OPTIMIZATION AND EVALUATION OF A VIRTUAL ARTIFICIAL HEAD FOR INDIVIDUAL DYNAMIC SPATIAL SOUND REPRODUCTION OVER HEADPHONES

von der Fakultät für Medizin und Gesundheitswissenschaften
der Carl von Ossietzky Universität Oldenburg
zur Erlangung des Grades und Titels eines
Doktors der Ingenieurwissenschaften (Dr.-Ing.)
angenommene Dissertation

von
Mina Fallahi
geboren am 28. Juni 1981
in Teheran (Iran)

Mina Fallahi: *Optimization and Evaluation of a Virtual Artificial Head for Individual Dynamic Spatial Sound Reproduction over Headphones*

ERSTGUTACHTER:
Prof. Dr. ir. Simon Doclo, *University of Oldenburg, Germany*

WEITERE GUTACHTER:
Prof. Dr.-Ing. Matthias Blau, *Jade University of Applied Sciences, Oldenburg, Germany*
Doz. Dr. Piotr Majdak, *Acoustics Research Institute of the Austrian Academy of Sciences, Vienna, Austria*

TAG DER DISPUTATION:
27. September 2021

ABSTRACT

The ability of humans to perceive sound spatially is based on binaural hearing, i.e. on signals arriving at the two ears which supply the listener with important spatial and spectral cues. The aim of binaural technology is to capture and reproduce the sound field in such a way that these cues are preserved. A well-known drawback of using artificial heads for this aim is that they exhibit different anthropometrical measures compared to individual listeners. When playing back the recorded signals over headphones, the non-individual design of artificial heads may lead to localization ambiguities such as front-back reversals and perception inside the head. Moreover, it is hardly possible to achieve dynamic signal playback, accounting for the listener's head movements. As an alternative, it has been proposed to use a Virtual Artificial Head (VAH), which is a microphone array where spectral weights are applied to the microphone signals, aiming at synthesizing the directivity pattern of Head Related Transfer Functions (HRTFs). By adjusting the spectral weights to HRTFs of individual listeners, the signals recorded with a VAH can be individualized *post-hoc* for different listeners. In addition, the spectral weights can be adapted to account for the listener's head movements during signal playback.

The aim of this thesis is to improve the performance of a state-of-the-art VAH approach for synthesizing individual HRTF directivity patterns and to evaluate it for situations which have not been considered before. The first focus is to improve the horizontal spatial resolution of the VAH synthesis using a limited number of microphones. The second focus is to investigate the impact of the microphone array topology on the VAH performance in the horizontal plane. The third focus is to evaluate the VAH approach in dynamic auralizations for horizontal and non-horizontal sources, both in anechoic as well as in reverberant environments.

First, we propose a new constrained optimization method to calculate the spectral weights, which allows to increase the spatial resolution of the VAH synthesis in the horizontal plane using a limited number of microphones. In addition to imposing a constraint on the mean White Noise Gain (WNG) to increase robustness, we propose to impose constraints on the monaural spectral error, referred to as spectral distortion, at a high number of directions. For a simulated planar microphone array with 24 microphones, we show that the frequency range, for which the synthesis accuracy can be considered acceptable, can be increased from 2 kHz to 5 kHz compared to imposing only the mean WNG constraint. The VAH synthesis with the additional spectral distortion constraints is also shown to perceptually outperform the synthesis where only the mean WNG constraint is imposed. Second, based on simulations with four different microphone array topologies, we investigate the impact of array extension and microphone distribution on the VAH performance. While smaller inter-microphone distances enable to satisfy the spectral distortion

constraints at higher frequencies, they may cause difficulties in satisfying the mean WNG constraint at low and mid-frequency ranges. For an array topology combining dense and sparse inter-microphone distances, we show that the mean WNG and spectral distortion constraints can be satisfied for frequencies up to 8 kHz without deteriorating the phase accuracy at low frequencies. In addition, the binaural signals generated using the mixed array topology result in the best perceptual ratings compared to the other considered topologies, which result in either more high-frequency spectral distortion or more low-frequency phase inaccuracy. Third, we investigate the performance of the VAH approach for dynamic auralizations with speech signals in two studies, both considering sources in and outside the horizontal plane. Individual Binaural Room Impulse Responses (BRIRs) for different head orientations are synthesized for two VAHs, i.e. a planar array with 24 microphones and a three-dimensional array with 31 microphones. In the first study, we evaluate dynamic auralizations with the synthesized BRIRs for the VAH with 24 microphones in comparison to real (visible) sound source presentations. We show that both in a reverberant as well as in an anechoic environment close-to-reality dynamic auralizations with speech signals can be achieved. In the second study, we evaluate the localization performance of virtual sources generated with both VAHs in the absence of visual cues and in comparison to real hidden sound sources. We show that even in the absence of visual cues, virtual sources generated with both VAHs can be localized with a similar accuracy with respect to azimuth, externalization and the occurrence of front-back reversals as real sources. Interestingly, including only horizontal directions in the calculation of the spectral weights results in a better localization performance compared to including horizontal and non-horizontal directions. Moreover, localization experiments with and without head tracking show the importance of the dynamic presentation on the localization accuracy of virtual sound sources generated with the VAHs. Although individualization is an important capability of the VAH approach, both studies show that the possibility of presenting binaural signals dynamically is the main advantage of the VAH approach over conventional artificial heads.

ZUSAMMENFASSUNG

Die Fähigkeit des Menschen zur räumlichen Schallwahrnehmung basiert auf dem binauralen Hören, d.h. auf Signalen, die an beiden Ohren ankommen und dem Zuhörer wichtige räumliche und spektrale Informationen liefern. Ziel der Binauraltechnik ist es, das Schallfeld so aufzunehmen und wiederzugeben, dass diese Informationen erhalten bleiben. Ein bekannter Nachteil der Verwendung von Kunstköpfen für dieses Ziel besteht darin, dass sie im Vergleich zu individuellen Zuhörern unterschiedliche anthropometrische Maße aufweisen. Bei der Wiedergabe der aufgezeichneten Signale über Kopfhörer kann es durch das nicht-individuelle Design der Kunstköpfe zu Lokalisationsfehlern wie Vorne-Hinten-Vertauschungen und im-Kopf Lokalisation kommen. Außerdem ist eine dynamische Signalwiedergabe zur Kompensation der Kopfbewegungen des Zuhörers in der Regel nicht möglich. Alternativ wurde vorgeschlagen, einen virtuellen Kunstkopf (englisch: Virtual Artificial Head (VAH)) zu verwenden. Beim VAH handelt es sich um ein Mikrofonarray, bei dem Spektralgewichte auf die Mikrofonsignale angewendet werden, um die Richtcharakteristiken der kopfbezogenen Übertragungsfunktionen (englisch: Head Related Transfer Functions (HRTFs)) zu synthetisieren. Durch Anpassen der Spektralgewichte an HRTFs einzelner Zuhörer können die mit einem VAH aufgezeichneten Signale im Nachhinein für verschiedene Zuhörer individualisiert werden. Zusätzlich können die Spektralgewichte angepasst werden, um die Kopfbewegungen des Zuhörers während der Signalwiedergabe zu kompensieren.

Ziel dieser Arbeit ist es, die Performance eines modernen VAH-Ansatzes für die Synthese individueller HRTF-Richtcharakteristiken zu verbessern und für bisher nicht berücksichtigte Situationen zu evaluieren. Der erste Schwerpunkt besteht darin, die horizontale räumliche Auflösung der VAH-Synthese mit einer begrenzten Anzahl von Mikrofonen zu verbessern. Der zweite Schwerpunkt besteht darin, den Einfluss der Mikrofonarray-Topologie auf die VAH-Performance in der Horizontalebene zu untersuchen. Der dritte Schwerpunkt ist die Evaluierung des VAH-Ansatzes in dynamischen Auralisationen für horizontale und nicht-horizontale Quellen, sowohl in reflexionsarmen als auch in halligen Umgebungen.

Zunächst wird eine neue Optimierungsmethode zur Berechnung der Spektralgewichte vorgeschlagen, die es ermöglicht, die räumliche Auflösung der VAH-Synthese in der Horizontalebene mit einer begrenzten Anzahl von Mikrofonen zu erhöhen. Zusätzlich zur bereits bisher benutzten Nebenbedingung für den gemittelten White Noise Gain (WNG) zur Erhöhung der Robustheit wird vorgeschlagen, den monauralen Spektralfehler in einer hohen Anzahl von Richtungen zu beschränken. Verglichen zu dem Fall, dass nur der gemittelte WNG beschränkt wird, kann für ein simuliertes planares Mikrofonarray mit 24 Mikrofonen gezeigt werden, dass der Frequenzbereich, für den die Synthesegenauigkeit als akzeptabel angenommen wird,

von 2 kHz auf 5 kHz erhöht wird. Es wird auch gezeigt, dass die VAH-Synthese mit den zusätzlichen Nebenbedingungen für den Spektralfehler die Synthese mit ausschließlicher Beschränkung des gemittelten WNG perzeptiv übertrifft. Zweitens wird, basierend auf Simulationen mit vier verschiedenen Mikrofonarray-Topologien der Einfluss der Array-Ausdehnung und der Mikrofonverteilung auf die VAH-Performance untersucht. Während kleinere Mikrofonabstände es ermöglichen, die Nebenbedingungen für den Spektralfehler bei höheren Frequenzen zu erfüllen, können sie Schwierigkeiten bei der Erfüllung der gemittelten WNG-Nebenbedingung bei niedrigen und mittleren Frequenzen verursachen. Für eine Mikrofonarray-Topologie, die dichte und spärliche Abstände zwischen den Mikrofonen kombiniert, kann gezeigt werden, dass die gemittelte WNG-Nebenbedingung und die Nebenbedingungen für den Spektralfehler für Frequenzen bis zu 8 kHz erfüllt werden, ohne die Phasengenauigkeit bei tiefen Frequenzen zu verschlechtern. Darüber hinaus führen die unter Verwendung der kombinierenden Mikrofonarray-Topologie erzeugten binauralen Signale zu den besten Wahrnehmungsergebnissen im Vergleich zu den anderen betrachteten Topologien, die entweder zu erhöhten hochfrequenten Spektralfehlern oder zu erhöhten niederfrequenten Phasenungenauigkeiten führen. Drittens wird die Performance des VAH-Ansatzes für dynamische Auralisationen mit Sprachsignalen in zwei Studien mit Quellen in und außerhalb der Horizontalebene untersucht. Für zwei VAHs, ein planares Array mit 24 Mikrofonen und ein dreidimensionales Array mit 31 Mikrofonen, werden dazu individuelle binaurale Raumimpulsantworten (englisch: Binaural Room Impulse Responses (BRIRs)) für verschiedene Kopforientierungen synthetisiert. In der ersten Studie werden dynamische Auralisationen mit den synthetisierten BRIRs für den VAH mit 24 Mikrofonen im Vergleich zu realen (sichtbaren) Schallquellen evaluiert. Es wird gezeigt, dass sowohl in einer halligen als auch in einer reflexionsarmen Umgebung realitätsnahe dynamische Auralisationen mit Sprachsignalen erreicht werden können. In der zweiten Studie wird die Lokalisierungsperformance von virtuellen Quellen, generiert mit beiden VAHs im Vergleich zu echten versteckten Schallquellen und unter Abwesenheit visueller Cues, evaluiert. Es wird gezeigt, dass virtuelle Quellen, generiert mit beiden VAHs, auch ohne visuelle Informationen mit ähnlicher Genauigkeit wie reale Quellen in Bezug auf Azimut, Externalisierung und das Auftreten von Vorne-Hinten-Vertauschungen lokalisiert werden können. Interessanterweise führt die Einbeziehung ausschließlich horizontaler Richtungen in die Berechnung der Spektralgewichte zu einer besseren Lokalisierungsperformance im Vergleich zur Einbeziehung horizontaler und nicht-horizontaler Richtungen. Darüber hinaus zeigen Lokalisierungsexperimente mit und ohne Headtracking die Bedeutung der dynamischen Wiedergabe für die Lokalisierungsgenauigkeit virtueller Schallquellen, generiert mit beiden VAHs. Obwohl die Individualisierung eine wichtige Fähigkeit des VAH-Ansatzes ist, zeigen beide Studien, dass die Möglichkeit der dynamischen Wiedergabe binauraler Signale der Hauptvorteil des VAH-Ansatzes gegenüber herkömmlichen Kunstköpfen ist.

GLOSSARY

Acronyms and abbreviations

ANOVA	Analysis of Variance
ASW	Apparent Source Width
BEM	Boundary Element Method
BRIR	Binaural Room Impulse Response
BRTF	Binaural Room Transfer Function
DOA	Direction Of Arrival
DRR	Direct-to-Reverberant Ratio
FEC	Free-air-Equivalent Coupling
FIR	Finite Impulse Response
GUI	Graphical User Interface
HPIR	Headphone Impulse Response
HPTF	Headphone Transfer Function
HRIR	Head Related Impulse Response
HRTF	Head Related Transfer Function
HTK	Head-Tracked KEMAR
HTS	Head-Tracked Sphere
ILD	Interaural Level Difference
ITD	Interaural Time Difference
JND	Just Noticeable Difference
LSEQ	Loudspeaker Equalization
MAA	Minimum Audible Angle
RIR	Room Impulse Response
$RL_E^{'}$	modified Room Level (Early)
SD	Spectral Distortion
SH	Spherical Harmonics
SOFA	Spatially Oriented Format for Acoustics
TD	Temporal Distortion
VAH	Virtual Artificial Head

WNG	White Noise Gain
$\mathrm{WNG_m}$	Mean White Noise Gain

Mathematical notation

a	Scalar a
$\mathbf{a}$	Vector $\mathbf{a}$
$\mathbf{A}$	Matrix $\mathbf{A}$
a^*	Complex conjugate of scalar a
$\mathbf{A}^T$	Transpose of matrix $\mathbf{A}$
$\mathbf{a}^H$	Hermitian transpose of vector $\mathbf{a}$
$\mathbf{A}^{-1}$	Inverse of matrix $\mathbf{A}$
$\otimes$	Convolution
$\mathcal{F}^{-1}$	Inverse Fourier transform
j	Imaginary unit
$\lvert.\rvert$	Magnitude
$\angle\{.\}$	Unwrapped phase
$\lVert.\rVert_2$	2-norm

Fixed Symbols

$\mathbf{A}(f)$	$N \times P$ matrix of steering vectors between N microphones and source at P directions at frequency f
c	Speed of sound propagation
$\mathbf{d}(f, \Theta)$	$N \times 1$ steering vector containing the free-field transfer functions at frequency f between N microphones and source at direction Θ
$\mathrm{D}(f, \Theta)$	Desired HRTF directivity pattern at frequency f and direction Θ
$\mathbf{D}(f)$	$1 \times P$ vector of HRTFs at P directions and frequency f
$\mathbf{e}$	Error vector
f	Frequency
f_s	Sampling frequency
F	Total number of frequency bins
$\mathrm{H}(f, \Theta)$	Resulting directivity pattern of the beamformer at frequency f and direction Θ
$\mathbf{I}_N$	$N \times N$ identity matrix
$\mathrm{J_{LS}}$	Least-squares cost function

$\mathrm{J_m}$	Least-squares cost function subject to a constraint on the mean white noise gain
$L_n(\Theta)$	Length of the direction-dependent path between the n^{th} microphone and the center of the microphone array for sound incidence from source at direction Θ
$\mathrm{L_{Up}}(\Theta)$	Upper boundary set to the spectral distortion at direction Θ
$\mathrm{L_{Low}}(\Theta)$	Lower boundary set to the spectral distortion at direction Θ
N	Number of microphones
P	Number of directions considered in the desired directivity pattern for the calculation of the spectral weights
$P^{'}$	Number of directions at which the synthesis is done
R	Distance to the sound source
$\overline{r}$	Mean Pearson correlation coefficient
$\mathrm{S}(f, \Theta)$	Source signal at frequency f and at direction Θ arriving at the center of the microphone array
t	Time
$w_n(f)$	Spectral weight for the n^{th} microphone at frequency f
$\mathbf{w}(f)$	$N \times 1$ vector, containing the spectral weights for the N microphones at frequency f
$Y_n(f, \Theta)$	Signal of the n^{th} microphone at frequency f and direction Θ
$\mathbf{Y}(f, \Theta)$	$N \times 1$ vector containing the signals of the N microphones at frequency f and direction Θ
$\mathrm{Z}(f, \Theta)$	Output signal of the beamformer at frequency f and direction Θ

Greek letters

α	Cronbach's standardized coefficient
α_R	Factor indicating how much $\mathrm{L_{Low}}$ should be reduced
β_{power}	Minimum desired value for the mean white noise gain
β	Minimum desired value for the mean white noise gain in dB
$\beta_{inversion}$	Regularization parameter for the inversion of transfer functions
ΔILD	Absolute deviation between desired and synthesized ILDs in dB.
ΔITD	Absolute deviation between desired and synthesized ITDs
Θ	Source direction ($\Theta = (\theta, \phi)$)
$\Theta^{'}$	Direction at which the synthesis is done ($\Theta^{'} = (\theta^{'}, \phi^{'})$)
Θ_h	Head orientation ($\Theta_h = (\theta_h, \phi_h)$)
θ	Azimuth angle
ϕ	Elevation angle

μ	Lagrange multiplier associated with the constraint on the mean white noise gain for the minimization of J_m
τ_n	Time delay between the n^{th} microphone and the center of the microphone array

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Motivation and main objective

The aim of spatial sound reproduction is to record and reproduce sounds such that the listener is provided with a realistic spatial impression of the auditory scene. It means that by listening to the signals, the listener is able to perceive a realistic auditory image of the auditory scene including the position and distance of the sound sources as well as the acoustical properties of the environment in which the auditory scene takes place [1].

Spatial sound can be presented by reproducing the sound field for an extended listening area using multiple loudspeakers installed in a room (see [2] for an overview). Binaural technology is another method for spatial sound reproduction, in which the sound field is reproduced only at the two ears of the listener, by using either loudspeakers (e.g. [3]) or more commonly, as also considered in this thesis, by using headphones (e.g. [4]). In natural listening situations, the signals arriving at the ears, referred to as binaural signals, include the reflections and diffraction caused by the listener's own body (head, torso, pinna). This acoustical influence of the body provides the listener with important cues which contribute to the spatial impression of an auditory scene [5]. The aim of binaural technology is to preserve these cues in the reproduction of the binaural signals. With this aim in mind, so-called artificial heads are commonly used to record the auditory scene. Artificial heads are replicas of an average human head, torso and pinnae, with microphones in the ears (see e.g. [6]). By playing back the signals recorded with an artificial head over headphones, the listener is provided to a high extent with the spatial impression of the sound field.

