in front of the virtual loudspeaker) represents the subject's view direction. The subject needs to rotate the head to reach these target positions. As soon as one of the target positions is visited, the green ball at this position disappears, indicating that the HRIR for this source direction has been measured.

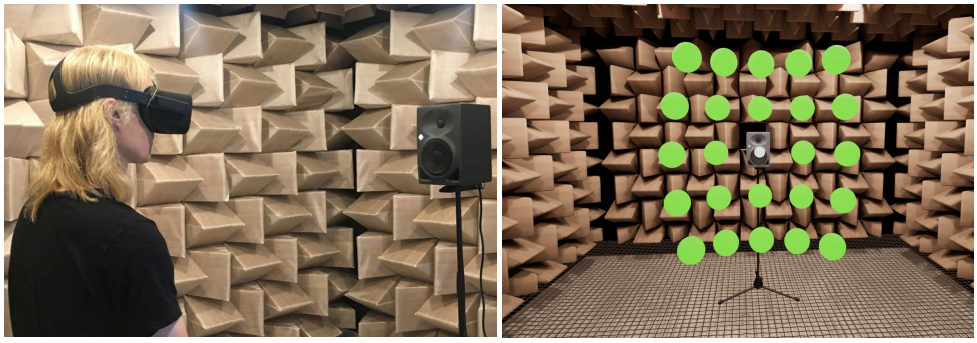


Figure 2.3: Mobile 2D HRTF measurement system based on a VR HMD (left panel). Visualization of the measurement environment and the desired measurement positions (right panel).

With this measurement system, subjects are constantly provided with the visual information about the head orientation and the visited/unvisited measurement positions (concurrent visual feedback) [89]. He et al. [40] performed

objective and subjective evaluations of HRTFs captured with the mobile HRTF measurement system. The quality of the measured HRTFs were high and comparable to the HRTFs measured with classical static measurement systems. Peksi et al. [90] have further developed the VR-based HRTF measurement system to enable the HRTF calculation in real-time. In this way, the recorded data, including the ear signals and the orientation data, no longer need to be stored. In addition, the HRTF quality is calculated at each measurement position and provided to subjects in real-time through the VR HMD.

2.3.3 *Multi-microphone setups*

Zotkin et al. [91] proposed a "reciprocity method" for rapid acquisition of HRTFs by exchanging microphone and speaker positions according to the Helmholtz principle of reciprocity [92]. A pair of miniature speakers is inserted into the subject's ear canals, and a microphone array surrounds the subject. The microphone positions represent the directions of the HRTFs to be measured. Excitation signals are played pack through the miniature speakers (left and right ear signals are played back consecutively), and HRTFs from all microphone directions can be captured at once. However, the measured HRTF has a poor performance at low frequencies because of the small-size speakers. Moreover, the playback level of the excitation signal is limited due to physiological safety, resulting in a relatively low SNR of the measurement result.

2.3.4 *HRTF measurements in non-anechoic environments*

HRTFs are typically measured in (semi-) anechoic chambers used to simulate free-field environments. Fast HRTF measurements in non-anechoic environments are a timely topic, because not all subjects have the opportunity to measure individual HRTFs in anechoic chambers.

The elimination of reflections and background noises is the challenge for measurements performed in ordinary rooms [41]. Truncating the measured HRIRs by applying a (frequency-dependent) window function [93, 94] is a typical method to remove reflections, but this approach may not be effective when performing measurements in complex acoustic environments. The influence of background noises on the measurement results, i.e., decreased SNR, can be reduced by repeating the measurement several times or using repeated excitation signals [93], but the measurement time increases accordingly.

Recently, two preliminary studies have proposed novel approaches to suppress the influence of background noises [95] and reflections [96] by analyzing

the sound field of the measurement environment captured by an additional microphone array.

He et al. [95] extracted the sound source signal, the ambisonic energy, the diffuseness, and the source direction from the local sound field recorded by an ambisonic microphone (Sennheiser AMBEO). For each measurement direction, a suitable time frame in which the diffuseness was minimum, the ambisonic energy was highest, or the ear signal energy was highest, was selected for the calculation of HRTFs using conventional deconvolution methods. The selection of time frame avoided the calculation of HRTFs when the energy of the ambient noise is high. Lopez et al. [96] proposed a method to measure HRTFs with a multi-loudspeaker-based measurement system in an ordinary room. Prior to the HRTF measurement, a set of room impulse responses between each loudspeaker and a spherical microphone array (custom-made) was measured, where the microphone array was positioned at the same place where the subject would be. Then, the measured impulse responses were decomposed in different directions to identify the reflection pattern between the positions of the loudspeaker and the microphone array using the Plane Wave Decomposition (PWD) approach. After the HRTF measurement, the detected reflection patterns were used to remove reflections contained in HRIRs.

2.4 TOWARDS MOBILE 3D HRTF MEASUREMENT

As mentioned above, most HRTF measurement systems are designed for measuring 2D HRTFs in the far-field, i.e., HRTFs in azimuth and elevation planes with a fixed source-listener distance. However, in the near-field, HRTF spectra change substantially with different distances. In addition, ILDs of lateral sound sources increase with decreasing distance, while ITDs remain unchanged for different distances [42]. As a result, direction- and distance-dependent HRTFs are required in binaural synthesis to create immersive VAEs.

To rapidly record 3D HRTFs, an MR-based mobile measurement system is developed in combination with a single loudspeaker. The depth camera integrated in the MR HMD can detect the distance between the loudspeaker and subjects, allowing subjects to move their bodies towards or away from the loudspeaker to measure HRTFs at different distances.

2.4.1 *Overview of the measurement system*

Figure 2.4 shows an overview of the MR-based mobile HRTF measurement system. A subject equipped with a pair of miniature microphones and an MR HMD is positioned in front of a loudspeaker. The loudspeaker plays back a

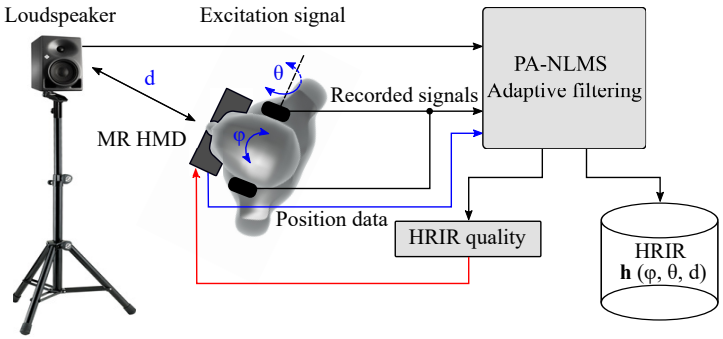


Figure 2.4: Overview of the MR-based mobile HRTF measurement system.

continuous excitation signal (e.g., white noise, perfect sweeps), and the emitted signal is recorded by the in-ear microphones. Through the inertial sensor and the depth camera integrated in the MR HMD, the position data (orientation and distance) of the subject is acquired and transmitted to a computer via the User Datagram Protocol (UDP). The position data, the excitation signal, and the recorded binaural signals are jointly used to estimate HRTFs frame by frame.

By applying the NLMS method, the algorithm for estimating the HRIR at discrete time $k+1$ $\left(\widehat{\mathbf{h}}(\varphi_{k+1}, \theta_{k+1}, d_{k+1})\right)$ can be formulated as:

$$\begin{aligned} e(k) &= y(k) - \widehat{\mathbf{h}}^{T}(\varphi_k, \theta_k, d_k)\, \mathbf{x}(k), \\ \widehat{\mathbf{h}}(\varphi_{k+1}, \theta_{k+1}, d_{k+1}) &= \widehat{\mathbf{h}}(\varphi_k, \theta_k, d_k) + \mu \frac{\mathbf{x}(k)}{||\mathbf{x}(k)||^2 + \epsilon}\, e(k), \end{aligned} \tag{2.8}$$

where d_k denotes the distance between the listener and the loudspeaker at discrete time k. This equation can further be extended based on the PA-NLMS method as described in Section 2.3.2. When measuring the HRIR at a new (unvisited) position ($\widehat{\mathbf{h}}(\varphi_{k+1}, \theta_{k+1}, d_{k+1}) = 0$), the HRIR update equation is the same as Equation 2.8. For an already measured position ($\widehat{\mathbf{h}}(\varphi_{k+1}, \theta_{k+1}, d_{k+1}) \neq 0$), the HRIR is updated based on its old weights for the same position ($\widehat{\mathbf{h}}_{\mathbf{old}}(\varphi_{k+1}, \theta_{k+1}, d_{k+1})$):

$$\widehat{\mathbf{h}}(\varphi_{k+1}, \theta_{k+1}, d_{k+1}) = \widehat{\mathbf{h}}_{\mathbf{old}}(\varphi_{k+1}, \theta_{k+1}, d_{k+1}) + \mu \frac{\mathbf{x}(k)}{||\mathbf{x}(k)||^2 + \epsilon}\, e(k). \tag{2.9}$$

$\widehat{\mathbf{h}}_{\mathbf{old}}(\varphi_{k+1}, \theta_{k+1}, d_{k+1})$ is then updated with the new filter coefficients obtained at time $k+1$. The estimation error of HRTFs is calculated and provided to subjects in real-time (see Section 2.4.2). Furthermore, a voice recognition function for starting and stopping the measurement is integrated into the measurement system so that subjects can control the measurement process themselves.

2.4.2 *Visual feedback*

Visualization of measurement positions

In the VR-based 2D HRTF measurement system, the 2D measurement points are visualized and appear on a spherical surface (see green balls in the right panel of Figure 2.3 or pink grids in the left panel of Figure 2.5). The measurement of 3D HRTFs is actually a multiple repetition of the 2D HRTF measurement procedure with different desired distances. The right panel of Figure 2.5 shows an example of measurement points distributed on three spherical surfaces at three measurement distances. Subjects can only see 2D measurement points on one spherical surface at each distance during the measurement. As soon as the estimated distance is one of the target distances, 2D measurement points appear on one spherical surface, while the measurement points on the other two spherical surfaces are not visible. Subjects are asked to perform head movements to visit 2D measurement points at each predefined source-listener distance. To cover different measurement points, subjects are asked not only to rotate their heads, but also to move their bodies towards or away from the loudspeaker. Note that the 3D measurement points, including measurement distances and 2D measurement points at each distance, should be predefined by the user. Figure 2.5 (right panel) shows an example of possible distributions of 3D measurement points. To test our proposed system, only the HRTFs in the horizontal plane with different distances were measured and objectively evaluated to reduce the amount of measurement positions (see Section 2.4.3).

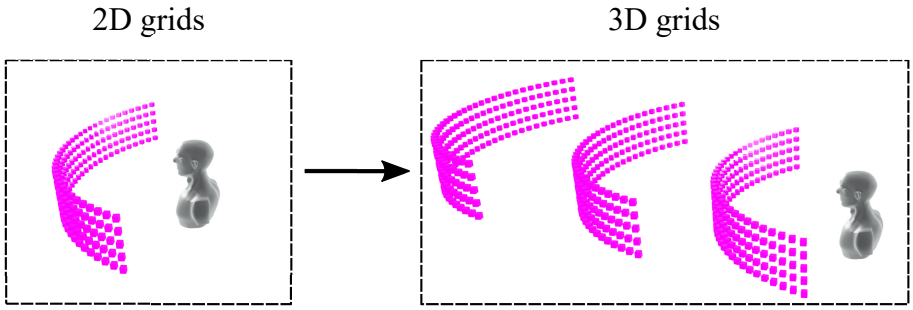


Figure 2.5: Virtual measurement points for the 3D HRTF acquisition.

Visualization of HRTF quality

A variety of factors, such as the rotational speed of head movements, impulsive noises, etc., may influence the quality of measured HRTFs. It is important that the HRTF quality can be calculated in real-time and provided to subjects during the measurement, allowing subjects to re-visit the positions where HRTFs have poor quality. The Normalized Mean Square Error (NMSE) is generally

applied as a metric to indicate the HRTF quality when using adaptive filtering approaches, which is expressed as (neglecting the subscripts designating the left and right ear):

$$\text{NMSE} = \frac{\|\mathbf{e}\|_2^2}{\|\mathbf{y}\|_2^2}, \tag{2.10}$$

where $\mathbf{e}$ and $\mathbf{y}$ are vectors consisting of the most recent N (HRIR length) samples of the error and recorded ear signals, respectively. The average NMSE of the left and the right ear is calculated and used as an indicator of the HRTF quality, i.e., the higher the NMSE, the lower the HRTF quality, and vice versa.

Since subjects may not be familiar with the meaning of NMSEs, the measurement points are marked in different colors based on the NMSE values. Figure 2.6 shows an example of the visual representation of HRTF quality with different colors. The grids with pink color represent unvisited measurement positions, and disappeared grids indicate sufficient quality of measured HRTFs (e.g., NMSEs < -25 dB). When the HRTF quality is poor or very poor, the grids are filled with blue (e.g., -25 dB < NMSEs < -5 dB) or red colors (e.g., NMSEs > -5 dB), and subjects are suggested to re-visit these measurement positions. In this way, subjects can easily be informed about the HRTF quality during the measurement.

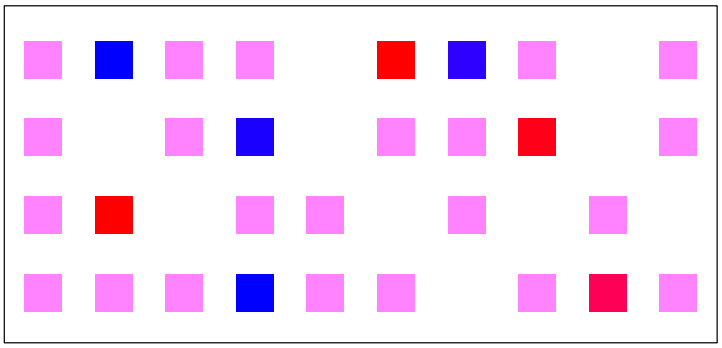


Figure 2.6: Visual feedback of the HRTF quality through the MR HMD.

2.4.3 *Measurement results*

A test measurement with a human subject was performed in an anechoic chamber (4.7 m × 4.3 m × 3.5 m) located in our institute (see Figure 2.7). A loudspeaker was positioned in front of the subject at a distance of 1.3 m. HRTFs were measured in the horizontal plane from $-90°$ to 90° (5° resolution) at three distances of 1.3 m, 1 m and 0.5 m. Target measurement positions in the horizontal plane are visible when the distance between the subject and the loudspeaker is one of the desired distances. Since it is not possible to exactly reach the target

distance, a tolerance was set for each target distance ($\pm$ 0.2 m). The main measurement equipment was an audio interface (Focusrite Scarlett), a pair of miniature microphones (Madness MM-BSM-8), a loudspeaker (Neumann KH 120A), a computer (Dell OptiPlex 5070), and an MR device (Microsoft HoloLens one). The sampling rate of the recorded and reproduced audio signals is 44.1 kHz.

Figure 2.7: MR-based mobile 3D HRTF measurement system.

Figure 2.8 shows an example of raw HRIRs measured at an azimuth angle of 60° (φ = 60°, θ = 0°) at three distances (1.3 m, 1 m, and 0.5 m). The left and right panels show the HRIRs of the ipsilateral and contralateral ears, respectively. It can be clearly seen that the onset delay (propagation time from the loudspeaker to binaural microphones) reduces and the relative amplitude of HRIRs increases as the measurement distance decreases. The measured results were then further post-processed, including truncating the HRIR lengths, and compensating for the microphone and loudspeaker transfer functions.

ILDs and ITDs were extracted from the measured HRTFs to show their changes with source directions and distances. For the estimation of ILDs, HRIRs were first filtered through a gammatone filter bank [97] with a bandwidth of one Equivalent Rectangular Bandwidth (ERB) [98], half-wave rectified and filtered with a 1 kHz first-order low-pass filter [99] (inner hair cell model) to approximate the cochlear filtering procedure, then the differences of root mean square values between the processed HRIRs of the left and right ears were calculated and averaged across frequency bands. Although ITDs vary with frequencies [100], these frequency-dependent variations are not relevant to the perception of source localization [101]. In this study, broadband ITDs were determined by calculating the Interaural Cross-Correlation Coefficient (IACC) between the low-pass filtered HRIRs (cut-off frequency at 3 kHz) of the left and right ears [102].

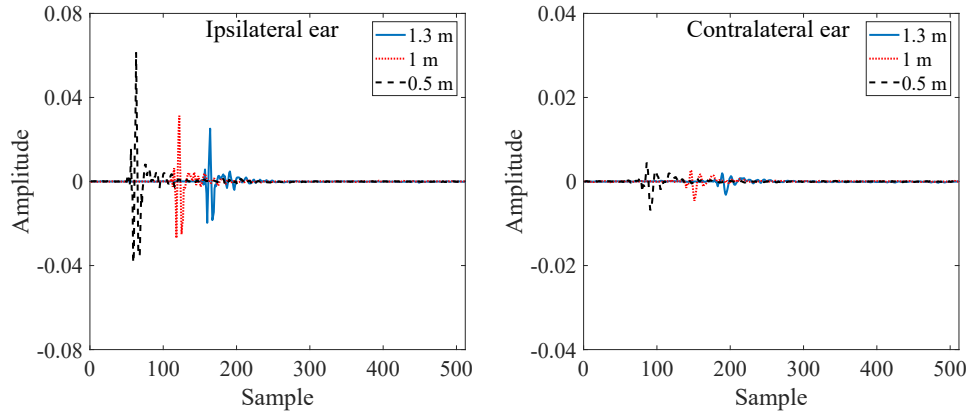


Figure 2.8: Raw HRIRs at an azimuth angle of 60° (φ = 60°, θ = 0°) at distances of 1.3 m (blue solid lines), 1 m (red dotted lines), and 0.5 m (black solid lines). The left and right panels show the HRIRs of the ipsilateral and contralateral ears, respectively. The sampling rate of HRIRs is 44.1 kHz.

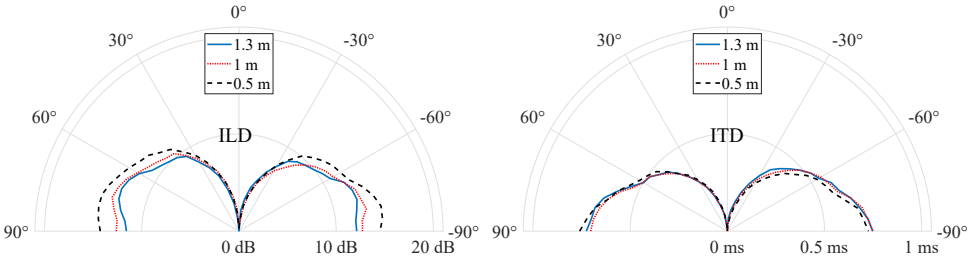


Figure 2.9: Absolute ILDs (left panel) and ITDs (right panel) of HRIRs measured in the horizontal plane ($-90° \leqslant \varphi \leqslant 90°$, θ = 0°) at three distances of 1.3 m (blue solid lines), 1 m (red dotted lines), and 0.5 m (black solid lines).

Figure 2.9 shows the absolute ILDs (averaged across frequency channels) and ITDs with different azimuth angles and source-listener distances. For each distance, ILDs and ITDs increase as the sound source moves laterally to the listener's head, and the maximum values of ILDs and ITDs are at the azimuth angles of around $\pm$ 70° and $\pm$ 90°, respectively. The ILDs of HRTFs measured at 1 m and 1.3 m are comparable for different directions, but a noticeable increase in ILDs of lateral sound sources can be observed as the distance decreases from 1 m to 0.5 m. In contrast, the ITDs remain almost unchanged with different distances. The changes in ILDs and ITDs with source directions and distances are comparable to the precise HRTF measurement results presented in [103], and are highly consistent with the observations by Brungart and Rabinowitz [42].

2.4.4 *Influences of the MR HMD on HRTFs*

The MR HMD used for the measurement might degrade the quality of HRTFs measured. To investigate its influence on the measurement results, HRTFs of a dummy head (Neumann KU-100) were measured with and without wearing the MR HMD (Microsoft Hololens one) in the horizontal plane from $-90°$ to $90°$ ($10°$ resolution). A loudspeaker (Neumann KH 120A) was fixed in a predefined position, and the dummy head was rotated to reach different orientations. An inertial sensor (MTw Xsens) was placed on the dummy head and used to control the head orientation. Two distances, 1.3 m (far-field) and 0.5 m (near-field), were considered for the measurement to further investigate whether or not the MR HMD has different influences on the near- and far-field HRTFs.

A 5 s-long exponential sweep [54] was used as the excitation signal to record HRIRs, and each measurement was repeated five times to increase the SNR of the results. The measured HRIRs were truncated/windowed to remove reflections, and equalized to compensate for the microphone and loudspeaker transfer functions. After the measurement, the deviations in the HRTF magnitude and binaural cues (ILDs and ITDs) caused by wearing the MR HMD were extracted and evaluated.

Influences of the MR HMD on HRTF magnitude

To analyze the deviations in the HRTF magnitude caused by wearing the MR HMD, the Spectral Deviation (SD) between two HRTF sets (with and without the MR HMD) was calculated for each measurement distance [104]:

$$\mathrm{SD}\,(f) = \frac{\mathrm{SD}_l\,(f) + \mathrm{SD}_r\,(f)}{2}, \tag{2.11}$$

with

$$\mathrm{SD}_i\,(f) = \frac{1}{N_\varphi} \sum_{\varphi=\varphi_1}^{\varphi_{N_\varphi}} \left| 20\,\log_{10} \frac{|\mathrm{HRTF}_{i,wo}(f,\varphi)|}{|\mathrm{HRTF}_{i,w}(f,\varphi)|} \right|, \quad i \in \{l, r\} \tag{2.12}$$

where N_φ is the number of azimuth angles ($\varphi \in \{\varphi_1, \varphi_2, ..., \varphi_{N_\varphi}\}$), and f denotes the frequency. The subscript i represents the left or the right ear. $\mathrm{HRTF}_{i,w}$ and $\mathrm{HRTF}_{i,wo}$ are HRTFs of the dummy head with and without the MR HMD, respectively.

Figure 2.10 shows the SDs (smoothed with a 1/12-octave filter) between HRTFs with and without the MR HMD ("$\mathrm{WO}_{\mathrm{HMD}}$ vs. $\mathrm{W}_{\mathrm{HMD}}$") at measurement distances of 1.3 m and 0.5 m. The SDs for both distances are comparable across frequencies. At low frequencies (below 1 kHz), the SDs are less than 1 dB, and at frequencies above 1 kHz, the SDs increase with increasing frequency. The

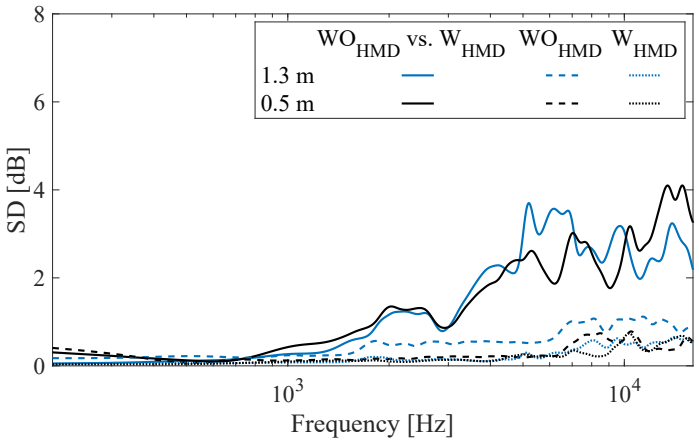


Figure 2.10: SDs caused by wearing the MR HMD at measurement distances of 1.3 m and 0.5 m over frequencies ("WO_{HMD} vs. W_{HMD}", blue and black solid lines). SDs introduced by repeated measurements of HRTFs over frequencies (without HMD: "WO_{HMD}", blue and black dashed lines; with HMD: "W_{HMD}", blue and black dotted lines).

highest SDs are about 3.7 dB and 4.0 dB for distances of 1.3 m and 0.5 m, respectively. The results show that the SD caused by wearing the MR HMD does not depend on the measurement distance.

Different factors might affect the measurement accuracy, such as drift of the inertial sensor and repositioning of the HMD on the artificial head. To validate the repeatability of the measurement system, the two HRTF sets (with and without the MR HMD) were re-measured, and the SDs of HRTFs between these two measurement sets were calculated [105]. The results are displayed in Figure 2.10, where W_{HMD} and WO_{HMD} are the SDs (smoothed with a 1/12-octave filter) caused by repeated measurements of HRTFs with and without HMD, respectively. Overall, the SDs are smaller than 1 dB across frequencies. For frequencies above 1 kHz, the SDs caused by the measurement uncertainty are clearly lower than those caused by wearing the MR HMD, indicating the high repeatability of the measurement system.

Influences of the MR HMD on binaural cues

The deviations in ILDs and ITDs caused by wearing the MR HMD were calculated for each distance and each azimuth angle. The methods to extract ILDs and ITDs from HRTFs were the same as described in Section 2.4.3. The results in Figure 2.11 show that the ILD and ITD deviations are less than about 2 dB and 0.05 ms, respectively. Further, no noticeable difference in ILD and ITD deviations between two distances can be observed. In comparison with the Just Noticeable Difference (JND) in ILDs (0.6 dB and 1.2 dB for anechoic and reverberant sounds, respectively) and ITDs (0.02 ms and 0.16 ms for anechoic and reverberant sounds, respectively) [106], the ITD deviations are less than JNDs

for most azimuth angles, but the ILD deviations are slightly larger than JNDs for lateral source directions.

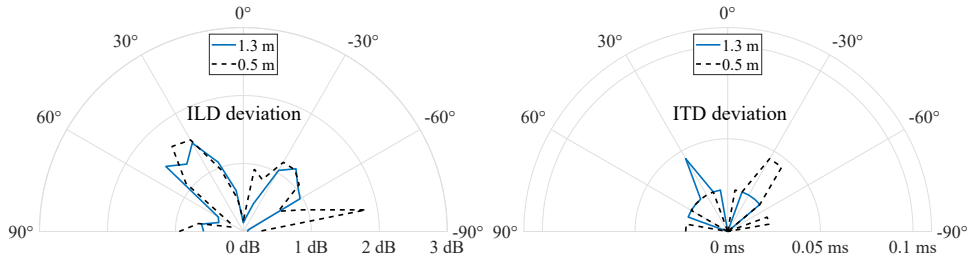


Figure 2.11: Absolute ILD (left panel) and ITD deviations (right panel) of HRTFs in the horizontal plane ($-90° \leqslant \varphi \leqslant 90°$, $\theta = 0°$) at the measurement distances of 1.3 m (blue solid lines) and 0.5 m (black dashed lines).