However, binaural reproduction with artificial heads is known to be subject to perceptual shortcomings such as front-back reversals (i.e. the sound source in front is perceived to be in back or vice versa) or errors in the perceived distance, often in the form of in-head-localization (i.e. perceiving the sound source in the head instead of outside the head) [7, 8]. These shortcomings are mainly due to the fact that artificial heads do not have the same anthropometric measures as individual listeners. To preserve the reflections in the same way as caused by individual listeners'

anatomies, artificial heads could be customized geometrically to match individual anthropometric measures, which is associated with enormous financial costs and therefore non-feasible. Another limitation of binaural technology using artificial heads is that the recorded signals can be presented only for a fixed head orientation, namely the orientation of the artificial head towards the auditory scene during the recording. Whereas by rotating the head in natural listening situations the sound source remains at its position, the auditory scene moves with the listener's head with headphone presentations made for a fixed head orientation. This markedly decreases the degree of realism of the reproduced spatial sound. In addition, head movements have been shown to play an important role in the accuracy of spatial sound perception in natural listening situations [9–11]. Accounting for head movements during headphone presentations can moreover reduce the front-back reversals and the in-head-localization [12–14]. Headphone signal playback for a fixed head orientation can therefore lead to an erroneous spatial perception of the sound source.

The main objective of this thesis is to overcome the shortcomings of conventional artificial heads using a Virtual Artificial Head (VAH). A VAH is a microphone array which employs beamforming methods to generate the binaural signals individually for each listener. More precisely, the recorded signals with a VAH can be individualized *post-hoc* for different listeners by applying individually calculated spectral weights to the microphone signals. Moreover, a VAH enables to account for the head movements of the listener during signal playback by adapting the spectral weights depending on the listener's head movements.

The purpose of this thesis is to optimize the VAH approach as already developed by Rasumow et al. [15–17] and evaluate it for new auditory scenarios. Specifically, the thesis aims at improving the spatial resolution of binaural signals generated with the VAH approach while maintaining the number of the VAH microphones limited as well as at investigating the impact of the array topology on the performance of a VAH. Moreover, the VAH approach is evaluated while accounting for head movements during signal playback for signals captured in different acoustical environments and considering sources distributed in three-dimensional space.

The introduction is structured as follows. **Section 1.2** reviews the spatial hearing ability of human listeners. **Section 1.3** introduces binaural technology and its application in spatial sound reproduction. In **section 1.4**, after giving a brief overview of some other microphone array-based approaches in binaural technology, the VAH approach is introduced in more detail, including the main motivation for its further optimization. **Section 1.5** presents the outline of this thesis and its main contributions.

1.2 Spatial hearing

The ability of humans to perceive sound spatially is based on binaural hearing, i.e. on signals arriving at the two ears which supply the listener with important cues

for sound source *localization*, i.e. for indicating the position of a sound source in three-dimensional space. In an anechoic environment, where no reflecting room boundaries or other objects exist, localization cues are created by the interaction of the sound signals with the head, torso, and external ears of the listener [5]. In reverberant environments, additional cues related to reflections and reverberation appear [18]. In addition, non-acoustical cues such as visual cues may impact spatial hearing [19, 20].

To discuss human spatial hearing in more detail, in section 1.2.1 conventions about the spatial position of the source with respect to the listener, as used in this thesis, are provided. Section 1.2.2 provides a summary of localization cues in anechoic environments, followed by measures of localization accuracy in section 1.2.3. Section 1.2.4 explains the impact of room reverberation on spatial hearing. Finally, section 1.2.5 offers a brief discussion about the impact of visual cues on spatial hearing.

1.2.1 *Angle convention*

Throughout this thesis, the spatial position of the sound source with respect to the listener is defined as depicted in the coordinate system in Figure 1.1. The origin of the coordinate system is located at the center of the head, exactly between the ears of the listener, and on the interaural axis, which connects the ear canal entrances. The position of a sound source is defined with respect to the origin by its distance (R) and direction (Θ). The direction $\Theta = (\theta, \phi)$ is described by a horizontal and a vertical part, indicated by azimuth θ and elevation ϕ, respectively. The direction in the front of the listener is defined as $\theta = 0°$ and $\phi = 0°$. Azimuth increases in the counterclockwise direction from $\theta = 0°$. Elevation changes in positive direction upwards from $\phi = 0°$ and in negative direction downwards from $\phi = 0°$. Accordingly, a source is located in the horizontal plane and at the same height as the ears if $\phi = 0°$ and it is located in the median plane if $\theta = 0°$.

The listener and source positions can also be defined using x-y-z coordinates within the depicted coordinate system. The conversion is similar to the conversion between the spherical and Cartesian coordinate systems, where the angle ϕ is defined here as the deviation from the x-y plane instead of from the z-axis.

1.2.2 *Localization cues*

For a sound source located away from the median plane, the sound reaches the ear closer to the source (the ipsilateral ear) earlier and with a higher level than the other ear (the contralateral ear). The time difference of arrival due to the distance between both ears and the level difference due to the shadowing effect of the head are known as Interaural Time Difference (ITD) and Interaural Level Difference (ILD), respectively. ITD and ILD are two important binaural cues for localization

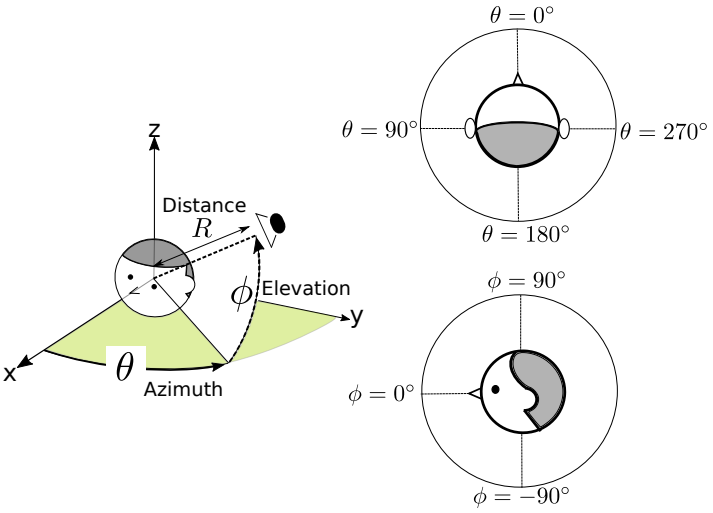

Fig. 1.1: Head-centered coordinate system for defining the sound source position with respect to the listener.

in horizontal directions [5]. Towards higher frequencies, where the shadowing effect of the head gets larger, ILD becomes more important and can increase up to 35 dB at 10 kHz for a sound source located at the side [21]. ITD varies with frequency as well and can be as high as 800μs at around 500 Hz for a sound source located at the side [22]. ITD is shown to be the dominant localization cue of signals with low-frequency energy [23].

For a sound source located in the median plane, the signals arriving at the two ears are very similar and both ILD and ITD are close to zero. In the absence of binaural cues, spectral cues take the dominant role. These spectral cues are mainly characterized by the listener's individual external ear shape, pinna, and become important at frequencies above approx. 3 kHz. Pinna reflections change with the elevation of the sound source and affect different parts of the spectrum, thus providing important cues to indicate the elevation of the sound source [24, 25]. Pinna cues are also important for elevation perception for sources off the median plane. However, with increasing lateral distance to the median plane, the contribution of the ear closer to the sound source gains more importance than the other ear [26]. Besides the high-frequency pinna cues, there are also low-frequency cues, originating from torso reflections and head diffraction, which contribute to the vertical localization, especially for low-pass filtered signals and for sources away from the median plane [27, 28].

A challenge in spatial hearing concerns the localization of sound sources in front and back which have the same distance difference to both ears. In this case, ITD and ILD can only partly indicate the source direction, i.e. whether the source is on

the left or the right side. The ambiguity of distinguishing between front or back remains however; a source in front may be perceived to be in back and vice versa, i.e. a front-back reversal occurs [5]. Due to the asymmetrical shape of the external ear, pinna reflections are different for sound incidence from front and back, which supply the listener with spectral cues that help against front-back reversals [9, 29]. However, the primary cue of disambiguation was shown to be head movements to the left and right. Changes in the binaural cues caused by head movements initiate new localization cues, which help resolve front-back ambiguity [9]. Head movements as small as 4° have been shown to be sufficient to significantly reduce front-back reversals [30]. In addition to this disambiguation, head movements also improve the accuracy of indicating the azimuth and elevation of a sound source [10, 11].

In comparison to direction perception, source distance perception is only related to the cues originated from the listener's anatomy for nearby sources. The binaural cues, especially the ILD, depend on the source distance when the source distance is below 1 m [31, 32]. For source distances beyond 1 m, where the binaural cues are nearly independent of the source distance, the sound pressure level at the ears is the primary acoustic cue for distance perception. In an anechoic environment, the sound pressure level decreases inversely proportional to the distance, which provides cues to perceive changes in the source distance. If no other cues such as room effects or familiarity with the sound source are available, it is hardly possible to perceive the absolute source distance [33]. For source distances beyond 15 m, additional spectral cues, related to the frequency-dependent attenuation of the air path between the source and listener, appear as well [5].

1.2.3 *Localization accuracy*

A crucial factor for the accuracy of sound localization is the bandwidth of the signal emitted from the sound source. Vertical localization, for example, is more accurate with signals containing high frequencies, which excite the pinna-related elevation cues [34, 35]. Horizontal localization, which relies more on binaural rather than spectral cues, is less sensitive to variations in signal bandwidth. However, a 2000-Hz pure tone, for example, can be too high in frequency to offer the ITD cue and too low to offer the ILD cue [36, 37]. In general, broadband signals, which provide the listener with different localization cues across frequencies, are the easiest to localize. Furthermore, when head movements are not allowed, the occurrence of front-back reversals was shown to decrease with increasing bandwidth [35, 38]. If head movements are allowed during listening, the duration of the signal should be long enough (at least 200 ms) to allow exploratory head movements to further reduce the occurrence of front-back reversals [39].

The human auditory system is capable of discriminating very small changes in the position of a sound source. The Minimum Audible Angle (MAA), which is the smallest distinguishable angle between two adjacent sound sources, has been reported to be about 1° in the horizontal direction [40, 41] and 3.6° in the vertical

direction [41] in anechoic environments. Instead of discriminating between relative source positions, the absolute localization ability can be assessed in a localization experiment by asking the listener to indicate the perceived direction of a sound source. The localization accuracy, defined as the difference between the target angle (i.e. the angle of the presented sound source) and the response angle (i.e. the angle perceived by the listener) has been reported in [42] to be about 2° in the horizontal direction and 3.5° in the vertical direction. Horizontal localization performance is best for source directions around the median plane and decreases for source directions off the median plane [43]. In general, vertical localization is less accurate and less consistent among listeners than horizontal localization [42,43].

In addition to localization accuracy, the thresholds of detecting changes in binaural cues have also been assessed, i.e. the smallest change in binaural cues which leads to a different spatial perception of a sound source. The Just Noticeable Differences (JNDs) for ILD have been reported to be between 0.6 dB and 2 dB, depending on the presented signals [44–46]. The JNDs for ITD have been reported to be between 10μs and 40μs, depending on the presented signals [47,48].

1.2.4 *Spatial hearing in reverberant environments*

The spatial impression of a sound source can be quite different in a reverberant environment compared to an anechoic environment. Room reflections interfere with the direct sound arriving at the two ears. The auditory system then needs to distinguish between the direct sound of the source and its reflected versions, which would represent competing sound sources in an anechoic environment. Due to the *precedence effect* [49], the signal arriving first dominates the localization perception, such that sound localization is also possible in the presence of room reflections. However, depending on the room geometry and the positions of the source and the listener, the order and intensity of room reflections can change the sense of localization. In fact, in a reverberant environment the similarity of the signals at the two ears, measured by the interaural coherence, is reduced by the room reflections. In case of reduced interaural coherence, the ITD and ILD don't describe the position of the sources the same way as in an anechoic environment [50, 51]. Similarly, the JNDs for the binaural cues increase in reverberant environments. According to [52], in a reverberant environment, the JNDs for ITD and ILD can be five to eight times and up to two times as high as in an anechoic environment, respectively. In general, the localization performance is poorer in reverberant than in anechoic environments, especially when listening to continuous non-transient signals [53].

The presence of reverberation is however beneficial for source distance perception. The Direct-to-Reverberant Ration (DRR), defined as the ratio between the energy of the direct sound (changing inversely proportional to the squared distance) and the reverberant energy (almost constant and independent of distance), is known to be an important distance cue in reverberant environments [54,55]. Furthermore, the

presence of reverberation leads to spatial impressions such as envelopment (a sense of being surrounded by the sound), mainly related to late reverberation, and apparent source width (spatial broadening of the source wider than its size), mainly related to early lateral reflections, which are both important attributes of a reverberant environment, e.g. a concert hall [56].

1.2.5 *The impact of visual cues on spatial hearing*

Despite the ability of the auditory system to utilize acoustical cues to accurately indicate the direction and/or distance of sound sources, visual cues can further improve the localization performance. In tests with blindfolded and sighted listeners, it was shown that the directional localization performance is better with vision than without, both for sources inside and outside the visual field [19]. Moreover, source distance estimation was shown to improve if the source is visible [57]. However, visual cues may also lead to a misperception of the position of a sound source, for example, in case of spatial disparity between the acoustical and visual stimuli of a sound source. If the listener is not explicitly asked to pay attention to the spatial origin of a sound source, he or she tends to ignore the spatial disparity between the acoustical and visual stimuli and localize based on the visual information only [20]. With horizontal displacements of even up to 20° between the acoustical and visual stimuli, the perceived position of the source can be drawn to the area around the position of the visual stimulus [58]. In line with the generally weaker localization performance of the auditory system in vertical directions, more spatial audio-visual disparity than in horizontal directions can be tolerated in vertical directions [59].

1.3 Binaural technology

The aim of binaural technology is to capture the signals arriving at the two ears, referred to as binaural signals, and reproduce them in a way that the spatial and spectral aspects of the sound field are preserved [4]. Binaural signals can be captured with the technique of *binaural recording*, using small microphones placed inside the ears of a listener or an artificial head. In this way, the information in the sound field including the spatial and spectral localization cues is inherently maintained in the binaural signals. As an alternative to binaural recording, in the technique of *auralization*, binaural signals are generated by filtering an anechoic single-channel audio signal with room reflections as well as with the reflections caused by the head, torso, and external ear of the listeners, to create a virtual auditory scene. The direction-dependent filtering effects of the listener's body are quantified as Head Related Transfer Functions (HRTFs), which play an important role in auralization. With auralization, it is possible to present the binaural signals dynamically, i.e. taking the effect of head movements into account.

In section 1.3.1, binaural recording with artificial heads and its limitations are explained. Section 1.3.2 introduces HRTFs, followed by an explanation of their application in auralization in section 1.3.3. Section 1.3.4 introduces dynamic auralization,

followed by a brief discussion of the quality assessment of dynamic auralization in section 1.3.5.

1.3.1 *Binaural recording*

The best way of making the full spatial and spectral information available to a listener is to record the sound signals with two microphones placed in his or her ear canals. For practical reasons, the listener is often replaced either by baffles, as done in the early decades of the 20^{th} century, or by manikins, which were later developed to what we know today as artificial heads: replicas of the human upper body, with similar acoustical properties as human listeners [6]. The most straightforward way to listen to binaural recordings is to play them over headphones. For the signals to sound naturally, headphones with Free-air-Equivalent Coupling (FEC) should be used in order not to disturb the radiation impedance as seen from the ear canal [60]. In addition, the frequency response of the headphones should be measured and compensated for [60], preferably individually for each listener [61]. Despite the correct compensation of headphone effects, binaural signals can still differ from the signals at the ears of a listener, if they are recorded with an artificial head or with a person other than the listener. The reason is the anatomical differences between different listeners, or the different geometrical measures of an artificial head compared to that of the listener. When listening to non-individual binaural recordings, localization performance has been shown to degrade in form of the increased number of front-back reversals or distance errors [7, 8]. Since for practical reasons binaural recordings are often made with artificial heads, non-individuality is a common issue in binaural recordings.

Another critical limitation of binaural recordings with artificial heads is that the effect of the listener's head movements can hardly be taken into account during playback. To compensate for the listener's head movements, motorized artificial heads have been proposed, which move synchronously with the listener's head (see for example [62]). However, such a binaural recording is limited to real-time transmission of the sound to one single listener. In practical applications, binaural recordings are usually performed with a fixed orientation of the artificial head, commonly directed to the auditory scene, and the binaural signals are played back for this specific scenario only. If the listener rotates his or her head, the presented acoustical scene follows the head. Therefore, the cues arising from head movements, which in normal listening situations help to resolve the localization ambiguities, are not present. Although the spatial information is preserved to a large extent in binaural recordings, the listening experience differs considerably from a natural listening situation due to the lacking head movement effects.

As an alternative, binaural signals can be generated via auralization. With auralization, binaural signals can be presented dynamically, i.e. the playback can be adapted to the head movements of the listener. Before continuing with auralization

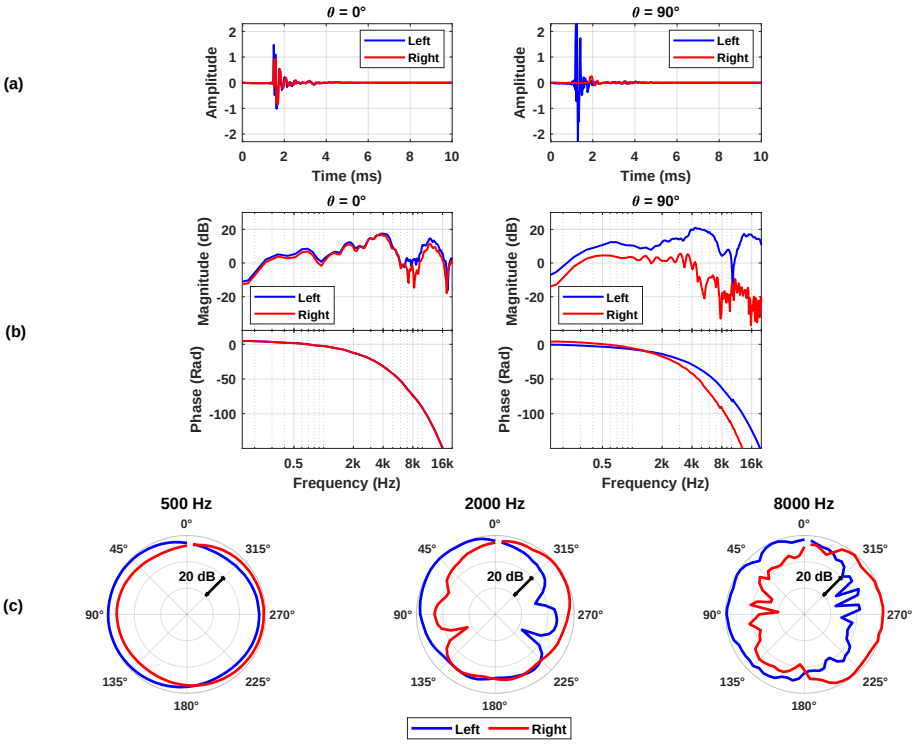

Fig. 1.2: **(a):** Horizontal left and right HRIRs for sound incidences from front ($\theta = 0°$) and from the side ($\theta = 90°$). **(b):** Magnitude (top) and phase (bottom) spectrum of horizontal left and right HRTFs for sound incidences from front ($\theta = 0°$) and from the side ($\theta = 90°$). **(c):** Directivity patterns of left and right HRTFs (magnitudes depicted over horizontal directions θ) for three exemplary frequencies (500 Hz, 2000 Hz and 8000 Hz)

in section 1.3.3, the next section briefly introduces HRTFs, since they play a key role in auralization.