2.4.5 *Summary*

The proposed mobile HRTF measurement system allows subjects to measure individual 3D HRTFs by rotating their heads and moving their bodies. The HRTF quality is calculated and provided to subjects in real-time. The change in binaural cues (ILD and ITD) extracted from the measured HRTFs with different directions and distances are in line with the results presented in [42, 103]. One limitation in the current measurement system is that the loudspeaker used can not be considered as a point sound source when measuring HRTFs in a close distance. The influences of the MR HMD (Microsoft HoloLens one) on the measured HRTFs have been further objectively evaluated. The deviations in HRTF magnitude and binaural cues (ILDs and ITDs) are overall small and not strongly dependent on distance. Compared to the JNDs, the deviations in binaural cues are slightly higher than the JNDs for lateral source directions. Nevertheless, our proposed measurement method can be considered as a potential solution for fast measuring 3D HRTFs with a flexible setup.

Future work includes performing psychoacoustic experiments to perceptually evaluate the quality of measured HRTFs, designing a suitable sound source to approximate the characteristics of an acoustic point source for measuring near-field HRTFs, and applying an equalization filter to compensate for the influences of the MR HMD on HRTFs. In addition, additional sensors are considered to measure HRTFs not only with different distances and directions, but also with head and torso orientations.

2.5 CONCLUDING REMARKS

Different measurement systems have been developed to acquire HRTFs of human subjects. By using the "reciprocity method" with a multi-microphone setup, HRTFs can be measured from multiple directions within a few seconds. However, the measured HRTFs have a poor performance at low frequencies because of the small-size speakers, and a low SNR due to the limited playback level of the excitation signal. Multi-loudspeaker setups are therefore preferred for the HRTF measurement.

The use of a large speaker setup (e.g., multiple loudspeakers distributed on a sphere) can speed up the measurement process, but high infrastructure costs are to be expected. With consideration of the measurement time and the cost-benefit, a loudspeaker array in combination with a single-axis positioning system, e.g., the combination of a vertical loudspeaker array and a turn table, is generally applied in acoustic laboratories. Compared to the stop & go measurement mechanism, the continuous mechanism shows its advantage in saving time for repositioning subjects or loudspeaker systems. Most multi-loudspeaker setups are only allowed for measuring HRTFs with a fixed distance (2D HRTFs). The measurement setup proposed in [82] enables to measure distance-dependent HRTFs (3D HRTFs), where multiple speakers are placed on a vertical locating loop with length-adjustable support rods [41].

Several VR/MR-based mobile measurement systems have recently been proposed to fast measure 2D HRTFs with a single loudspeaker [40, 51, 89, 107]. The advantages of these mobile measurement systems are the flexible hardware setups and the low infrastructure costs. In this thesis, a novel MR-based mobile measurement system is developed to capture distance- and direction-dependent HRTFs (3D HRTFs). The HRIRs and the estimation errors are calculated in real-time to ensure the measurement quality. However, the influence of the MR HMD on the measurement results can not be neglected and has to be further compensated, and a suitable point sound source is required for measuring HRTFs in the near-field.

HRTF measurements in complex acoustic environments are of great interest, since not all listeners can measure their HRTFs in (semi-) anechoic chambers. Some studies have shown the possibility to remove reflections and ambient noises from HRTFs measured in ordinary rooms by analyzing the acoustics of the measurement environment [95, 96, 108]. These methods, in combination with our proposed mobile HRTF measurement system, show the potential to rapidly measure individual 3D HRTFs in ordinary home environments in the future.

3

THE ROLE OF REVERBERATION IN CONTRALATERAL AND IPSILATERAL EAR SIGNALS ON PERCEIVED EXTERNALIZATION OF A LATERAL SOUND SOURCE

3.1 INTRODUCTION

Reverberation plays an important role in perceived externalization of virtual sound images presented over headphones. It is clear that reverberation heard by both ears is equally important for perceived externalization of a frontal sound image, but little is known about whether reverberation at the left and right ear has the same contribution to externalization of a lateral sound source.

In this Chapter, a pair of BRIRs is measured at an azimuth angle of -45° in a listening room, and the BRIR of each ear is truncated separately to reduce the amount of reverberation. Listeners are asked to rate externalization of test signals generated with these modified BRIRs to investigate the relative influence of reverberation in the contralateral versus ipsilateral ear signal on externalization. Different acoustic cues extracted from the test signals are compared to the subjective results obtained in listening experiments. Furthermore, the influence of lateralized reverberation on perceived externalization is evaluated for different source directions in the horizontal plane. Parts of this Chapter have been published in [109, 110].

The rest of this Chapter is organized as follows. Section 3.2 shows the effect of reverberation on BRIRs. The experimental paradigm and the results are described in Section 3.3. Then, a comparison between acoustic cues and externalization results is presented in Section 3.4. In Section 3.5, the experimental results and the externalization models based on different acoustic cues are discussed, and the influence of lateralized reverberation on externalization is tested for different source directions. Finally, conclusions and future work are drawn in Section 3.6.

3.2 INFLUENCE OF REVERBERATION ON BRIRS

To illustrate the effect of reverberation on BRIRs of the left and right ear at different source directions, two pairs of BRIRs were measured with a dummy head KEMAR 45BC-12 at azimuth angles of 0° and -45° in a listening room (6.7 m × 4.8 m × 3.2 m) located in our institute [111]. The listening room is designed under the ITU-R BS.1116 standard and has a reverberation time T_{60} of about 260 ms. The distance between the loudspeaker (Neumann KH 120A) and the KEMAR was 1.7 m. A 5 s-long exponential sweep [54] from 20 Hz to 20 kHz was used as the excitation signal to measure BRIRs, and each measurement was repeated five times. The BRIRs measured were truncated to a length of 260 ms with a 10 ms-long half raised-cosine fall time. The measurements were performed at a sampling rate of 44.1 kHz.

The direct parts of the measured BRIRs were extracted by applying a time window, which has a constant amplitude for up to 2.5 ms after the onset delay (the propagation time from the loudspeaker to the KEMAR's ipsilateral ear) and a 0.5 ms-long half raised-cosine fall time. The direct parts obtained are not exactly the same as HRIRs, because it is difficult to perfectly split the direct and reverberant components in BRIRs, especially at low frequencies. Hence, the obtained direct components from measured BRIRs are referred to as pseudo HRIRs [23].

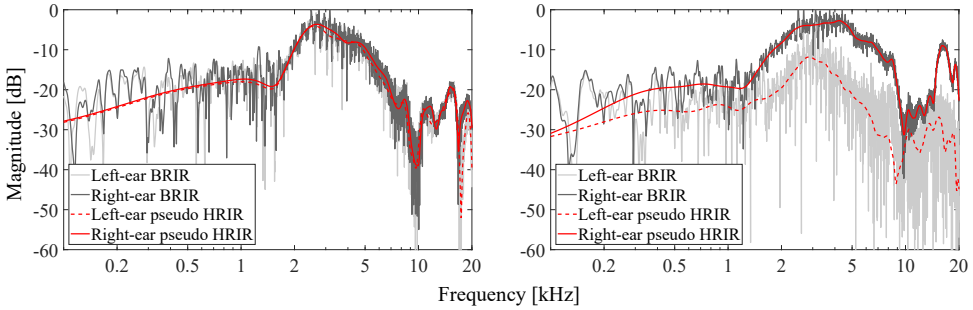


Figure 3.1: Magnitude spectra of BRIRs measured at 0° (left panel) and -45° (right panel) relative to the KEMAR. Light and dark gray solid lines represent the magnitude spectra of BRIRs of left and right ears, respectively. The red solid and dashed lines show the magnitude spectra of extracted direct components (pseudo HRIRs) of left and right ears, respectively.

Figure 3.1 shows the magnitude spectra of BRIRs and the extracted direct components (pseudo HRIRs) at the azimuth angle of 0° (left panel) and -45° (right panel). The magnitude spectra in each panel were normalized to the maximal magnitude level. Additionally, to show the relative change in direct and reverberant sound energy at the left and right ear for different source directions, the ratio of direct sound energy at the left versus the right ear (D_l/D_r),

and the ratio of reverberant sound energy at the left versus the right ear (R_l/R_r) were calculated in dB:

$$D_l/D_r = 10\log_{10}\frac{\int_0^T \mathrm{BRIR}_l(\tau)^2 d\tau}{\int_0^T \mathrm{BRIR}_r(\tau)^2 d\tau}, \tag{3.1}$$

$$R_l/R_r = 10\log_{10}\frac{\int_T^\infty \mathrm{BRIR}_l(\tau)^2 d\tau}{\int_T^\infty \mathrm{BRIR}_r(\tau)^2 d\tau}, \tag{3.2}$$

where T represents the length of direct parts in BRIRs, and is defined as 2.5 ms after the initial delay. BRIR_l and BRIR_r indicate the BRIR of the left and right ear, respectively.

For the frontal sound source (0°), both the direct sound and the reverberation reaching the left and right ear have almost the same energy (D_{left}/D_{right}= −0.7 dB, R_{left}/R_{right}= 0.04 dB). In the case of the lateral sound source (-45°), the energy of the direct sound is much higher at the ipsilateral ear (right ear) than at the contralateral ear (left ear) because of the head shadow effect, which is well reflected in the D_l/D_r value (−12.3 dB). On the other hand, the energy of reverberation is almost the same at both ears (R_{left}/R_{right}= −1.0 dB), resulting in a higher DRR value for the right ear than for the left ear.

Further, the absolute magnitude differences between each pair of adjacent frequency bins (frequency resolution is 1 Hz/bin) were calculated to represent the FFV in the BRIR of each ear (c.f. Section 3.4.1). The average FFV values across frequencies are approximately 0.6 dB/Hz and 0.1 dB/Hz for the contralateral and ipsilateral ear BRIR, respectively. The results demonstrate that the FFV is more pronounced in the contralateral ear than in the ipsilateral ear, which is in agreement with the observations in [23]. Since FFV and DRR are two cues related to distance perception, the difference in these two cues between two ears suggests that reverberation at the ipsilateral and contralateral ear may have different influences on perceived externalization of a lateral sound source. The pair of BRIRs measured at -45° was used in the experiment (see Section 3.3).

3.3 EXPERIMENT

3.3.1 *Experimental paradigm*

In order to investigate the relative influence of reverberation at the contralateral versus the ipsilateral ear on externalization, the amount of reverberation at each ear was removed separately with different degrees. In the experiment, the amount of reverberation at each ear was reduced by truncating the BRIR of each ear to durations of 2.5 ms, 5 ms, 10 ms, 20 ms, 40 ms, 80 ms, 120 ms and 200 ms

with a 0.5 ms-long half raised-cosine fall time. Note that such modification was not intended to simulate naturally occurring conditions. Rather, this artificial modification could potentially inform the use of reverberation in the virtual sound environment.

Hence, three conditions of modified BRIRs could be observed: (a) BRIRs of both ears were truncated ("both truncated" condition). (b) The BRIR of the contralateral ear was truncated with different window durations, while the BRIR of the ipsilateral ear was not truncated ("truncated contralaterally" condition). (c) The BRIR of the ipsilateral ear was truncated to different lengths, while the BRIR of the contralateral ear was kept unchanged ("truncated ipsilaterally" condition). All truncated BRIRs were zero-padded to the same length of the original/reference BRIRs (260 ms). With different BRIR durations, this resulted in 9 test signals to be evaluated in each condition.

Eight subjects (two females and six males) with normal hearing aged from 25 to 29 participated in the experiment. The listening test was performed in the same room where the BRIRs were measured before. Each subject sat in a chair and listened to the test signals with a pair of compensated headphones (Sennheiser HD800). The Headphone Transfer Function (HpTF) was measured with KEMAR, and the compensation filter was calculated by applying the least-squares inversion approach combined with a frequency-dependent regularization to the HpTF [112]. A loudspeaker was placed at the measurement position (-45° relative to the listener) to serve as a visual cue. As shown in Table 3.1, a four-point subjective rating scale was used to let subjects rate the stimuli by using a slider with a step-size of 0.1 between 0 and 3, which was similar to the scales used in [12] and [18].

Table 3.1: A subjective rating scale to rate perceived externalization.

Degree	Meaning of the degree
3	The sound is externalized and at the position of the loudspeaker.
2	The sound is externalized but not as far as the loudspeaker.
1	The sound is externalized but very close to me.
0	The sound is in my head.

Before the experiment, subjects were asked to listen to all test signals once to become familiar with each stimulus. In addition, subjects were able to listen to the stimulus played back through the loudspeaker, and they were informed that this stimulus came from the loudspeaker and such should act as a well externalized sound (externalization rating = 3). During the listening test, listeners could repeat every sequence and also listen to the reference signal played back over the loudspeaker (headphones should be taken off to hear the speaker signal). Note that this experiment was designed to evaluate the degree of externaliza-

tion of virtual sound images, other perceptual attributes, such as coloration [17] and plausibility [113] were not evaluated. Subjects were not allowed to move their heads during the experiment, since externalization of static sound images could be degraded if listeners moved their heads without head tracking [28]. During the experiment, subjects' head movements were monitored by the supervisor through a transparent window. A white noise of 1 s duration was applied as the sound stimulus in the experiment because of its uniform energy distribution across frequencies. The audio signals generated with the unmodified BRIRs were presented via headphones at a level of 64 dBA. For calibration, the headphones were put on the dummy head KEMAR and reproduced binaural signals. The playback level of the headphones was then adjusted based on the averaged measured sound level between the left and right ears of KEMAR.

3.3.2 *Experimental results*

Figure 3.2 shows the median externalization ratings with non-parametric 95 % Confidence Intervals (CIs) (notch-edges) [114] for the "both truncated", "truncated contralaterally" and "truncated ipsilaterally" conditions. When the BRIR lengths are longer than 10 ms, the generated virtual sound images are perceived as externalized for all three conditions (median externalization ratings > 1). When the truncated window durations are below 10 ms, the sound source is perceived within the subject's head for "both truncated" and "truncated contralaterally" conditions (median externalization ratings < 1), but is still perceived as externalized for the "truncated ipsilaterally" condition (median externalization ratings ≈ 2.5).

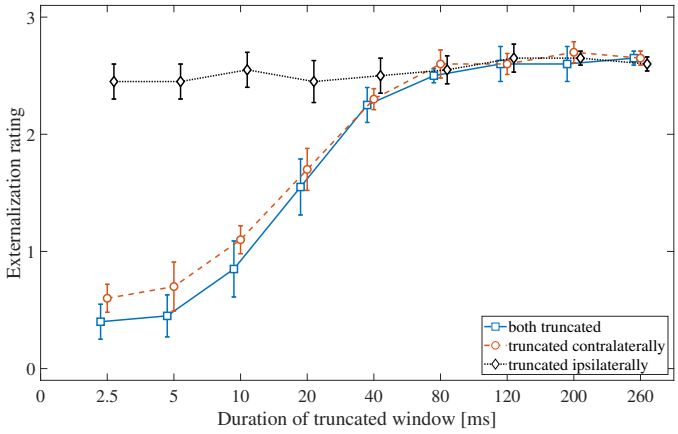


Figure 3.2: Median externalization ratings with non-parametric 95 % CIs (notch-edges) across subjects for the "both truncated" (solid line and squares), "truncated contralaterally" (dashed line and circles) and "truncated ipsilaterally" (dotted line and diamonds) conditions.

When the BRIR lengths are longer than 80 ms, the median externalization ratings for the three conditions are high and almost the same. The overall patterns of externalization ratings across different window durations are similar for the "both truncated" and "truncated contralaterally" conditions. In the case of short truncated window durations below 10 ms, the externalization ratings are low. A noticeable improvement in the externalization ratings is observed with increased window durations from 10 ms to 80 ms. Then, they do not change substantially for longer window durations (above 80 ms). The median externalization ratings for the "truncated contralaterally" condition are slightly higher than those for the "both truncated" condition below window lengths of 80 ms. However, for the "truncated ipsilaterally" condition, the median externalization ratings are always high (median externalization ratings $\approx$ 2.5) and remain almost unchanged across different window durations, corresponding to a sound source being externalized and close to the loudspeaker's position.

A Friedman test was performed for each condition. The results confirmed a significant effect by changing the window durations on externalization ratings for "truncated contralaterally" and "both truncated" conditions ($p \ll 0.05$). As expected, in the case of the "truncated ipsilaterally" condition, the externalization ratings were not affected by different window durations ($\chi^2(8) = 12.7$, $p = 0.12$).

3.4 ANALYSIS OF ACOUSTIC CUES

Perceived externalization of virtual sound images is usually assessed by listening experiments using a given rating scale (see Section 3.3). Some important acoustic cues mentioned in literature (see Section 1.1), including DRR [22], FFV [23] and reverberation-related binaural cues (IC, temporal fluctuations of ILD and IC) [18], have the potential to indicate perceived externalization. In this study, these cues are extracted from BRIRs and binaural signals and compared with the externalization results obtained in the experiment. It is important to know whether or not perceived externalization can be roughly predicted by these measured parameters instead of time-consuming subjective listening experiments.

3.4.1 *Monaural acoustic cues of the modified BRIRs*

DRR

DRR describes the ratio of direct and reverberant sound energy (in dB), and is expressed as:

$$\mathrm{DRR} = 10 \log_{10} \frac{\int_0^T h(\tau)^2 d\tau}{\int_T^\infty h(\tau)^2 d\tau}, \tag{3.3}$$

where h is the impulse response, and T represents the duration of the direct part in the impulse response, which is defined as 2.5 ms after the initial delay in the present study.

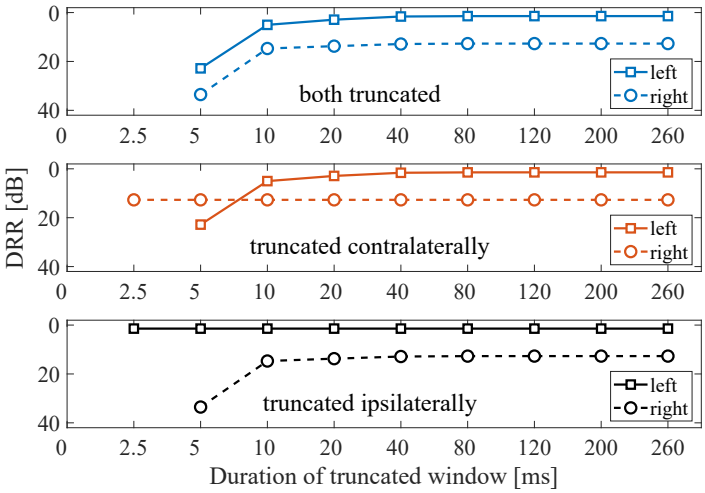


Figure 3.3: DRR of the left (solid lines) and right ear (dashed lines) BRIRs for the "both truncated" (top), "truncated contralaterally" (middle) and "truncated ipsilaterally" (bottom) conditions.

Figure 3.3 shows the DRR values for all experimental conditions in the experiment. For the window duration of 2.5 ms, the DRR value is infinite because no reverberation is present. For the "both truncated" condition, the DRR values are always higher for the ipsilateral ear than for the contralateral ear due to the head shadow effect. A large decrease in DRR can be observed for truncated window durations from 2.5 ms to 10 ms, then DRR slowly decreases up to 40 ms. For window durations above 40 ms, the DRR value shows no noticeable change. Furthermore, the decrease in DRR is more pronounced for the contralateral ear than for the ipsilateral ear. In comparison with the perceptual data, the large change in DRR values from 2.5 ms to 10 ms (from ∞ to 22 dB and 34 dB for the left and right ear, respectively) can not be observed in the externalization ratings. For truncated window durations between 80 and 260 ms, the DRR values are almost constant, which is consistent with the externalization ratings.

In the case of "truncated contralaterally" and "truncated ipsilaterally" conditions, the DRR values change over the window durations only for one ear. In comparison with the experimental data, it can be concluded that the change in DRR values for the ipsilateral ear does not substantially affect perceived externalization.

FFV

FFV describes the variability in the magnitude spectra of BRIRs from frequency to frequency, and is quantified by the absolute magnitude differences (in dB) between every pair of adjacent frequency bins (frequency resolution is 1 Hz/bin) in this study. The average values of the changes in the spectral magnitude across frequencies are shown in Figure 3.4.

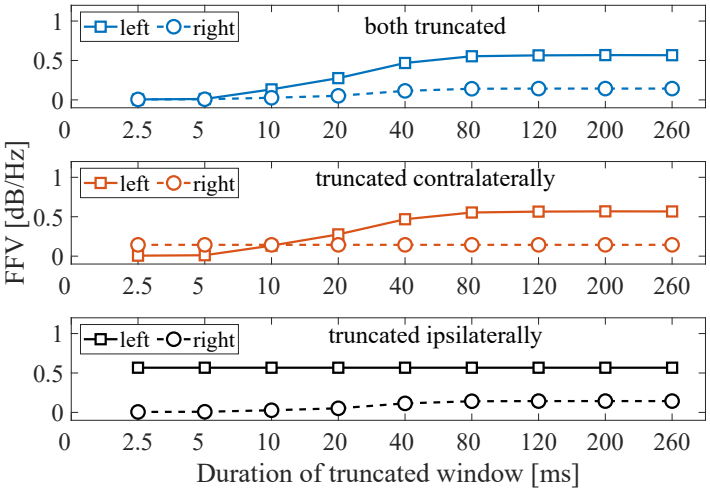


Figure 3.4: FFV of the left (solid lines) and right ear (dashed lines) BRIRs for the "both truncated" (top), "truncated contralaterally" (middle) and "truncated ipsilaterally" (bottom) conditions.

The FFV caused by interaction between direct sound and reverberation is higher in the BRIR of the contralateral ear than in that of the ipsilateral ear (the FFV in the unprocessed BRIR is about 0.6 dB/Hz and 0.1 dB/Hz for the contralateral and ipsilateral ear, respectively). In the case of the "both truncated" condition, the FFV values in both ears are almost 0 dB/Hz for truncated window durations below 5 ms due to the lack of sufficient reverberant energy. Then, FFV increases with increasing window durations up to 80 ms, and the change in FFV is more pronounced for the contralateral ear than for the ipsilateral ear. When window durations are above 80 ms, the FFV values do not change noticeably.

For "truncated contralaterally" and "truncated ipsilaterally" conditions, the FFV values change across different truncated window durations only in one ear. It can be seen that the change in FFV in the contralateral ear corresponds well

with the change in externalization ratings over truncated window durations and for different conditions, suggesting that FFV in the contralateral ear is more important for externalization than that in the ipsilateral ear.

3.4.2 *Reverberation-related binaural acoustic cues*

Previous studies have shown that the reverberation-related binaural cues, i.e., IC and IC temporal fluctuations at low frequencies, and ILD temporal fluctuations at high frequencies, are highly correlated with externalization ratings [18, 24]. It is hypothesized that high IC temporal fluctuations, high ILD temporal fluctuations, and low correlated binaural signals (low IC) correspond to high externalization ratings in the experiment.

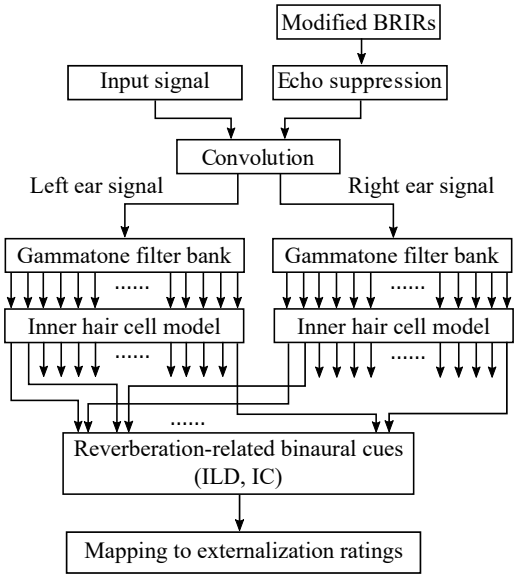


Figure 3.5: Structure of the model to obtain the reverberation-related binaural cues from binaural signals for different experimental conditions, consisting of an echo-suppression mechanism, a binaural rendering model (convolution process), and an auditory periphery model (gammatone filter bank and inner hair cell model).

Figure 3.5 shows the structure of the model to obtain the reverberation-based short-term binaural cues (ILD and IC) from binaural signals generated with modified BRIRs [18]. At first, the modified BRIRs under different experimental conditions are processed through an echo-suppression mechanism motivated by the "precedence effect" [7], which is realized by multiplying BRIRs with a time window that has a value of one from 0 ms to 2.5 ms (direct part), followed by zeros up to 10 ms (echo-suppression process) and a transition from zero to one from 10 ms to 15 ms [18]. After that, the input signal (white noise) is convolved with the echo-suppressed BRIRs. Then, the generated binaural signals are filtered through a gammatone filter bank [97] with a bandwidth of

one ERB [98], half-wave rectified and filtered with a 1 kHz low-pass filter [99] (simulation of human auditory periphery). Finally, in each frequency channel, the short-term binaural cues are calculated in a 20 ms-long Hann window with a 50 % overlap over the 1 s-long binaural signal.

ILD temporal fluctuations

In each frequency band centered at f_c, the standard deviation of short-term ILDs collected over the 1 s-long binaural signal, namely ILD Temporal Standard Deviation (ILD TSD), is calculated to characterize ILD temporal fluctuations:

$$\mathrm{ILD}_{\mathrm{TSD}}(f_c) = \sqrt{\frac{1}{N_{\mathrm{frame}} - 1} \sum_{n=1}^{N_{\mathrm{frame}}} (\mathrm{ILD}(f_c, n) - \overline{\mathrm{ILD}(f_c)})^2}, \tag{3.4}$$

where N_{frame} represents the number of frames in the binaural signal. $\mathrm{ILD}(f_c, n)$ and $\overline{\mathrm{ILD}(f_c)}$ are the ILD in the n^{th} frame and the average ILD over all frames, respectively.

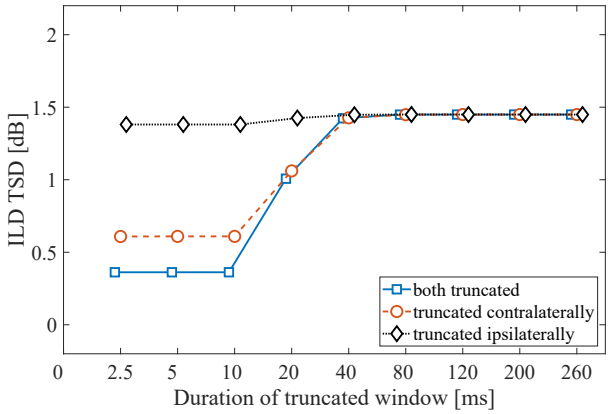


Figure 3.6: Average ILD SDs across frequency channels for the "both truncated" (solid line and squares), "truncated contralaterally" (dashed line and circles) and "truncated ipsilaterally" (dotted line and diamonds) conditions.