1.3.2 *Head Related Transfer Functions*

Head Related Transfer Functions (HRTFs) describe the free-field sound transmission between a sound source and the two ears of the listener, including the important reflections from the listener's body. This transmission consists of a direction-dependent part (propagation from the sound source to the ear canal entrance) and a part which can be considered as independent of the source direction (propagation through the ear canal to the eardrum) [63]. The direction-dependent part of HRTFs is defined as the ratio between the sound pressure at the entrance of the (blocked) ear canal to the sound pressure measured at the center position of

the head in absence of the listener [4]. For a sound source at azimuth θ, elevation ϕ and distance R, the sound propagation path to the left and right ear canals is described by a pair of complex-valued, frequency- and direction-dependent left and right transfer functions, $\mathrm{HRTF_L}(f,\theta,\phi,R)$ and $\mathrm{HRTF_R}(f,\theta,\phi,R)$, where f denotes frequency. The time-domain equivalent of the HRTFs is given as their inverse Fourier transform, known as Head Related Impulse Responses (HRIRs), $\mathrm{HRIR_L}(t,\theta,\phi,R)$ and $\mathrm{HRIR_R}(t,\theta,\phi,R)$, where t denotes time.

HRTFs are commonly measured in anechoic environments, although reflective environments under elimination of room reflections are also applicable, see e.g. [64] as well as Appendix A. Measurements are often performed at a fixed, sufficiently large distance ($R > 1$ m) to the sources such that the dependency of HRTFs on distance can be neglected [31]. Miniature microphones are commonly positioned at the entrance of the blocked left and right ear canals of the listener or an artificial head. A broadband excitation signal, e.g. white noise or sweeps, is emitted from the source at direction θ and ϕ. The HRIRs or equivalently the HRTFs are identified by processing the signals captured by the miniature microphones and the source signal (see for example [65–68]). If morphological measures of head and pinna are available, HRTFs can also be calculated numerically using, e.g., Boundary Element Methods (BEM) [69, 70].

HRIRs or HRTFs inherently contain the binaural localization cues. Figure 1.2a shows exemplary measured HRIRs for two directions ($\theta = 0°$ and $\theta = 90°$) in the horizontal plane ($\phi = 0°$). While for the sound incidence from front ($\theta = 0°$), the left and right HRIRs have almost the same onset time, for sound incidence from the side ($\theta = 90°$), there is a time difference of arrival between the left and right HRIRs, which illustrates the effect of the ITD. The time difference of arrival at $\theta = 90°$ can also be seen in the difference between the left and right HRTF phases shown in the lower part of Figure 1.2b, which are known to be perceptually relevant at low frequencies ($f < 2$ kHz) [71]. Due to the shadowing effect of the head, the ILD at $\theta = 90°$ can be observed as the amplitude difference between the left and right HRIRs in Figure 1.2a, as well as the difference in the left and right HRTF magnitudes in the upper part of Figure 1.2b. The frequency-dependent shadowing effect of the head can be better observed in the HRTF directivity patterns, which illustrate the HRTF magnitudes at a fixed frequency over horizontal directions. Figure 1.2c shows the HRTF directivity patterns at three exemplary frequencies. At 500 Hz, the HRTF directivity patterns are quite simple and almost omnidirectional. With increasing frequency (2000 Hz and 8000 Hz), the shadowing effect of the head results in lower magnitudes at the contralateral directions (directions at the opposite side of the ear, i.e. $180° < \theta < 360°$ for the left ear) compared to the ipsilateral directions (directions at the same side as the ear). In addition to binaural cues, the effect of different parts of the listener's body has been shown to be represented at different frequencies in the HRTF magnitudes. Torso and shoulder reflection cues, for example, appear at frequencies below about 2-3 kHz [27, 28] and the important peaks and notches due to pinna reflections appear at frequencies

above about 5 kHz, which are known to be crucial for elevation perception [72–74].

Due to the fact that human listeners are morphologically different, it is evident that HRTFs vary from person to person. The analysis of individually measured HRTFs and individual anthropometric measures has shown correlations between, e.g., the maximum ITD and the head size, or between the pinna dimensions (concha height) and the main pinna-related notch [75].

While HRTFs describe the free-field sound transmission between a sound source and the ears, in reverberant environments the room reflections are added to the direct sound arriving at the listener's ears. In this case, the sound transmission between the source and the ears is represented by BRIRs (Binaural Room Impulse Responses) or BRTFs (Binaural Room Transfer Functions), which combine the information contained in HRIRs/HRTFs with the acoustical information of the room. BRIRs/BRTFs can be acquired in the same way as HRIRs/HRTFs, but in reverberant environments.

1.3.3 *Auralization*

According to [76], *Auralization is the process of rendering audible, by physical or mathematical modeling, the sound field of a source in a space, in such a way as to simulate the binaural listening experience at a given position in the modeled space.* As one possible method of auralization, any given direction of sound incidence can be synthesized by convolving the anechoic source signal with a pair of left and right HRIRs (free-field auralization) or BRIRs (room auralization) of this direction to represent the sound source virtually [77]. Although the convolution with HRIRs or BRIRs for headphone playback is not the only possible method of auralization [76], the term *auralization* is used for this method throughout the thesis.

Auralization can be done with measured or numerically simulated HRIRs/BRIRs. In simulation-based room auralization, the Room Impulse Response (RIR), which describes the path between the source and the listener position (center of the head), is predicted either using geometrical methods such as ray tracing or the image source method [78], wave-based methods such as BEM or Finite-Difference Time-Domain [79], or hybrid methods using e.g. the image source method to generate early room reflections and feedback delay networks to generate late reverberation [80, 81]. The directional information of the listener, i.e. the HRIRs, is integrated into the process of RIR estimation to result in BRIRs.

In order for auralizations to be realistic, the virtually presented sound sources should also be externalized, i.e. be perceived outside of the head [82]. According to [83], both spectral as well as binaural cues should be preserved well in order to have externalized virtual sources. This can be achieved with individual HRIRs or BRIRs and correct compensation of the headphones. However, due to the missing distance cues in HRIRs, free-field auralizations can still fail to

be externalized, whereas room auralizations can be better externalized due to the reverberation information included in the BRIRs [84]. Despite the fact that reverberation-based distance cues support source distance perception, it should be realized that externalization and distance perception are not the same, since they do not rely on the same cues [85]. While in reverberant environments the main cue for distance perception is the DRR and the binaural cues rarely play a role, the improved externalization with reverberation relies on binaural cues such as ILD fluctuation over time or reverberation-based interaural differences [86, 87]. Nevertheless, distance perception and externalization are often linked and distance measures with respect to the head have been used in many studies to investigate the externalization of virtual sound sources [83, 86, 88].

In addition, the room in which the binaural signals are presented also plays a role in the externalization of virtual sources. It has been shown that if one listens to auralizations of a room while being in an acoustically different room, the perceived externalization can be negatively affected [89, 90].

1.3.4 *Dynamic auralization*

Compared to binaural recordings, the advantage of auralization is that the binaural signals can be presented dynamically, i.e. in response to the head movements of the listener. Assuming that the head movements can be captured, e.g. with a head tracker, HRIRs or BRIRs used in the auralization can be adapted in real-time depending on the listener's current head orientation. Such a dynamic auralization requires, however, the HRIRs/BRIRs for different head orientations. This poses a challenge since HRIRs/BRIRs are commonly measured only with the head oriented to the frontal direction. For free-field auralizations, HRIRs for the head rotated to a non-frontal orientation Θ_h can be approximately replaced by the HRIRs for the source at the direction $-\Theta_h$ contrary to the head orientation. For room auralizations, a similar approximation can be performed by modifying the measured omnidirectional RIRs with the HRIRs for the source at the direction contrary to the head orientation. However, the direct, early reflections and the reverberant parts of the RIR should be separated first and only the direct part and early reflections are filtered with HRIRs [91]. An alternative way of acquiring the BRIRs is to measure them for different head orientations of individual listeners or artificial heads [92–94], which can however be very tedious and time-consuming if different environments with sources at different positions are to be auralized.

Due to the fact that HRIRs are individual, auralizations using individual HRIRs or BRIRs are commonly recommended. Previous studies have shown that the localization performance reduces when listening to non-individual auralizations, especially with respect to front-back reversals or externalization [95–98]. These results, however, apply to listening situations with static signal playback, i.e. without considering the effects of head movement via head tracking. In contrast, it was shown that with dynamic auralizations, a similar localization performance can

be achieved with individual and with non-individual auralizations [12, 92, 98]. With dynamic auralizations, the non-individuality is less critical, front-back reversals are reduced, and externalization is supported by the interaction between binaural cues and head movements [12–14, 98]. To maintain the quality of the listening experience with dynamic auralizations, certain criteria with respect to the resolution of impulse responses and system latency should be satisfied. For an artifact-free dynamic auralization, impulse responses (HRIRs/BRIRs) should be available for a fine grid of head orientations. In [99], the minimum audible grid resolution in azimuth and elevation was reported as $2^\circ \times 2^\circ$ with broadband noise signals and $5^\circ \times 5^\circ$ with musical stimuli. In [98], a minimum grid resolution of 5° in horizontal directions was suggested with broadband noise signals. There are in addition some delays between the movement of the listener's head and the response of the system due to, e.g., tracker device latency, the time required to select impulse responses and the signal processing. In order for this delay not to cause audible artifacts, the maximum overall latency should not exceed 100 ms [100, 101].

1.3.5 *Quality assessment of dynamic auralization*

The focus of auralization has often been on localization performance. The quality assessment of auralizations goes much further than localization performance and includes many other aspects. Two well-known criteria for quality assessment of virtually presented auditory scenes are *plausibility* and *authenticity*. While plausibility is defined as the agreement between the presented auditory scene and the listener's expectation, i.e. with respect to an internal reference [93], authenticity is defined as the agreement between the presented and the actual auditory scene, i.e. with respect to an external reference [5]. Accordingly, an auralization is plausible when it is in accordance with the listener's expectations regarding the perception of the presented scene. A plausible auralization can however be inauthentic, i.e. it can sound different from the real auditory scene. For example, there can be audible differences in form of spectral coloration, caused e.g. by the non-individuality of the auralization or the positioning of headphones, even when using FEC headphones with individual headphone equalization [102]. Positioning artifacts, e.g. different positioning of subjects during HRIR/BRIR measurements and binaural signal playback, may also impact the authenticity [103]. A detailed investigation on the authenticity of dynamic auralizations was performed in [94], based on a wide range of attributes including spectral coloration (e.g. timbre changes), room-related attributes (e.g. level of reverberance), source-related attributes (e.g. source width) or possible audible artifacts. Despite careful considerations for an error-free auralization, it was shown that when listening to broadband noise signals, the difference between auralization and reality could be perceived. For speech signals, the difference was detected to a lesser degree, especially in reverberant environments. In general, plausible auralizations are easier to achieve than authentic auralizations. However, authentic auralizations are often not required, since in many practical applications, no reference for direct comparison is available. The quality assessment of an auralization hence depends particularly on the application requirements

which signal is used and which environment is auralized.

1.4 Microphone arrays in binaural technology

As discussed in section 1.3.1, critical limitations of binaural recording using artificial heads are the non-individuality of the recordings and the fact that the effect of the listener's head movements can hardly be taken into account during playback. As an alternative to artificial heads, it has hence been proposed to use microphone arrays, consisting of multiple microphones at different spatial positions, both for capturing the sound field as well as for dynamic auralizations. To create binaural signals from the microphone signals, the spatial information in the recorded signals can either be used directly, as in the Motion-Tracked Binaural sound (MTB) [104,105], or further processed in combination with HRTFs, as in approaches based on Spherical Harmonics (SH) [106–114] and beamforming [15, 115–117]. As will be discussed in this section, in general the quality of binaural signals increases with the number of microphones, which on the other hand is associated with increased system complexity and cost. Therefore, reducing the number of microphones while maintaining the quality is an important objective.

After briefly reviewing MTB and SH approaches in sections 1.4.1 and 1.4.2, in section 1.4.3 we introduce the Virtual Artificial Head (VAH), which is a microphone array-based system applying beamforming. In this thesis we will only focus on binaural recordings and auralizations using the VAH approach.

1.4.1 *Motion-Tracked Binaural sound (MTB)*

The MTB system [104] consists of a rigid sphere of typical head diameter, with several microphones distributed on the equator of the sphere. The rigid sphere acts as a simplified model of the head, using which the effects of important binaural cues (ILD and ITD) can be approximated. For a given head orientation of the listener, the signals captured by the microphones nearest to the ear positions are interpolated and presented as binaural signals. Similarly, RIRs measured by the microphones nearest to the ear positions are interpolated to result in the BRIRs for dynamic auralizations. Approximating the head by a simple rigid sphere leads to some shortcomings of the MTB system, such as spectral coloration and poor localization performance caused by the different diameters of the sphere and the listener's head (leading to ITD mismatch). In addition, the interpolation algorithm can also impact the quality of the binaural signals [105].

A perceptual evaluation of the MTB system was performed in [118], considering different numbers of microphones (8, 16, 24 and 32 microphones), interpolation algorithms (the five suggested in [104]), and test signals (noise, speech, music). The plausibility of the binaural MTB signals was shown to be interdependent on these three parameters. Recently, the MTB system was evaluated in [119] with respect to

plausibility against a physical sound source and with respect to specific attributes such as localization (elevation), externalization, or tone color in comparison to a reference signal. The ITDs in the binaural signals were adjusted individually for the subjects, based on their head sizes. Despite missing pinna cues, the perceived elevation and the degree of externalization were well in accordance with the reference signal. This good performance of the MTB system was related to the dynamic presentation, i.e. the presence of head movement cues [119].

1.4.2 *Approaches based on Spherical Harmonics*

Other well-known approaches to represent binaural signals with microphone arrays use Spherical Harmonics (SH) [106–110]. These approaches are based on the fact that a sound field in three-dimensional space can be represented as the sum of plane waves using the SH decomposition. If the sound field is evaluated on the surface of a sphere, it can be directly described by SH coefficients [120]. The sound field is therefore commonly sampled with microphones distributed on the surface of a sphere, although non-spherical microphone arrays can also be used to represent the three-dimensional sound field using the SH decomposition (see e.g. [121,122]). Binaural signals can be generated by multiplying the SH coefficients of the sound field with the SH coefficients of the HRTFs [123, 124]. By choosing the appropriate HRTFs for a given head orientation, signals can be played back dynamically. Equivalently, measured RIRs can be processed to BRIRs for different head orientations. In addition, by including the HRTFs of individual listeners, the recordings can be individualized for different listeners.

To represent the whole acoustic frequency range up to 20 kHz, a huge number of microphones (a total of 1936 according to [111]) would be required on the surface of the sphere. Due to the fact that the sound field on the surface of the sphere is sampled with a limited number of microphones in practice, the SH decomposition is limited to a maximum order. Order-truncated HRTFs show a magnitude roll-off at higher frequencies. Although it has been shown that directional sound source localization is possible with HRTFs truncated to low orders [125], the order truncation of HRTFs leads to shortcomings with respect to externalization and spectral coloration, especially for signals containing energy at high frequencies [112]. Solutions have been suggested to overcome the order truncation problems, e.g. by applying timbre correction [126], order-dependent compensation filters [127] or SH domain tapering for coloration compensation [128] (see also [129] for the perceptual evaluation of state-of-the-art algorithms). It was shown in [107] that over 100 microphones are required in order to achieve near-to-authentic auralizations with respect to timbre and space- and source-related attributes such as source distance, source width or level of reverberance.

As another application of the SH approach, first-order ambisonic microphone arrays, known as B-format microphones, are used, based on measured RIRs [113, 130, 131]. A B-format microphone consists of four microphone capsules, with one omnidirec-

tional channel (representing the zeroth-order SH) and three figure-of-eight channels directed to the orthogonal axes of the three-dimensional space (representing the first-order SH). To each discrete-time sample of the measured omnidirectional RIR, a Direction Of Arrival (DOA) is assigned which is estimated from the three directional microphone signals. The omnidirectional RIR is transformed into a higher-order RIR using the SHs evaluated at the estimated DOA [113, 131]. The higher-order RIR is equalized in separate spectral sub-bands and for each SH order to avoid an increased spectral brightness in the reverberant part of the impulse responses [130, 131].

B-format microphones can also be used with Directional Audio Coding (DirAC) [114], where based on a parametric time-frequency decomposition the signal in each frequency band is decomposed into a direct and a reverberant part, and the direct part is convolved with the HRIRs of a given direction [132]. Using the mixing method proposed in [133], the DirAC signals can be extended to higher-order signals. It was shown in [134] that plausible dynamic auralizations are possible with first- or third-order DirAC signals with respect to overall quality and spatial accuracy.

1.4.3 *Virtual Artificial Head (VAH)*

The Virtual Artificial Head (VAH) approach uses a (planar or three-dimensional) microphone array in combination with a filter-and-sum beamformer to mimic the HRTF directivity patterns. In a filter-and-sum beamformer, the microphone signals are filtered and combined in order to achieve a desired spatial selectivity [135]. Although filter-and-sum beamformers are often designed to enhance a certain direction [136, 137], they can also be optimized to approximate an arbitrary spatial directivity pattern [138], e.g. of individual HRTFs. Although such an approach has been used in several studies [15, 115–117], the term Virtual Artificial Head (VAH) was coined by Rasumow et al. [15–17] and this thesis will focus particularly on the VAH approach as developed in these papers.