Figure 3.6 shows the average ILD TSDs over frequency channels (from 1 kHz to 18 kHz) in the gammatone filter bank for different experimental conditions. For the "truncated ipsilaterally" condition, the ILD TSDs are high and nearly constant with different truncated window durations (about 1.4 dB). In the case of "both truncated" and "truncated contralaterally" conditions, the ILD TSDs are constant below truncated window durations of 10 ms due to the echo-suppression process, then they increase noticeably with window durations until 80 ms. For longer window durations above 80 ms, they remain unchanged. For window durations below 40 ms, the ILD TSDs are slightly higher for the "truncated contralaterally" condition than for the "both truncated" condition.

It can be seen that reverberation at the contralateral ear has a larger effect on the ILD temporal fluctuations than that at the ipsilateral ear, and the change in ILD TSDs for different experimental conditions corresponds well to the externalization ratings.

IC and IC temporal fluctuations

The correlation between the left and right ear signals can be calculated by the normalized interaural cross-correlation function ($\rho(\tau)$):

$$\rho(\tau) = \frac{\int_{t_1}^{t_2} x_l(t)\, x_r(t+\tau)dt}{\sqrt{\int_{t_1}^{t_2} (x_l(t))^2 dt \int_{t_1}^{t_2} (x_r(t))^2 dt}}, \tag{3.5}$$

where $x_l(t)$ and $x_r(t)$ represent left and right ear signals in a time window between t_1 and t_2, respectively. τ denotes the time difference between the left and right ear signals, which is normally limited between -1 ms and 1 ms because of the plausible range of ITDs. IC describes the degree of coherence between the left and right ear signals and can be represented by IACC which is defined as the maximum of the absolute value of $\rho(\tau)$ in each time frame of the binaural signals [115]:

$$\text{IACC} = \max\{|\rho(\tau)|\}. \tag{3.6}$$

Similar to the approach in [18], IC and IC temporal fluctuations are used to indicate externalization of sounds at low to mid frequencies, together with ILD temporal fluctuations that are used to indicate externalization of sounds at mid to high frequencies. In this study, the short-term ICs are collected over the binaural signal in each frequency channel of the gammatone filter bank, and the average IC 10th and 90th percentiles are calculated over frequency channels centered from 150 Hz to 1.7 kHz to characterize the IC temporal fluctuations. Actually, there is no exact rule for selecting the frequency range of IC with respect to perceived externalization. Similar shapes of IC over experimental conditions can be observed in different frequency bands.

By this means, not only the absolute IC value but also the IC temporal fluctuations can be observed: IC 10th and 90th percentiles represent the absolute IC values when IC over the 1 s-long binaural signal is low and high, respectively; the difference between IC 10th and 90th percentiles indicates the size of IC temporal fluctuations.

Figure 3.7 shows the mean IC 10th and 90th percentiles for all experimental conditions. The mean IC 90th percentiles are displayed additionally to show their magnified details. Two black arrows are plotted in the left panel as an example to demonstrate the size of IC temporal fluctuations (the difference

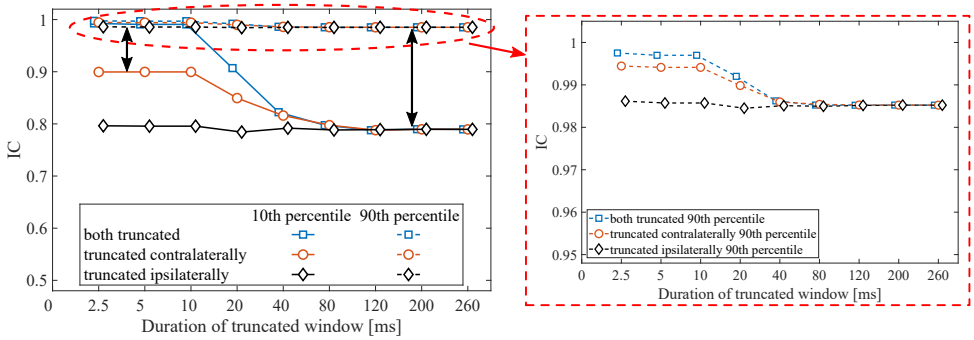


Figure 3.7: Average IC 10th and 90th percentiles across frequency channels for the "both truncated" (squares), "truncated contralaterally" (circles) and "truncated ipsilaterally" (diamonds) conditions (left panel). The mean IC 90th percentiles for all conditions are displayed additionally in the right panel to show their magnified details. Two black arrows are plotted in the left panel as an example to show the size of IC temporal fluctuations for short and large window durations under the "truncated contralaterally" condition.

between the mean IC 10th and 90th percentiles) for short and large window durations under the "truncated contralaterally" condition.

For the "truncated ipsilaterally" condition, the average IC values are nearly constant (0.8 for the IC 10th percentile, and 0.985 for the IC 90th percentile) across different truncated window durations, and the mean IC values are lower than for the other two conditions when the window durations are below 80 ms. The change in IC 10th and 90th percentiles is similar for "both truncated" and "truncated contralaterally" conditions across window durations. A large decrease in the mean IC 10th and 90th percentiles is observed for truncated window durations from 10 ms and 80 ms. Additionally, the mean IC 10th percentile decreases more than the mean IC 90th percentile, resulting in an increased extent of IC temporal fluctuations. For window durations shorter than 40 ms, the mean IC 10th and 90th percentiles are smaller and the size of IC temporal fluctuations is larger for the "truncated contralaterally" condition than for the "both truncated". Overall, the change in mean IC values and IC temporal fluctuations across window durations are well reflected in the experimental data.

3.5 DISCUSSION

As shown in Figure 3.1, for a lateral sound source, the effect of reverberation on the magnitude spectra of BRIRs is not identical for the left and right ear, with FFV higher and DRR lower at the contralateral ear than the ipsilateral ear. Because FFV and DRR are relevant cues in the context of distance perception, it is hypothesized that reverberation at each ear does not have the same effect on externalization of a lateral sound source.

A listening experiment has been performed to study the relative influence of reverberation at the contralateral versus ipsilateral ear on externalization. The experimental results show that perceived externalization is degraded when reverberation is reduced in both ears ("both truncated" condition) or only in the contralateral ear ("truncated contralaterally" condition), whereas externalization ratings are not substantially affected by reducing reverberation in the ipsilateral ear ("truncated ipsilaterally" condition). Various acoustic cues are extracted from BRIRs (DRR and FFV) and binaural signals (IC and temporal fluctuations of ILD and IC) and compared with the subjective data. Both subjective and objective data indicate that reverberation presented in the contralateral ear signal has a greater effect on perceived externalization than that presented in the ipsilateral ear signal.

3.5.1 *The relation between acoustic parameters and perceptual data*

Increasing or decreasing the reverberant energy leads to a change in DRR. When no reverberation is present in the ear signals, DRR has a value of infinity. On the other hand, if the reverberant energy is higher than the direct sound energy, DRR has a negative value. For truncated window durations above 10 ms, the change in DRR of the contralateral ear corresponds well to the externalization ratings. However, a noticeable change in contralateral DRR for window durations between 2.5 ms and 10 ms is not observed from externalization ratings ("both truncated" and "truncated contralaterally" conditions).

FFV increases with increasing reverberant energy, and a saturation of the FFV is apparent as the reverberant energy increases further (0.6 dB/Hz in this study). If no reverberation is present in the ear signals, the value of FFV is about 0 dB/Hz. It can be seen that FFV in the ipsilateral ear is lower than that in the contralateral ear. Moreover, the change in ipsilateral FFV is not as pronounced as that in contralateral FFV over truncated window durations. As mentioned above, the change in DRR does not correspond to the externalization ratings obtained for window durations between 2.5 ms and 40 ms. In contrast, the change in FFV is well reflected in the externalization results in this region. Similar to the impact of DRR on perceived externalization, the change in FFV in the ipsilateral ear signal does not noticeably affect externalization.

The reverberation-related binaural cues (ILD TSDs, and IC 10th and 90th percentiles) are calculated from binaural signals to compare to the experimental results. Higher ILD TSD values are observed for the "truncated ipsilaterally" condition compared to the "both truncated" and "truncated contralaterally" conditions for window durations up to 80 ms. For longer window durations (above 80 ms), the ILD TSDs are almost the same for these three conditions.

Though the IC 90th percentile is overall high for all experimental conditions (around 0.99), the change in IC 90th percentile can still be observed. Similar to the observation in ILD TSDs, lower IC 10th and 90th percentiles, and higher IC temporal fluctuations are found for the "truncated ipsilaterally" condition compared to the "both truncated" and "truncated contralaterally" conditions for window durations up to 80 ms. Overall, the ILD TSDs, IC 10th and 90th percentiles, and the size of IC temporal fluctuations are highly consistent with the externalization results for different experimental conditions. These analyses suggest that manipulating the reverberant energy in the contralateral ear has a stronger effect on the change in reverberation-related binaural cues than that in the ipsilateral ear, and these binaural cues can be used for indicating perceived externalization.

3.5.2 *Externalization model based on acoustic parameters*

Through the direct comparison of acoustic cues with perceptual data obtained in the listening experiments, the change in DRR and FFV in the contralateral ear signal, ILD TSDs, IC 10th and 90th percentiles, and IC temporal fluctuations (difference between IC 10th and 90th percentiles) corresponds well to the change in externalization for most experimental conditions. To evaluate whether or not these objective data can be used to predict externalization results, the deviations of these measured parameters between the target (modified signal) and the template (reference signal) are mapped onto the externalization rating (E), and the mapping is realized through an exponential function according to Hassager et al. [15]:

$$E = ae^{b\cdot\Delta m} + c, \tag{3.7}$$

where a, b and c are the mapping parameters, and Δm represents the normalized variation of each measured parameter between the target and the template signal:

$$\Delta m = \frac{|m_{signal} - m_{ref}|}{m_{ref}}, \tag{3.8}$$

where m_{signal} and m_{ref} denote the measured parameters of the modified signal and the reference signal, respectively. If Δm is zero, the externalization result of the modified signal is the same as that of the reference signal. Therefore, the sum of the parameter a and c should be equal to 2.6 (median externalization rating of the reference signal generated by unprocessed BRIRs), and the mapping function (Equation 3.7) can be rewritten as:

$$E = ae^{b\cdot\Delta m} + 2.6 - a. \tag{3.9}$$

The median externalization ratings in the "both truncated" condition were used to fit the mapping parameters a and b for each acoustic cue using the least-squares method, and the median externalization results from the other two conditions ("truncated contralaterally" and "truncated ipsilaterally" conditions) were applied for testing the model.

Table 3.2 shows the mapping parameters calculated for different acoustic cues. Since the shapes of IC 10th and 90th percentiles are similar across truncated window durations (c.f. Figure 3.7), only the IC 10th percentile was applied for the externalization model to reduce the amount of data. To calculate the mapping parameters for DRR, the subjective data for the 2.5 ms window duration was ignored due to the infinite value of DRR.

Table 3.2: Mapping parameters for different acoustic cues.

	a	b
Contralateral DRR	2.1	-0.7
Contralateral FFV	8.4	-0.3
ILD temporal fluctuations	2.3	-2.6
IC 10th percentile	5.1	-2.0
IC temporal fluctuations	2.5	-1.5

Figure 3.8 shows the perceptual data and the simulated results (open and filled symbols for mapped and predicted results, respectively). Overall, the predicted data matches well with the perceptual data obtained in most experimental conditions. Although the change in DRR for the contralateral ear does not correspond well to the change in externalization ratings for window durations between 2.5 ms and 10 ms, the DRR-based simulation results agree well with the externalization results over truncated window durations (convergence of the mapping function). Some slight deviations can be observed between the measured externalization ratings and the IC-based (IC 10th percentile and IC fluctuations) simulation results in "truncated contralaterally" and "truncated ipsilaterally" conditions, but the deviations are always within one level of the externalization rating.

To quantify how well the predicted data match the perceptual data, the Normalized Root Mean Square Deviation (NRMSD) between the predicted and observed values is calculated (in percentage):

$$\mathrm{NRMSD} = 100 \times \frac{1}{O_{range}} \sqrt{\frac{\sum_{i=1}^{N} (O_i - P_i)^2}{N}}, \tag{3.10}$$

where P_i and O_i represent the simulated and the observed values for each truncated window duration, respectively. N denotes the number of truncated window durations for each condition. O_{range} is a normalization factor and is

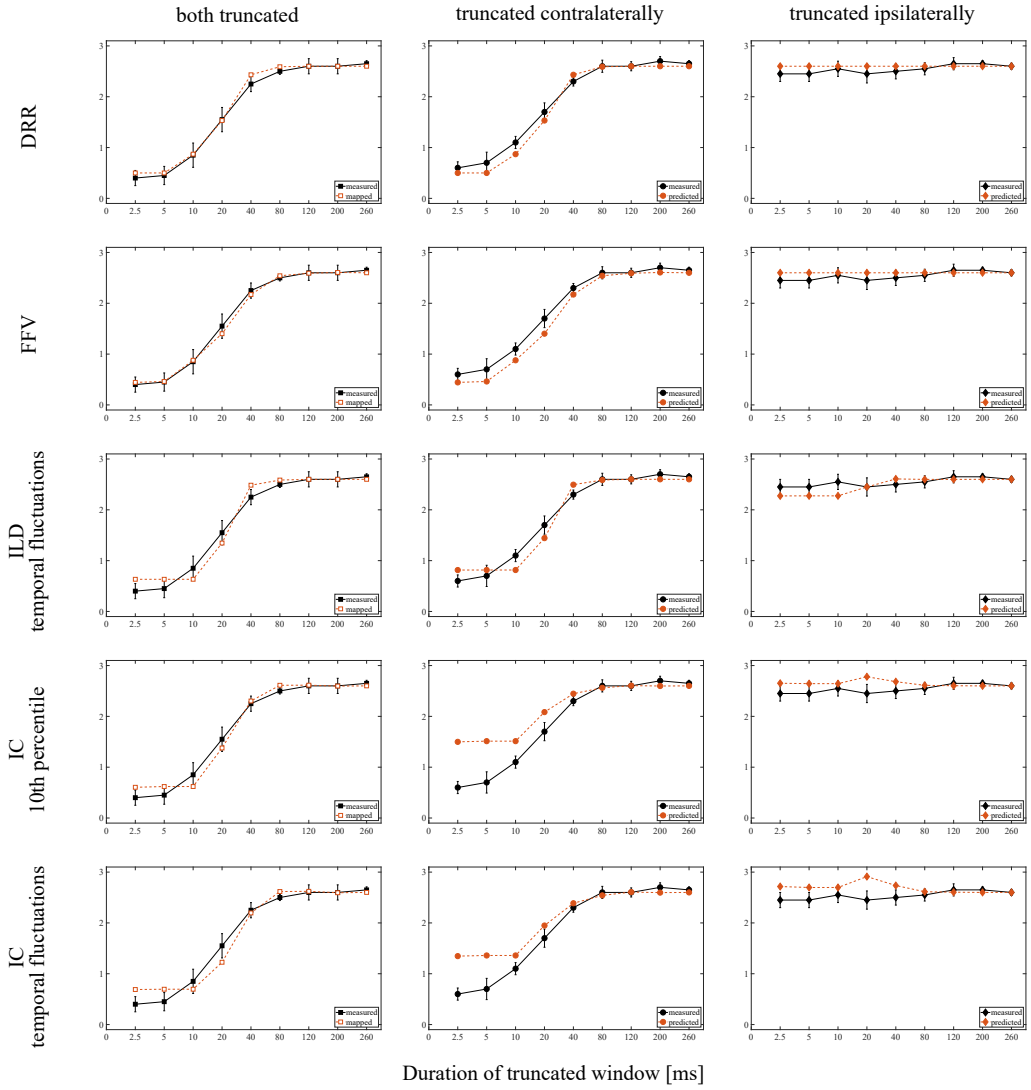


Figure 3.8: Measured (solid lines) and predicted (dashed lines, open and filled symbols for mapped and predicted results, respectively) median externalization ratings for different experimental conditions. Rows represent the predicted results calculated with different model parameters. From top to bottom: contralateral DRR, contralateral FFV, ILD temporal fluctuations, IC 10th percentile and IC temporal fluctuations. The left, middle and right column represent the "both truncated", "truncated contralaterally" and "truncated ipsilaterally" condition, respectively.

defined as the maximal range of the externalization ratings in the experiment, i.e., O_{range} is equal to 3.

The NRMSD results (see Table 3.3) show a high prediction accuracy of the externalization model based on different acoustic cues. The prediction errors are overall smaller than 10 % of the rating range in most experimental conditions. Only for the "truncated contralaterally" condition, about 12-15 % prediction error is accounted for the IC-based externalization model, in line with the observations in Figure 3.8.

Table 3.3: Normalized root mean square deviation (NRMSD) between the predicted and perceptual data.

	both truncated	truncated contralaterally	truncated ipsilaterally
Contralateral DRR	3 %	5 %	3 %
Contralateral FFV	2 %	6 %	3 %
ILD temporal fluctuation	5 %	6 %	4 %
IC 10th percentile	5 %	15 %	5 %
IC temporal fluctuation	6 %	12 %	7 %

3.5.3 *The effect of lateralized reverberation on externalization for different source directions*

Both subjective and objective results show that reverberation at the contralateral ear has more influence on externalization of a -45° sound source than that of the ipsilateral ear. However, it is unknown how this effect changes as a sound source moves from lateral to frontal directions. Therefore, we have further investigated the influence of lateralized reverberation on externalization for different source directions.

BRIR measurement and modification

Seven pairs of BRIRs were measured with the dummy head KEMAR at azimuth angles of −90°, −60°, −30°, 0°, 30°, 60°, and 90° in the listening room, and the distance between each loudspeaker and the KEMAR was 1.9 m (see Figure 3.9). The measurement procedure was the same as described in Section 3.2.

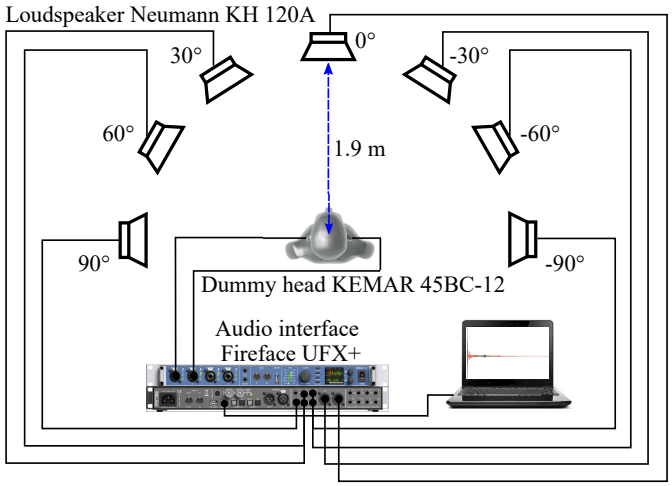


Figure 3.9: Illustration of the setup for the measurement of BRIRs.

In order to reduce the amount of data to be tested, the reverberant part was "maximally" (duration of truncated window = 2.5 ms) removed in (i) BRIRs of the left ear (“RL” condition), (ii) BRIRs of the right ear (“RR” condition),

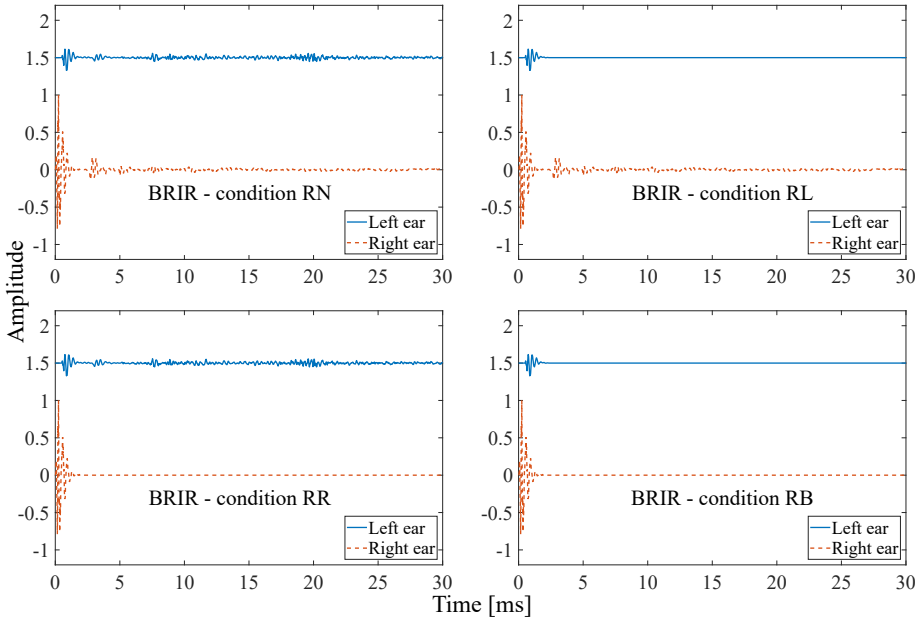


Figure 3.10: BRIRs with modified reverberant parts at an azimuth angle of -60° for the "RN" (top left), "RL" (top right), "RR" (bottom left), and "RB" (bottom right) conditions in the time domain. The solid and dashed lines represent the modified BRIRs of the left and right ear, respectively (Left ear BRIR is offset by 1.5 for better visibility).

and (iii) BRIRs of both ears ("RB" condition). The binaural signals generated with unprocessed BRIRs were considered as reference signals ("RN" condition). Note that for sound sources located at azimuth angles of $-30°$, $-60°$, $-90°$, the left ear is the contralateral ear, and the right ear is the ipsilateral ear. On the contrary, for 30°, 60°, and 90° sound sources, the left and right ears are the ipsilateral and contralateral ears, respectively. In the case of a frontal sound source (0°), both ears are facing the loudspeaker, and they are neither ipsilateral nor contralateral ears.

Figure 3.10 shows the modified BRIRs at an azimuth angle of -60° for the "RN", "RL", "RR", and "RB" conditions in the time domain. For better visibility, only the first 30 ms impulse responses are displayed, and an offset (amplitude of 1.5) is added to the BRIRs of the left ear.

Experimental paradigm

Seven subjects (two females and five males, aged between 26 and 31) participated in the listening experiment. Each subject sat in a chair, and listened to the test signals presented over a pair of compensated headphones (Sennheiser HD800). The rating scale for assessing the degree of externalization was the same as that used in the previous experiment (see Table 3.1). As shown in Figure 3.11, seven loudspeakers were placed at the measurement positions ($-90°$,

−60°, −30°, 0°, 30°, 60°, and 90° azimuth angles relative to the listener) to serve as visual cues. Subjects rated perceived externalization for each stimulus according to Table 3.1. In the experiment, four audio sequences generated by modified BRIRs, i.e., "RN", "RL", "RR", and "RB" conditions, were to be evaluated for each source direction. The experiment was repeated once. The stimulus used in the listening experiments was a speech sentence with a length of 1.3 s taken from the European Broadcasting Union (EBU) Sound Quality Assessment Material (SQAM) [116]. The audio signals generated with unprocessed BRIRs were presented at a level of 64 dBA over headphones. During the listening test, listeners were able to repeat every sequence. In addition, they could listen to the original stimulus played back through loudspeakers, and were informed that such stimulus should act as a well externalized sound (externalization rating = 3). Other experimental paradigms were the same as described in Section 3.3.

Figure 3.11: A photograph of the experimental setup. Seven loudspeakers are positioned at −90°, −60°, −30°, 0°, 30°, 60°, and 90° relative to the subject with a distance of 1.9 m.

Experimental results

Figure 3.12 shows the median externalization ratings with non-parametric 95 % CIs (notch-edges) of the test stimuli at different azimuth angles under the "RN", "RL", "RR", and "RB" conditions. Overall, the median externalization ratings for the unprocessed BRIRs ("RN" condition) are high across different source directions, and increase as the sound source moves from frontal to lateral directions, which is consistent with the outcomes of previous studies [117, 118]. A Friedman test was conducted to the experimental data at each azimuth angle and confirmed the main effect of window truncation on the externalization results ($p \ll 0.05$). Wilcoxon tests (5% significance level with Bonferroni adjustment) were further performed to compare externalization ratings between experimental conditions at each azimuth angle.

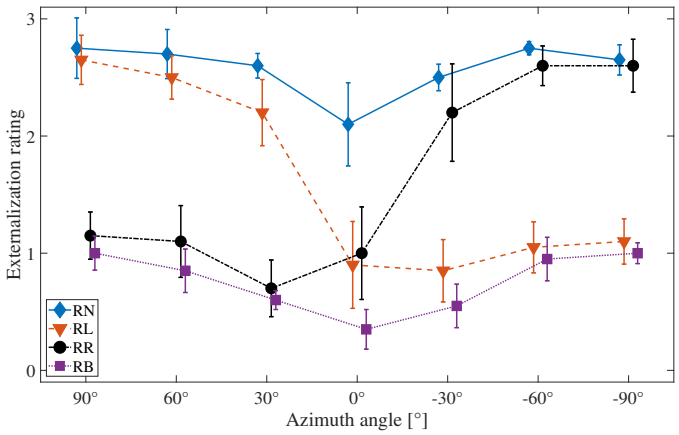


Figure 3.12: Median externalization ratings with non-parametric 95 % CIs (notch-edges) across subjects for sound sources at different azimuth angles under the "RN" (diamonds), "RL" (triangles), "RR" (circles), and "RB" (squares) conditions.

The externalization ratings reduce significantly by removing the reverberant parts in BRIRs of both ears ("RB" condition) for all azimuth angles ($p \ll 0.05$). For the "RB" condition, the median externalization rating increases as the sound source moves from frontal to lateral directions, but they are overall lower than 1, corresponding to sound sources being internalized. For a frontal sound source, the externalization ratings are almost the same for "RL" and "RR" conditions ($p = 0.13$) due to the symmetry of the human head. In the case of sound sources located at the front left (positive azimuth angles), the externalization ratings are significantly higher for the "RL" condition than for "RB" and "RR" conditions ($p \ll 0.05$). In addition, the ratings are slightly lower for the "RL" condition than for the "RN" condition ($p > 0.06$) for azimuth angles larger than 30°. For sound sources located on the front right side and larger than -30°, the externalization ratings are significantly higher for the "RR" condition than for the "RB" and "RL" conditions ($p \ll 0.05$), and slightly lower than for the "RN" condition ($p > 0.06$).

A speech signal was used in this experiment, which resembled a real life source. It seems that the effect of lateralized reverberation on externalization does not depend on the stimulus type, as similar results are observed for the speech and noise signals. Furthermore, the results show that the contribution of reverberation at the contralateral ear to externalization increases as the sound source moves laterally. For azimuth angles greater than $\pm$ 30°, reverberation at the ipsilateral ear has only a little contribution to externalization of sound images, and reverberation at the contralateral ear dominates the assessment of perceived externalization.

Recently, Băcilă and Lee [119] hypothesized that perceived reverberation loudness (PRL) would be greater at the contralateral ear than at the ipsilateral ear due to the binaural unmasking effect [120, 121], which might lead to differences in the perceived magnitudes of PRL with different head orientations. This hypothesis implies that reverberation received by the contralateral ear contributes more to PRL than that received by the ipsilateral ear. Since reverberation is highly related to perceived externalization, this analysis may explain the effect of lateralized reverberation on externalization. This assumption should be further objectively verified by analyzing sounds with a binaural model, taking into account the binaural unmasking effect [122].

3.6 CONCLUDING REMARKS

Reverberation is relevant to perceived externalization of headphone-based 3D audio. This study has investigated the role of reverberation in contralateral versus ipsilateral ear signals on externalization of a lateral sound source.

The experimental results demonstrate that reverberation at the contralateral ear has more influence on externalization than that at the ipsilateral ear. DRR, FFV and reverberation-related short-term binaural cues (ILD and IC) have been analyzed, and compared to the perceptual data for each experimental condition. The results illustrate that the ILD temporal fluctuations, IC temporal fluctuations, and IC values change noticeably when the amount of reverberation is changed at both ears or only at the contralateral ear. In contrast, these cues do not change substantially and are almost constant when the reverberant energy is changed only at the ipsilateral ear. This means that reverberation at the contralateral ear has more influence on the reverberation-related binaural cues and thus on perceived externalization. In addition, the change in contralateral DRR and FFV corresponds well to the change in externalization ratings over truncated window durations. A mapping function is applied for each measured parameter to build a model for predicting externalization ratings. The simulation results suggest that perceived externalization can be well predicted based on the deviations of these measured parameters from the reference signal.

A similar experiment has been performed by scaling (increasing or decreasing) the reverberant energy on each ear (see Experiment II in [109]). The results obtained in that experiment are consistent with the findings presented in this Chapter, that modifying (increasing or decreasing) the reverberant energy in the contralateral ear signal has more influence on externalization than that in the ipsilateral ear signal. The perceptual results can also be well predicted by using models based on different acoustic cues, and the NRMSD values are smaller than 8 % for all experimental conditions.

The effect of lateralized reverberation on externalization has been studied for sound sources at different azimuth angles in the horizontal plane. The experimental results demonstrate that the contribution of reverberation at the contralateral ear to perceived externalization increases as the source moves laterally. For azimuth angles larger than $\pm$ 30°, reverberation at the ipsilateral ear has no appreciable contributions to externalization of sound images.

The finding in this study can be used in the design of binaural rendering systems. In the case of lateral sound sources (azimuth angles greater than $\pm$ 30°), the amount of reverberation at the ipsilateral ear can be reduced appropriately to reduce the computational complexity, taking into account the perceptual attributes such as listener envelopment, localization accuracy, naturalness.

Future work is to investigate this effect in different listening environments and with different source-listener distances. In addition, the perceptual consequences on sound coloration, perceived naturalness, auditory source width and listener envelopment of the audio signal caused by the removal of lateralized reverberation would have to be investigated.

4

MODELING PERCEIVED EXTERNALIZATION OF A STATIC, LATERAL SOUND SOURCE

4.1 INTRODUCTION

Some recent studies have shown that the short-term binaural cues contained in the reverberant parts, and the spectral details of the direct sound components are relevant to perceived externalization [14, 15, 18–20]. Hassager et al. [15] developed an externalization model based on the deviations of ILDs from the reference signal, and it was suitable for explaining the externalization results obtained in their experiments. Unfortunately, they did not consider other auditory cues related to externalization even though it is known that the correct ILD alone is not enough to externalize virtual sound sources [12]. Li et al. [109] contrasted various reverberation-related binaural cues including IC, ILD and IC temporal fluctuations to predict externalization of reverberant sound images. The model based on ILD temporal fluctuations showed better performance in terms of prediction accuracy compared to the one based on IC and IC temporal fluctuations. Although DRR and FFV could be used to predict externalization in [109], some other research showed that DRR and FFV were only partially related to externalization data obtained in their studies [15, 18, 20]. Baumgartner and Majdak [36] compared the performance of models based on various acoustic cues regarding externalization prediction. The results revealed that the monaural spectral cues represented by SGs were relevant for predicting externalization of anechoic sounds (c.f. Section 1.2).

The present study aims to build a quantitative model to explain the interplay of important acoustic cues in perceived externalization based on a template-matching procedure [36] (c.f. Figure 1.3). The proposed model extends the above-mentioned approaches [15, 36, 109] by incorporating all relevant acoustic cues including ILDs, monaural spectral cues, and ILD temporal fluctuations to jointly predict externalization of anechoic and reverberant lateral sounds. The perceptual weights of the different acoustic cues and the binaural weighting

between the ipsi- and contralaterally processed cues are derived from a variety of experiments. Parts of this Chapter have been published in [123].

This Chapter is organized as follows. Section 4.2 describes the proposed externalization model. In Section 4.3, five experiments are designed to study the role of ILD, ILD temporal fluctuations, and monaural spectral cues on externalization and to parameterize the model. The calculation of the weighting factors for the model parameters and the prediction results are presented in Section 4.4. The experimental results and the model components are discussed in Section 4.5. At last, Section 4.6 concludes this study.

4.2 EXTERNALIZATION MODEL

4.2.1 *Concept and overview*

Figure 4.1 shows the structure of the proposed externalization model according to Plenge [34], consisting of a long-term and a short-term memory. A virtual sound image is expected to be perceived as well externalized if the target sounds provide auditory cues similar to those stored in long- and short-term memory (c.f. Section 1.2).

Our proposed model is based on a template-matching procedure [36]. The "template" signal represents the internal representation of a perfectly externalized sound source, which is synthesized by convolving the individual BRIRs with an input signal. The "target" signal represents the sound reproduced through headphones and heard by listeners. If there is no deviation between the acoustic cues provided by the "target" signal and the "template" signal, the sound should be perceived as well externalized. In contrast, the externalization of sound source is degraded if there is a mismatch between the acoustic cues contained in the "target" and the "template" signal.

As introduced in Section 1.2, an adaptation to HRTF-related acoustic cues takes a long time, while the adaptation of reverberation-related auditory information is relatively fast and is required every time the listening environment changes. In this model, the SGs and ILDs are used to represent the acoustic cues extracted from HRTFs, which are stored in the long-term memory. The ILD temporal fluctuations are considered as the reverberation-related acoustic cues stored in the short-term memory. To compare the information stored in the long-term memory, the direct sound component is first extracted from binaural signals. In the present study, the direct sound is simulated by convolving the direct part of BRIRs with the input signal. Subsequently, the obtained direct sound is processed through the auditory periphery model to approximate the

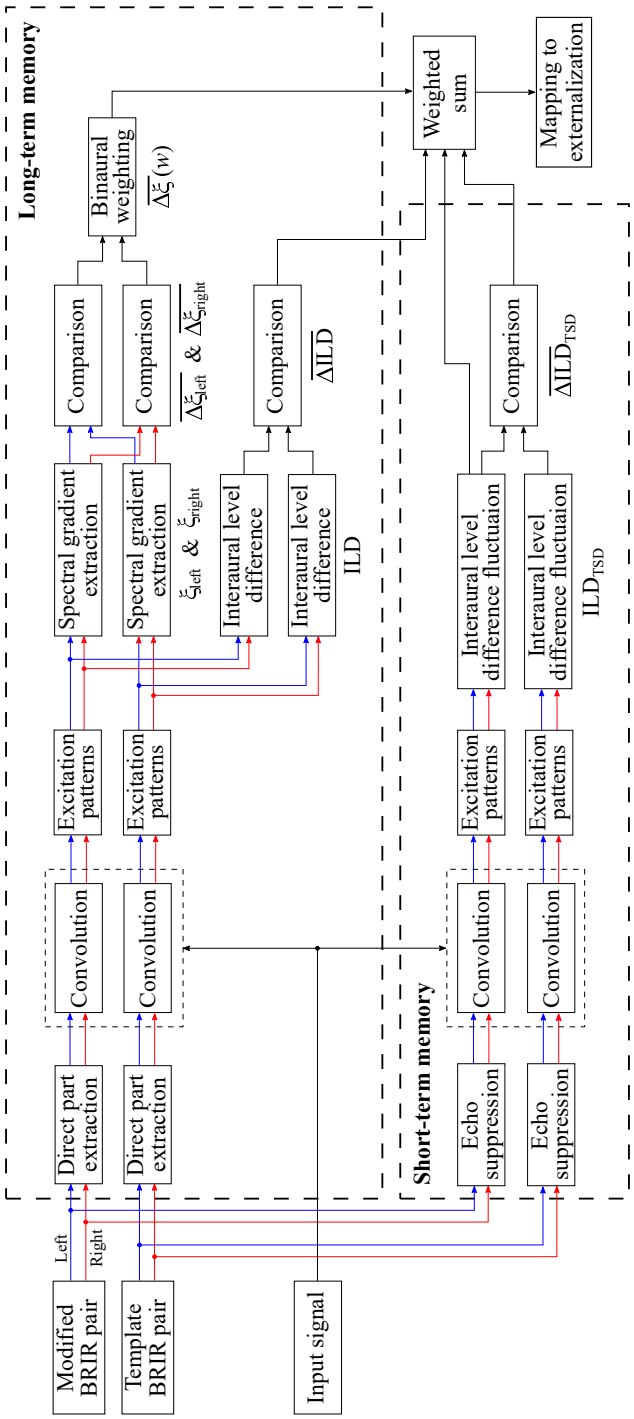


Figure 4.1: Structure of the proposed externalization model, consisting of a short-term and a long-term memory. In the long-term memory, SGs and ILDs are extracted from the direct sound components in each frequency channel of a gammatone filter bank. In the short-term memory, ILD TSDs are calculated from the echo-suppressed reverberant signals in each frequency channel. The deviations of these three acoustic cues from the template signals are summed up with different weighting factors and mapped to perceived externalization of virtual sound images.

cochlear filtering procedure as described in Section 3.4.2. Afterwards, the SGs and ILDs of the target signal are extracted from the "excitation patterns" in each frequency band and compared with those stored in the long-term memory ("template"). The deviations are weighted based on the relevance of the SGs and ILDs on perceived externalization in different listening environments. For the comparison of information in the short-term memory, the ILD temporal fluctuations are extracted from the echo-suppressed reverberant sounds and compared with those stored in the short-term memory. The deviations of these three acoustic cues from the template are summed up with different weightings and mapped to externalization results.

4.2.2 *Processing stages*

SG comparison

SG is applied to characterize the monaural spectral information of the sound source in each ear. It is defined as the excitation differences (in dB) between each pair of adjacent frequency channels ($\xi_k(i)$) [36]:

$$\xi_k(i) = M_k(f_{c,i}) - M_k(f_{c,i-1}), \text{ for } i = 2, 3, \ldots N, \tag{4.1}$$

where the index k stands for the left or right ear ($k \in \{\text{left}, \text{right}\}$). $M_k(f_{c,i})$ represents the excitation in the i^{th} frequency channel centered at $f_{c,i}$, and N is the number of frequency channels allocated from 200 Hz to 16 kHz (the frequency range of the stimulus). The absolute differences in SGs between the "target" and "template" signals are averaged over the frequency bands, and further normalized by dividing the average SG of the template signal:

$$\overline{\Delta\xi}_k = \frac{\sum_{i=2}^{N} |\xi_{k,\text{target}}(i) - \xi_{k,\text{template}}(i)|}{\sum_{i=2}^{N} |\xi_{k,\text{template}}(i)|}, \tag{4.2}$$

where $\xi_{k,\text{target}}(i)$ and $\xi_{k,\text{template}}(i)$ represent the SGs of the target and the template signal in the i^{th} frequency channel, respectively. It should be noted that the normalization procedure for the SG deviation is not performed in each frequency band, since the zero gradient points are arbitrarily distributed over frequencies and we do not want to introduce any bias by normalization. After that, the normalized SG deviation calculated for the left ($\overline{\Delta\xi}_{\text{left}}$) and right ear ($\overline{\Delta\xi}_{\text{right}}$) is weighted with a binaural weighting factor, w ($0 < w < 1$), which determines the relative contribution of the monaural spectral information of the left and right ear to externalization:

$$\overline{\Delta\xi}(w) = w\,\overline{\Delta\xi}_{\text{left}} + (1-w)\,\overline{\Delta\xi}_{\text{right}}. \tag{4.3}$$

ILD comparison

In each frequency band, the ILD is calculated as the difference between the left and right ear excitation. Since ILDs are naturally larger at high frequencies [4], the same absolute ILD deviation results in a smaller relative ILD change when applied to a high frequency versus a low frequency sound. Therefore, in the comparison phase, unlike the normalization procedure used for the SG deviations (c.f. Equation 4.2), the absolute ILD deviations are first normalized within each frequency band and then averaged over the frequency bands [15]:

$$\overline{\Delta \mathrm{ILD}} = \frac{1}{N} \sum_{i=1}^{N} \frac{|\mathrm{ILD}_{\mathrm{target}}\,(f_{c,i}) - \mathrm{ILD}_{\mathrm{template}}\,(f_{c,i})|}{|\mathrm{ILD}_{\mathrm{template}}\,(f_{c,i})|}, \tag{4.4}$$

where $\mathrm{ILD}_{\mathrm{target}}\,(f_{c,i})$ and $\mathrm{ILD}_{\mathrm{template}}\,(f_{c,i})$ denote the ILD of the target and the template signal centered at $f_{c,i}$, respectively.

Temporal fluctuation comparison

ILD temporal fluctuations are used to represent reverberation-related auditory cues and can be expressed as the standard deviation of short-term ILDs collected over the binaural signal (ILD TSD) according to Equation 3.4 (c.f. Section 3.4.2). However, under anechoic conditions, random fluctuations in the sound source cause the ILD TSDs to be slightly larger than zero. Consequently, even small changes in ILD TSDs caused by HRTF modifications can result in large undesirable relative deviations. For this reason, the absolute ILD TSD deviation between the target and the template signal is calculated in the comparison phase. In order to maintain consistency of the unit-free deviation metrics, the absolute ILD TSD deviation is divided by the average ILD TSD over frequencies of a reference acoustic environment ($\overline{\mathrm{ILD}}_{\mathrm{TSD,\ reference}}$):

$$\overline{\Delta \mathrm{ILD}}_{\mathrm{TSD}} = \frac{\sum_{i=1}^{N} \left|\mathrm{ILD}_{\mathrm{TSD,target}}(f_{c,i}) - \mathrm{ILD}_{\mathrm{TSD,template}}(f_{c,i})\right|}{\overline{\mathrm{ILD}}_{\mathrm{TSD,\ reference}}}, \tag{4.5}$$

where $\mathrm{ILD}_{\mathrm{TSD,target}}(f_{c,i})$ and $\mathrm{ILD}_{\mathrm{TSD,template}}(f_{c,i})$ are the ILD TSD of the target and the template signal centered at $f_{c,i}$, respectively. Note that $\overline{\mathrm{ILD}}_{\mathrm{TSD,\ reference}}$ is actually a scaling factor and the reference acoustic environment can be chosen arbitrarily. In this study, the listening room (see Experiment E) is considered as the reference acoustic environment.

Reverberation-related weighting

The impact of SGs and ILDs on perceived externalization is investigated mainly under anechoic conditions in this study (see Experiments A - D). Since reverberation reduces the relevance of spectral details [23], the influence of SGs and

ILDs on externalization in reverberant environments may be reduced compared to that in anechoic environments. To represent this effect with the externalization model, the comparison results from the long-term memory are weighted by a factor γ ($0 < \gamma < 1$) depending on the presented amount of reverberation-related features:

$$\gamma = 1 - b_\gamma \frac{\overline{\mathrm{ILD}}_{\mathrm{TSD,\ template}}}{\overline{\mathrm{ILD}}_{\mathrm{TSD,\ reference}}}, \tag{4.6}$$

where b_γ is a weighting parameter. $\overline{\mathrm{ILD}}_{\mathrm{TSD,\ template}}$ is the average ILD TSD over frequencies of the template signal, indicating the current acoustic environment. $\overline{\mathrm{ILD}}_{\mathrm{TSD,\ reference}}$ is the average ILD TSD of a reference acoustic environment as described above (c.f. Equation 4.5).

The factor γ is used to adjust the influence of SGs and ILDs on externalization according to the current acoustic environment ($\overline{\mathrm{ILD}}_{\mathrm{TSD,\ template}}$). γ is close to one and smaller than one for anechoic and reverberant conditions, respectively. Since γ is limited between 0 and 1 in this study, it is referred to as a reduction term.

Mapping to externalization

After the comparison stage in short- and long-term memory, the normalized deviations of SGs, ILDs and ILD TSDs between the target and the template signal are summed up with corresponding weighting factors [124]:

$$\Delta m = \gamma \left(b_\xi \overline{\Delta\xi}(w) + b_{\mathrm{ILD}} \overline{\Delta\mathrm{ILD}} \right) + b_{\mathrm{ILD\,TSD}} \overline{\Delta\mathrm{ILD}}_{\mathrm{TSD}}, \tag{4.7}$$

where b_ξ, b_{ILD} and $b_{\mathrm{ILD\,TSD}}$ are weighting factors for normalized deviations of SGs, ILDs and ILD TSDs, respectively. An exponential function is used to map the objective measures to the subjective externalization ratings [15, 109]:

$$E = a\, e^{-\Delta m} + c, \tag{4.8}$$

where a and c are mapping parameters. The derivation of weighting factors for model components is explained in section 4.4.

4.3 EXPERIMENTS

4.3.1 *General methods*

Measurement of individual impulse responses

Individual HRIRs and BRIRs were measured in an anechoic room (c.f. Section 2.4.3) and a listening room (c.f. Section 3.2), respectively. A loudspeaker

(Neumann KH 120A), which served as a sound source, was placed at an azimuth angle of 90° with a distance of 1.5 m from the subject. A pair of miniature microphones (Madness MM-BSM-8) was placed at the entrance of the subject's ear canals. A 5 s-long exponential sweep was applied as the excitation signal to measure individual impulse responses, and the measurement was repeated 10 times. After the HRIR measurement, the impulse responses were truncated by a 2.5 ms-long time window with a 0.5 ms-long half raised-cosine decay time (after the onset delay). Then, the HRIRs were equalized by a reference measurement in which the miniature microphones were placed at the position of the center of the subject's head without the subject being present [41]. After the BRIR measurement, the impulse responses were truncated by a 260 ms-long time window including a 10 ms-long half raised-cosine decay time (after the onset delay) [109].

Experimental paradigm

Five subjects (one female and four males, aged between 24 and 30) with normal hearing participated in the experiments. Five listening tests were designed to study how perceived externalization was influenced by (A) changes in ILDs, (B) changes in monaural spectral information while preserving original ILDs, (C) interaural reductions in spectral details, (D) deviations in both ILDs and spectral details, and (E) deviations in ILDs, spectral information, and reverberation.

Also, these experiments aimed to determine the model weightings for SGs, ILDs and ILD TSDs with respect to perceived externalization while attempting to isolate individual acoustic cues. Note that not all experiments/experimental conditions were designed to fit the model parameters, some were used only to validate the performance of the externalization model. The experimental results from Experiments A ("BB" condition, see Section 4.3.2), B (see Section 4.3.3), C (see Section 4.3.4), and E ("B = 0" and "0 % reverberation reduction" conditions, see Section 4.3.6) were used to calculate the weighting factors for acoustic cues, and the results from the rest conditions in Experiments A and E, as well as all conditions in Experiment D were used to validate the externalization model. In each experiment considered for fitting the model, only one acoustic cue changed substantially. By this means, the initial fitting process could be simplified since only one acoustic cue was changed and there was no need to consider the weightings for other acoustic cues (details can be found in Section 4.4.1). The mapped results are represent with unfilled symbols while the validated results are denoted with filled symbols (see Figures 4.2 to 4.7). The experiments (Experiment A, B, C and D) concerning the influence of spectral information and ILDs on externalization were performed in the anechoic chamber where individual HRTFs were measured before, and the experiment (Experiment E)

regarding the influence of reverberation on externalization was conducted in the listening room where individual BRIRs were previously measured.

The HRIRs and the direct parts of the BRIRs were represented as minimum-phase components, followed by all-pass filters [15, 125]. In this study, the magnitude spectra of the minimum-phase components were modified while retaining the phase information. The test signals used in the experiments were synthesized by convolving a 1 s-long white Gaussian noise (200 Hz–16 kHz) with modified HRIRs/BRIRs. The audio signals were presented via headphones at a SPL of about 67 dB.

Each listener sat in a chair, listened to the test stimuli with a pair of individually compensated headphones (Sennheiser HD800) according to Schärer and Lindau [112]. A rating scale with four externalization levels from 0 to 3 was used to evaluate externalization, which was the same as what we used in our previous study (see Table 3.1). Subjects were asked to rate test stimuli by using a slider with a step-size of 0.1 between 0 and 3, and to ignore audible artifacts that did not influence externalization perception. During each listening test, subjects were not allowed to turn their heads. Friedman test was performed to verify the significant effects of different experimental conditions on externalization results.

Before each listening test, subjects were asked to listen to all stimuli once to familiarize themselves with these test signals to be presented. Each experiment was performed four times and the stimuli were presented in random order. The loudspeaker was always present during all experiments. As anchors, subjects listened to the original sound played back over the loudspeaker and were informed that such stimulus should act as a well externalized sound (externalization rating = 3). Headphones should be taken off to listen to the speaker signal. In addition, diotic playback of the signal acted as a fully internalized sound (externalization rating = 0). Subjects were able to listen to the anchor sounds at any time during all experiments.

This study focuses on perceived externalization of a 90° virtual sound source. Such extreme lateralization allows the range of the binaural weighting factor w (c.f. Equation 4.3) applied to the monaural spectral information to be determined over different azimuth angles. w is assumed to be 0.5 for a frontal sound source due to the symmetry of the human head, and can be interpolated for other azimuth angles [126]. Additionally, a 90° sound source allows to manipulate broadband ILDs without affecting the source direction (c.f. Experiment A). Moreover, perceived externalization of a lateral sound source is hardly influenced by potential head movements, which could not be perfectly controlled during the experiment [21, 28].

4.3.2 *Experiment A: Influence of ILDs*

Experiment A was designed to study the role of ILDs on externalization and to obtain the weighting factor for the ILD while maximally isolating the changes in the HRTF magnitude spectra. Since it is not possible to manipulate the frequency-dependent ILDs (ILD contrasts over frequencies) without changing the magnitude spectra of HRTFs, the ILD information was modified by controlling the level of sounds delivered to each ear, i.e., manipulating the broadband ILDs [127]. By increasing the sound level at the contralateral ear, the generated sound image might be perceived as coming from a more central lateral direction, or listeners could even hear two split sound images due to the inconsistent ILD and ITD information. Therefore, the ILD was expanded by reducing the level of the signal delivered to the contralateral ear to keep the perceived lateralization of the sound source at 90°.

In this experiment, the sound level at the contralateral ear was decreased by 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB at the broadband ("BB" condition: 200 Hz - 16 kHz), at the low frequency range ("LO" condition: 200 Hz - 3 kHz), and at the high frequency range ("HI" condition: 3 kHz - 16 kHz), while the SPL of the ipsilateral (left) ear signal remained unchanged. The magnitude spectrum of the contralateral (right) ear HRTF, $|\mathrm{HRTF}_{\mathrm{right,\,mod}}(f)|$, was expressed as:

$$|\mathrm{HRTF}_{\mathrm{right,\,mod}}(f)| = \frac{|\mathrm{HRTF}_{\mathrm{right}}(f)|}{10^{\frac{A}{20}}}, \text{ with } f \in \{\mathrm{BB, LO, HI}\}, \tag{4.9}$$

where A denotes the attenuation level in dB.

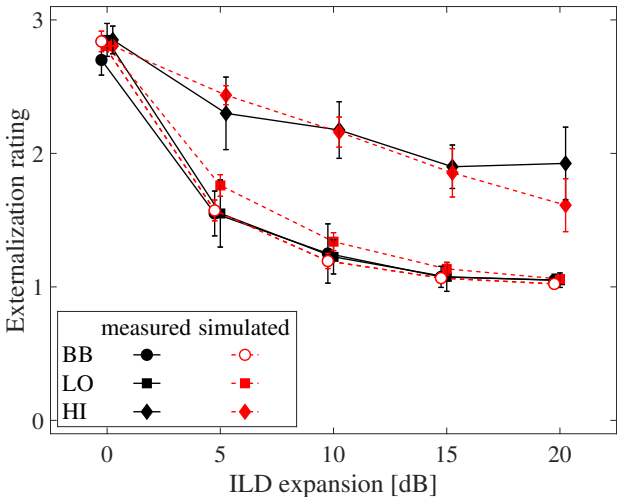


Figure 4.2: Median values of externalization ratings (solid lines) and model simulations (dashed lines, open and filled symbols for mapped and predicted results, respectively) with non-parametric 95 % CIs (notch-edges) for ILD expansions in three different frequency ranges ("BB", "LO" and "HI" conditions).

Figure 4.2 shows the median externalization ratings for ILD expansions in three frequency ranges. For visual comparison, the simulated externalization results based on the proposed model (see Section 4.4) are also presented. Overall, the externalization ratings decrease with increasing ILD expansions for all three conditions ("BB", "LO" and "HI"). Both the ILD expansions ($\chi^2(4)$ =52.5, $p \ll 0.05$) and the manipulated frequency ranges ($\chi^2(2)$ =32.1, $p \ll 0.05$) have significant effects on the externalization results. The externalization ratings decrease substantially by an ILD extension of 5 dB for the "BB" and "LO" conditions. The sound image is perceived as being at the ear by further increasing the ILD (externalization ratings ≈ 1), and there are no significant differences in the externalization ratings between "BB" and "LO" conditions ($\chi^2(1) = 0.89$, $p = 0.4$). In contrast, the reduction in externalization results is small when the ILD expands only at high frequencies. The sound image is still perceived as externalized even if the ILD increases by 20 dB at high frequencies.

4.3.3 *Experiment B: Influence of spectral details with unchanged ILDs*

Experiment B was designed to study the influence of spectral information of HRTFs on perceived externalization. The ILD remained unchanged across frequencies, while the magnitude spectrum of the ipsilateral (left) HRTF, $|\mathrm{HRTF}_{\mathrm{left}}(f)|$, was smoothed by using a gammatone filter according to Hassager et al. [15]:

$$|\mathrm{HRTF}_{\mathrm{left,\,mod}}(f_c)| = \sqrt{\frac{\int_0^\infty |\mathrm{HRTF}_{\mathrm{left}}(f)|^2\,|H(f,f_c)|^2\,df}{\int_0^\infty |H(f,f_c)|^2\,df}}\,, \tag{4.10}$$

where $|H(f,f_c)|$ denotes the spectral magnitude of a 4^{th} order gammatone filter centered at f_c with a bandwidth of $b(f_c)$ [128]:

$$|H(f,f_c)| = \left|\left(\frac{b(f_c)}{j(f-f_c)+b(f_c)}\right)^4\right|\,, \tag{4.11}$$

with

$$b(f_c) = B \times \frac{24.7\,(0.00437\times f_c)}{2\,\sqrt{2^{1/4}-1}} = 0.1241\;B\times f_c\,, \tag{4.12}$$

where j is the imaginary unit ($j = \sqrt{-1}$), and B represents the bandwidth factor relative to a value of one ERB. The smoothing level of the magnitude spectrum depends on the bandwidth of the gammatone filters, and the spectral magnitude of the ipsilateral HRTF was smoothed with a bandwidth factor $B \in \{0,1,4,16,64\}$ in this study. The magnitude spectrum of the contralateral HRTF was adjusted accordingly to preserve the original ILD. $B = 0$ indicated the unprocessed HRTF.