The VAH concept is depicted in Figure 1.3. For a microphone array with N microphones, at each frequency f two sets of spectral weights, one for the left ear, $\mathbf{w}_{\mathrm{L}}(f)$, and one for the right ear, $\mathbf{w}_{\mathrm{R}}(f)$, are applied to the microphone signals. Each vector of spectral weights $\mathbf{w}_{\mathrm{L}}(f)$ and $\mathbf{w}_{\mathrm{R}}(f)$ contains N complex-valued coefficients. The weighted sum of the microphone signals with $\mathbf{w}_{\mathrm{L}}(f)$ and $\mathbf{w}_{\mathrm{R}}(f)$ results in the left and right binaural signals at frequency f, respectively. By carefully choosing $\mathbf{w}_{\mathrm{L}}(f)$ and $\mathbf{w}_{\mathrm{R}}(f)$, the resulting directivity pattern of the microphone array at frequency f can be made similar to the directivity pattern of individual HRTFs. Thus, the VAH can replace any individual listener during the recordings by mimicking the acoustical effects of the head, torso, and external ears of this listener. In addition, the spectral weights can be adjusted to mimic the HRTF directivity patterns of different head orientations, enabling dynamic playback of the recordings. Based on the

same concept, individual BRIRs for different head orientations can be synthesized from measured RIRs.

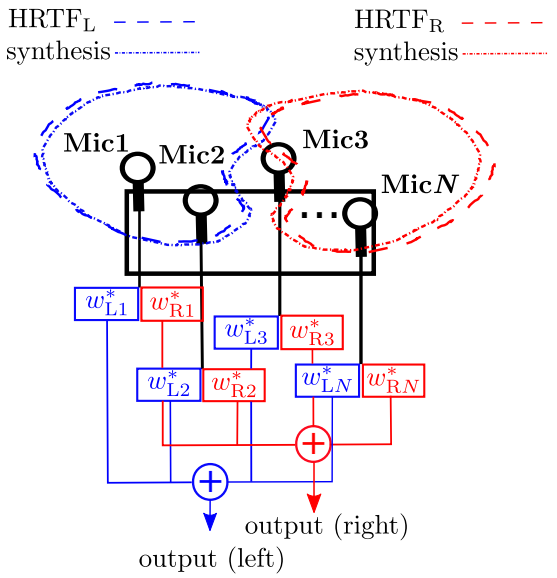

Fig. 1.3: Desired directivity patterns of the left and right HRTFs (HRTF$_L$, HRTF$_R$), and the left and right directivity patterns synthesized with a VAH with N microphones and N complex-valued spectral weights $\mathbf{w}_L(f)=[w_{L1}(f), w_{L2}(f), ..., w_{LN}(f)]^T$ for the left ear and $\mathbf{w}_R(f)=[w_{R1}(f), w_{R2}(f), ..., w_{RN}(f)]^T$ for the right ear.

It has been proposed in [17, 139, 140] to calculate the VAH spectral weights by either minimizing a least-squares cost function or a non-linear cost function, where the cost function is a measure for the difference between the desired and the synthesized HRTF directivity patterns at a selected number of discrete directions. Minimizing these cost functions may, however, result in non-robust spectral weights, causing a large sensitivity to even small deviations in the microphone positions and/or characteristics and an undesirable amplification of spatially uncorrelated self-noise of the microphones [136,139,141]. In order to increase the robustness, regularization constraints are typically incorporated into the optimization problem. A commonly used measure for the robustness of beamformers is the White Noise Gain (WNG), which is defined as the ratio between the output power of the microphone array in the look direction of the beamformer and the output power for spatially uncorrelated noise [108,136,141]. The larger the WNG, the larger the attenuation of spatially uncorrelated noise and therefore the higher the robustness. Rasumow et al. proposed to consider the mean WNG, averaged over all directions [17, 142], since for synthesizing HRTF directivity patterns all directions should be considered as a look direction. In combination with the least-squares cost function and employing the method of Lagrange multipliers, a closed-form solution was derived in [142,143] to calculate the spectral weights subject to the mean WNG constraint, either

separately for each frequency bin or for frequencies grouped in perceptually relevant bandwidths. The optimal value of the Lagrange multiplier was chosen iteratively until the desired mean WNG was achieved. More details will be provided in Chapter 2.

In addition, it was proposed to smooth the desired HRTF directivity patterns in spectral and spatial domains prior to calculating the spectral weights. This spectro-spatial smoothing was shown to introduce no perceptible degradation to the HRTFs and was performed in order to facilitate the synthesis with a VAH [16].

Rasumow et al. used a planar microphone array with N=24 microphones (Analog Devices ADMP504 Ultralow Noise) [15], shown in Figure 1.4. The microphones were arranged in a Golomb-based topology [144, 145] on a 20 cm × 20 cm plate covered with absorbing material. This VAH was perceptually evaluated in [15]. The main focus was on synthesizing HRTF directivity patterns at horizontal directions for static presentations, i.e. only for the frontal head orientation and without considering head movements during playback. A total of 24 equiangular horizontal directions (i.e. 15° azimuthal resolution) were included in the desired directivity pattern. Recordings were performed with the VAH in an anechoic environment for sources at directions coinciding with the 24 horizontal directions included in the desired directivity pattern as well as at intermediate directions. The recorded VAH signals were subsequently individualized for the subjects with individually calculated spectral weights. In a listening test, subjects evaluated binaural signals generated with the VAH as well as binaural recordings with conventional artificial heads in comparison to loudspeaker signals with respect to localization, spectral coloration, and overall quality. For sources at directions coinciding with the 15°-resolution grid included in the desired directivity pattern, binaural signals generated with the VAH outperformed the non-individual binaural recordings of conventional artificial heads. For sources at intermediate directions, however, the quality of binaural signals generated with the VAH recordings was comparable to or below the non-individual recordings of one of the considered conventional artificial heads in that study.

Although the synthesized directivity pattern of the VAH implicitly interpolates between the directions included in the desired directivity pattern, the optimal performance can be expected only at these directions and may degrade at other directions. In order to increase the spatial resolution of a VAH by including more directions in the desired directivity pattern, the number of microphones should be increased as well. An example is the SENZI system developed in [117, 146–148], consisting of a head-sized rigid sphere with 252 microphones on its surface. With SENZI, a spatial resolution of 5° in both horizontal and vertical directions can be achieved. One of the main objectives of the thesis is to increase the spatial resolution compared to the VAH of Rasumow et al. while keeping the number of microphones far below the number of microphones used in SENZI.

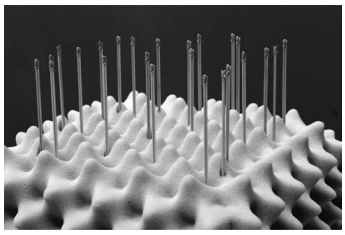

Fig. 1.4: Virtual artificial head (VAH) as developed by Rasumow et al. [15]: planar microphone array with 24 microphones and an extension of 20 cm×20 cm.

1.5 Outline of the thesis and main contributions

The main objective of this thesis is to improve the performance of existing VAH approach for HRTF synthesis and to evaluate it for situations which were not considered before. First, we aim at **improving the horizontal spatial resolution using a limited number of microphones,** similarly to the VAH of Rasumow et al. (Figure 1.4). Second, we aim at investigating **the impact of the array topology**, i.e. the array extension and the distribution of microphones, on the VAH synthesis performance in the horizontal plane. Third, we aim at **evaluating the VAH approach in dynamic auralizations**, i.e. for different head orientations of the listener using head tracking, both in anechoic and reverberant environments and for horizontal and non-horizontal sources. For all investigations, the quality of VAH synthesis is assessed both physically, i.e. based on objective measures such as the spectral differences or deviations in binaural aspects, as well as perceptually in listening tests, where the perceptual results are discussed in relation to the objective measures. The spectral weights are always calculated for each subject individually by considering individually measured HRTFs. The comparison to conventional artificial heads is considered throughout all perceptual evaluations.

As the first contribution of this thesis, we **propose a new constrained optimization method to calculate the VAH spectral weights, which allows to increase the spatial resolution of the synthesis in the horizontal plane using a limited number of microphones**. In addition to the constraint on the mean WNG, new constraints are introduced as upper and lower boundaries on the monaural spectral magnitude error at a high number of directions. This spectral error is referred to as spectral distortion in this thesis and is defined as the spectral difference between the desired and synthesized HRTFs. It is shown that by applying constraints on the spectral distortion, **the frequency range, up to which the synthesis accuracy can be considered acceptable, is increased from 2 kHz to 5 kHz** compared to imposing only the mean WNG constraint.

As the second contribution, we **investigate the impact of the microphone array topology on the VAH performance when imposing both mean**

WNG and spectral distortion constraints. Smaller inter-microphone distances enable to satisfy the spectral distortion constraints at higher frequencies. However, for smaller microphone arrays, the mean WNG constraint is more difficult to satisfy. When aiming at higher mean WNG values to increase robustness, the phase accuracy may be deteriorated. Based on evaluations of simulated VAHs with four different microphone array topologies (different array extension and distribution of the microphones), it is shown that **for topologies combining dense and sparse inter-microphone distances, the mean WNG and spectral distortion constraints can be satisfied for frequencies up to 8 kHz without deteriorating the phase accuracy.**

As the third contribution, we **evaluate the performance of the VAH approach for dynamic auralizations with speech signals** within two studies. Both studies employ the proposed constrained optimization method with individually synthesized BRIRs for different head orientations. Sources in and outside the horizontal plane are considered and the BRIRs are synthesized for different parameters concerning the directions included in the desired directivity pattern and values for the minimum desired mean WNG. The results indicate the **good perceptual performance of the VAH approach for practical applications with speech signals, even in case that the VAH contains a limited number of 31 microphones and less**. In the first study, dynamic auralizations with different BRIRs synthesized with a VAH consisting of 24 microphones is perceptually evaluated in comparison to real (visible) sound source presentations. It is shown that close-to-reality dynamic auralizations of speech signals can be realized with the VAH consisting of only 24 microphones. The results apply to anechoic as well as to reverberant environments. In the second study, the localization performance with binaural signals generated with two VAHs consisting of 24 and 31 microphones is assessed in the absence of the visual cues and in comparison to the localization performance of real sound sources. Moreover, the impact of head tracking on the localization performance of virtual sources is investigated. It is shown that even within a more challenging localization experiment in the absence of visual cues, azimuth and externalization results as well as front-back reversal rates similar to real sources can be achieved with virtual sources generated with VAHs consisting of only 24 or 31 microphones. Furthermore, the importance of the dynamic presentation of binaural signals on the accuracy of localizing virtual sound sources generated with the VAHs is confirmed. In addition, both studies confirm **the advantage of the VAH approach over conventional artificial heads**.

In the remainder of this section, a chapter-by-chapter overview of this thesis is provided, which offers a summary of the contents and the contributions of each chapter.

In **chapter 2**, a more detailed overview of the methods employed by Rasumow et al. in previous VAH studies is presented. The synthesis of HRTF directivity patterns using the VAH as a filter-and-sum beamformer is introduced and the calculation of the spectral weights based on the minimization of a least-squares cost function is described. Furthermore, the need for regularization to increase the robustness

is explained, which is implemented by adding a mean WNG constraint to the optimization problem. Finally, spectro-spatial smoothing of HRTFs prior to calculating the spectral weights is reviewed.

In **chapter 3**, we discuss the first and second contributions of the thesis, i.e. increasing the spatial resolution of the VAH synthesis and the impact of the microphone array topology. All investigations in this chapter are based on simulated microphone arrays and consider HRTF synthesis in the horizontal plane. First, we explain the limitation of the least-squares-based optimization subject to the mean WNG constraint with respect to the spatial resolution of the VAH approach. To overcome this limitation, we propose to impose additional constraints on the spectral distortion at a high number of directions included in the desired directivity pattern (72 horizontal directions, i.e. 5° resolution). Compared to only imposing the mean WNG constraint, it is shown that the additional spectral distortion constraints not only improve the synthesis accuracy at more directions but also up to higher frequencies. Next, it is discussed that the ability to satisfy the mean WNG and spectral distortion constraints depends on the microphone array topology, i.e. the array extension and the distribution of the microphones. Four VAHs are simulated with different topologies, and the impact of these topologies on the resulting synthesis accuracy is investigated based on objective measures. While small inter-microphone distances are advantageous for satisfying the spectral distortion constraints at higher frequencies, a small array extension can cause difficulties in satisfying the mean WNG constraint at low and mid-frequency ranges. Objective measures indicate that aiming at higher mean WNG values to increase robustness may deteriorate the phase accuracy for smaller arrays. It is shown that a combination of dense and sparse microphone distances offers an appropriate trade-off between satisfying spectral distortion and mean WNG constraints, without losing the phase accuracy. Finally, the perceptual quality of static (i.e. without head tracking) free-field auralizations with individually synthesized HRTFs with the four considered VAHs as well as non-individual HRTFs of a conventional artificial head is assessed with respect to localization, spectral coloration and overall quality in a listening test with eleven subjects. It is shown that the binaural signals generated with the VAH with the additional spectral distortion constraints perceptually outperform both the binaural signals generated with the VAH where only the mean WNG constraint is imposed as well as the non-individual binaural signals of the conventional artificial head. In addition, perceptual evaluations indicate that the binaural signals generated with the microphone array topology combining dense and sparse inter-microphone distances perceptually outperform the binaural signals generated with the other evaluated microphone array topologies. The publications related to this chapter are [149, 150].

In **chapter 4**, we present the general methodology for dynamic auralizations with the VAH approach in this thesis. First, the calculation of the spectral weights for different head orientations of the listener is explained. Then, it is discussed how these spectral weights are applied to measured RIRs in order to achieve individually

synthesized BRIRs for different head orientations. Finally, a short review of the technical implementations for presenting the binaural signals dynamically is provided, including the head tracker device and the employed algorithms for the real-time head-tracked signal playback.

In **chapter 5**, the first study related to the third contribution is presented, in which the VAH approach is evaluated for dynamic auralizations of a reverberant and an anechoic environment, with sources in and outside the horizontal plane. The investigations are based on the 24-channel planar microphone array of Rasumow et al. [15] in both environments. Individual spectral weights, calculated for different head orientations of the listener, are applied to the RIRs measured with the VAH to result in individually synthesized BRIRs for different head orientations. The spectral weights are calculated with two different values for the minimum desired mean WNG and different selected directions considered in the desired directivity pattern, either 72 directions from the horizontal plane or $3{\times}72{=}216$ directions from the horizontal as well as two non-horizontal planes. Accordingly, different versions of individual BRIRs are synthesized with the VAH, which are used for the auralizations. Dynamic auralizations are also presented with measured BRIRs of a conventional artificial head as well as a simplified model of the head (a rigid sphere with two microphones at the ear positions), for which the BRIRs were measured for different head orientations in order to enable dynamic presentations. The auralizations with different BRIRs are evaluated perceptually with respect to room- and source-related attributes as well as overall quality in comparison to real sound source presentations in listening tests with ten subjects. Despite sources both in and outside the horizontal plane in the auralizations, the results show that it is advantageous to consider only horizontal directions in the calculation of the spectral weights. The smaller value for the minimum desired mean WNG is shown to lead to generally lower perceptual ratings, indicating the possibly increased susceptibility of the VAH synthesis to deviations in microphone characteristics. In addition, the auralization of the reverberant environment indicates that the presence of reverberation helps to cover up synthesis inaccuracies and thus improve the perceived quality of the VAH. Furthermore, the evaluation of dynamic auralizations of the two acoustical environments with non-individual BRIRs of conventional artificial heads indicate that if binaural speech signals can be presented dynamically, individual auralizations are not necessary for a perceptually convincing spatial impression. However, it should be realized that the acquisition of BRIRs for different head orientations of conventional artificial heads necessitates repeated measurements with different head orientations, whereas with the VAH, BRIRs for different head orientations are synthesized with measurements performed with a single orientation of the VAH. The VAH approach is evaluated as a well-suited solution for close-to-reality dynamic auralizations of speech signals. The publications related to this chapter are [151–153].

In **chapter 6**, the second study related to the third contribution is presented, in which two VAHs with different topologies (the 24-channel VAH of Rasumow et al. [15] and a three-dimensional array with 31 microphones) are used

to auralize an anechoic environment with sources in and outside the horizontal plane. Similar to chapter 5, dynamic auralizations are also presented with measured BRIRs of a conventional artificial head, for which the BRIRs were measured for different head orientations. The first part of the study concerns the localization performance when listening to virtual sources generated with BRIRs synthesized with both VAHs. The aim is to indicate to which extent the visual information about the sound sources in the investigations in chapter 5 contributed to the good performance of the VAH. Therefore, the localization experiment with 14 subjects takes place without providing any visual cues about the sound sources to the listener. Repeating the similar localization experiment with real (hidden) sound sources, the results indicate that a similar localization accuracy with respect to azimuth, externalization and front-back reversal occurrence can be achieved with virtual sources generated with the VAHs as with real sources. In line with the results of chapter 5, for both VAHs, the localization accuracy is higher when including only horizontal directions in the calculation of the spectral weights than when including both horizontal and non-horizontal directions. The second part of the study concerns the impact of head tracking on the localization performance with virtual sources. The aim is to investigate the possibly positive impact of head tracking on the perceptual results in chapter 5. Two localization experiments are performed, one with and one without head tracking. It is shown that without head tracking, the localization accuracy degrades drastically with respect to azimuth, externalization and front-back reversal occurrence, confirming the importance of dynamic presentation for the localization accuracy when listening to virtual sources generated with both VAHs. Moreover, when dynamically presented, virtual sources generated with non-individual BRIRs of a conventional artificial head can lead as well to convincing localization performances. Taking into account the fact that in practical applications, binaural recordings with conventional artificial heads cannot be presented dynamically, the results indicate that the possibility of presenting binaural signals dynamically is the major advantage of the VAH approach over conventional artificial heads. The publications related to this chapter are [154–156]

In **chapter 7**, the main results of the thesis are summarized and an outlook on potential further research is provided.

2

SYNTHESIZING HRTF DIRECTIVITY PATTERNS WITH A VIRTUAL ARTIFICIAL HEAD

2.1 Introduction

This chapter provides a summary of methods employed for the calculation of the Virtual Artificial Head (VAH) spectral weights, mainly based on the previous studies on the VAH approach by Rasumow et al. [16, 17, 142]. The chapter starts in section 2.2 with the VAH as filter-and-sum beamformer and the calculation of the spectral weights by minimizing a least-squares cost function with the aim of synthesizing HRTF directivity patterns. Section 2.3 deals with the need for regularization constraints to increase the robustness. Section 2.4 reviews the spectro-spatial smoothing of HRTF directivity patterns prior to calculating the spectral weights.