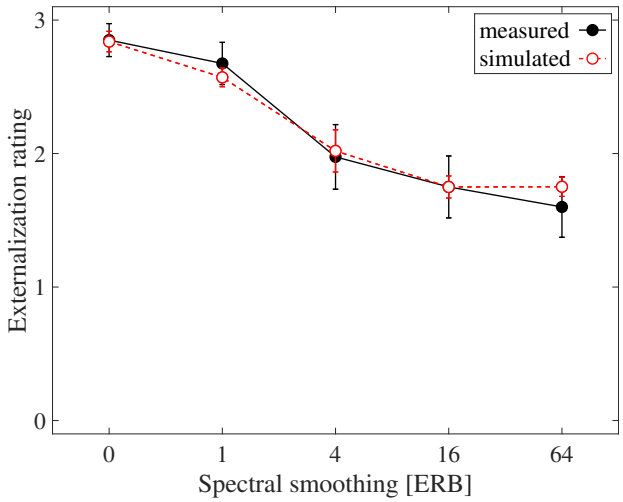


Figure 4.3: Median values of externalization ratings and mapped results for smoothed spectral magnitude in the HRTF of the ipsilateral ear while maintaining the original ILD. All other conventions are as in Figure 4.2.

Figure 4.3 shows the externalization ratings by spectrally smoothing the HRTF of the ipsilateral ear while retaining the ILD information. The externalization ratings decrease significantly for bandwidth factors larger than one ($\chi^2(4) = 19.3$, $p \ll 0.05$). The median value of the externalization rating is about 1.6 for the highest smoothing level ($B = 64$). The result illustrates that maintaining the correct ILD may be sufficient to externalize a lateral sound source, but not to externalize it well.

4.3.4 *Experiment C: Influence of interaural spectral details*

Experiment C aimed to investigate the role of interaural spectral information in HRTFs on perceived externalization. Unfortunately, it is not possible to modify the interaural spectral contrast while preserving the original HRTF magnitude spectra in both ears. The spectral details in HRTFs are more pronounced at high frequencies (above 3 kHz) than at low frequencies due to multiple reflections and diffractions of the pinnae, head and torso. Hence, in this experiment, the ILD spectral contrast was compressed to different degrees at high frequencies (between 3 kHz and 16 kHz), while the magnitude spectrum of the original HRTF was preserved in one ear. The ILD spectral contrast in dB was compressed with a compression factor $C \in \{0\,\%, 25\,\%, 50\,\%, 75\,\%, 100\,\%\}$ according to Baumgartner et al. [14]:

$$\mathrm{ILD}_{\mathrm{mod}}(f) = (1 - C)\ \mathrm{ILD}(f) + C\ \frac{1}{\sum_{k \in f} w(k)} \sum_{k \in f} w(k)\ \mathrm{ILD}(k), \tag{4.13}$$

where $w(k)$ represents a weighting factor that approximates the resolution of auditory filters by the across-frequency derivative of ERB frequencies [14].

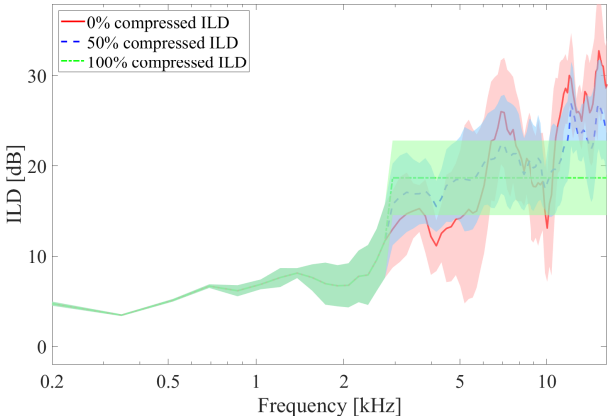


Figure 4.4: Median ILD across subjects with 0 % (red solid line), 50 % (blue dotted line) and 100 % (green dashed line) compression factors. Shaded areas denote non-parametric 95 % CIs (notch-edges) of the median values.

Figure 4.4 shows an example of the compressed ILDs averaged across subjects with C of 0 %, 50 % and 100 %. The ILD is unprocessed when C is equal to 0 %. For C equals to 100 %, the ILD is constant over frequencies between 3 kHz and 16 kHz. Two conditions were considered in this experiment: (i) The magnitude spectrum of the contralateral HRTF was changed according to the compressed ILD spectral contrast ("contra" condition). (ii) The magnitude spectrum of the ipsilateral HRTF was changed based on the compressed ILD spectral contrast ("ipsi" condition).

Figure 4.5 shows the externalization ratings by compressing the ILD spectral contrast at high frequencies while changing the spectral details in the ipsilateral (circles) or the contralateral ear (squares). Both the ILD contrast ($\chi^2(4) = 36.3$, $p \ll 0.05$) and the spectral information in the HRTF of one ear ($\chi^2(1) = 14.4$, $p \ll 0.05$) have significant effects on externalization ratings. The externalization ratings decrease noticeably for C above 25 %. Also, the degree of externalization reduces more for the "ipsi" condition than for the "contra" condition, indicating that the spectral information contained in the HRTF of the ipsilateral ear is more relevant to externalization than that of the contralateral ear.

4.3.5 *Experiment D: Influences of ILDs and spectral details*

Experiment D aimed to study the additive influences of monaural spectral information and ILDs on perceived externalization. For this purpose, the magnitude spectra were smoothed with a bandwidth factor $B \in \{0, 1, 4, 16, 64\}$, in HRTFs of (i) both ears ("bi" condition), (ii) the ipsilateral ear ("ipsi" condition) and (iii)

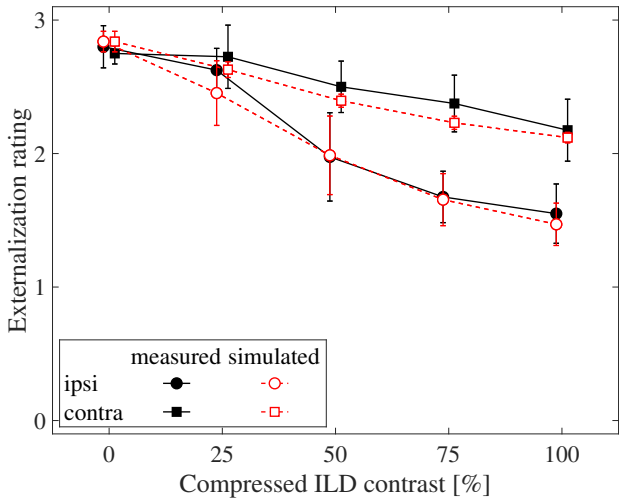


Figure 4.5: Median values of externalization ratings and the mapped results for compressed ILD contrasts with ipsilateral ("ipsi" condition) versus contralateral ("contra" condition) spectral distortions. All other conventions are as in Figure 4.2.

the contralateral ear ("contra" condition). The smoothing approach was the same as used in Experiment B, but unlike the manipulation in Experiment B, the ILD was not maintained by spectral compensation in the other ear.

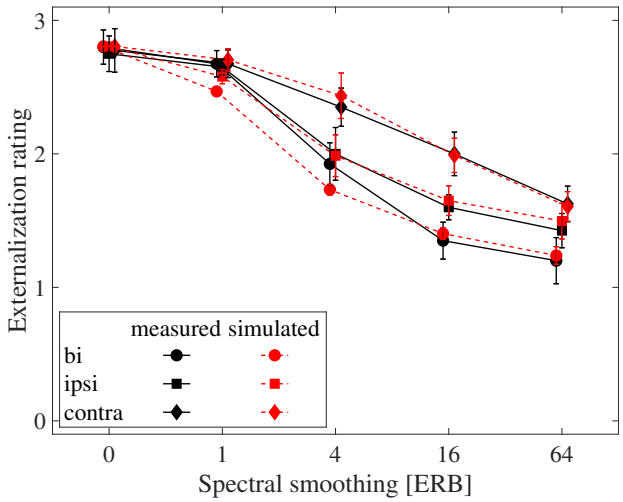


Figure 4.6: Median values of externalization ratings and predicted results by reducing spectral details in HRTFs of both ears ("bi" condition), the ipsilateral ear ("ipsi" condition) or the contralateral ear ("contra" condition). All other conventions are as in Figure 4.2.

Figure 4.6 illustrates the externalization ratings for "bi", "ipsi", and "contra" conditions. Both the spectral smoothing levels ($\chi^2(4) = 62.9$, $p \ll 0.05$) and the smoothing conditions ($\chi^2(2) = 15.8$, $p \ll 0.05$) have significant effects on externalization results. For bandwidth factors larger than one, the externalization ratings decrease noticeably for all three conditions. Furthermore, the external-

ization ratings for the "contra" condition reach a low degree of about 1.5 later than those for the "bi" and "ipsi" conditions. This result indicates that smoothing the magnitude spectrum of the HRTF at the ipsilateral ear is more effective than at the contralateral ear in affecting externalization.

4.3.6 *Experiment E: Influences of ILDs, spectral information and reverberation*

Experiment E was designed to study the influences of reverberation, ILDs and the monaural spectral cues on externalization. For this purpose, the magnitude spectra of direct parts in BRIRs ($\mathrm{BRIR}_{\mathrm{direct,\,mod}}(t)$) were bilaterally smoothed with a bandwidth factor $B \in \{0, 1, 4, 16, 64\}$, where the direct part was obtained with a 2.5 ms-long window including a 0.5 ms-long half raised-cosine decay time [109]. The reverberant part was extracted by subtracting the direct part from the BRIR. Also, the amount of reverberation was modified with a scaling factor $\alpha \in \{0\,\%, 25\,\%, 50\,\%, 75\,\%, 100\,\%\}$. The modified BRIR ($\mathrm{BRIR}_{\mathrm{mod}}(t)$) can be expressed as:

$$\mathrm{BRIR}_{\mathrm{mod}}(t) = \mathrm{BRIR}_{\mathrm{direct,\,mod}}(t) + (1-\alpha)\,\mathrm{BRIR}_{\mathrm{reverb}}(t). \tag{4.14}$$

For α equals to 0 %, the original reverberant component was present. On the contrary, for α equals to 100 %, only the direct sound component was present.

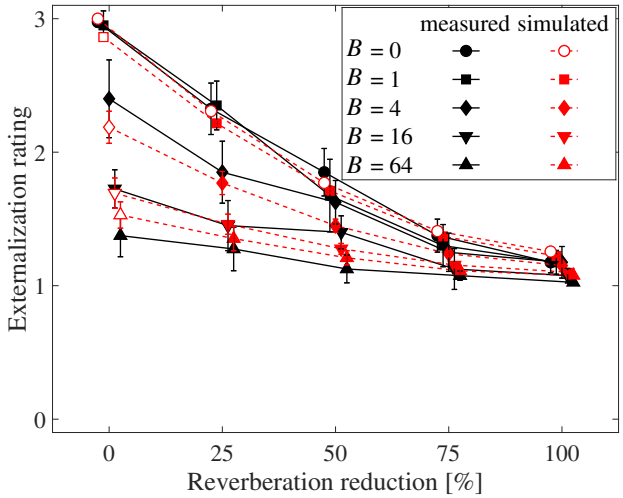


Figure 4.7: Median values of externalization ratings and the model predictions for different bilateral spectral smoothing (B) and reverberation reduction levels (α) across subjects. All other conventions are as in Figure 4.2. Note that the model outputs for the "α = 0" condition are mapped results.

Figure 4.7 shows the externalization ratings for different spectral smoothing and reverberation reduction levels. Both the compression of the reverberant part ($\chi^2(4) = 106.2$, $p \ll 0.05$) and the smoothing of the spectral details ($\chi^2(4) = 90.1$, $p \ll 0.05$) significantly affect the externalization results. When the rever-

beration is fully present ($\alpha = 0\,\%$), the externalization ratings decrease dramatically with increasing bandwidth factors above one. For the highest smoothing level ($B = 64$), the median value of the externalization rating is about 1.4, which corresponds to the sound source being perceived as slightly externalized. When the reverberation is fully absent ($\alpha = 100\,\%$), the externalization ratings are low with all smoothing levels, and only slight differences in externalization results are observed by reducing the spectral details in the direct part. It should be noted that the reproduced sound for this condition ($\alpha = 100\,\%$) is almost the same as for the "bi" condition in Experiment D. The differences in externalization results between these two experimental conditions ($\alpha = 100\,\%$ condition in Experiment E and "bi" condition in Experiment D) are mainly caused by the different experimental contexts with reference signals (with or without reverberation).

4.4 MODEL FITTING AND EVALUATION

4.4.1 *Model fitting*

The results of the listening experiments confirm that all the three acoustic cues, namely SGs, ILDs and ILD TSDs, have influences on externalization. If one of these cues is distorted, the sound image can not be perceived as well externalized. A set of the experimental results was used to fit the weighting factors in the externalization model (see Figure 4.1), and the remaining experimental results were used to validate the prediction performance of the proposed model in terms of externalization.

When the acoustic cues provided by the target and the template signal have the same values ($\overline{\Delta\mathrm{ILD}}$, $\overline{\Delta\xi}(w)$ and $\overline{\Delta\mathrm{ILD}}_{\mathrm{TSD}}$ are equal to zero), the target signal reaches the highest externalization level, resulting in the sum of the mapping factors a and c being equal to 3 (see Equation 4.8). To simplify the mapping function, the mapping factor c was defined as the minimal externalization rating obtained from the five experiments, i.e., c was equal to 1. Hence, the mapping factors a and c were set to 2 and 1, respectively. The experimental results from Experiments A ("BB" condition), B, C, and E ("B = 0" and "0 % reverberation reduction" conditions) were used to calculate the weighting factors for acoustic cues by fitting the normalized deviations of these cues from the template to the individual externalization ratings using the least-squares method.

Table 4.1 shows the iterative steps to determine unknown weighting factors ($b_{\mathrm{ILD\,TSD}}$, b_{ILD}, b_{ξ}, w and b_{γ}) for model components based on subjective experimental results. For better visibility, these steps are also illustrated in Figure 4.8. The weighting factors were first determined individually for each subject to

Table 4.1: The steps of model fitting for each subject. N_d represents the number of data points per subject.

Step	Experiment	Condition	N_d	Fitting parameters	Initial value	Fixed parameters
1	E	*"B=0"*	5	$b_{ILD\,TSD}$	Random	$b_{ILD} = b_{\xi}(w) = b_{\gamma} = 0$
2	A	"BB"	5	b_{ILD}	Random	$b_{ILD\,TSD}$ (step 1), $b_{\xi}(w) = 0$, $b_{\gamma} = 1$
3	B	-	5	b_{ξ} & w	b_{ξ} is random; $w \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1\}$	$b_{ILD\,TSD}$ (step 1), b_{ILD} (step 2), b_{γ} =1
4	C	"ipsi" &"contra"	10	b_{ξ} & w	Taken from step 3	$b_{ILD\,TSD}$ (step 1), b_{ILD} (step 2), b_{γ} =1
5	E	"0% reverb. reduction"	5	b_{γ}	Random	$b_{ILD\,TSD}$ (step 1), b_{ILD} (step 2), $b_{\xi}(w)$ (step 4)
6	A,B,C,E	"BB", "ipsi", "contra" & "0% reverb. reduction"	25	b_{ILD}, b_{ξ}, w and b_{γ}	Taken from steps 2, 3 and 4	$b_{ILD\,TSD}$ (step 1)

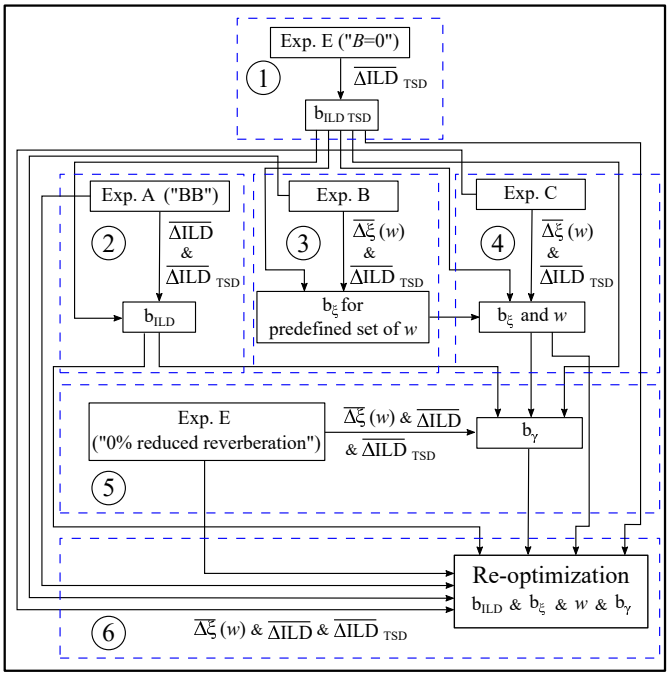


Figure 4.8: Illustration of the iterative steps of model parametrization for each subject.

check the non-parametric 95 % CIs for the average weighting factors (see Table 4.2). Then the weighting factors were averaged across subjects and the average/generic weights were used in model validation.

In the first step, the weighting factor $b_{ILD\,TSD}$ for the model component ILD TSD was calculated based on the results from the condition "$B = 0$" in Experiment E (5 data points per subject and parameter). Since the ILD TSD was the only affected acoustic cue in this experimental condition and could be decoupled from other model components (c.f. Equation 4.7), the determined weighting factor $b_{ILD\,TSD}$ did not need to be re-optimized in the last step.

The weighting factors b_{ILD}, b_{ξ} and w were pre-optimized based on the externalization results under anechoic conditions, assuming that $\overline{\Delta ILD}_{TSD}$ was

small and γ was close to one (steps 2-4). These determined pre-optimal weightings were further used as initial values to jointly estimate the optimal weights (step 6), since the model components in Experiments A, B and C were weakly coupled.

In the second step, b_{ILD} was calculated by fitting $\overline{\Delta ILD}$ to the externalization results from "BB" conditions in Experiment A (5 data points, γ and $\overline{\Delta\xi}(w)$ were set to one and zero, respectively).

In steps three and four, the weighting factors b_ξ and w were jointly calculated and optimized according to the results from Experiments B and C. First, the weighting factors b_ξ were determined to pair a predefined set of binaural weighting factors $w \in \{0.5,\ 0.6,\ 0.7,\ 0.8,\ 0.9,\ 1\}$ with the results from Experiment B (5 data points, γ was set to one). After that, the best pair of w and b_ξ was chosen for which the summed square of errors between the simulated and perceptually measured ratings over experimental conditions in Experiment C reached its minimum (10 data points).

In the fifth step, b_γ was adjusted according to the results from "0 % reverberation reduction" condition in Experiment E (5 data points) combined with pre-optimal b_{ILD}, w and b_ξ.

At last, b_{ILD}, w, b_ξ, and b_γ were jointly re-optimized by minimizing the simulated and perceptually measured results (25 data points) from Experiments A ("BB" condition), B, C, and E ("0 % reverberation reduction" conditions). Their pre-optimal results were used as the initial values for the final optimization.

In addition to the above-mentioned weights and mapping parameters, there has been another unknown parameter, namely the expected ILD temporal fluctuations under anechoic listening conditions. This parameter was considered because the maximal externalization ratings depended on the listening environment. For the anechoic condition, the virtual sound sources rendered with individually measured HRTFs were not perceived as fully externalized, the mean ratings were about 2.7 across experiments (c.f. Experiments A-D). In contrast, the virtual sound source was perceived as almost completely externalized in the reverberant condition (c.f. Experiment E). The reason for this could be that the subjects were more familiar with reverberant environments due to their daily life experience. This result may indicate that the information stored in the short-term memory, i.e., reverberation-related acoustic cues, was not completely adapted by changing the acoustic environment from a listening room to an anechoic chamber. To represent such an incomplete adaptation in the model, we added an offset in the template ILD TSDs (about 0.07 dB) under the anechoic conditions ("mal-adapted ILD temporal fluctuations").

The average weighting factors across subjects and with their corresponding non-parametric 95 % CIs are shown in Table 4.2. Overall, the weighting factors

Table 4.2: Mean weighting factors with ± non-parametric 95 % CIs for different acoustic cues.

Factors	b_{ILD}	b_{ξ}	w	$b_{ILD\ TSD}$	b_{γ}
Values	2.1 ± 0.8	1.7 ± 0.5	0.9 ± 0.1	2.8 ± 0.5	0.5 ± 0.1

are highly consistent and do not show large individual differences, especially w and b_{γ}.

4.4.2 *Model evaluation*

As introduced in Section 4.2, the proposed model is based on the a template-matching process. Hence, the knowledge of the "template" information, i.e., individual BRIR (individual HRIR and room) is required for assessing the externalization of headphone-based virtual sounds. In this study, the performance of the model was validated with the same listeners but with experimental results that were not used to fit the model.

The simulated results are shown together with the measured externalization ratings in Figures 4.2 to 4.7 (except Figure 4.4). The mapped and predicted results are represented with open and filled symbols, respectively. The results from Experiments B ("LO" and "HI" conditions), D (all three conditions) and E ("*B*=1", "*B*=4", "*B*=16", and "*B*=64" conditions except for the 0 % reverberation reduction condition) were used to evaluate the performance of the model regarding externalization.

The NRMSD was calculated between the simulated and perceptually measured data to quantify the prediction performance of the model according to Equation 3.10 (see Section 3.5.2). Table 4.3 shows the average NRMSD result across subjects for each experimental condition (mapping and prediction errors are shown in italic and bold, respectively). It can be seen that the mapping and prediction errors are less than 10 % for all experimental conditions, indicating the high prediction accuracy of the proposed model.

4.5 DISCUSSION

A series of experiments was performed to show the relevance of three important acoustic cues to perceived externalization and to determine the weighting factors for different components in the proposed externalization model. The model components, interpretation of individual externalization results, and model limitations are discussed below.

Table 4.3: Average NRMSD between the simulated and perceptually measured data. The mapping and prediction errors of calculated results are shown in italic and bold, respectively.

Experiment	Modification	Condition	Fitting parameter	NRMSD
A	ILD expansion	BB	b_{ILD}	*4.8%*
		LO		**5.5%**
		HI		**8.1%**
B	spectral magnitude smoothing	ipsilateral, constant ILD	b_ξ & w	*4.7%*
C	ILD contrast compression	ipsi	b_ξ & w	*4.3%*
		contra	b_ξ & w	*3.9%*
D	spectral magnitude smoothing	bi		**5.2%**
		ipsi		**3.9%**
		contra		**4.3%**
E	reverberation reduction and spectral magnitude smoothing	high spectral detail ($B = 0$)	$b_{ILD\,TSD}$ & b_γ	*3.3%*
		$B = 1$		**4.9%**
		$B = 4$		**7.0%**
		$B = 16$		**3.5%**
		low spectral detail ($B = 64$)		**2.5%**

4.5.1 *Model components*

The results from Experiment A demonstrate that the ILD is relevant to externalization perception, and the correct monaural spectral information alone is not sufficient to well externalize virtual sound images. Furthermore, the ILD expansion at low frequencies has more influence on externalization than that at high frequencies, indicating that the deviation of the average ILD across the whole frequency range (broadband ILD) may not be able to explain the externalization ratings for the "LO" and "HI" condition. In order to illustrate this, the deviation of the broadband ILD instead of the frequency-dependent ILD is used in the comparison stage to map to externalization ratings. As shown in Figure 4.9, the model outcomes show high prediction errors for the "LO" (NRMSD = 12.9 %) and "HI" (NRMSD = 10.7 %) conditions, suggesting that the broadband ILD is not suitable for predicting externalization results caused by the frequency-dependent ILD expansion.

In experiment B, the magnitude spectrum of the HRTF was smoothed at the ipsilateral ear and adjusted at the contralateral ear to preserve the original ILD across frequencies. The experimental results show that the externalization rating is decreased by the deterioration of the HRTF magnitude spectrum, although the correct ILD information is maintained, indicating that the spectral information of binaural sounds in both ears is relevant to perceived externalization. These results are generally in line with the findings in [12], but the reduction in externalization is less pronounced because a more lateral sound source was tested in our experiment.

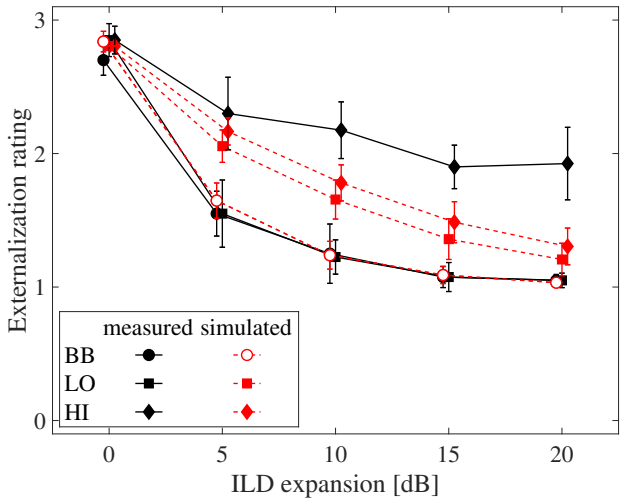


Figure 4.9: Median values of predicted externalization ratings based on deviation of broadband ILDs for "BB", "LO" and "HI" conditions. All other conventions are as in Figure 4.2.

Experiment C studied the relative influence of the spectral magnitude in HRTFs of the contralateral versus ipsilateral ear on externalization. The experimental results illustrate that the spectral information in the HRTFs of the ipsilateral ear is more relevant to externalization than that of the contralateral ear. This finding is closely related to the results of localization studies in [129, 130]. Morimoto [129] illustrated that the contribution of the ipsilateral ear to sound source localization increased as the source moved laterally. For azimuth angles larger than 60°, the contralateral ear had almost no contribution to the determination of source localization. This result reveals that the spectral information in the ipsilateral ear signal is more relevant to sound localization than that in the contralateral ear signal. Macpherson and Sabin [130] further quantified the binaural weighting of spectral cues of each ear as a function of source directions. However, the binaural weighting factor determined in our externalization study is not equal to one (weight for the ipsilateral spectrum) at the extreme lateralization, implying that the spectral information in the HRTFs of the contralateral ear can not be ignored to generate well externalized virtual sound images.

The results from Experiment D show that perceived externalization is reduced by smoothing the spectral magnitude with bandwidth factors above one. The smoothing process applied to HRTFs leads not only to a change in the ILDs, but also to a change in the spectral information of HRTFs. The predicted externalization ratings match well with the perceptually measured results, since all relevant acoustic cues are incorporated in the proposed model. To test whether a single cue is sufficient to explain the experimental results for different conditions, the externalization ratings are additionally predicted by using only one

acoustic cue (either SG or ILD is optimized). For the SG-based model, the weighting factor for normalized SG deviations, b_ξ, and the binaural weighting factor, w, are re-optimized according to the results from Experiment C. As a result, the estimated average values of b_ξ and w are 1.9 and 0.8, respectively. For the ILD-based model, the average value of b_{ILD} is 2.5.

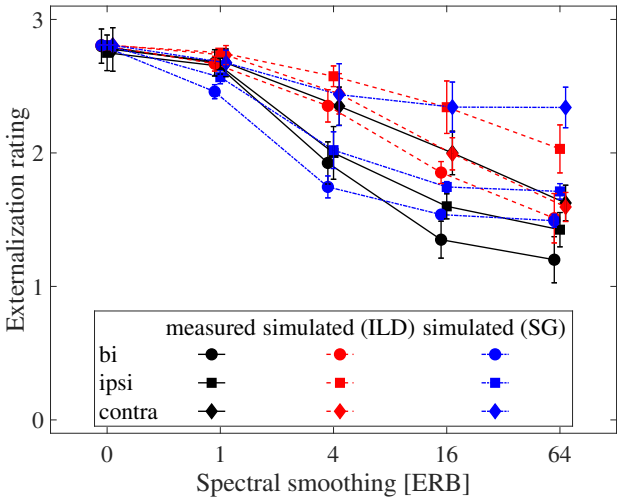


Figure 4.10: Median values of predicted externalization ratings using a single acoustic cue (ILD or SG deviations) for "bi", "ipsi" and "contra" condition conditions. ILD- and SG-based prediction results are represented with dashed and dash-dotted lines, respectively. All other conventions are as in Figure 4.2.

Figure 4.10 shows the predicted externalization ratings based on either ILD deviations (dashed lines) or SG deviations (dash-dotted lines). For the SG-based model [36], the calculated ratings for both "bi" and "ipsi" conditions match well with the perceptually measured results with increasing bandwidth factors. Unfortunately, due to the low weighting for SG at the contralateral ear, the calculated externalization results for the "contra" condition change only slightly with different smoothing levels. In the case of the ILD-based model [15], the change in predicted externalization ratings is comparable to the measured data as the degree of smoothing increases, but the relative externalization results between the "ipsi" and "contra" conditions do not match the subjective data. Also, the computed results are overall higher than the measured data. The simulation results point out that both monaural and interaural spectral cues are necessary for explaining the change in externalization ratings caused by spectral smoothing of the HRTFs.

In Experiment E, when the reverberation is fully present (α =0 %), the externalization ratings decrease with increasing smoothing level, as a result of destroyed SGs and ILDs. On the contrary, if the reverberant part is completely absent (α =100 %), the externalization ratings for different smoothing levels are low, even if the direct part is unmodified.

These results suggest that all three acoustic cues (SGs, ILDs, and ILD TSDs) are important for perceived externalization, but only one of them, preserved in binaural signals, is not sufficient to well externalize virtual sound sources. In the developed model, the influence of the SG and ILD deviations on externalization is adjusted based on the amount of reverberation. To illustrate how the influence of SG and ILD on predicted results changes from an anechoic to a reverberant environment, the weighting factors (b_{ILD}, b_{ξ}, w, $b_{ILD\,TSD}$ and b_{γ}) are re-optimized without considering the reduction term (γ is set to one), and the externalization results for Experiment E are re-calculated.

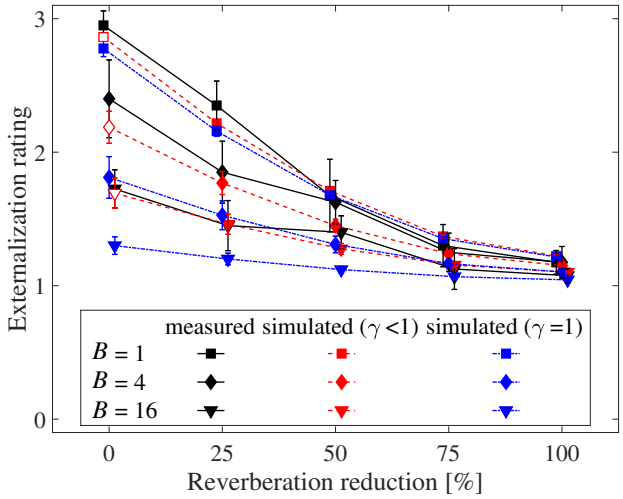


Figure 4.11: Median values of simulated externalization ratings with (dashed lines, "$\gamma < 1$", the model outputs for the "$\alpha = 0$" condition are mapped results) and without (dash-dotted lines, "$\gamma = 1$") the reduction term in the model for different spectral smoothing (B) and reverberation compression levels (α). All other conventions are as in Figure 4.2.

Figure 4.11 shows the prediction results for some experimental conditions ("*B*=1", "*B*=4", and "*B*=16") with and without considering the reduction term. When the reduction term is not included in the model, the predicted results clearly underestimate the perceptually measured data for bandwidth factors above one, especially when the reverberant part is fully present ($\alpha = 0\,\%$). As expected, the prediction errors (NRMSD) become larger, increasing to 5.8%, 12.3%, 7.8% and 5.1% for "*B*=1", "*B*=4", "*B*=16", and "*B*=64" conditions, respectively. The comparison results reveal that the relevance of SGs and ILDs to externalization is reduced when reverberation is present, and the proposed model is valid for explaining this effect.

4.5.2 *Individual differences*

Due to listener-specific acoustics, the spectral smoothing of individual HRTFs may cause different amounts of changes in ILDs and SGs. As a result, the ILD can change more than the SG for one listener and less for another. For instance, the externalization results assessed by subjects 2 and 3 for "contra" and "ipsi" conditions in Experiment D are shown in Figure 4.12.

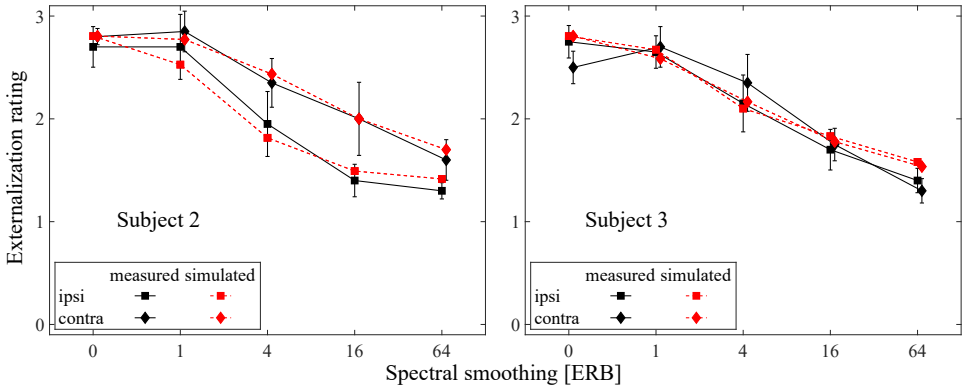


Figure 4.12: Median values of externalization ratings and the predicted results of subject 2 (left panel) and subject 3 (right panel) for smoothed spectral information in the HRTF ("contra" and "ipsi" conditions in Experiment D).

It can be seen that different results were reported for these two conditions. The median values of externalization ratings given by subject 2 are higher for the "contra" than for the "ipsi" condition with bandwidth factors above one. In contrast, subject 3 reported similar externalization ratings for both conditions with large bandwidth factors.

To investigate the reason for the different externalization ratings reported by these two listeners, normalized ILD and SG deviations are extracted and compared with the perceptual data (see Figure 4.13). For subject 2 and large spectral smoothing levels, the normalized ILD deviation is slightly higher for the "contra" condition than for the "ipsi" condition, while the normalized SG deviation is much larger for the "ipsi" condition than for the "contra" condition. Thus, the high contribution of normalized SG deviations dominates the opposite deviations in ILD, leading to lower externalization results for the "ipsi" condition. On the contrary, for subject 3, the normalized ILD and SG deviations are similar in extent and also act in opposite directions, effectively balancing the two conditions.

Spectral smoothing of the HRTFs results in different levels of ILD and SG deviations and thus individual externalization ratings. Since these two cues are incorporated in the proposed model, the predicted results are highly consistent with the perceptual data for both subjects (dashed lines in Figure 4.12). This

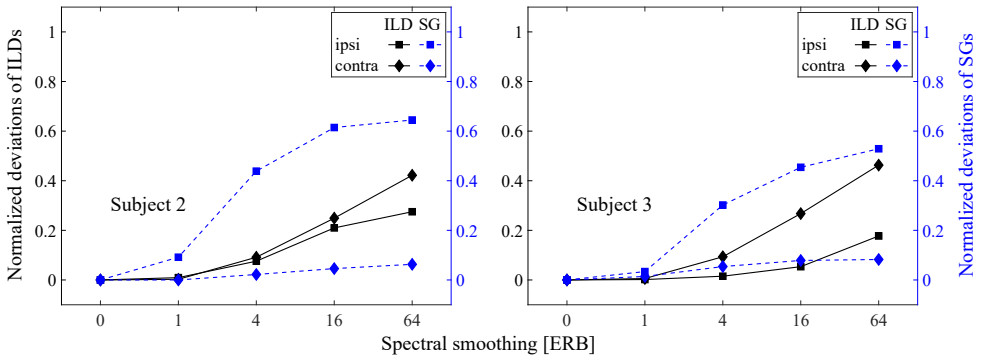


Figure 4.13: Deviation of normalized ILDs (solid lines) and SGs (dashed lines) of the individually synthesized binaural signals for subject 2 (upper panel) and subject 3 (lower panel) under different conditions.

simulation result demonstrates the importance of accounting for differences in individual acoustics for accurate modeling of perceived externalization.

4.5.3 *Model limitations*

In this study, different weighting parameters for the proposed model were determined based on the analysis of a 90° sound source and two rooms (anechoic chamber and listening room). However, for more frontal sound sources, the effects of these acoustic cues (ILDs, SGs, and ILD TSDs) on perceived externalization may change, as represented by the weighting factors and mapping parameters in the model. For instance, the binaural weighting factor, w, for monaural spectral cues is assumed to be similar to that determined for sound localization in sagittal planes [126]. Baumgartner et al. [126] applied a sigmoid function to interpolate w for an azimuth angle of φ: $w = \left(1 + e^{(-\varphi/\Phi)}\right)^{-1}$, where Φ is a scaling factor, and was chosen to be 13° to match the results of localization studies. Our findings (w = 0.9 for a 90° sound source) imply a lower dependence of w on the lateral angle for externalization perception compared to localization perception (Φ = 41°). The mapping parameter c is expected to have directional properties, which can be formulated as $c = |\sin(\varphi)|$ based on a spherical head and lateralization between the left and right ear [11]. These interpolation approaches need to be validated in future externalization tests. In addition, the weighting factors, b_γ and $b_{\text{ILD TSD}}$, were obtained based on a reference value of ILD TSDs, and the existence of a general pre-expectation of the room-related information is a matter of ongoing discussion (c.f. [31, 131]).

Next, in our experiments, a loudspeaker was located at the HRTF/BRIR measurement position to serve as a reference position, and listeners could also listen to the reference sounds played back through the loudspeaker. Some studies have shown that the visual [31, 32] and auditory information [33] of the lis-

tening environment could potentially influence distance perception. The externalization ratings obtained were therefore task-specific, and the results might change without providing the room-related visual and auditory information. This problem is more serious for frontal sound sources, because lateral sound images can be perceived as more externalized than frontal sound images, even without a real reference source and visual information [19].

Furthermore, the spectral characteristics of stimuli were not considered in the proposed model. White noise was used as the stimulus because it has a uniform energy distribution across frequencies and time, which is advantageous for the stable extraction of acoustic cues. However, specific stimulus types may reduce the accessibility of acoustic cues, triggering an adjustment of the weighting factors. The proposed model can serve as a valuable tool to analyze such adaption processes for weighting factors based on stimulus types in future experiments.

At last, the contribution of head or source movements to externalization was not included in the model. Though these dynamic cues can not dramatically affect externalization results of a lateral sound source, they are important for frontal sound sources [28]. Therefore, the current model is to be extended to include the temporal integration processes to represent dynamic cues.

4.6 CONCLUDING REMARKS

This study has investigated the contribution of monaural spectral information, ILDs, and ILD temporal fluctuations to externalization of a 90° sound source. The results of five psychoacoustic experiments show that all of these acoustic cues are relevant and that a single cue alone is not sufficient to well externalize sound images. Moreover, the spectral information in the contralateral HRTF can not be neglected to create well externalized virtual sound sources, although it contributes less to externalization than that in the ipsilateral HRTF. In addition, the contribution of monaural spectral information and ILDs to externalization is reduced when reverberation is present.

A model is proposed to predict externalization results on anechoic and reverberant lateral sound sources based on a template-matching process. The predicted results match well with the perceptually measured data. Unlike the model developed by Hassager et al. [15] (ILD-based model), not only ILDs but also monaural spectral cues and ILD temporal fluctuations are incorporated into the model to jointly predict externalization ratings. The developed externalization model can be considered as a starting point for further extensions mentioned above, e.g., taking into account for different source directions, listening environments, stimulus types, and dynamic scenes, etc. Further, the influences of other acoustic cues such as inconsistency between interaural cues (ILD

and ITD) and ITD temporal fluctuations, on perceived externalization need to be studied, and incorporated into the model. In the future, the quantitative model should be used to generate hypotheses for externalization experiments.

5

EXTERNALIZATION ENHANCEMENT OF VIRTUAL FRONTAL AND REAR SOUND SOURCES

5.1 INTRODUCTION

A binaural rendering system aims at artificially generating virtual 3D sound images reproduced over headphones in real-time. As described in section 1.3, instead of directly convolving with BRIRs, a commonly used strategy in binaural synthesis is to separately synthesize the direct sound, the early reflections, and the late reverberation based on HRTFs and room models. Non-individual HRTFs from publicly available databases (e.g., CIPIC [3], SADIE II [71], IRCAM [132], and THK [133]) are often used in binaural rendering systems, since measuring personal HRTFs is time-consuming and impractical for each listener, especially in consumer scenarios.

Several studies have shown that the use of non-individual HRTFs in binaural synthesis may reduce externalization of auditory images, especially for frontal and rear sound sources [118, 134, 135]. This Chapter proposes a novel rendering system to improve perceived externalization of frontal and rear sound images, consisting mainly of direction-dependent peak and notch filters for direct parts, and filter-bank-based decorrelation filters for early reflections. The system is further subjectively evaluated concerning externalization of frontal and rear sound sources. Parts of this Chapter have been published in [136].

The remainder of this Chapter is divided into the following parts. Section 5.2 introduces the proposed binaural rendering system. To evaluate the performance of the developed system, a listening experiment is designed. The experimental paradigm and the results are presented in Section 5.3. Section 5.4 discusses the proposed binaural rendering system. Finally, conclusions and future work are drawn in Section 5.5.

5.2 OVERVIEW OF THE PROPOSED BINAURAL RENDERING SYSTEM

Figure 5.1 shows the block diagram of our proposed binaural rendering system, which is based on the framework described in Section 1.3.

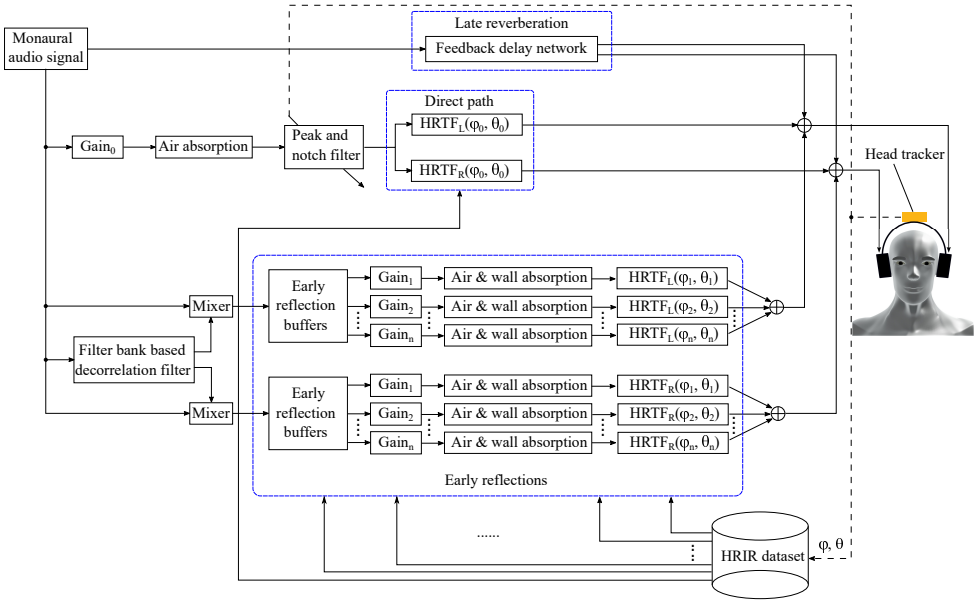


Figure 5.1: Proposed binaural rendering system.

The direct sound component is simulated by filtering the input signal with a pair of HRTFs. For frontal or rear sound sources, a peak and notch filter is applied to improve their localization accuracy. The direction of each early reflection relative to the listener is determined based on a rectangular (shoe-box) room model using the image source method [38], and the locations of image sources (early reflections) along the x-, y- and z-coordinate $\{x', y', z'\}$ of the room are expressed as:

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \pm x_s + 2nL_x \\ \pm y_s + 2lL_y \\ \pm z_s + 2mL_z \end{pmatrix}, \tag{5.1}$$

where $\{L_x, L_y, L_z\}$ and $\{x_s, y_s, z_s\}$ represent the room sizes and coordinates of the sound source, respectively. $\{n, l, m\}$ are the integer vector triplet. The number of image sources increases exponentially with the reflection order. To reduce the computational complexity, only the 1^{st} and 2^{nd} order early reflections are calculated, which are sufficient to create externalized sound images [137]. The attenuations and delays of early reflections due to the distances between image sources and the listener are realized with gain factors (referred to as "Gain" in Figure 5.1) and tapped delay lines, respectively. Low-pass filters

with different cut-off frequencies and attenuations are used to represent the frequency-dependent reflection factors of walls, ceiling, and floor (referred to as "wall absorption" in Figure 5.1). Distance-dependent low-pass filters are applied to the direct sound and early reflections to simulate the effect of air absorption [138]. Additionally, the input signals used to generate sparse early reflections are decorrelated between the left and right ears to improve them spaciousness of the virtual sound images and thus increase the degree of externalization [139]. An FDN with eight internal feedback channels [39] is applied to synthesize the late reverberation, which produces a natural reverberation effect and is widely used in binaural synthesis.

For dynamic scenarios, a head tracker placed on the headphones is used to detect the listener's head movement so that the VAE can be rotated accordingly to fix the absolute position of the virtual sound images. The HRTFs are represented as minimum-phase systems, followed by all-pass filters [125]. The minimum-phase components of the HRTFs are linearly interpolated in 1° azimuth and elevation off-line to ensure the dynamic binaural reproduction without resolution artifacts [140]. The minimum-phase components are updated by directly switching the interpolated sets (1° resolution) according to the directional information received from the head tracker. Additionally, the current and previous minimum-phase components are cross-faded to eliminate audible artifacts. The all-pass filters are considered as pure delays, and their changes (updates) are realized by linear interpolation of tapped delay lines. In the following sections, the important system components are described in details.

5.2.1 *Peak and notch filter*

Several studies have shown that some specific frequency components of ear signals are closely related to the subjective impression for source localization in the median plane [2, 141, 142].

Hebrank and Wright [142] demonstrated that a sound source was perceived as frontal, when it contained a 1-octave notch having a lower cut-off frequency between 4 kHz and 10 kHz and increased energy above 13 kHz; the perception of a rear sound source was cued by a peak between 10 kHz and 12 kHz; a 1/4-octave peak between 7 kHz and 9 kHz was relevant to the perception of a sound source above the head.

Blauert [2] investigated the relationship between the center frequency of a 1/3-octave band noise and the perceived direction in the median plane. The "directional band" was proposed based on the experimental results, indicating that the narrow band noise signals centered at 500 Hz and 4 kHz were perceived as frontal, and those centered at 1 kHz and 8 kHz were perceived as rear and

above, respectively. Wallis and Lee [141] revised the "directional band", and confirmed the Blauert's findings for 1 kHz, 4 kHz, and 8 kHz but not for 500 Hz.

Yao and Chen [143] adjusted HRTFs with additional peak and notch filters to enhance localization accuracy of virtual sound sources in the median plane. In their study, two peak filters (1/4-octave bandwidth) centered at 4 kHz and 14 kHz, and a notch filter (1-octave bandwidth) centered at 7.5 kHz were applied for the 0° (frontal) HRTFs. Two peak filters (1/4-octave bandwidth) centered at 4 kHz and 11 kHz and two notch filters (1/4-octave bandwidth) centered at 9 kHz and 16 kHz were used to adjust the 180° (rear) HRTFs. However, the information of the "directional band" was not included in the designed filters for the rear sound source. Further, the spectral characteristic of HRTFs itself was also included for designing the filters, e.g., a peak filter centered at 4 kHz [144] was applied for both 0° and 180° HRTFs.

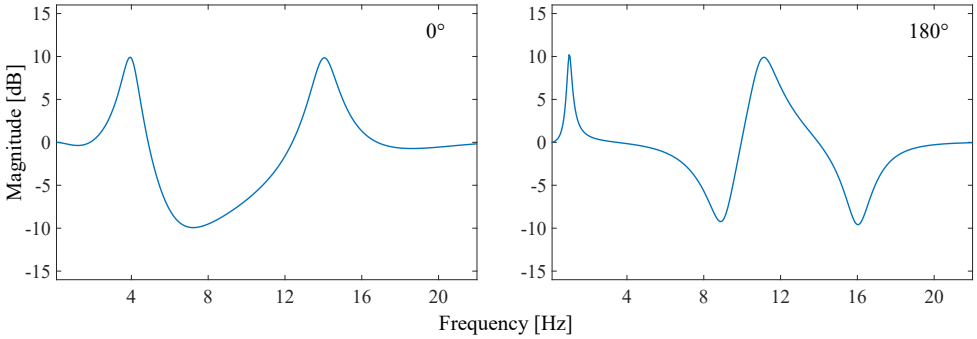


Figure 5.2: Magnitude spectra of designed equalizers (peak and notch filters) for frontal (left panel) and rear (right panel) sound sources.

In the present study, two equalizers, consisting of cascaded 2nd order peak and notch filters, are designed for frontal and rear sound sources (HRTFs) to improve their localization performance based on the method proposed in [143], while considering results from previous psychoacoustic experiments [2, 141, 142, 144]. As a result, a notch filter centered at 7 kHz with a bandwidth of 1-octave, a peak filter centered at 4 kHz with a bandwidth of 1/3-octave, and a peak filter centered at 14 kHz with a bandwidth of 1/4-octave are cascaded and applied for the frontal sound source. A peak filter centered at 1 kHz with a bandwidth of 1/3-octave, a notch filter centered at 9 kHz with a bandwidth of 1/4-octave, a peak filter centered at 11 kHz with a bandwidth of 1/4-octave, and a notch filter centered at 16 kHz with a bandwidth of 1/4-octave are cascaded and used for the rear sound source. The coefficients for peak and notch filters are determined according to Zölzer [145]. Through some informal listening tests, approximately +/- 10 dB gains are applied to peak and notch filters to account for the trade-off between sound quality and localization accuracy.

Figure 5.2 shows the magnitude spectra of designed peak and notch filters for frontal (left panel) and rear (right panel) sound sources, respectively.

5.2.2 *Decorrelation for early reflections*

Decorrelated reverberation can improve perceived externalization for 3D audio reproduction over headphones [139]. Catic et al. [18] showed that the binaural cues contained in reverberation were important for perceived externalization, and IC could be used as an indicator of the degree of externalization. Hence, in the proposed system, we attempt to artificially reduce the IC of early reflections between the left and right ears to increase perceived externalization of sound sources rendered by non-individual HRTFs. A common decorrelation method is to use all-pass filters with random phase responses [139]. However, the uniform magnitude spectrum over the whole frequency can not be guaranteed due to the high variations in the phase response of the decorrelation filter [146]. To avoid this issued, the filter-bank-based decorrelation method introduced by Bouéri and Kyriakakis [146] is applied to early reflection parts in this study.

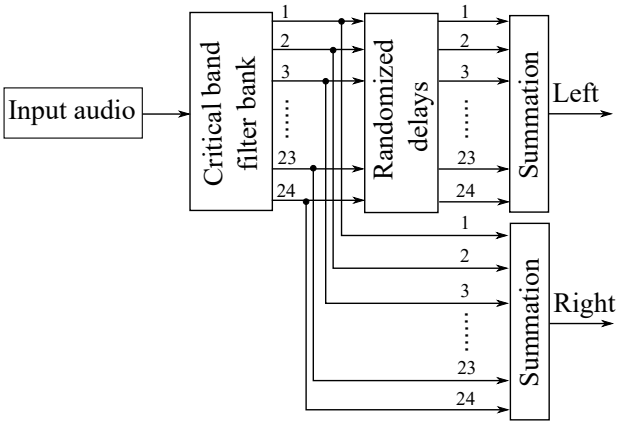


Figure 5.3: Filter-bank-based decorrelation filter according to [146].

As illustrated in Figure 5.3, the input audio signal is divided into 24 critical bands and delivered to the left and right channels. After that, a random delay is introduced into each frequency band for the left channel signals. Finally, the signals of the frequency bands are summed up in the left and right channels. Since the audio signals in the high frequencies are more sensitive to the time shifts than those in the low frequencies, the maximum values of the delays introduced into the frequency bands are limited based on the wavelengths of the bands (see Figure 4 in [146]). The decorrelated signals are then mixed with the input signal to adjust the degree of correlation.

5.3 EXPERIMENT

5.3.1 *Experimental paradigm*

A listening test was performed to evaluate the performance of the proposed binaural rendering system regarding perceived externalization of frontal and rear sound sources. The non-individual HRTFs taken from the THK database [133] were used in binaural synthesis. The experiment was performed in a listening room located in our institute (see Section 3.4), and the early reflections were simulated based on the geometry of this room. The reflection factors and the parameters for the artificial late reverberator were adjusted to match the acoustics of the real room. The stimuli used in the listening experiment were a guitar recording [147] and a speech signal [116] with a duration of 10 s.