2.2 Filter-and-sum beamformer with least-squares solution for synthesizing HRTF directivity patterns

Consider the microphone array shown in Figure 2.1a as a two-dimensional array with N microphones distributed in the x-y plane and the sound source at direction $\Theta = (\theta, \phi = 0°)$ in the horizontal plane[1]. The sound source is assumed to be in the far-field of the microphone array (plane wave propagation). The output $\mathrm{Z}(f, \Theta)$ of this microphone array is the weighted sum of the microphone signals $\mathrm{Y}_n(f, \Theta)$, each weighted with a complex-valued *spectral weight* $w_n(f)$

$$\mathrm{Z}(f, \Theta) = \sum_{n=1}^{N} w_n^*(f)\mathrm{Y}_n(f, \Theta) = \mathbf{w}^H \mathbf{Y}(f, \Theta), \tag{2.1}$$

1 The two-dimensional array and the sound source in the horizontal plane (ϕ=0°) are considered here for the sake of simplicity. All results can be extended to three-dimensional arrays and the source at any elevation ϕ.

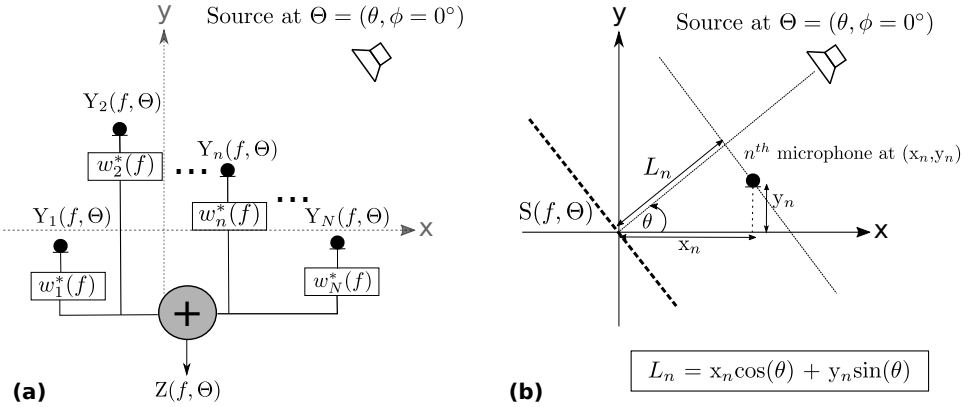

Fig. 2.1: **(a)**: Filter-and-sum beamformer as a two-dimensional array with N microphones distributed in the x-y plane. $Y_n(f,\Theta)$: signal of the n^{th} microphone at frequency f and direction Θ, $w_n(f)$: complex-valued spectral weight for the n^{th} microphone at frequency f, and $Z(f,\Theta)$: output signal of the beamformer at frequency f and direction Θ as the weighted sum of the microphone signals. **(b)**: Direction dependent path L_n between the n^{th} microphone at (x_n,y_n) and the center of the microphone array (origin of the coordinate system). $S(f,\Theta)$ indicates the source signal at frequency f and at direction Θ arriving at the center of the microphone array.

where $(.)^H$ denotes the Hermitian transpose and $(.)^*$ denotes the complex conjugate. The vector $\mathbf{w}(f) = [w_1(f), w_2(f), ...w_N(f)]^T$ contains the complex-valued spectral weights for the N microphones. Consider $S(f,\Theta)$ as the source signal arriving at the center of the array (origin of the coordinate system). Assuming omni-directional microphones and under the assumed far-field conditions, the signal of the n^{th} microphone can be considered the same as $S(f,\Theta)$, shifted by the time delay between this microphone and the center of the array

$$Y_n(f,\Theta) = e^{-j2\pi f \tau_n(\Theta)} S(f,\Theta). \tag{2.2}$$

The relative delay $\tau_n(\Theta)$ is given by

$$\tau_n(\Theta) = L_n(\Theta)/c, \tag{2.3}$$

with c the speed of sound and $L_n(\Theta)$ the length of the direction-dependent path which the plane wave travels between the n^{th} microphone and the center of the array as shown in Figure 2.1b. The directivity pattern $H(f,\Theta)$ of the beamformer is given as the relation between output $Z(f,\Theta)$ and signal $S(f,\Theta)$

$$H(f,\Theta) = \frac{Z(f,\Theta)}{S(f,\Theta)}, \tag{2.4}$$

which in combination with Eqs. (2.1) and (2.2) can be written as

$$\mathrm{H}(f,\Theta) = \sum_{n=1}^{N} w_n^* e^{-j2\pi f \tau_n(\Theta)} = \mathbf{w}^H(f)\mathbf{d}(f,\Theta). \tag{2.5}$$

The $N \times 1$ steering vector $\mathbf{d}(f,\Theta)$ indicates the direction- and frequency-dependent free-field transfer function between a source at direction Θ and the N microphones. According to Eq. (2.5), under far-field conditions, steering vectors can be modeled as pure relative delays between the microphones and the center of the array

$$\mathbf{d}(f,\Theta) = [e^{-j2\pi f \tau_1(\Theta)}, e^{-j2\pi f \tau_2(\Theta)}, ...e^{-j2\pi f \tau_n(\Theta)}, ..., e^{-j2\pi f \tau_N(\Theta)}]^T, \tag{2.6}$$

with $\tau_n(\Theta)$ given in Eq. (2.3).

Intending to synthesize HRTF directivity patterns, the aim is to calculate the spectral weights $\mathbf{w}(f)$ such that the directivity pattern $\mathrm{H}(f,\Theta)$ of the beamformer resembles the desired directivity pattern $\mathrm{D}(f,\Theta)$ of the left or right HRTFs. Assume that these HRTFs are known at a number of P discrete directions Θ_k, $k = 1, 2, ..., P$, such that the desired directivity pattern at frequency f is given as $\mathbf{D}(f) = [\mathrm{D}(f,\Theta_1), \mathrm{D}(f,\Theta_2), ..., \mathrm{D}(f,\Theta_P)]$. The calculation of the spectral weights is considered as the problem of minimizing the error vector $\mathbf{e}$ in

$$\mathbf{D}(f) = \mathbf{w}^H(f)\mathbf{A}(f) + \mathbf{e}, \tag{2.7}$$

where $\mathbf{A}(f) = [\mathbf{d}(f,\Theta_1), \mathbf{d}(f,\Theta_2), ...\mathbf{d}(f,\Theta_P)]$ is the $N \times P$ matrix of steering vectors with $\mathbf{d}(f,\Theta_k)$ given by Eq. (2.6). One well-known solution for the problem in Eq. (2.7) is given by minimizing a least-squares cost function $\mathrm{J_{LS}}$, defined as the sum over P directions of the squared absolute deviations between synthesized and desired directivity patterns

$$\mathrm{J_{LS}}(\mathbf{w}(f)) = ||\mathbf{e}||_2^2 = \sum_{k=1}^{P} |\mathbf{w}^H(f)\mathbf{d}(f,\Theta_k) - \mathrm{D}(f,\Theta_k)|^2. \tag{2.8}$$

This cost function can be written as [139, 143]

$$\mathrm{J_{LS}}(\mathbf{w}(f)) = \mathbf{w}^H(f)\mathbf{Q}(f)\mathbf{w}(f) - \mathbf{w}^H(f)\mathbf{a}(f) - \mathbf{a}^H(f)\mathbf{w}(f) + d(f), \tag{2.9}$$

with

$$\begin{aligned} \mathbf{Q}(f) &= \sum_{k=1}^{P} \mathbf{d}(f,\Theta_k)\mathbf{d}^H(f,\Theta_k) \\ \mathbf{a}(f) &= \sum_{k=1}^{P} \mathbf{d}(f,\Theta_k)\mathrm{D}^*(f,\Theta_k) \\ d(f) &= \sum_{k=1}^{P} |\mathrm{D}(f,\Theta_k)|^2. \end{aligned} \tag{2.10}$$

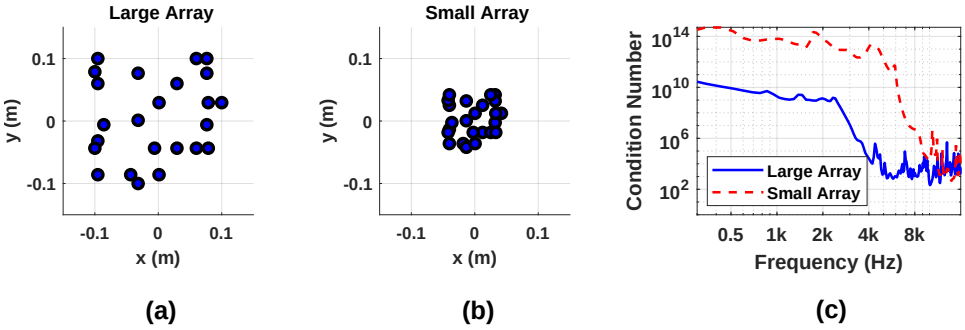

Fig. 2.2: Microphone positions for **(a)**: planar microphone array with 24 microphones [15], and **(b)**: 42% donw-sized copy of array shown in (a). **(c)**: Condition number of matrix $\mathbf{Q}$ defined in Eq. (2.10), over frequency, for the arrays in (a) and (b).

By setting the gradient of $\mathrm{J_{LS}}(\mathbf{w}(f)$ to zero, spectral weights minimizing the cost function $\mathrm{J_{LS}}(\mathbf{w}(f))$ are derived as [139, 143]

$$\mathbf{w}(f) = \mathbf{Q}^{-1}(f)\mathbf{a}(f). \tag{2.11}$$

2.3 Regularization

In general, matrix $\mathbf{Q}$ in Eq. (2.11) can become ill-conditioned, leading to numerical errors during its inversion. In such a case, even small deviations in the steering vector $\mathbf{d}$ can lead to large errors in the synthesized directivity pattern. This may lead to an undesirable amplification of the microphone self-noise. Deviations in steering vector $\mathbf{d}$ can be caused by small displacements of microphones, changes in microphone characteristics, or temperature fluctuations. The sensitivity to these deviations is higher if inter-microphone distances are small and the signals arriving at microphones exhibit high similarity to each other. Figure 2.2c shows the condition number[2] of matrix $\mathbf{Q}$ for two exemplary microphone arrays of different sizes shown in Figures 2.2a and 2.2b. Steering vectors for these arrays were calculated using Eq. (2.6) for P=24 equiangular directions from the horizontal plane ($\theta_k = 0°, 15°, ..., 345°, \phi_k = 0°$). As the condition numbers in Figure 2.2c show, numerical errors during the inversion of matrix $\mathbf{Q}$ can be more critical especially at low frequencies and for the smaller array. Therefore, the impact of microphone self-noise is expected to be more prominent at lower frequencies and for the smaller array. The deviations in steering vectors can be accounted for by including the probability density function of microphone characteristics [139, 141] or by joint optimization, considering multiple sets of measured steering vectors [17]. Other commonly used

2 Condition number of a non-singular matrix $\mathbf{A}$ is given by $\|\mathbf{A}\|_2 . \|\mathbf{A}^{-1}\|_2$ and is a measure for the potential sensitivity of the solution $\mathbf{x}$ of equation $\mathbf{Ax}$=$\mathbf{b}$ to changes in $\mathbf{b}$ [157].

methods to handle the ill-conditioned problem are singular value truncation [117] or Tikhonov regularization [145,158,159]. A commonly used measure of the robustness is the White Noise Gain (WNG), defined as the output power of the beamformer for a desired look direction Θ_{look} compared to the output power for spatially uncorrelated noise [136,141,160]

$$\mathrm{WNG}(\mathbf{w}(f), \Theta_{look}) = \frac{|\mathbf{w}^H(f)\mathbf{d}(f, \Theta_{look})|^2}{\mathbf{w}^H(f)\mathbf{w}(f)}. \tag{2.12}$$

The larger the WNG, the larger will be the attenuation of the spatially uncorrelated noise compared to the output power at the look direction. The robustness can be accounted for by imposing a WNG constraint on spectral weights. However, when synthesizing HRTF directivity patterns, all P directions must be considered as a desired look direction. In [17, 142], it was shown that for synthesizing HRTF directivity patterns, the *mean* WNG over P directions ($\mathrm{WNG_m}$), defined as

$$\mathrm{WNG_m}(\mathbf{w}(f)) = \frac{1}{P}\sum_{k=1}^{P}\mathrm{WNG}(\mathbf{w}(f), \Theta_k), \tag{2.13}$$

should be constrained. In combination with Eq. (2.12), the $\mathrm{WNG_m}$ in Eq. (2.13) can be written as

$$\mathrm{WNG_m}(\mathbf{w}(f)) = \frac{\mathbf{w}^H(f)\mathbf{Q}_m(f)\mathbf{w}(f)}{\mathbf{w}^H(f)\mathbf{w}(f)}, \tag{2.14}$$

with

$$\mathbf{Q}_m(f) = \frac{1}{P}\sum_{k=1}^{P}\mathbf{d}(f, \Theta_k)\mathbf{d}^H(f, \Theta_k). \tag{2.15}$$

By setting a minimum desired value β for the $\mathrm{WNG_m}$ in dB, the regularized spectral weights can be calculated by solving the new constrained minimization problem

$$\min \mathrm{J_{LS}}(\mathbf{w}(f)) \quad \text{subject to} \quad 10\log_{10}(\mathrm{WNG_m}(\mathbf{w}(f)))\mathrm{dB} \geq \beta. \tag{2.16}$$

Note that the minimum desired value of $\mathrm{WNG_m}$, i.e. β, is defined in dB in Eq. (2.16). When applying the $\mathrm{WNG_m}$ constraint in the calculation of the spectral weights, β is recalculated into β_{power}, with $\beta_{power} = 10^{\frac{\beta}{10}}$. In [17, 142], the method of Lagrange multipliers was employed to formulate the constrained optimization problem in Eq.(2.16) as

$$\mathrm{J_m}(\mathbf{w}(f), \mu) = \mathrm{J_{LS}}(\mathbf{w}(f)) + \mu(\mathbf{w}^H(f)\mathbf{w}(f) - \frac{1}{\beta_{power}}\mathbf{w}^H(f)\mathbf{Q}_m(f)\mathbf{w}(f)), \tag{2.17}$$

with μ the Lagrange multiplier. Setting the gradient of $\mathrm{J_m}(\mathbf{w}(f), \mu)$ to zero, spectral weights minimizing the cost function in Eq. (2.17) are derived as [143]

$$\mathbf{w}(f, \mu) = (\mathbf{Q}(f) + \mu(\mathbf{I}_N - \frac{1}{\beta_{power}}\mathbf{Q}_m(f)))^{-1}\mathbf{a}(f). \tag{2.18}$$

In Eq. (2.18), $\mathbf{I}_N$ is the $N \times N$ identity matrix, and $\mathbf{Q}(f)$ and $\mathbf{a}(f)$ are given in Eq. (2.10). The inclusion of the Lagrange multiplier μ in the calculation of the spectral weights introduces unavoidable errors in the synthesized HRTFs. In order to find the optimal value for the Lagrange multiplier as a trade-off between synthesis accuracy and robustness, μ is increased gradually in steps of e.g. $\Delta\mu = \frac{1}{100}$, until the minimum desired value for $\mathrm{WNG_m}$ or an upper limit of μ_{max} is reached [142].

2.4 Spectro-spatial smoothing of HRTFs

It is easier for a VAH to resemble a smoother directivity pattern, especially if the number of microphones is limited. Rasumow et al. [16] proposed a method for HRTF smoothing in spectral and spatial domains and showed that this smoothing is imperceptible. The spectral smoothing consists of a complex smoothing of HRTFs, presupposing the phase spectra to be substituted by linear phases above a certain cutoff frequency. It was shown that this cutoff frequency can be as low as 1 kHz without being detectable. After phase linearizations above this cutoff frequency, the magnitude and phase spectra of HRTFs are simultaneously smoothed into constant relative bandwidths of $\frac{1}{5}$ octaves. In the next step, in the spatial smoothing, the HRTF directivity patterns are smoothed by leveling out the spatial notches. It was shown in [16] that there is no need to retain spatial notches if they are less than 29 dB below the maximum level in the directivity pattern. It was also shown in [143] that the synthesis error with a VAH is reduced when the desired HRTF directivity patterns are smoothed.

The spectro-spatial smoothing proposed by Rasumow et al. [16] was perceptually evaluated only for horizontal HRTFs. Nevertheless, in this thesis, non-horizontal HRTFs were smoothed the same as horizontal HRTFs, i.e. HRTF phase spectra were linearized for frequencies above 2 kHz, and the HRTFs were smoothed complexly into constant relative bandwidths of $\frac{1}{5}$ octave. The spatial notches were smoothed if they were less than 29 dB below the maximum level in the directivity pattern at each elevation. Without perceptual experiments, it cannot be assessed to which extent the smoothed non-horizontal HRTFs are perceptually distinguishable from original ones e.g. with respect to coloration, detectability of phase linearization, or variation in the notch depths or peak heights in the spectrum. However, the potential shift of pinna-related spectral notches, which contribute to the elevation perception, can be roughly estimated. These notches appear in the HRTF magnitude spectrum in the frequency range above approx. 5 kHz and vary depending on the elevation of the incoming sound [24, 72–74]. Figure 2.3 shows original and smoothed left HRTFs spectra at elevations $\phi = -30°$ to $+30°$ in the median plane for exemplary three subjects in this thesis. The frequency of the first pinna notch, which has been shown to be the most important notch for the vertical localization accuracy [161], is marked both for original and smoothed HRTF spectra in Figure 2.3. The notches were indicated at the first local minimum above 5 kHz in the spectrum of HRTFs. According to Moore et al. [162], a shift of 4° in the elevation of a source between

$-20°$ and $+20°$ corresponds to a shift of about 4% in the center frequency of the prominent notch in the HRTF spectrum. Since the notch frequency shifts in the smoothed HRTF spectra in Figure 2.3 were below 4%, the spectral smoothing of HRTFs was assumed to lead only to changes of the perceived sound source below the minimum audible angle in the vertical direction (around 3.6° [41]). In addition, similar to horizontal HRTFs, the spatial smoothing of non-horizontal HRTFs was assumed to be noncritical since the spatial notches in the HRTF directivity pattern occur mostly at contralateral directions and the role of the contralateral ear on the elevation perception is known to decrease for sound sources away from the median plane [163, 164].

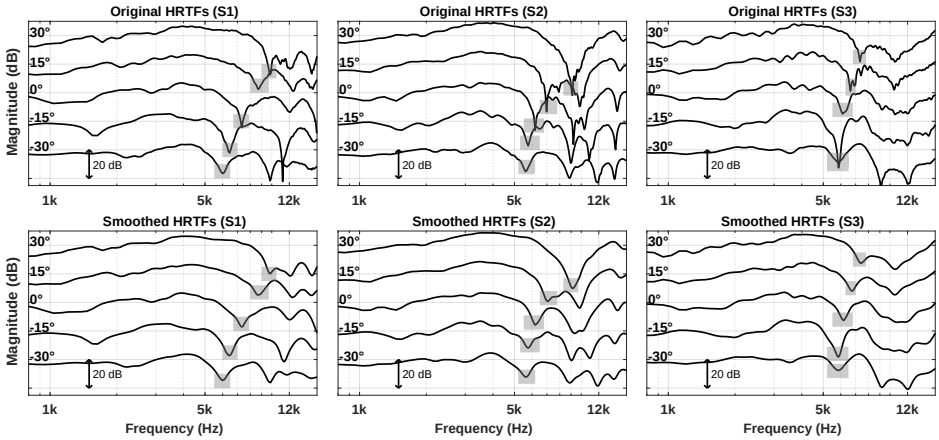

Fig. 2.3: Magnitude spectra of left HRTFs of three exemplary subjects (S1, S2, and S3) in the median plane at elevations $-30°$, $-15°$, $0°$, $+15°$, and $+30°$ (shifted vertically for convenience). The first prominent notch in the spectrum is marked with the grey box. **Top:** Originally measured HRTFs. **Bottom:** Spectro-spatially smoothed HRTFs according to Rasumow et al. [16].

2.5 Summary

In this chapter, an overview of the methods employed for the calculation of the VAH spectral weights was presented. The VAH as a filter-and-sum beamformer was introduced and the calculation of the regularized spectral weights by minimizing a narrow-band least-squares cost function subject to a constraint on the mean White Noise Gain ($\mathrm{WNG_m}$) was explained. Finally, the spectro-spatial smoothing of HRTFs was briefly reviewed. In the remainder of the thesis, the spectro-spatial smoothing of HRTFs and the least-squares cost function $\mathrm{J_{LS}}$ in Eq. (2.8) will remain the same as discussed in this chapter. The $\mathrm{WNG_m}$ constraint in Eq. (2.16) will be applied to the minimization of the cost function $\mathrm{J_{LS}}$, together with other additional

constraints within a constrained optimization problem, as will be discussed in the next chapter.

3

IMPROVEMENT OF SPATIAL RESOLUTION AND BANDWIDTH OF HRTF SYNTHESIS IN THE HORIZONTAL PLANE USING A VIRTUAL ARTIFICIAL HEAD

3.1 Introduction

The reflection and diffraction of the sound caused by the listener's torso, head, and external ear provide the listener with important cues for the spatial impression of the sound field. The objective of binaural technology is to preserve these cues as described by the Head Related Transfer Functions (HRTFs) in order to recreate the spatial impression of the sound field. So-called artificial heads are used as an established binaural recording method to resemble the information encompassed in the HRTFs. However, a well-known drawback of artificial heads is their non-individual design, leading to problems such as in-head-localization or front-back reversals [83, 95]. In addition, when presenting signals recorded with an artificial head, it is not easily possible to include head movements of the listener via head tracking during signal playback.

As an alternative to conventional artificial heads, a Virtual Artificial Head (VAH), designed as a microphone array-based filter-and-sum beamformer, can be used to synthesize the directivity patterns of individual HRTFs. This is done by applying the complex-valued spectral weights, calculated individually for each listener,

This chapter is based on:
[149] M. Fallahi, M. Hansen, S. Doclo, S. van de Par, D. Püschel, and M. Blau, "Individual binaural reproduction with high spatial resolution using a virtual artificial head with a moderate number of microphones", in preparation.
[150] M. Fallahi, M. Hansen, S. Doclo, S. van de Par, V. Mellert, D. Püschel, and M. Blau, "High spatial resolution binaural sound reproduction using a virtual artificial head", *Proc. of Fortschritte der Akustik - DAGA*, Kiel, Germany, pp. 1061-1064, 2017.

to the microphone signals. By doing so, not only the recorded signals can be individualized for different listeners, but also the listener's head movements can be accounted for during signal playback by adjusting the spectral weights according to the orientation of the listener's head.

The calculation of the VAH spectral weights following Rasumow et al. [16,17,142] was presented in chapter 2, according to which, spectral weights are calculated by minimizing a narrow-band least-squares cost function with a constraint on the mean White Noise Gain ($\mathrm{WNG_m}$). The $\mathrm{WNG_m}$ constraint is applied in order to increase robustness against small deviations in microphone positions and characteristics and to limit the microphone self-noise amplification [17]. However, following these methods, the spatial resolution of the synthesis is limited by the number of microphones. In [15], Rasumow et al. used the VAH as a planar microphone array with 24 microphones to synthesize 24 equiangular horizontal HRTFs. The evaluation was performed with respect to localization, spectral coloration, and overall performance at six relevant directions; three directions coincided with directions explicitly considered in the optimization, at which perceptually comparable or better results compared to a conventional artificial head could be achieved. The other three evaluated directions were at intermediate directions, at which the synthesis accuracy degraded.

Aiming at achieving a higher spatial resolution than in [15] using the same number of microphones, in this chapter, a novel constrained optimization method is applied and evaluated to compute the spectral weights. In addition to imposing a $\mathrm{WNG_m}$ constraint, this method imposes upper and lower boundaries on the spectral error at a large number of directions. The main research questions to be answered are: (1) can the performance of a VAH be improved with the proposed optimization method? and (2) what is the impact of the array topology on the performance of a VAH using the proposed optimization method? These questions are addressed based on simulations with different array topologies and perceptual evaluations.

The remainder of this chapter is structured as follows. In section 3.2, the VAH as a narrow-band filter-and-sum beamformer is reviewed again based on the method of Rasumow et al. [15] as also introduced in sections 2.2 and 2.3. In section 3.3, the spatial and spectral performance of the synthesis with a VAH using the methods of Rasumow et al. [15] is addressed. In section 3.4, the method proposed to increase the spatial resolution of the synthesis is introduced. In section 3.5, the impact of array topology on the performance of a VAH using the proposed method is discussed. The chapter continues with perceptual evaluations in section 3.6. Finally, section 3.7 offers a brief discussion about the relaxation of the constraints and its impact on the feasibility of satisfying them.

3.2 Optimization method with a White Noise Gain constraint

A VAH can be considered as a filter-and-sum beamformer with N microphones, as shown in Figure 3.1. The synthesized directivity pattern of this filter-and-sum beamformer at frequency f and direction $\Theta_k = (\theta_k, \phi_k)$, with θ the azimuth angle and ϕ the elevation angle, can be expressed as the scalar product

$$\mathrm{H}(f, \Theta_k) = \mathbf{w}^H(f)\mathbf{d}(f, \Theta_k) = \sum_{n=1}^{N} w_n^*(f)\mathrm{d}_n(f, \Theta_k), \tag{3.1}$$

where $(.)^H$ denotes the Hermitian transpose and $(.)^*$ denotes the complex conjugate. The $N \times 1$ steering vector $\mathbf{d}(f, \Theta_k) = [\mathrm{d}_1(f, \Theta_k), ..., \mathrm{d}_N(f, \Theta_k)]^T$ contains the free-field acoustic transfer functions between a source at direction Θ_k and the N microphones of the microphone array. The $N \times 1$ vector $\mathbf{w}(f) = [w_1(f), ..., w_N(f)]^T$ contains the complex-valued spectral weights at frequency f for each of the N microphones. These spectral weights can be transformed to filter coefficients of FIR filters using the inverse Fourier transform. The goal is to calculate spectral weights $\mathbf{w}(f)$ such that the resulting directivity pattern $\mathrm{H}(f, \Theta_k)$ resembles a desired directivity pattern $\mathrm{D}(f, \Theta_k)$ at P discrete directions $\Theta_k, k = 1, 2, ..., P$. In our case, D is derived from the individual HRTF directivity pattern of the left or the right ear. Once the spectral weights $\mathbf{w}_\mathrm{L}(f)$ and $\mathbf{w}_\mathrm{R}(f)$ for the left and the right ear are calculated, the corresponding FIR filters are applied to the recorded signals of the VAH microphones and added up to result in the binaural left and right signals. In other words, by applying the spectral weights, the HRTF directivity patterns of individual listeners are included in the recording.

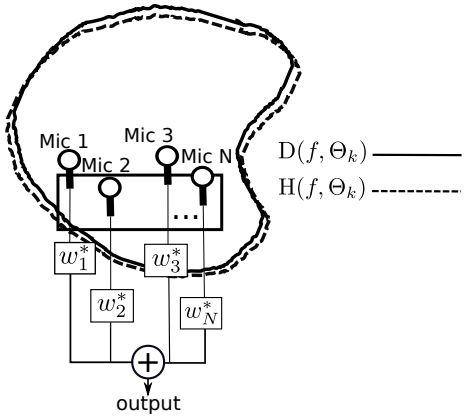

Fig. 3.1: Synthesized directivity pattern $\mathrm{H}(f, \Theta_k)$ of the VAH as a filter-and-sum beamformer with N microphones and N complex-valued spectral weights aiming at resembling the desired directivity pattern $\mathrm{D}(f, \Theta_k)$.

The spectral weights can be calculated, separately for the left and the right ear, by minimizing a narrow-band cost function, i.e. independently for each frequency bin. A least-squares-based cost function was proposed in [17,139], i.e. as the sum of the squared absolute differences between the desired and the synthesized directivity patterns over P discrete directions, for $\zeta \in$ {L,R} (Left or Right)

$$\begin{aligned} \mathrm{J_{LS}}(\mathbf{w}_\zeta(f)) &= \sum_{k=1}^{P} |\mathrm{H}_\zeta(f,\Theta_k) - \mathrm{D}_\zeta(f,\Theta_k)|^2 \\ &= \sum_{k=1}^{P} |\mathbf{w}_\zeta^H(f)\mathbf{d}(f,\Theta_k) - \mathrm{D}_\zeta(f,\Theta_k)|^2. \end{aligned} \tag{3.2}$$

Since the minimization of $\mathrm{J_{LS}}$ in Eq. (3.2) can become ill-conditioned, especially at low frequencies, small deviations in the steering vector $\mathbf{d}$ (e.g. due to small microphone displacements, aging, or temperature fluctuations) may lead to large deviations in the spectral weights $\mathbf{w}$. This in turn may lead to amplification of the self-noise of the microphones, which is a well-known issue in beamforming [108,136,141,165,166]. Different approaches have been suggested to deal with the ill-conditioned problem, e.g. using singular value truncation [117] or by applying Tikhonov regularization [17,158], thereby increasing the robustness of the beamformer against deviations in the steering vectors. A commonly used measure for the robustness of a beamformer is the White Noise Gain (WNG), defined as the ratio between the output power of the array for a source at one direction (typically the look direction of the beamformer) and the output power for spatially uncorrelated noise [108,136,141]. Correspondingly, the spectral weights can be calculated by constraining the WNG to improve robustness. In contrast to beamformers with one desired look direction, in [17,142] it was shown that for synthesizing HRTF directivity patterns the *mean* White Noise Gain ($\mathrm{WNG_m}$), averaged over all considered P directions, defined as

$$\mathrm{WNG_m}(\mathbf{w}_\zeta(f)) = \frac{1}{P}\sum_{k=1}^{P} \frac{|\mathbf{w}_\zeta^H(f)\mathbf{d}(f,\Theta_k)|^2}{\mathbf{w}_\zeta^H(f)\mathbf{w}_\zeta(f)} \tag{3.3}$$

should be constrained. Imposing a minimum desired value β in dB onto the $\mathrm{WNG_m}$, the regularized spectral weights can be calculated by solving the constrained optimization problem

$$\min \mathrm{J_{LS}}(\mathbf{w}_\zeta(f)) \quad \text{s.t.} \quad 10\log_{10}(\mathrm{WNG_m}(\mathbf{w}_\zeta(f)))\mathrm{dB} \geq \beta. \tag{3.4}$$

Rasumow et al. [17] employed the method of Lagrange multipliers to solve the problem in Eq. (3.4), where the Lagrange multiplier was increased gradually until the minimum desired value of $\mathrm{WNG_m}$ was reached. This way, the smallest possible value of the Lagrange multiplier satisfying Eq. (3.4) could be found. At the same time, the non-zero Lagrange multiplier introduced unavoidable deviations in the synthesized

HRTF directivity patterns from the desired directivity patterns. The limitations of this optimization approach will be discussed in more detail in the following section.

3.3 Spatial and spectral performance of HRTF synthesis with a VAH

As can be seen from Eq. (3.2), a selected number of P directions is included in the cost function used to calculate the spectral weights for a VAH. These directions are referred to as calibration directions. As can be intuitively expected, a larger synthesis error in the resulting HRTF directivity pattern may occur for directions that were not included in the calculation of the spectral weights than for the calibration directions.

As an example, for the planar microphone array of Rasumow et al. with N=24 microphones (see Figure 1.4) the steering vectors were simulated as pure relative delays $\tau_n(\Theta_k)$ between the n^{th} microphone and the center of the array as

$$\mathbf{d}(f,\Theta_k) = [e^{-2\pi f \tau_1(\Theta_k) j}, ..., e^{-2\pi f \tau_n(\Theta_k) j}, ...e^{-2\pi f \tau_N(\Theta_k) j}]^T, \tag{3.5}$$

i.e. assuming perfectly matched microphones and far-field conditions. P=24 calibration directions with 15° resolution in the horizontal plane ($\theta_k = 0°, 15°, ...345°, \phi_k = 0°$) were considered. Left and right horizontal HRTFs measured at these directions were considered as desired directivity pattern, with 0° referring to the frontal direction and 90° to the left side. In this example and in the following sections, HRTFs measured for one of the subjects in this study (subject 1, see section 3.6) were considered to be synthesized with the VAH. HRTFs were spectro-spatially smoothed according to [16], as this smoothing was shown to cause no perceptible degradation. Spectral weights were calculated by solving Eq. (3.4) for the left and right ear separately. These spectral weights were then used to synthesize the HRTFs at $P^{'}$=72 horizontal directions with a resolution of 5° ($\theta_k^{'} = 0°, 5°, ..., 355°, \phi_k^{'} = 0°$).

As a measure for the spectral accuracy of the results, Spectral Distortion (SD), defined as the spectral level difference in dB between the desired and synthesized HRTFs, was calculated for each frequency and at each synthesis direction $\Theta_k^{'}$,

$$\mathrm{SD}_\zeta(f,\Theta_k^{'}) = 10\log_{10}(\frac{|\mathbf{w}_\zeta^H(f)\mathbf{d}(f,\Theta_k^{'})|^2}{|\mathrm{D}_\zeta(f,\Theta_k^{'})|^2})\mathrm{dB}. \tag{3.6}$$

To evaluate the phase accuracy in the frequency range of $f \leq 2$ kHz, Temporal Distortion (TD), defined as the error in the timing of a single frequency component of the HRTF was derived from the phase angle as

$$\mathrm{TD}_\zeta(f,\Theta_k^{'}) = \frac{\angle\mathbf{w}_\zeta^H(f)\mathbf{d}(f,\Theta_k^{'}) - \angle\mathrm{D}_\zeta(f,\Theta_k^{'})}{2\pi f}. \tag{3.7}$$

In Eqs. (3.6) and (3.7), D corresponds to the spectro-spatially smoothed desired directivity pattern according to [16] as also described in section 2.4. All evaluations were performed with respect to this smoothed desired directivity pattern. Figure 3.2 shows the resulting SD, TD, $\mathrm{WNG_m}$, the directivity patterns of synthesized and desired HRTFs at 6 kHz, as well as the absolute value of the calculated complex-valued spectral weights for four exemplary microphones. The upper row shows the results for the case where no $\mathrm{WNG_m}$ regularization was applied, leading to a perfect synthesis at synthesis directions coinciding with the calibration directions ($\theta_k^{'} = 0°, 15°, ...$), but large spectral distortions at the intermediate synthesis directions ($\theta_k^{'} = 5°, 10°, 20°, 25°...$). The middle row shows the results for the case where $\mathrm{WNG_m}$ regularization with β=0 dB was applied. Choosing β=0 dB was based on the perceptual evaluations in [142], where this value was shown to be appropriate for the considered microphone array. As can be observed, constraining

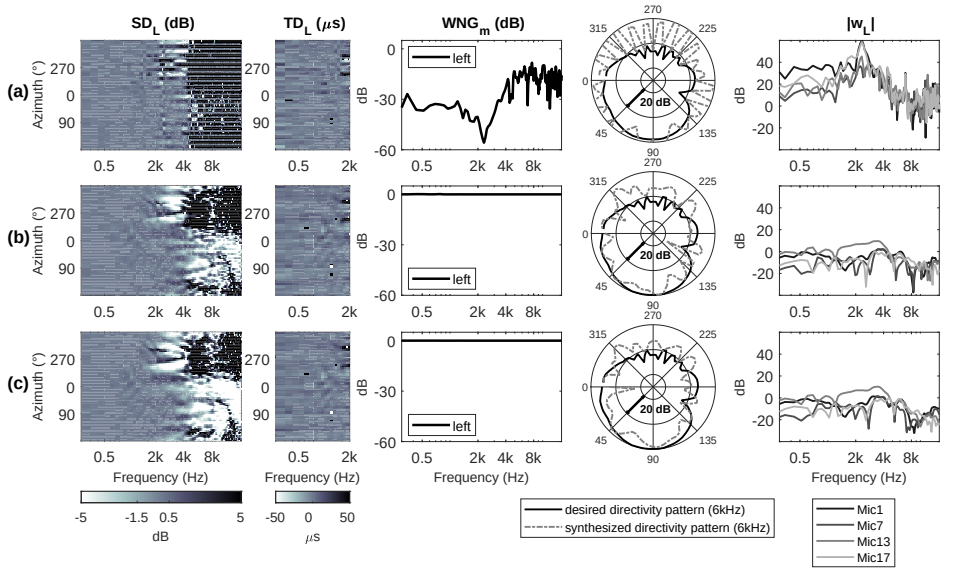

Fig. 3.2: From left to right: Resulting Spectral Distortion (SD), Temporal Distortion (TD) up to 2 kHz, mean WNG ($\mathrm{WNG_m}$), directivity patterns of the desired and synthesized HRTFs at 6 kHz, and spectral weights (shown for four exemplary microphones). Synthesis was done at P'=72 horizontal synthesis directions with spectral weights calculated with **(a)**: P=24 horizontal directions (subset of the 72 synthesis directions) without regularization, **(b)**: P=24 horizontal directions with $\mathrm{WNG_m}$ regularization with β=0 dB, and **(c)**: P=72 horizontal directions with $\mathrm{WNG_m}$ regularization with β=0 dB. Results are shown for the left ear of subject 1 in this chapter.

the $\mathrm{WNG_m}$ led to a re-distribution of the error over the synthesis directions, improving the synthesis accuracy at intermediate directions, but impairing the synthesis accuracy at the calibration directions. The synthesis accuracy is limited

to frequencies below 2 kHz at the contralateral directions ($180° < \theta_k^{'} < 360°$ for the left ear) and below about 4 kHz at the ipsilateral directions. In addition, it can be observed that constraining the $\mathrm{WNG_m}$ resulted in smaller magnitudes of the spectral weights, which is the expected effect of regularization: in case of deviations in the steering vectors, these steering vectors are multiplied by smaller spectral weights in Eq. (3.1), which increases the robustness. As shown in the lower row, simply increasing the number of calibration directions to $P{=}P^{'}{=}72$ (while applying $\mathrm{WNG_m}$ regularization with $\beta{=}0$ dB) does not drastically improve the synthesis accuracy, which can be explained by the fact that the number of microphones is much smaller than the number of calibration directions.