Six normal-hearing listeners (one female and five males, aged between 25 and 30) participated in the experiment. Each subject sat in a chair, and listened to the test signals presented by a pair of headphones (Sennheiser HD800). Two real loudspeakers (0° and 180° relative to the listener) were placed in the listening room as references for the externalization evaluation. The real loudspeaker positions were matched with the "theoretical" positions of synthesized virtual loudspeakers. Listeners compared the stimuli produced by the proposed and conventional rendering systems against the loudspeaker positions, and gave their assessment using a four-point rating scale that we have used in previous studies (see Table 3.1). The conventional binaural rendering system did not contain equalizers (peak and notch filters) for the direct sound component and decorrelation filters for early reflections. To avoid the enhancement of externalization caused by head movements [28], the head tracker device was deactivated during the experiment, and subjects were not allowed to turn their heads. The rendered audio signals presented through headphones had a level of about 64 dBA.

5.3.2 *Experimental results*

Figure 5.4 shows the median externalization ratings with non-parametric 95 % CIs (notch-edges) for "guitar" and "speech" signals generated with the proposed and the conventional rendering systems. For the stimulus "guitar", the average externalization ratings provided by both systems are slightly higher for the frontal direction than for the rear direction, which is not observed for the "speech" signal. For each stimulus type and source direction, the externalization rating of the sound source synthesized by the proposed rendering system is significantly higher than that synthesized by the conventional system

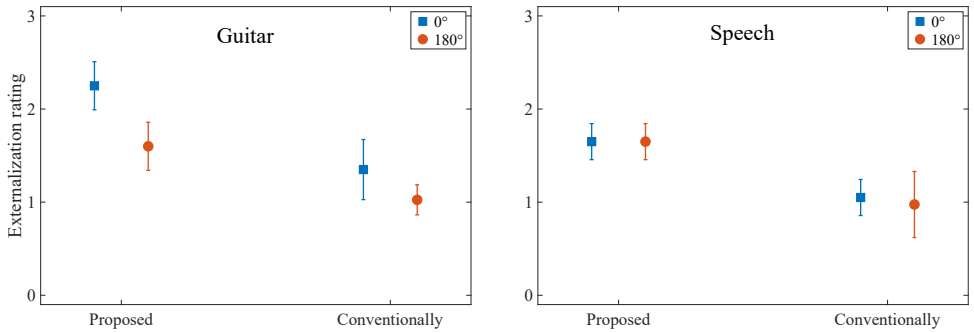


Figure 5.4: Median externalization ratings with non-parametric 95 % CIs (notch-edges) for guitar (left panel) and speech signals (right panel) using proposed and conventional binaural rendering systems across subjects. The results for frontal and rear sound sources are shown with squares and circles, respectively.

(Wilcoxon tests: $p \ll 0.05$). Unfortunately, even with the proposed rendering system, the sound source still can not be perceived as being at the position of the real loudspeaker.

5.4 DISCUSSION

This study aims to enhance externalization of virtual frontal and rear sound images synthesized with non-individual HRTFs.

Kim et al. [148] have developed a similar rendering system to enhance perceived externalization of a mono audio signal. A pair of decorrection filters was applied to the direct part to increase the spaciousness. However, the localization accuracy of the frontal and rear sound sources might be reduced. To increase the localization accuracy of a frontal sound source, a pair of notch filters was designed based on the spectral magnitude of average 0° HRTFs taken from the CIPIC database [3]. In addition, in their system, HRTFs were simulated with a head and pinna model according to Brown and Duda [149]. However, some specific parameters for the pinna model have to be determined through listening tests. Our idea differs in that the non-individual HRTFs were directly used in the binaural rendering system; the decorrelation filter was applied only to the early reflections to increase the spaciousness without affecting the localization accuracy; the design of peak and notch filter for the direct part was according to the results of past psychoacoustic experiments.

Localization accuracy of binaurally synthesized frontal and rear sound sources is commonly low with non-individual HRTFs due to the lack of correct spectral cues from individual ears [117]. Applying the equalizer (cascaded peak and notch filters) to the direct parts aims at increasing the localization accuracy of

frontal and rear sound sources. An informal listening test confirms the reduction in the front-back confusion when applying the equalizer. In this experiment, the head tracker is deactivated to avoid the impact of head movements on externalization [21, 28]. For dynamic scenarios (head tracker is active and head movements are allowed), the gain factors for cascaded peak and notch filters should be designed as direction-dependent to avoid audible artifacts between frontal/rear and lateral sound sources (smooth transition of gain factors between the frontal/rear and lateral regions). For instance, in our current implementation, the gain factors change linearly from 10 dB to 0 dB with increasing angles in azimuth or elevation from 0° to $\pm$ 20°.

Catic et al. [18] concluded that lower IC between left and right ear signals corresponded to higher externalization ratings based on the analysis of modified and individually measured BRIRs. Reverberation is relevant to perceived externalization. For the binaural rendering system, the use of non-individual HRTFs and a simple room model is known to create a great number of wrong cues for externalization perception. In the present study, decorrelation filters are applied to the early reflections to artificially decrease the IC between left and right ear signals and hence to increase externalization. Note that this does not mean that the early reflections between the left and right ear must be completely decorrelated, as the direction of each early reflection should be discernible. The experimental results show that such processing, in combination with peak and notch filters applied to the direct parts, can effectively increase perceived externalization of frontal and rear sound sources.

5.5 CONCLUDING REMARKS

This study has developed an advanced binaural rendering system to enhance externalization of virtual frontal and rear sound sources synthesized with non-individual HRTFs. The rendering system consists mainly of an early reflection module based on the image source method, an FDN-based artificial late reverberator, cascaded peak and notch filters applied to direct sounds, and a pair of decorrelation filters applied to early reflections.

Compared to the conventional binaural playback system, perceived externalization of frontal and rear sound sources is substantially improved with the proposed system. However, due to the lack of listener-specific acoustic transfer characteristics and the acoustics of the real room, the sound source still can not be perceived as well externalized with the proposed method. Further work is to adapt the binaural rendering system to the real room acoustics to further improve plausibility and externalization of sound sources.

6

CONCLUSIONS AND FUTURE WORK

6.1 CONCLUSIONS

HRTFs are essential for creating headphone-based virtual 3D sound images. Different measurement systems have been developed for fast measuring high-resolution individual HRTFs. Considering the measurement time and infrastructure costs, the hybrid setup consisting of a loudspeaker array and a single-axis positioning system is widely used in acoustic laboratories. However, multiple loudspeakers are required to measure HRTFs from different directions and most setups are only able to measure HRTFs with a fixed distance (2D HRTFs). In the first part of this thesis (Chapter 2), a detailed overview of the state of the art in HRTF measurement systems/approaches is provided. Further, an MR-based mobile system has been proposed to record individual 3D HRTFs with only a single loudspeaker. The depth camera and inertial sensor integrated in the MR device are used to detect the source (loudspeaker)-listener distance and head orientation, respectively. With this system, subjects can rotate their heads and move their bodies towards or away from the loudspeaker to cover desired measurement positions. The information of the visited and unvisited measurement positions, and the quality of measured HRTFs are visually displayed through the MR HMD. Although the influence of the HMD on the measurement results can not be ignored and a suitable sound source is required to measure HRTFs at close distances, the proposed system shows the potential to fast measure personal 3D HRTFs with a flexible setup.

Perceived externalization plays an important role in creating immersive VAEs. We have investigated relevant auditory cues contained in binaural sounds in the context of perceived externalization, and proposed a model to predict externalization of anechoic and reverberant lateral sound sources.

The second part of this thesis (Chapter 3) has investigated the effect of removing reverberation in contralateral versus ipsilateral ear signals on perceived externalization of a lateral sound source. The BRIRs of the contralateral and ipsilateral ears were separately truncated, and such modified BRIRs were used to

synthesize virtual sound sources and played back to listeners via headphones. The experimental results suggest that reverberation at the contralateral ear has more influence on perceived externalization of a lateral sound source than that at the ipsilateral ear. Perceived externalization is slightly changed by removing the reverberation in the ipsilateral ear signals. The change in externalization ratings caused by the removal of lateral reverberation can be well explained by the reverberation-related binaural cues. The effect of lateralized reverberation on externalization was further tested for different source directions. The results show that the contribution of reverberation at the contralateral ear to perceived externalization increases as the source moves laterally. For azimuth angles larger than $\pm$ 30°, reverberation at the contralateral ear dominates the assessment of externalization.

The third part of this thesis (Chapter 4) has further investigated the additive influences of relevant acoustic cues on externalization, and has developed a model to predict externalization. A series of experiments was designed to investigate the relevance of ILDs, SGs, and ILD temporal fluctuations on externalization of a 90° sound source. The experimental results indicate that the spectral details in the HRTFs of the ipsilateral ear are more important to perceived externalization than that of the contralateral ear, but maintaining the correct magnitude spectra in HRTFs only at the ipsilateral ear is not sufficient for well externalizing sound sources. Additionally, the relevance of the spectral information provided by HRTFs for externalization decreases with increasing reverberation. Furthermore, all these three cues (ILDs, SGs, and ILD temporal fluctuations) are relevant to externalization, and that a single acoustic cue alone is not sufficient for well externalizing virtual sound images. These three acoustic cues are then used jointly to predict externalization of anechoic and reverberant lateral sound sources based on a template-matching procedure. The predicted results match well with the externalization ratings obtained in listening experiments.

The use of non-individual HRTFs and a simple room model in binaural synthesis reduces perceived externalization of auditory images, especially for frontal and rear sound sources. The last part of this thesis (Chapter 5) has proposed an advanced binaural rendering system to enhance externalization of frontal and rear sound sources synthesized with non-individual HRTFs. Cascaded peak and notch filters are applied for direct sound components to improve the localization accuracy, which are designed according to the results of past psychoacoustic experiments. In addition, filter-bank-based decorrelation filters are used to decorrelate the early reflections to artificially reduce the IC between two ear signals. The subjective experimental results confirm the ex-

ternalization enhancement of virtual frontal and rear sound sources with the proposed approach.

6.2 FUTURE WORK

HRTF individualization/calculation is a timely topic, since not all listeners have the possibility to measure their personal HRTFs [150]. However, the acquisition of highly accurate personal HRTFs is still a challenge with these methods. The proposed method shows the possibility to rapidly measure personal HRTFs with a flexible setup. It is interesting to investigate whether the accuracy of HRTF individualization can be improved with recorded HRTFs at some specific measurement positions.

The current model is limited to predict externalization of a 90° sound source, and it needs to be further extended with the consideration of different source directions, listening environments, stimulus types, and dynamic scenes. Further, the model is based on a template-matching procedure, it would be interesting to study whether it is possible to "blindly" predict externalization without the "template" information.

The congruent room-related auditory information between the synthesized and real listening environment is important for perceived externalization. Therefore, prior knowledge of the real room acoustic (e.g., reverberation time) is required for binaural synthesis. It can be further investigated how to quickly adapt the binaural rendering system to the real room acoustic to create well externalized sound images.

Part II

APPENDIX

BIBLIOGRAPHY

[1] N. I. Durlach, A. Rigopulos, X. D. Pang, W. S. Woods, A. Kulkarni, H. S. Colburn, and E. M. Wenzel. "On the Externalization of Auditory Images." In: *Presence: Teleoperators and Virtual Environments* 1.2 (1992), pp. 251–257. ISSN: 1054-7460. DOI: 10.1162/pres.1992.1.2.251.

[2] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, England: MIT Press, 1997. ISBN: 0262024136.

[3] V. R. Algazi, R. O. Duda, D. M. Thompson, and Carlos Avendano. "The CIPIC HRTF Database." In: *IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics* (2001). DOI: 10.1109/ASPAA.2001.969552.

[4] L. Rayleigh. "XII. On our perception of sound direction." In: *Philos. Mag. Series 6* 13.74 (1907), pp. 214–232. DOI: 10.1080/14786440709463595.

[5] K. Sunder, J. J. He, E. L. Tan, and W.-S. Gan. "Natural Sound Rendering for Headphones: Integration of signal processing techniques." In: *IEEE Signal Processing Magazine* 32.2 (2015), pp. 100–113. ISSN: 1558-0792. DOI: 10.1109/msp.2014.2372062.

[6] F. L. Wightman and D. J. Kistler. "Headphone simulation of free-field listening. I: Stimulus synthesis." In: *J. Acoust. Soc. Am.* 85.2 (1989), pp. 858–867. DOI: 10.1121/1.397557.

[7] H. Wallach, E. B. Newman, and M. R. Rosenzweig. "The precedence effect in sound localization." In: *The American Journal of Psychology* 62.3 (1949), p. 315. ISSN: 00029556. DOI: 10.2307/1418275.

[8] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman. "The precedence effect." In: *J. Acoust. Soc. Am.* 106.4 Pt 1 (1999), pp. 1633–1654. DOI: 10.1121/1.427914.

[9] H. Haas. "Über den Einflub eines Einfachechos auf die Hörsamkeit von Sprache: ("On the influence of a single echo on the audibility of speech")." In: *Acustica* 1 (1951), pp. 49–58.

[10] D. M. Howard and J. Angus. *Acoustics and psychoacoustics*. Fifth edition. New York, NY: Routledge, 2017. ISBN: 1317508297.

[11] V. Best, R. Baumgartner, M. Lavandier, P. Majdak, and N. Kopčo. "Sound Externalization: A Review of Recent Research." In: *Trends in hearing* 24 (2020), pp. 1–14. DOI: 10.1177/2331216520948390.

[12] W. M. Hartmann and A. Wittenberg. "On the externalization of sound images." In: *J. Acoust. Soc. Am.* 99.6 (1996), pp. 3678–3688. DOI: 10.1121/1.414965.

[13] A. Kulkarni and H. S. Colburn. "Role of spectral detail in sound-source localization." In: *Nature* 396.6713 (1998), pp. 747–749. ISSN: 0028-0836. DOI: 10.1038/25526.

[14] R. Baumgartner, D. K. Reed, B. Tóth, V. Best, P. Majdak, H. S. Colburn, and B. Shinn-Cunningham. "Asymmetries in behavioral and neural responses to spectral cues demonstrate the generality of auditory looming bias." In: *Proceedings of the National Academy of Sciences of the United States of America* 114.36 (2017), pp. 9743–9748. DOI: 10.1073/pnas.1703247114.

[15] H. G. Hassager, F. Gran, and T. Dau. "The role of spectral detail in the binaural transfer function on perceived externalization in a reverberant environment." In: *J. Acoust. Soc. Am.* 139.5 (2016), p. 2992. DOI: 10.1121/1.4950847.

[16] D. R. Begault, E. M. Wenzel, and M. R. Anderson. "Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source." In: *J. Audio Eng. Soc.* 49.10 (2001), pp. 904–916.

[17] R. Crawford-Emery and H. Lee. "The Subjective Effect of BRIR Length on Perceived Headphone Sound Externalization and Tonal Coloration." In: *136th Convention of the Audio Engineering Society* (2014).

[18] J. Catic, S. Santurette, and T. Dau. "The role of reverberation-related binaural cues in the externalization of speech." In: *J. Acoust. Soc. Am.* 138.2 (2015), pp. 1154–1167. DOI: 10.1121/1.4928132.

[19] T. Leclère, M. Lavandier, and F. Perrin. "On the externalization of sound sources with headphones without reference to a real source." In: *J. Acoust. Soc. Am.* 146.4 (2019), p. 2309. DOI: 10.1121/1.5128325.

[20] Z. R. Jiang, J. Q. Sang, C. S. Zheng, and X. D. Li. "The effect of pinna filtering in binaural transfer functions on externalization in a reverberant environment." In: *Applied Acoustics* 164 (2020), pp. 1–10. DOI: 10.1016/j.apacoust.2020.107257.

[21] W. O. Brimijoin, A. W. Boyd, and M. A. Akeroyd. "The Contribution of Head Movement to the Externalization and Internalization of Sounds." In: *Plos One* 8.12 (2013), e83068. DOI: 10.1371/journal.pone.0083068.

[22] A. J. Kolarik, B. C. J. Moore, P. Zahorik, S. Cirstea, and S. Pardhan. "Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss." In: *Attention, perception & psychophysics* 78.2 (2016), pp. 373–395. DOI: 10.3758/s13414-015-1015-1.

[23] B. G. Shinn-Cunningham, N. Kopco, and T. J. Martin. "Localizing nearby sound sources in a classroom: binaural room impulse responses." In: *J. Acoust. Soc. Am.* 117.5 (2005). DOI: 10.1121/1.1872572.

[24] J. Catic, S. Santurette, J. M. Buchholz, F. Gran, and T. Dau. "The effect of interaural-level-difference fluctuations on the externalization of sound." In: *J. Acoust. Soc. Am.* 134.2 (2013), pp. 1232–1241. DOI: 10.1121/1.4812264.

[25] H. Wallach. "The role of head movements and vestibular and visual cues in sound localization." In: *Journal of Experimental Psychology* 27.4 (1940), pp. 339–368. ISSN: 0022-1015. DOI: 10.1037/h0054629.

[26] F. L. Wightman and D. J. Kistler. "Resolution of front-back ambiguity in spatial hearing by listener and source movement." In: *J. Acoust. Soc. Am.* 105.5 (1999), pp. 2841–2853. DOI: 10.1121/1.426899.

[27] E. Hendrickx, P. Stitt, J.-C. Messonnier, J.-M. Lyzwa, B. F. G. Katz, and C. de Boishéraud. "Improvement of Externalization by Listener and Source Movement Using a "Binauralized" Microphone Array." In: *J. Audio Eng. Soc.* 65.7/8 (2017), pp. 589–599. DOI: 10.17743/jaes.2017.0018.

[28] E. Hendrickx, P. Stitt, J.-C. Messonnier, J.-M. Lyzwa, B. F. G. Katz, and C. de Boishéraud. "Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis." In: *J. Acoust. Soc. Am.* 141.3 (2017), pp. 2011–2023. DOI: 10.1121/1.4978612.

[29] S. Li, R. Schlieper, and J. Peissig. "The effect of variation of reverberation parameters in contralateral versus ipsilateral ear signals on perceived externalization of a lateral sound source in a listening room." In: *J. Acoust. Soc. Am.* 144.2 (2018), p. 966. DOI: 10.1121/1.5051632.

[30] S. Li, R. Schlieper, and J. Peissig. "The Role of Reverberation and Magnitude Spectra of Direct Parts in Contralateral and Ipsilateral Ear Signals on Perceived Externalization." In: *Applied Sciences* 9.3 (2019), p. 460. DOI: 10.3390/app9030460.

[31] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg. "A summary on acoustic room divergence and its effect on externalization of auditory events." In: *8th International Conference on Quality of Multimedia Experience (QoMEX)* (2016). DOI: 10.1109/QoMEX.2016.7498973.

[32] J. Udesen, T. Piechowiak, and F. Gran. "The Effect of Vision on Psychoacoustic Testing with Headphone-Based Virtual Sound." In: *J. Audio Eng. Soc.* 63.7/8 (2015), pp. 552–561. DOI: 10.17743/jaes.2015.0061.

[33] J. C. Gil-Carvajal, J. Cubick, S. Santurette, and T. Dau. "Spatial Hearing with Incongruent Visual or Auditory Room Cues." In: *Scientific Reports* 6.1 (2016), pp. 1–10. DOI: 10.1038/srep37342.

[34] G. Plenge. "On the problem of "In Head Localization"." In: *Acta Acustica united with Acustica* 26.5 (1972), pp. 241–252. ISSN: 16101928.

[35] C. Mendonça. "A review on auditory space adaptations to altered head-related cues." In: *Frontiers in Neuroscience* 8 (2014). ISSN: 1662-4548. DOI: 10.3389/fnins.2014.00219.

[36] R. Baumgartner and P. Majdak. "Predicting Externalization of Anechoic Sounds." In: *23rd International Congress on Acoustics* (2019).

[37] A. W. Boyd, W. M. Whitmer, J. J. Soraghan, and M. A. Akeroyd. "Auditory externalization in hearing-impaired listeners: the effect of pinna cues and number of talkers." In: *J. Acoust. Soc. Am.* 131.3 (2012), EL268–74. DOI: 10.1121/1.3687015.

[38] J. B. Allen and D. A. Berkley. "Image method for efficiently simulating small–room acoustics." In: *J. Acoust. Soc. Am.* 65.4 (1979), pp. 943–950. DOI: 10.1121/1.382599.

[39] J.-M. Jot and A. Chaigne. "Digital Delay Networks for Designing Artificial Reverberators." In: *90th Convention of the Audio Engineering Society* (1991).

[40] J. J. He, R. Ranjan, W.-S. Gan, N. K. Chaudhary, N. D. Hai, and R. Gupta. "Fast Continuous Measurement of HRTFs with Unconstrained Head Movements for 3D Audio." In: *J. Audio Eng. Soc.* 66.11 (2018), pp. 884–900.

[41] S. Li and J. Peissig. "Measurement of Head-Related Transfer Functions: A Review." In: *Applied Sciences* 10.14 (2020), p. 5014. DOI: 10.3390/app10145014.

[42] D. S. Brungart and W. M. Rabinowitz. "Auditory localization of nearby sources. Head-related transfer functions." In: *J. Acoust. Soc. Am.* 106.3 Pt 1 (1999), pp. 1465–1479. DOI: 10.1121/1.427180.

[43] K. Hartung, J. Braasch, and S. J. Sterbing. "Comparison of Different Methods for the Interpolation of Head-Related Transfer Functions." In: *16th International Conference: Spatial Sound Reproduction* (1999).

[44] V. Larcher, O. Warusfel, J.-M. Jot, and J. Guyard. "Study and Comparison of Efficient Methods for 3-D Audio Spatialization Based on Linear Decomposition of HRTF Data." In: *108th Convention of the Audio Engineering Society* (2000).

[45] Z. Ben-Hur, D. L. Alon, R. Mehra, and B. Rafaely. "Efficient Representation and Sparse Sampling of Head-Related Transfer Functions Using Phase-Correction Based on Ear Alignment." In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.12 (2019), pp. 2249–2262. ISSN: 2329-9304. DOI: 10.1109/TASLP.2019.2945479.

[46] A. Kan, C. Jin, and A. van Schaik. "A psychophysical evaluation of near-field head-related transfer functions synthesized using a distance variation function." In: *J. Acoust. Soc. Am.* 125.4 (2009), pp. 2233–2242. DOI: 10.1121/1.3081395.

[47] R. Duraiswaini, D. N. Zotkin, and N. A. Gumerov. "Interpolation and range extrapolation of HRTFs." In: *IEEE International Conference on Acoustics, Speech, and Signal Processing* (2004). DOI: 10.1109/ICASSP.2004.1326759.

[48] P. Minnaar, J. Plogsties, and F. Christensen. "Directional Resolution of Head-Related Transfer Functions Required in Binaural Synthesis." In: *J. Audio Eng. Soc.* 53.10 (2005), pp. 919–929.

[49] W. Zhang, M. Zhang, R. A. Kennedy, and T. D. Abhayapala. "On High-Resolution Head-Related Transfer Function Measurements: An Efficient Sampling Scheme." In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.2 (2012), pp. 575–584. DOI: 10.1109/TASL.2011.2162404.

[50] S. Li and J. Peissig. "Fast estimation of 2D individual HRTFs with arbitrary head movements." In: *22nd IEEE International Conference on Digital Signal Processing* (2017). DOI: 10.1109/ICDSP.2017.8096086.

[51] S. Li, A. Tobbala, and J. Peissig. "Towards Mobile 3D HRTF Measurement." In: *148th Convention of the Audio Engineering Society* (2020).

[52] M. Vorländer. *Auralization*. 1st. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. ISBN: 978-3-540-48829-3. DOI: 10.1007/978-3-540-48830-9.

[53] O. Kirkeby, P. A. Nelson, H. Hamada, and F. Orduna-Bustamante. "Fast deconvolution of multichannel systems using regularization." In: *IEEE Transactions on Speech and Audio Processing* 6.2 (1998), pp. 189–194. DOI: 10.1109/89.661479.

[54] S. Müller and P. Massarani. "Transfer-Function Measurement with Sweeps." In: *J. Audio Eng. Soc.* 49.6 (2001), pp. 443–471.

[55] S. Müller. "Measuring Transfer-Functions and Impulse Responses." In: *Handbook of signal processing in acoustics*. Ed. by D. I. Havelock, S. Kuwano, and M. Vorländer. New York: Springer, 2008, pp. 65–85. ISBN: 978-0-387-77698-9. DOI: 10.1007/978-0-387-30441-0_5.

[56] F. M. Wiener and D. A. Ross. "The Pressure Distribution in the Auditory Canal in a Progressive Sound Field." In: *J. Acoust. Soc. Am.* 18.2 (1946), pp. 401–408. DOI: 10.1121/1.1916378.

[57] S. Mehrgardt and V. Mellert. "Transformation characteristics of the external human ear." In: *J. Acoust. Soc. Am.* 61.6 (1977), pp. 1567–1576. DOI: 10.1121/1.381470.

[58] J. C. Middlebrooks, J. C. Makous, and D. M. Green. "Directional sensitivity of sound-pressure levels in the human ear canal." In: *J. Acoust. Soc. Am.* 86.1 (1989), pp. 89–108. DOI: 10.1121/1.398224.

[59] H. Møller. "Fundamentals of binaural technology." In: *Applied Acoustics* 36.3/4 (1992), pp. 171–218. DOI: 10.1016/0003-682X(92)90046-U.

[60] D. Hammershøi and H. Møller. "Sound transmission to and within the human ear canal." In: *J. Acoust. Soc. Am.* 100.1 (1996), pp. 408–427. DOI: 10.1121/1.415856.

[61] V. R. Algazi, C. Avendano, and D. Thompson. "Dependence of Subject and Measurement Position in Binaural Signal Acquisition." In: *J. Audio Eng. Soc.* 47.11 (1999), pp. 937–947.

[62] A. Farina. "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique." In: *108th Convention of the Audio Engineering Society* (2000).

[63] A. Farina. "Advancements in Impulse Response Measurements by Sine Sweeps." In: *122nd Convention of the Audio Engineering Society* (2007).

[64] B. S. Xie. *Head-related transfer function and virtual auditory display*. 2nd. Plantation, FL.: J. Ross Publishing, 2013. ISBN: 9781604270709.

[65] F. J. MacWilliams and N. J. A. Sloane. "Pseudo-random sequences and arrays." In: *Proceedings of the IEEE* 64.12 (1976), pp. 1715–1729. ISSN: 1558-2256. DOI: 10.1109/PROC.1976.10411.

[66] C. Dunn and M. J. Hawksford. "Distortion Immunity of MLS-Derived Impulse Response Measurements." In: *J. Audio Eng. Soc.* 41.5 (1993), pp. 314–335.