3.4 Proposed constrained optimization method

Aiming at increasing the synthesis accuracy at a large number of directions while not increasing the number of microphones, in this section, it is proposed to add additional constraints to the least-squares-based optimization problem in Eq. (3.4). Since in this chapter only horizontal calibration and synthesis directions are used, i.e. $\phi_k = 0°$ and $\phi_k^{'} = 0°$, these directions are indicated only with the azimuth angle. Without loss of generality, it will be assumed in the remainder of the chapter that the calibration and the synthesis directions are the same (i.e. $\theta_k{=}\theta_k^{'}$ and $P{=}P^{'}$).

Constraints can be applied to the monaural and/or binaural aspects of the synthesized HRTFs. Interaural Level Differences (ILDs) and Interaural Time Differences (ITDs) are the two well-known binaural cues for localization. Deviations in the resulting ILDs and ITDs between synthesized and desired HRTFs can lead to the noticeable displacement of the virtual sound source [5]. Nevertheless, imposing constraints solely onto binaural cues would be insufficient since monaural errors would not be controlled for. Besides audible colorations, the lack of monaural spectral accuracy may also impact the perceived externalization of signals presented with headphones [83].

Constraints can also be applied to the monaural SDs as defined in Eq. (3.6). Spectral distortion has been defined and evaluated differently in different studies. Studying the effect of HRTF interpolation, Minnaar et al. [167] calculated the absolute magnitude differences between interpolated and reference HRTFs at 94 logarithmically distributed frequency points between 100 Hz and 20 kHz. The sum of the errors at the left and right sides was averaged over frequency. Mean errors exceeding 1 dB were reported to be audible. Brinkmann et al. [168] used a similar measure to study the effect of head-above-torso orientation in HRTFs, but instead of 94 frequency points, they calculated the error in 39 auditory bands between 70 Hz and 20 kHz. Their results showed mean errors of 0.5 dB to 1 dB to be audible in an ABX test paradigm. In [169], level differences between reference and pre-processed HRTFs were calculated in auditory filters between 50 Hz and 20 kHz. Giving more relevance to the ear side which contains more energy, the binaurally weighted sum

of the error at the left and right sides was considered as a measure for coloration error. The Just Noticeable Difference (JND) for this measure was assumed to be 1 dB.

In the current study, we propose to set frequency-independent upper and lower boundaries, $\mathrm{L_{Up}}$ and $\mathrm{L_{Low}}$, to the SD at the left and the right side, i.e.

$$\mathrm{L_{Low}}(\theta_k) \leq \mathrm{SD}_\zeta(f,\theta_k) \leq \mathrm{L_{Up}}(\theta_k), \tag{3.8}$$

for all synthesis directions θ_k, $k = 1, 2, ..., P$. The upper and lower boundaries $\mathrm{L_{Up}}$ and $\mathrm{L_{Low}}$ were chosen such that the deviation between the ILDs of the synthesized and the desired HRTFs, defined as

$$\begin{aligned}&\Delta\mathrm{ILD}(f,\theta_k) = \\ &|10\log_{10}\frac{|\mathbf{w}_\mathrm{L}^H(f)\mathbf{d}(f,\theta_k)|^2}{|\mathbf{w}_\mathrm{R}^H(f)\mathbf{d}(f,\theta_k)|^2} - 10\log_{10}\frac{|\mathrm{D_L}(f,\theta_k)|^2}{|\mathrm{D_R}(f,\theta_k)|^2}|\mathrm{dB} \\ &= |\mathrm{SD_L}(f,\theta_k) - \mathrm{SD_R}(f,\theta_k)|\mathrm{dB},\end{aligned} \tag{3.9}$$

is kept in a reasonable range. In different studies, the measured JND for ILD deviation has been shown to be between 0.6 dB and 2 dB (for sinusoids of different frequencies), and approximately 1.5 dB for broadband signals (clicks) in the median plane [5]. Although the JND for ILD deviations can get larger with increasing lateralization, here, $\mathrm{L_{Up}}$ and $\mathrm{L_{Low}}$ were chosen independent of synthesis direction[1]. More in particular, monaural SD constraints

$$-1.5\mathrm{dB} \leq \mathrm{SD}_\zeta(f,\theta_k) \leq 0.5\mathrm{dB}, \tag{3.10}$$

were used separately for $\zeta = \mathrm{L,R}$ and for all synthesis directions θ_k, $k = 1, 2, ..., P$, such that the ILD deviation does not exceed 2 dB at any of the synthesis directions. The choice of a tighter boundary for positive SDs (0.5 dB) compared to negative SDs (-1.5 dB) was motivated by the fact that spectral peaks are known to be more detectable than notches [170]. Adding the SD constraints to Eq. (3.4) yields the constrained optimization problem

1 Whereas in this chapter as well as in further investigations in chapters 5 and 6, the upper and lower boundaries for SD were direction-independent, these constraint parameters could be modified to assign more importance to certain directions. As discussed in Appendix B, reducing the lower boundary $\mathrm{L_{Low}}$ at contralateral directions and thus allowing more synthesis error at these directions introduces perceptually non-significant changes, while at the same time satisfying more constraints on spectral distortion with unchanged lower and upper boundaries at ipsilateral directions.

$$\min \mathrm{J_{LS}}(\mathbf{w}_\zeta(f)) \quad \text{s.t.}$$

$$\begin{cases} 10\log_{10}(\mathrm{WNG_m}(\mathbf{w}_\zeta(f)))\mathrm{dB} \geq \beta \\ \\ -1.5\mathrm{dB} \leq \mathrm{SD}_\zeta(f,\theta_k) \leq 0.5\mathrm{dB} \quad k = 1,2,...,P \end{cases}, \tag{3.11}$$

with a total number of $P+1$ constraints. To solve this optimization problem, an Interior-Point algorithm was used [171]. The implementation of the Interior-Point algorithm provided by the function fmincon in the MATLAB optimization toolbox (ver. R2018b) was used, where the solutions of Eq. (3.4), obtained with the method of Lagrange multipliers described in [142], were used as initial values.

For the same planar microphone array as in section 3.3 with N=24 microphones, the spectral weights were calculated by solving the constrained optimization problem in Eq. (3.11) with spectral distortion constraints at P=72 horizontal directions ($\theta_k = 0°, 5°, 10°, ...$) and $\mathrm{WNG_m}$ constraint with β=0 dB. Figure 3.3 shows for both ears the resulting SD, TD, $\mathrm{WNG_m}$, directivity patterns of synthesized and desired HRTFs, the calculated spectral weights, as well as the resulting ILD and ITD deviations (ΔILD and ΔITD). The ITD deviation is defined as the absolute difference between the ITDs of the synthesized and the desired HRTFs

$$\begin{aligned} \Delta\mathrm{ITD}(f,\theta_k) &= |\mathrm{ITD_{synth}}(f,\theta_k) - \mathrm{ITD_{desired}}(f,\theta_k)| \\ &= |\mathrm{TD_L}(f,\theta_k) - \mathrm{TD_R}(f,\theta_k)|, \end{aligned} \tag{3.12}$$

with TD defined in Eq. (3.7), i.e.

$$\mathrm{ITD_{synth}}(f,\theta_k) = \frac{\angle \mathbf{w}_\mathrm{L}^H(f)\mathbf{d}(f,\theta_k) - \angle \mathbf{w}_\mathrm{R}^H(f)\mathbf{d}(f,\theta_k)}{2\pi f}, \tag{3.13}$$

and

$$\mathrm{ITD_{desired}}(f,\theta_k) = \frac{\angle \mathrm{D_L}(f,\theta_k) - \angle \mathrm{D_R}(f,\theta_k)}{2\pi f}. \tag{3.14}$$

As can be observed, SD at all synthesis directions as well as $\mathrm{WNG_m}$ could be kept within the desired boundaries of -1.5 dB $\leq$ SD $\leq$ 0.5 dB and $\mathrm{WNG_m} \geq 0$ dB up to about 5 kHz. As a direct consequence, ΔILD could be kept in the desired range of below 2 dB at all synthesis directions up to about 5 kHz. In addition, even though ΔITD is not directly controlled by the optimization problem in Eq. (3.11), it can be observed that the resulting ΔITD could be kept below 20μs up to 2 kHz for all synthesis directions, which is in accordance with the reported JNDs for ITD deviations [5].

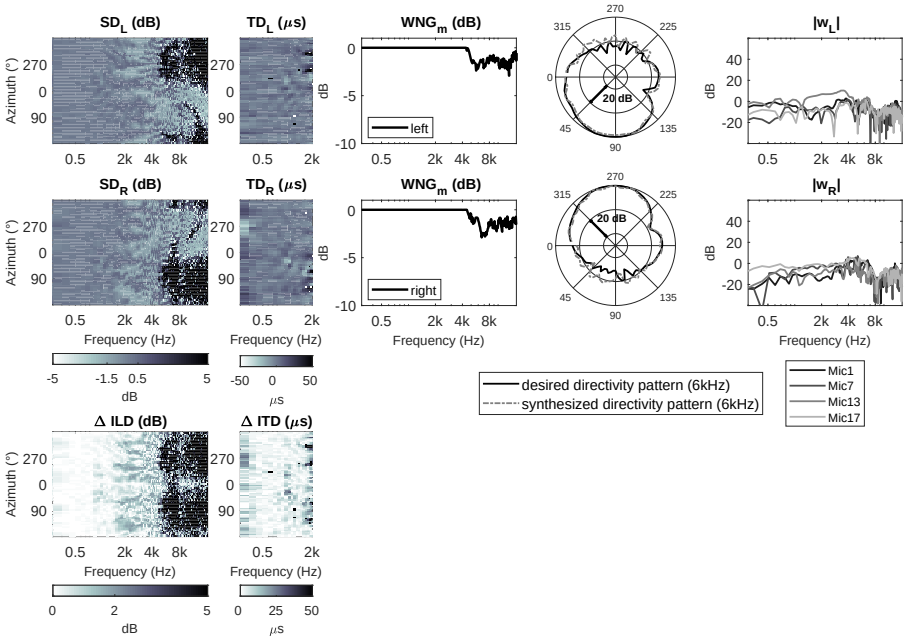

Fig. 3.3: Resulting left and right SD, TD, $\mathrm{WNG_m}$, directivity patterns of the desired and synthesized HRTFs at 6 kHz, spectral weights (shown for four exemplary microphones) and resulting ΔILD and ΔITD, with spectral distortion constraints at P=72 horizontal directions and $\mathrm{WNG_m}$ constraint with β=0 dB.

When comparing the results in Figure 3.3 (with the proposed SD constraints) with the results in Figure 3.2c (without SD constraints), it can be observed that the frequency range for which the SD can be considered acceptable was extended from about 2 kHz to up to 5 kHz for contralateral directions or even to higher frequencies for the ipsilateral directions. However, it was not feasible to satisfy all constraints for frequencies above 5 kHz, which can be explained by aliasing effects as well as more spatial details contained in the HRTF directivity patterns at high frequencies, especially at contralateral directions. The frequency range for which the constraints can be satisfied depends on the microphone array topology, which will be investigated in the next section.

3.5 The impact of microphone array topology

Intuitively, the inter-microphone distances influence the frequency range for which the $\mathrm{WNG_m}$ and SD constraints in Eq. (3.11) can be satisfied. On the one hand, since smaller inter-microphone distances shift spatial aliasing effects to higher frequencies, it is expected that the frequency range for which the SD constraints can be satisfied can be extended. On the other hand, since smaller inter-microphone

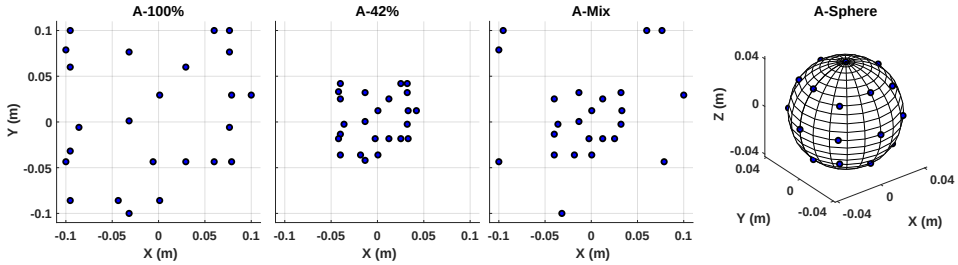

Fig. 3.4: Microphone positions for the four array topologies considered in this study: A-100%, (planar microphone array with 24 microphones in Figure 1.4), A-42% (42% down-sized version of A-100%), A-Mix (combination of A-100% and A-42%), and A-Sphere (spherical microphone array with 26 microphones distributed according to Lebedev grid on the surface of a rigid sphere with 4.2 cm radius).

distances cause the steering vectors to be more similar to each other, especially at low frequencies, the minimization of the cost function in Eq. (3.2) becomes more ill-conditioned. This may render satisfying the $\mathrm{WNG}_{\mathrm{m}}$ constraint more difficult, which may also negatively impact the resulting spectral and temporal distortion.

To investigate this, four array topologies, as shown in Figure 3.4, were considered:

- The first array topology, referred to as A-100%, corresponded to the planar microphone array of Rasumow et al. [15], which was already used for the simulations in the previous sections. In this array topology, the microphones were arranged according to a two-dimensional Golomb ruler [144], such that the inter-microphone distances are as different as possible in all possible directions.
- The second array topology was a scaled version of A-100%, down-sized to 42% of the original size, which is referred to as A-42%.
- In order to examine whether it is possible to profit from small inter-microphone distances while not suffering from robustness issues, the third array topology was a combination of A-100% and A-42%. This array topology, which is referred to as A-Mix, combined the eight outermost microphone positions of A-100% and the 16 innermost microphone positions of A-42%.
- The fourth array topology corresponded to a rigid spherical microphone array with a radius of 4.2 cm (the same as the Eigenmike [172]), referred to as A-Sphere. An equidistant microphone distribution based on a Lebedev quadrature was chosen as the most straightforward solution for many possible positionings on the surface of a sphere. Since the Lebedev quadratures are defined only for a certain number of points, the 26-point Lebedev grid was chosen [173] to keep the number of microphones of A-Sphere comparable to the 24 microphones of the other three topologies. The steering vectors of A-Sphere were calculated using the method proposed by [174].

For these four simulated microphone arrays, the spectral weights were calculated by solving the constrained optimization problem in Eq. (3.11) for two values of β=0 dB and β=−10 dB, in order to investigate the interaction between the $\mathrm{WNG_m}$ constraint parameter β and the array topology.

Figures 3.5 and 3.6 show the resulting SD, TD, and $\mathrm{WNG_m}$ obtained with the four considered array topologies for β=0 dB and β=−10 dB, respectively. Comparing the results for A-100% between β=0 dB (repetition of the results for the left ear shown in Figure 3.3) and β=−10 dB, it can be observed that the $\mathrm{WNG_m}$ constraints could be satisfied for all frequencies for the easier case of β=−10 dB. However, this does not appear to have been a large influence on the resulting SD and TD, which were similar for both values of β.

When comparing the results for the arrays with a smaller extension (A-42% and A-Sphere) with the results for A-100%, it can be observed that, as expected, the frequency range for which the $\mathrm{WNG_m}$ constraint could be satisfied was smaller ($f < 2$ kHz), while the SD constraint could be satisfied up to higher frequencies for the majority of directions. At lower frequencies ($f < 2$ kHz), where the desired directivity pattern is rather simple, both SD and $\mathrm{WNG_m}$ constraints could be satisfied. For β=−10 dB, also the resulting TD remained in a moderate range similar to A-100%, where for β=0 dB the resulting TD drastically increased (however, higher robustness is obtained). At higher frequencies, where the desired directivity pattern is getting more complicated, the SD was increased, especially at some contralateral directions. This effect was more prominent for β=0 dB than for β=−10 dB.

Similarly as for the smaller arrays, the results for A-Mix show that the SD constraints could be satisfied up to higher frequencies than for A-100%, while also the frequency range for which the $\mathrm{WNG_m}$ constraint could be satisfied could be extended (up to about 8 kHz for β=0 dB and for all frequencies for β=−10 dB) without considerably affecting the TD. The combination of sparse and dense microphone distances in A-Mix combined the advantage of a larger array extension in terms of robustness ($\mathrm{WNG_m}$) and the advantage of smaller inter-microphone distances in terms of synthesis accuracy (SD) at higher frequencies. With β=−10 dB, similar SD and TD compared to smaller arrays could be achieved with A-Mix.

3.6 Perceptual evaluation

In this section, the perceptual quality of synthesized HRTFs using the proposed constrained optimization method in section 3.4 is investigated for the simulated microphone array topologies considered in section 3.5. In the listening test, subjects listened to stimuli, generated either with individually measured HRTFs or different individual synthesized HRTFs:

- **A-100%***, individual HRTFs synthesized with A-100%, without SD constraints, and with β=0 dB, corresponding to the original method in [15, 17].

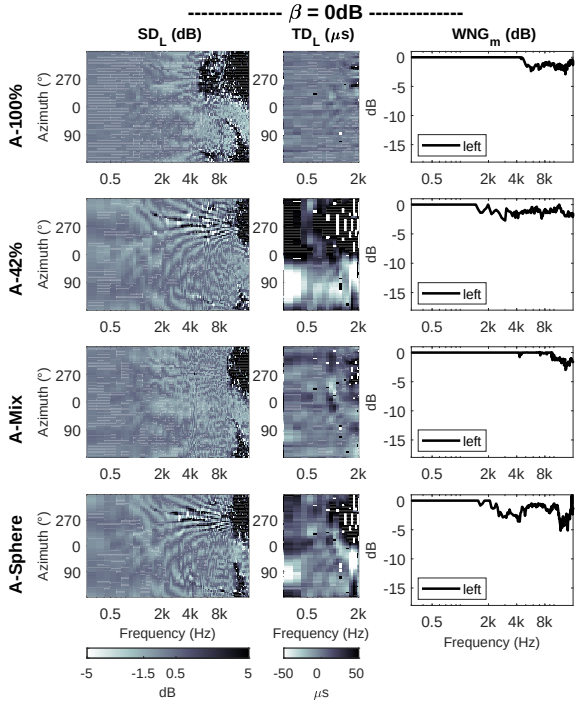

Fig. 3.5: Resulting Spectral Distortion (SD), Temporal Distortion (TD), and mean WNG ($\mathrm{WNG_m}$), when using -1.5 dB $\leq$ SD $\leq$ 0.5 dB at P=72 horizontal directions and $\mathrm{WNG_m} \geq \beta$ as constraints, with β=0 dB. Results are shown for the left ear of subject 1, from top to bottom for the array topologies shown in Figure 3.4.