[67] G. Enzner, C. Antweiler, and S. Spors. "Trends in Acquisition of Individual Head-Related Transfer Functions." In: *The Technology of Binaural Listening*. Ed. by J. Blauert. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 57–92. ISBN: 978-3-642-37762-4. DOI: 10.1007/978-3-642-37762-4_3.

[68] G. D. Romigh, D. S. Brungart, R. M. Stern, and B. D. Simpson. "Efficient Real Spherical Harmonic Representation of Head-Related Transfer Functions." In: *IEEE Journal of Selected Topics in Signal Processing* 9.5 (2015), pp. 921–930. DOI: 10.1109/JSTSP.2015.2421876.

[69] F. Denk, S. M. A. Ernst, J. Heeren, S. D. Ewert, and B. Kollmeier. "The Oldenburg Hearing Device(OlHeaD) HRTF Database." Technical Report. University of Oldenburg, 2018.

[70] J.-G.t Richter and J. Fels. "On the Influence of Continuous Subject Rotation During High-Resolution Head-Related Transfer Function Measurements." In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.4 (2019), pp. 730–741. ISSN: 2329-9304. DOI: 10.1109/TASLP.2019.2894329.

[71] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney. "A Perceptual Evaluation of Individual and Non-Individual HRTFs: A Case Study of the SADIE II Database." In: *Applied Sciences* 8.11 (2018), p. 2029. DOI: 10.3390/app8112029.

[72] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen. "Transfer Characteristics of Headphones Measured on Human Ears." In: *J. Audio Eng. Soc.* 43.4 (1995), pp. 203–217.

[73] P. Bilinski, J. Ahrens, M. R. P. Thomas, I. J. Tashev, and J. C. Platt. "HRTF magnitude synthesis via sparse representation of anthropometric features." In: *IEEE International Conference on Acoustics, Speech and Signal Processing* (2014). DOI: 10.1109/ICASSP.2014.6854447.

[74] P. Majdak, P. Balazs, and B. Laback. "Multiple Exponential Sweep Method for Fast Measurement of Head-Related Transfer Functions." In: *J. Audio Eng. Soc.* 55.7/8 (2007), pp. 623–637.

[75] P. Dietrich. "Uncertainties in acoustical transfer functions : modeling, measurement and derivation of parameters for airborne and structure-borne sound." phD thesis. RWTH Aachen University, 2013.

[76] S. Haykin. *Adaptive filter theory*. 4. ed., international ed. Prentice Hall informations and system sciences series. Upper Saddle River, NJ: Prentice Hall, 2002. ISBN: 0130484342.

[77] G. Enzner. "Analysis and optimal control of LMS-type adaptive filtering for continuous-azimuth acquisition of head related impulse responses." In: *IEEE International Conference on Acoustics, Speech and Signal Processing* (2008). DOI: 10.1109/ICASSP.2008.4517629.

[78] C. Antweiler, A. Telle, P. Vary, and G. Enzner. "Perfect-sweep NLMS for time-variant acoustic system identification." In: *IEEE International Conference on Acoustics, Speech and Signal Processing* (2012). DOI: 10.1109/ICASSP.2012.6287930.

[79] H. D. Lüke. "Sequences and arrays with perfect periodic correlation." In: *IEEE Transactions on Aerospace and Electronic Systems* 24.3 (1988), pp. 287–294. ISSN: 1557-9603. DOI: 10.1109/7.192096.

[80] D. Jungnickel and A. Pott. "Perfect and almost perfect sequences." In: *Discrete Applied Mathematics* 95.1-3 (1999), pp. 331–359. ISSN: 0166218X. DOI: 10.1016/S0166-218X(99)00085-2.

[81] G. Enzner. "3D-continuous-azimuth acquisition of head-related impulse responses using multi-channel adaptive filtering." In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (2009). DOI: 10.1109/ASPAA.2009.5346532.

[82] G. Z. Yu and B. S. Xie. "Multiple Sound Sources Solution for Near-Field Head-Related Transfer Function Measurements." In: *AES International Conference on Audio for Virtual and Augmented Reality* (2018).

[83] R. Ranjan, J. J. He, and W.-S. Gan. "Fast Continuous Acquisition of HRTF for Human Subjects with Unconstrained Random Head Movements in Azimuth and Elevation." In: *AES International Conference on Headphone Technology* (2016).

[84] T. Aboulnasr and K. Mayyas. "A robust variable step-size LMS-type algorithm: analysis and simulations." In: *IEEE Transactions on Signal Processing* 45.3 (1997), pp. 631–639. ISSN: 1941-0476. DOI: 10.1109/78.558478.

[85] C. K. Correa, S. Li, and J. Peissig. "Analysis and Comparison of different Adaptive Filtering Algorithms for Fast Continuous HRTF Measurement." In: *Tagungsband Fortschritte der Akustik-DAGA* (2017).

[86] K. Fukudome, T. Suetsugu, T. Ueshin, R. Idegami, and K. Takeya. "The fast measurement of head related impulse responses for all azimuthal directions using the continuous measurement method with a servo-swiveled chair." In: *Applied Acoustics* 68.8 (2007), pp. 864–884. DOI: 10.1016/j.apacoust.2006.09.009.

[87] V. Pulkki, M.-V. Laitinen, and V. Sivonen. "HRTF Measurements with a Continuously Moving Loudspeaker and Swept Sines." In: *128th Convention of the Audio Engineering Society* (2010).

[88] J. Reijniers, B. Partoens, J. Steckel, and H. Peremans. "HRTF measurement by means of unsupervised head movements with respect to a single fixed speaker." In: *IEEE Access* 8 (2020), pp. 92287–92300. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2994932.

[89] D. Heer, F. de Mey, J. Reijniers, S. Demeyer, and H. Peremans. "Evaluating Intermittent and Concurrent Feedback during an HRTF Measurement." In: *AES International Conference on Headphone Technology* (2019).

[90] S. Peksi, N. D. Hai, R. Ranjan, R. Gupta, J. J. He, and W.-S. Gan. "A Unity Based Platform for Individualized HRTF Research and Development: From On-the-Fly Fast Acquisition to Spatial Audio Renderer." In: *AES International Conference on Headphone Technology* (2019).

[91] D. N. Zotkin, R. Duraiswami, E. Grassi, and N. A. Gumerov. "Fast head-related transfer function measurement via reciprocity." In: *J. Acoust. Soc. Am.* 120.4 (2006), pp. 2202–2215. DOI: 10.1121/1.2207578.

[92] P. M. C. Morse and K. U. Ingard. *Theoretical acoustics*. Repr., 1. Princeton University Press ed. Princeton, NJ: Princeton University Press, 1986. ISBN: 0691024014.

[93] Q. H. Ye, Q. J. Dong, Y. Zhang, and X. D. Li. "Fast Head-Related Transfer Function Measurement in Complex Environments." In: *20th International Congress on Acoustics* (2010).

[94] F. Denk, B. Kollmeier, and S. D. Ewert. "Removing Reflections in Semi-anechoic Impulse Responses by Frequency-Dependent Truncation." In: *J. Audio Eng. Soc.* 66.3 (2018), pp. 146–153.

[95] J. J. He, R. Gupta, R. Ranjan, and W.-S. Gan. "Non-Invasive Parametric HRTF Measurement for Human Subjects Using Binaural and Ambisonic Recording of Existing Sound Field." In: *AES International Conference on Headphone Technology* (2019).

[96] J. J. Lopez, S. Martinez-Sanchez, and P. Gutierrez-Parera. "Array processing for echo cancellation in the measurement of Head-Related Transfer Functions." In: *Euronoise* (2018).

[97] R. F. Lyon. "All-pole models of auditory filtering." In: *Diversity in Auditory Mechanics*. Ed. by E. R. Lewis, G. R. Long, R. F. Lyon, P. M. Narins, C. R. Steele, and E. Hecht-Poinar. World Scientific Publishing, 1997.

[98] B. R. Glasberg and B. C. J. Moore. "Derivation of auditory filter shapes from notched-noise data." In: *Hearing Research* 47.1 (1990), pp. 103–138. ISSN: 0378-5955. DOI: 10.1016/0378-5955(90)90170-T.

[99] T. Dau, B. Kollmeier, and A. Kohlrausch. "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers." In: *J. Acoust. Soc. Am.* 102.5 Pt 1 (1997), pp. 2892–2905. DOI: 10.1121/1.420344.

[100] G. F. Kuhn. "Model for the interaural time differences in the azimuthal plane." In: *J. Acoust. Soc. Am.* 62.1 (1977), pp. 157–167.

[101] D. J. Kistler and F. L. Wightman. "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction." In: *J. Acoust. Soc. Am.* 91.3 (1992), pp. 1637–1647. DOI: 10.1121/1.402444.

[102] A. Andreopoulou and B. F. G. Katz. "Identification of perceptually relevant methods of inter-aural time difference estimation." In: *J. Acoust. Soc. Am.* 142.2 (2017), p. 588. DOI: 10.1121/1.4996457.

[103] J. M. Arend, A. Neidhardt, and C. Pörschmann. "Measurement and Perceptual Evaluation of a Spherical Near-Field HRTF Set." In: *29th Tonmeistertagung - VDT International Convention* (2016).

[104] J. M. Arend and C. Pörschmann. "Synthesis of Near-Field HRTFs by Directional Equalization of Far-Field Datasets." In: *Tagungsband Fortschritte der Akustik-DAGA* (2019).

[105] M. Cuevas-Rodriguez, D. L. Alon, P. W. Robinson, and R. Mehra. "Evaluation of the effect of head-mounted display on individualized head-related transfer functions." In: *23rd International Congress on Acoustics* (2019).

[106] S. Klockgether and S. van de Par. "Just noticeable differences of spatial cues in echoic and anechoic acoustical environments." In: *J. Acoust. Soc. Am.* 140.4 (2016), EL352. DOI: 10.1121/1.4964844.

[107] R. Braun, S. Li, and J. Peissig. "A Measurement System for Fast Estimation of 2D Individual HRTFs with Arbitrary Head Movements." In: *4th International Conference on Spatial Audio* (2017).

[108] K. Diepold, M. Durkovic, and F. Sagstetter. "HRTF Measurements with Recorded Reference Signal." In: *129th Convention of the Audio Engineering Society* (2010).

[109] S. Li, J. X. E, R. Schlieper, and J. Peissig. "The Impact of Trajectories of Head and Source Movements on Perceived Externalization of a Frontal Sound Source." In: *144th Convention of the Audio Engineering Society* (2018).

[110] S. Li, R. Schlieper, and J. Peissig. "The Impact of Head Movement on Perceived Externalization of a Virtual Sound Source with Different BRIR Lengths." In: *AES International Conference on Immersive and Interactive Audio* (2019).

[111] R. Hupke, M. Nophut, S. Li, R. Schlieper, S. Preihs, and J. Peissig. "The Immersive Media Laboratory: Installation of a Novel Multichannel Audio Laboratory for Immersive Media Applications." In: *144th Convention of the Audio Engineering Society* (2018).

[112] Z. Schärer and A. Lindau. "Evaluation of Equalization Methods for Binaural Signals." In: *126th Convention of the Audio Engineering Society* (2009).

[113] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl. "A Spatial Audio Quality Inventory (SAQI)." In: *Acta Acustica united with Acustica* 100.5 (2014), pp. 984–994. ISSN: 16101928. DOI: 10.3813/AAA.918778.

[114] R. McGill, J. W. Tukey, and W. A. Larsen. "Variations of Box Plots." In: *The American Statistician* 32.1 (1978), p. 12. ISSN: 00031305. DOI: 10.2307/2683468.

[115] C. Faller and J. Merimaa. "Source localization in complex listening situations: selection of binaural cues based on interaural coherence." In: *J. Acoust. Soc. Am.* 116.5 (2004), pp. 3075–3089. DOI: 10.1121/1.1791872.

[116] EBU. *3253: Sound Quality assessment material Recordings for subjective tests*. 2008. URL: https://tech.ebu.ch/publications/sqamcd.

[117] D. R. Begault and E. M. Wenzel. "Headphone localization of speech." In: *Human factors* 35.2 (1993), pp. 361–376. ISSN: 0018-7208. DOI: 10.1177/001872089303500210.

[118] S.-M. Kim and W. Choi. "On the externalization of virtual sound images in headphone reproduction: a Wiener filter approach." In: *J. Acoust. Soc. Am.* 117.6 (2005), pp. 3657–3665. DOI: 10.1121/1.1921548.

[119] B. I. Băcilă and H. Lee. "Listener-Position and Orientation Dependency of Auditory Perception in an Enclosed Space: Elicitation of Salient Attributes." In: *Applied Sciences* 11.4 (2021), p. 1570. DOI: 10.3390/app11041570.

[120] S. van de Par and A. Kohlrausch. "Dependence of binaural masking level differences on center frequency, masker bandwidth, and interaural parameters." In: *J. Acoust. Soc. Am.* 106.4 Pt 1 (1999), pp. 1940–1947. DOI: 10.1121/1.427942.

[121] R. L. Freyman, K. S. Helfer, D. D. McCall, and R. K. Clifton. "The role of perceived spatial separation in the unmasking of speech." In: *J. Acoust. Soc. Am.* 106.6 (1999), pp. 3578–3588. DOI: 10.1121/1.428211.

[122] N. I. Durlach. "Equalization and Cancellation Theory of Binaural Masking–Level Differences." In: *J. Acoust. Soc. Am.* 35.8 (1963), pp. 1206–1218. DOI: 10.1121/1.1918675.

[123] S. Li, R. Baumgartner, and J. Peissig. "Modeling perceived externalization of a static, lateral sound image." In: *Acta Acust.* 4.5 (2020), p. 21. DOI: 10.1051/aacus/2020020.

[124] R. Baumgartner and P. Majdak. "Decision making in auditory externalization perception." In: *bioRxiv preprint* (2020). DOI: 10.1101/2020.04.30.068817.

[125] A. Kulkarni, S. K. Isabelle, and H. S. Colburn. "On the minimum-phase approximation of head-related transfer functions." In: *IEEE Workshop on Applications of Signal Processing to Audio and Accoustics* (1995). DOI: 10.1109/ASPAA.1995.482964.

[126] R. Baumgartner, P. Majdak, and B. Laback. "Modeling sound-source localization in sagittal planes for human listeners." In: *J. Acoust. Soc. Am.* 136.2 (2014), p. 791. DOI: 10.1121/1.4887447.

[127] F. L. Wightman and D. J. Kistler. "Monaural sound localization revisited." In: *J. Acoust. Soc. Am.* 101.2 (1997), pp. 1050–1063. DOI: 10.1121/1.418029.

[128] A. Kohlrausch and J. Breebaart. "Perceptual (ir)relevance of HRTF magnitude and phase spectra." In: *110th Convention of the Audio Engineering Society* (2001).

[129] M. Morimoto. "The contribution of two ears to the perception of vertical angle in sagittal planes." In: *J. Acoust. Soc. Am.* 109.4 (2001), pp. 1596–1603. DOI: 10.1121/1.1352084.

[130] E. A. Macpherson and A. T. Sabin. "Binaural weighting of monaural spectral cues for sound localization." In: *J. Acoust. Soc. Am.* 121.6 (2007), p. 3677. DOI: 10.1121/1.2722048.

[131] F. Klein, S. Werner, and T. Mayenfels. "Influences of Training on Externalization of Binaural Synthesis in Situations of Room Divergence." In: *J. Audio Eng. Soc.* 65.3 (2017), pp. 178–187. DOI: 10.17743/jaes.2016.0072.

[132] LISTEN HRTF Database. 2002. URL: http://recherche.ircam.fr/equipes/salles/listen/.

[133] B. Bernschütz. "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU100." In: *Tagungsband Fortschritte der Akustik-DAGA* (2013).

[134] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman. "Localization using nonindividualized head-related transfer functions." In: *J. Acoust. Soc. Am.* 94.1 (1993), pp. 111–123. DOI: 10.1121/1.407089.

[135] F. Völk, F. Heinemann, and H. Fastl. "Externalization in binaural synthesis: effects of recording environment and measurement procedure." In: *J. Acoust. Soc. Am.* 123.5 (2008), p. 3935. DOI: 10.1121/1.2936001.

[136] S. Li, R. Schlieper, and J. Peissig. "Externalization Enhancement for Headphone-Reproduced Virtual Frontal and Rear Sound Images." In: *AES International Conference on Headphone Technology* (2019).

[137] Y. G. Yuan, Z. H. Fu, M. Xu, L. Xie, and Q. Cong. "Externalization improvement in a real-time binaural sound image rendering system." In: *International Conference on Orange Technologies (ICOT)* (2015), pp. 165–168. DOI: 10.1109/ICOT.2015.7498514.

[138] R. Kronland-Martinet and T. Voinier. "Real-Time Perceptual Simulation of Moving Sources: Application to the Leslie Cabinet and 3D Sound Immersion." In: *EURASIP Journal on Audio, Speech, and Music Processing* 2008 (2008), pp. 1–10. ISSN: 1687-4714. DOI: 10.1155/2008/849696.

[139] G. S. Kendall. "The decorrelation of audio signals and its impact on spatial imagery." In: *Computer Music Journal* 19.4 (1995), pp. 71–87. ISSN: 01489267. DOI: 10.2307/3680992.

[140] A. Lindau and S. Weinzierl. "On the spatial resolution of virtual acoustic environments for head movements in horizontal, vertical, and lateral direction." In: *European Acoustics Association (EAA) Symposium on Auralization* (2009).

[141] R. Wallis and H. Lee. "Directional bands revisited." In: *138th Convention of the Audio Engineering Society* (2015).

[142] J. Hebrank and D. Wright. "Spectral cues used in the localization of sound sources on the median plane." In: *J. Acoust. Soc. Am.* 56.6 (1974), pp. 1829–1834. DOI: 10.1121/1.1903520.

[143] S.-N. Yao and L. J. Chen. "HRTF adjustments with audio quality assessments." In: *Archives of Acoustics* 38.1 (2013), pp. 55–62.

[144] K. Iida. "Measurement Method for HRTF." In: *Head-related transfer function and acoustic virtual reality*. Ed. by K. Iida. Singapore: Springer, 2019, pp. 149–156. ISBN: 978-981-13-9744-8. DOI: 10.1007/978-981-13-9745-5_9.

[145] U. Zölzer, ed. *DAFX: Digital audio effects*. Second edition. Chichester: Wiley, 2011. ISBN: 0470665998. DOI: 10.1002/9781119991298.

[146] M. Bouéri and C. Kyriakakis. "Audio signal decorrelation based on a critical band approach." In: *117th Convention of the Audio Engineering Society* (2004).

[147] M. Woirgard, P. Stade, J. Amankwor, B. Bernschütz, and J. Arend. *Cologne University of Applied Sciences - Anechoic Recordings*. 2012. URL: http://www.audiogroup.web.fh-koeln.de.

[148] Y. G. Kim, C. J. Chun, H. K. Kim, Y. J. Lee, D. Y. Jang, and K. Kang. "An Integrated Approach of 3D Sound Rendering Techniques for Sound Externalization." In: *Pacific-Rim Conference on Multimedia* (2010), pp. 682–693. DOI: 10.1007/978-3-642-15696-0_63.

[149] C. P. Brown and R. O. Duda. "An efficient HRTF model for 3-D sound." In: *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics* (1997), p. 4. DOI: 10.1109/ASPAA.1997.625596.

[150] C. Guezenoc and R. Seguier. "HRTF Individualization: A Survey." In: *145th Convention of the Audio Engineering Society* (2018).

LIST OF FIGURES

LIST OF TABLES

CURRICULUM VITAE

Name	Song Li
Day of birth	08 December 1989

Education

03/2016 to present	**Ph.D.** student Leibniz Universität Hannover Thesis Title: On externalization of virtual sound images presented via headphones
10/2013 to 12/2015	**M.Sc.** in Electrical and Computer Engineering Leibniz Universität Hannover Thesis Title: Impact of Doppler effect, echo and reverberation on the externalization and plausibility of binaurally rendered moving sound sources presented via headphones
09/2010 to 07/2013	**B.Eng.** in Electrical and Computer Engineering University of Applied Sciences and Arts Hannover Thesis Title: Development and testing of a sensor system for motion analysis
09/2008 to 08/2010	**B.Eng.** in Electrical and Computer Engineering Zhejiang University of Science and Technology, China Since 09/2010 as a program student at the University of Applied Sciences and Arts Hannover

Work Experience

03/2016 to present	**Institute of Communications Technology**, Leibniz Universität Hannover *Research Assistant*
09/2020 to 12/2020	**Facebook Reality Lab (FRL)**, Remote *PhD Research internship*
10/2014 to 03/2015	**Daimler AG**, Ulm, Germany *Internship*
06/2014 to 09/2014	**Institute of Production Engineering and Machine Tools**, Leibniz Universität Hannover *Student Assistant*
01/2013 to 06/2013	**N-Transfer GmbH / ITI**, Hannover, Germany *Internship and Bachelor thesis*

Teaching Experience

03/2016 to present	**Leibniz Universität Hannover**, Faculty of Electrical Engineering and Computer Science
Academic year of 03/2016 to present	Exercises on "Fundamentals of acoustics"
Academic year of 2014 to 2016	Laboratory on "Audio Communication and Acoustics"

PUBLICATIONS

Patent:

Liyun Pang, Fons Adriaensen, Roman Schlieper, and Song Li (2020). "Method and apparatus for processing an audio signal based on equalization filter," WO Patent WO2020164746A1

Liyun Pang, Fons Adriaensen, Roman Schlieper, and Song Li (2020). "System and method for evaluating an acoustic characteristic of an electronic device," WO Patent WO2020177845A1.

Liyun Pang, Fons Adriaensen, Song Li, and Roman Schlieper (2020). "Device and Method for Adaptation of Virtual 3D Audio to a Real Room," WO Patent WO2020057727A1.

Liyun Pang, Fons Adriaensen, Song Li, and Roman Schlieper (2020). "Method and Apparatus for Processing a Stereo Signal," WO Patent WO2020151837.

Journal:

Song Li and Jürgen Peissig (2020). "Measurement of Head-Related Transfer Functions: A Review." In: *Applied Sciences*, 2020, 10(14), 5014. DOI: 10.3390/app10145014.

Song Li, Robert Baumgartner and Jürgen Peissig (2020). "Modeling perceived externalization of a static, lateral sound image." In: *Acta Acustica*, 2020, 4, 21. DOI: 10.1051/aacus/2020020.

Song Li, Roman Schlieper, and Jürgen Peissig (2019). "The Role of Reverberation and Magnitude Spectra of Direct Parts in Contralateral and Ipsilateral Ear Signals on Perceived Externalization." In: *Applied Sciences*, 2019, 9(3), 460. DOI: 10.3390/app9030460.

Song Li, Roman Schlieper, and Jürgen Peissig (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." In: *The Journal of the Acoustical Society of America (JASA)*, 144(2), 966-980. DOI: 10.1121/1.5051632.

Conference paper:

Song Li, Aly Tobbala, and Jürgen Peissig (2020). "Towards Mobile 3D HRTF Measurement." In: *148th Convention of the Audio Engineering Society, Online Virtual Conference.*

Song Li, Roman Schlieper, and Jürgen Peissig (2019). "The Impact of Head Movement on Perceived Externalization of a Virtual Sound Source with Different BRIR Lengths." In: *2019 AES International Conference on Immersive and Interactive Audio, York, UK.*

Song Li, Roman Schlieper, and Jürgen Peissig (2019). "A Hybrid Method for Blind Estimation of Frequency Dependent Reverberation Time Using Speech Signals." In: *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK. DOI: 10.1109/ICASSP.2019.8682661.*

Song Li, Roman Schlieper, and Jürgen Peissig (2019). "Externalization Enhancement for Headphone-Reproduced Virtual Frontal and Rear Sound Images." In: *2019 AES International Conference on Headphone Technology, San Francisco, CA, USA.*

Roman Schlieper, Song Li, Stephan Preihs, and Jürgen Peissig (2019). "The Effect of Active Noise Cancellation on the Acoustic Impedance of Headphones." In: *2019 AES International Conference on Headphone Technology, San Francisco, CA, USA.*

Roman Schlieper, Song Li, Stephan Preihs, and Jürgen Peissig (2019). "The Relationship between the Acoustic Impedance of Headphones and the Occlusion Effect, 2019 AES International Conference on Headphone Technology." In: *2019 AES International Conference on Headphone Technology, San Francisco, CA, USA.*

Song Li, Roman Schlieper, Stephan Preihs, and Jürgen Peissig (2019). "The Relative Influence of Reverberation at the Contralateral versus Ipsilateral Ear on Perceived Externalization of a Lateral Sound Source." In: *Fortschritte der*

Akustik - DAGA, Rostock, Germany.

Robert Hupke, Marcel Nophut, Song Li, Roman Schlieper, Stephan Preihs, and Jürgen Peissig (2018). "The Immersive Media Laboratory: Installation of a Novel Multichannel Audio Laboratory for Immersive Media Applications." In: *144th Convention of the Audio Engineering Society, Milan, Italy.*

Roman Schlieper, Song Li, and Jürgen Peissig (2018). "Development and Validation of a Full Range Acoustic Impedance Tube." In: *144th Convention of the Audio Engineering Society, Milan, Italy.*

Roman Schlieper, Song Li, Stephan Preihs, and Jürgen Peissig (2018). "Estimation of the Headphone "Openness" Based on Measurements of Pressure Division Ratio, Headphone Selection Criterion, and Acoustic Impedance." In: *145th Convention of the Audio Engineering Society, New York, NY, USA.*

Song Li, Jiaxiang E, Roman Schlieper, and Jürgen Peissig (2018). "The Impact of Trajectories of Head and Source Movements on Perceived Externalization of a Frontal Sound Source." In: *144th Convention of the Audio Engineering Society, Milan, Italy.*

Camilo Klinkert Correa, Song Li and Jürgen Peissig (2017). "Analysis and Comparison of different Adaptive Filtering Algorithms for Fast Continuous HRTF Measurement." In: *Fortschritte der Akustik – DAGA, Kiel, Germany.*

Ruben Braun, Song Li and Jürgen Peissig (2017). "A Measurement System for Fast Estimation of 2D Individual HRTFs with Arbitrary Head Movements." In: *4th International Conference on Spatial Audio, Graz, Austria.*

Song Li and Jürgen Peissig (2017). "Fast Estimation of 2D Individual HRTFs with Arbitrary Head Movements." In: *22nd International Conference on Digital Signal Processing (DSP), London, UK. DOI: 10.1109/ICDSP.2017.8096086.*

Song Li, Sanam Moghaddamnia and Jürgen Peissig (2016). "Impact of Doppler Effect, Echo and Reverberation on the Externalization and Plausibility of Binaural Rendered Moving Sound Sources Presented via Headphones." In: *Fortschritte der Akustik – DAGA, Aachen, Germany.*