- **A-100%, A-42%, A-Mix**, individual HRTFs synthesized with A-100%, A-42% and A-Mix, respectively, with the proposed SD constraints in Eq. (3.11) with β=0 dB.
- **A-Sphere$_{\beta 0}$** and **A-Sphere$_{\beta -10}$**, individual HRTFs synthesized with A-Sphere, with the proposed SD constraints in Eq. (3.11) with β=0 dB and β=-10 dB, respectively. The case β=-10 dB was considered in order to investigate the perceptual effect of a smaller β value on the perceptual quality.

For all synthesized HRTFs, spectral weights were calculated with P=72 equiangular directions in the horizontal plane (5° resolution). It should be noted again that the simulated microphone arrays were assumed to be perfectly robust, i.e. neither microphone self-noise nor deviations in microphone characteristics and positions were considered. The stimulus was also generated with non-individual HRTFs, measured with the KEMAR artificial head, which was presented as a hidden anchor (labeled as **Anchor**). This Anchor, together with the individually synthesized HRTFs resulted in a total of seven HRTFs sets, which were evaluated in the

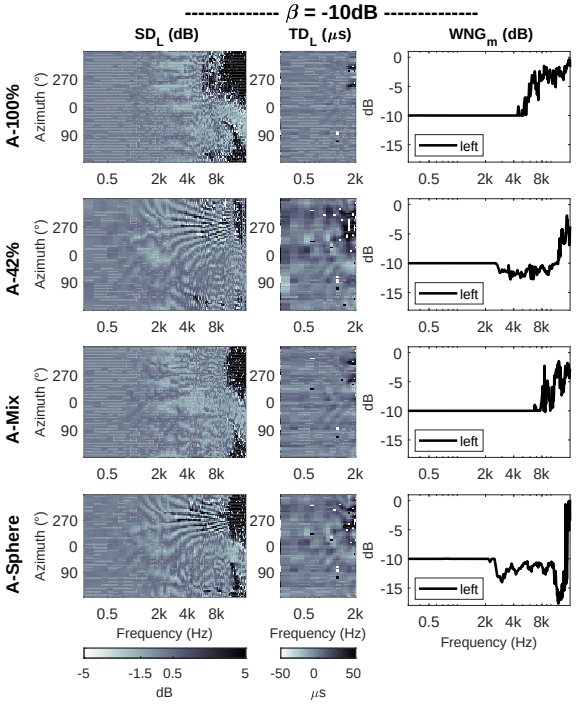

Fig. 3.6: The same as Figure 3.5, with β=−10 dB.

listening test.

3.6.1 *Methods*

3.6.1.1 *HRTF measurement and test signal*

Individual HRTFs of subjects were measured at 72 horizontal directions (5° resolution) using the methods and measurement setup described in Appendix A. Similar to [15], the test signal for the listening test consisted of three bursts of pink noise with a spectral content of 300 Hz $\leq f \leq$ 16 kHz, each lasting $\frac{1}{3}$s, followed by $\frac{1}{6}$s of silence. The stimuli were generated by filtering the test signal either with the individually measured original HRTFs or with the six versions of individually synthesized HRTFs, or with the non-individual HRTFs measured using the KEMAR artificial head. The stimuli were presented over Sennheiser HD800 headphones. Although the headphone transfer function was not compensated for, the perceptual evaluation was considered as valid, since solely different headphone signals were compared to each other without any comparison to real sound sources.

3.6.1.2 *Procedure*

Subjects were asked to rate the stimuli generated with the six different versions of synthesized HRTFs as well as the hidden anchor signal, compared to the reference signal (generated with the individually measured HRTFs). This resulted in seven test stimuli and one reference stimulus. Subjects evaluated the different stimuli on a 9-point scale between 1 and 5 in 0.5 steps with five German labels "schlecht" (bad), "dürftig" (poor), "ordentlich" (fair), "gut" (good) and "ausgezeichnet" (excellent) and four unlabeled intermediate points. The scale point labels and their English translations originate from the recommendations in [175]. Subjects could listen to the different headphone stimuli and the reference stimulus as often as they liked and could sort the test stimuli according to their current rating by clicking a sort button on the GUI, in order to facilitate the comparison [176]. A total of eleven (self-reported) normal hearing subjects (seven male, four female, aged 22 to 54 years) with experience in binaural psychoacoustic listening tests took part in the listening test. Subjects were asked to rate three perceptual attributes: Overall Quality, Spectral Coloration, and Localization. For all subjects, the experiment started with the evaluation of Overall Quality. Five subjects continued the experiment with the evaluation of Spectral Coloration as the second attribute, whereas the other six subjects evaluated Localization as the second attribute. The frontal direction $\theta = 0°$, as well as two lateral directions $\theta = 90°$ and $\theta = 220°$ were considered as nominal azimuthal positions for the virtual source. For all perceptual attributes, each of the three nominal source positions was presented three times in a randomized order, resulting in nine rounds.

The listening test took place with subjects seated in an empty anechoic room, wearing headphones. The headphone signals were calibrated with a G.R.A.S type 43AA artificial ear to have 75 dB SPL for the left ear ($\theta = 90°$) and were played back over an audio interface (RME Fireface UC) and headphone amplifier (Lake People Phone-Amp G103). The evaluation of each attribute lasted on average 30 minutes. Subjects were encouraged to take a break after completing one attribute.

All procedures were approved by the ethics committee of the Carl von Ossietzky University of Oldenburg.

3.6.2 *Results*

As mentioned in section 3.6.1.2, each of the three nominal source positions (0°, 90°, 220°) was presented three times for each attribute and subject. The consistency of the ratings over the three repetitions was assessed by calculating the Pearson correlation coefficients between the presentation pairs (1-2, 1-3, 2-3), see also [92]. As a measure for consistency, the mean correlation coefficient $\overline{r}$ was computed as

$$\overline{r} = \frac{\alpha}{n_{\text{repeat}} + (1 - n_{\text{repeat}})\alpha}, \tag{3.15}$$

with α the Cronbach's standardized coefficient [177] and n_{repeat} the number of repetitions. Considering $\alpha > 0.8$ as *good* and with n_{repeat}=3, the ratings of a subject were considered consistent and repeatable if $\overline{r} > 0.57$. According to the results shown in Figure 3.7, only Localization ratings of subject 5 were excluded from the analysis, since the $\overline{r}$ was markedly below the lower threshold of 0.57. All other ratings were considered consistent and were averaged over the three repetitions.

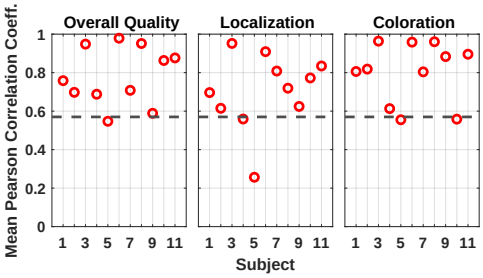

Fig. 3.7: Mean Pearson correlation coefficients $\overline{r}$ for the three presentation pairs (1-2, 1-3, 2-3) for each attribute and subject. The dashed horizontal line indicates the lower threshold for $\overline{r}$.

Figure 3.8 shows the averaged ratings for the three perceptual attributes and the three evaluated nominal source positions. For statistical analysis, the Shapiro-Wilk test of normality was applied to the averaged ratings, separately for each source position and for each perceptual attribute. The results indicated that for some HRTF sets, the gathered data could not be assumed to be normally distributed ($p<0.05$). Therefore, the Friedman test was applied, which confirmed a significant effect of the HRTF set ($p<1e-4$) for all perceptual attributes and source positions. The analysis was followed by multiple comparisons after Friedman test (function friedmanmc in the statistical software R [178]). Significantly different HRTF sets ($p<0.05$) are indicated with horizontal lines in Figure 3.8.

It can be observed that for all attributes and source positions the median ratings given to A-100% and Anchor were lower than ratings given to the other HRTF sets, with partly significant differences. For all attributes and source positions, the median ratings given to A-100% were higher than the median ratings given to A-100%*. For A-42%, Localization ratings were similar to or lower than A-100%, whereas Coloration and Overall Quality ratings were higher than for A-100%. Among all planar array topologies, the best median ratings were given to A-Mix for all attributes (except for the Coloration ratings at $\theta = 0°$), with partly significant differences to A-100% and A-42%. For all attributes and source positions, the ratings given to A-Sphere$_{\beta-10}$ were similarly high as the ratings given to A-Mix, with no significant differences. It should however be noted that the effect of poorer robustness of A-

Sphere$_{\beta-10}$ (β=−10 dB) compared to A-Mix (β=0 dB) could not be demonstrated here with this listening test. For A-Sphere$_{\beta 0}$, the median of Coloration and Overall Quality ratings were slightly lower than the median ratings given to A-Sphere$_{\beta-10}$ and A-Mix; for the Localization ratings, these differences were more prominent but not significant.

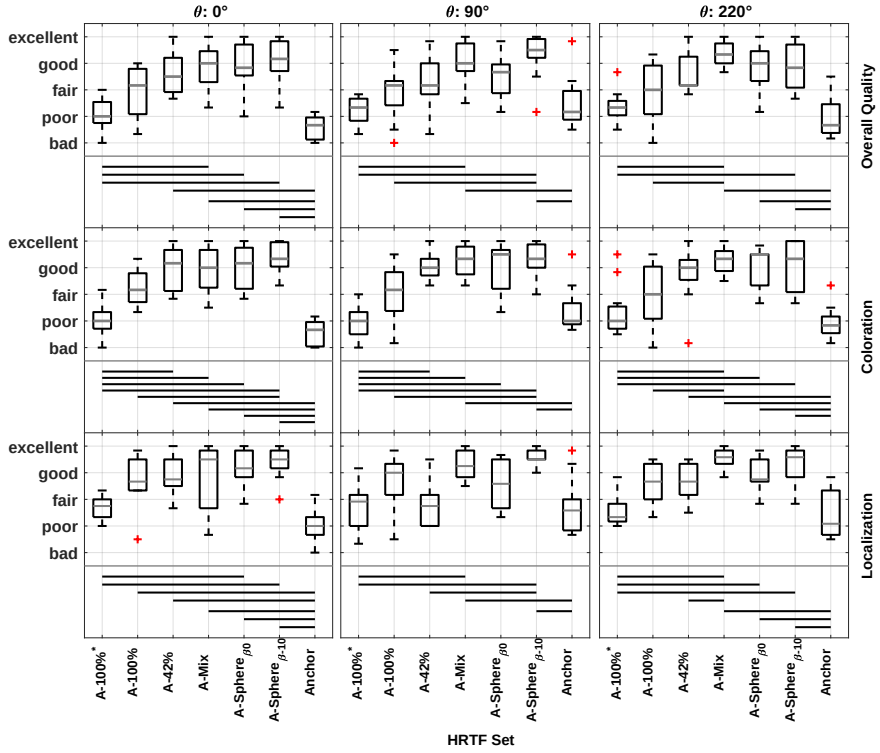

Fig. 3.8: Results of perceptual evaluations with respect to Overall Quality, Spectral Coloration and Localization for three nominal source positions (0°, 90°, 220°). Results were averaged over three repetitions of 11 subjects for Overall Quality and Coloration and 10 subjects for Localization. HRTF sets with significant differences are indicated with horizontal lines ($p<0.05$).

3.6.3 *Discussion*

Since the median ratings given to A-100% were consistently higher than the ratings given to A-100%*, the advantage of the proposed method with constraints on the spectral distortion could be confirmed perceptually. This is in accordance with the objective results in Figures 3.5 and 3.6, where it was shown that not only the spatial resolution of the synthesis could be improved but also the synthesis error could be reduced for a larger frequency range compared to the method of Rasumow et al. [15].

Among the arrays for which the method with SD constraints was applied, better Localization ratings were obtained at $\theta = 90°$ and $\theta = 220°$ for A-100%, A-Mix and A-Sphere$_{\beta-10}$ compared to A-42% and A-Sphere$_{\beta0}$. This is in accordance with the objective results in Figure 3.5, where the resulting TD was larger for A-42% and A-Sphere$_{\beta0}$.

For A-Mix, the Localization and Coloration ratings were almost everywhere higher than the ratings for A-100% and A-42%. As discussed in section 3.5, the combination of a large array extension and small inter-microphone distances yielded an improved synthesis accuracy at higher frequencies while maintaining a low TD at lower frequencies for A-Mix. The perceptual evaluation results confirmed this advantage. This is also visible in the comparison between A-Mix and A-Sphere$_{\beta0}$: the higher TD associated with A-Sphere$_{\beta0}$ compared to that of A-Mix led to better Localization ratings for A-Mix (albeit not statistically significant).

In general, for all HRTFs, the Overall Quality ratings showed higher similarity to Coloration than to Localization ratings. It seemed that subjects paid more attention to spectral coloration than to localization to give their ratings for the Overall Quality.

The relatively low ratings for A-100%* in this study were not in accordance with the perceptual results of Rasumow et al. [15], where the signals generated with the VAH were rated significantly higher than the signals generated with non-individual HRTFs of the KEMAR artificial head. Although the spectral weights for A-100%* in this study and in [15] were calculated with the same method (using Eq. (3.4)) and for the same microphone array topology, in this study P=72 calibration directions were used instead of P=24 calibration directions in [15]. However, the results in Figure 3.2b for P=24 and in Figure 3.2c for P=72 showed that the resulting SD and TD are very similar. Instead, the presence of the other VAH versions with better performance was suspected to be responsible for the low ratings of A-100%* in the current study, as subjects were instructed to compare each VAH signal to the reference signal as well as to other VAH signals. As a result, the perceived difference between A-100%* and Anchor was reduced.

As already mentioned, the robustness of the VAHs due to the resulting WNG$_{\mathrm{m}}$ was not assessed in this study. Effects such as microphone self-noise or deviations in microphone positions can be considered either in real recordings or via simulations as in [17]. Perceptual results would then be expected to show a trade-off between self-noise amplification and synthesis accuracy.

The spectral weights in this study were calculated for each listener for a fixed head orientation to the frontal direction. The calculation of the spectral weights can be repeated for different orientations of the listener's head. This constitutes an important feature of the VAH approach, which enables to include head tracking into the playback.

3.7 The effect of constraint relaxation

Finally, it should be mentioned that besides the microphone array topology, the feasibility of satisfying the SD and $\mathrm{WNG_m}$ constraints depends also on the constraints themselves. Relaxing the values given to $\mathrm{L_{Up}}$, $\mathrm{L_{Low}}$ and β can make the constraints easier to satisfy. In addition, the effect of constraint relaxation can be different for different microphone array topologies. The impact of constraint relaxation was studied using two different simulated microphone arrays. Three different cases of constraint relaxation were compared to the constraints in Eq. (3.11) with β=0 dB, either by relaxing the lower boundary of SD ($\mathrm{L_{Low}}$) at contralateral directions, or by relaxing the minimum desired value of $\mathrm{WNG_m}$ (β), or a combination of relaxing β and reducing the number P of directions considered in the calculation of the spectral weights. It was shown that constraint relaxation can increase the chance of satisfying the constraints and that this increase is differently effective when applied to different microphone array topologies. Perceptual evaluations with respect to Localization, Spectral Coloration, and Overall Quality showed that with a proper constraint relaxation and a properly chosen microphone array topology, more constraints can be satisfied without degrading the synthesis accuracy. The VAH syntheses with different constraint relaxations also outperformed the binaural signals generated with non-individual HRTFs measured for a conventional artificial head. A thorough description of the study of constraint relaxation and the results of objective and subjective evaluations is provided in Appendix B.

3.8 Summary

In this chapter, a method for high spatial resolution HRTF synthesis with a Virtual Artificial Head (VAH) was presented and evaluated. In order to maintain the synthesis accuracy at a high number of synthesis directions without increasing the number of microphones, a method based on constrained optimization was proposed. In addition to the constraint on the mean WNG ($\mathrm{WNG_m}$), an upper and a lower boundary of 0.5 dB and -1.5 dB, respectively, were set as constraints for the Spectral Distortion (SD) between desired and synthesized HRTF directivity patterns. Objective and perceptual results showed that by imposing the SD constraints, the performance of the VAH (planar microphone array with 24 microphones) could be improved compared to the original method of Rasumow et al. [15] by maintaining the synthesis accuracy at an increased number of horizontal synthesis directions (5° azimuthal resolution) for frequencies up to 5 kHz.

Simulation results with different microphone array topologies indicated that smaller inter-microphone distances were advantageous for satisfying the SD constraints at higher frequencies. On the other hand, array topologies with a smaller extension failed to satisfy the $\mathrm{WNG_m}$ constraint in the mid-frequency range. In particular, if a higher robustness was intended by increasing the minimum desired value of $\mathrm{WNG_m}$, the resulting temporal distortion at $f < 2$ kHz increased, which was shown to impact the Localization ratings negatively. With an array topology combining

sparse and dense inter-microphone distances, the SD constraints were satisfied up to higher frequencies. At the same time, the $\mathrm{WNG_m}$ constraint was satisfied at the low and mid-frequency range while avoiding an increased temporal distortion at lower frequencies. The advantage of this array topology over the other considered array topologies was furthermore confirmed by perceptual evaluations.

Further investigations are required to investigate the potential effect of head tracking on the quality of auralizations with respect to plausibility or localization accuracy. Moreover, besides evaluating the VAH approach at non-horizontal directions, other investigations are required to examine applications in a reverberant environment with realistic test signals such as speech and music.

4

DYNAMIC AURALIZATION WITH A VAH - GENERAL METHODOLOGY

4.1 Introduction

In static auralizations, i.e. without considering the listener's head movements, the source signal is convolved with HRIRs or BRIRs of the direction at which the sound source is located. Moving the head during listening does not cause any changes in the presented virtual scene. This is contrary to real listening situations and against the expectation of the listener. The compensation of head movements during signal playback has been shown to decrease front-back reversals and improve the localization accuracy and externalization when listening to virtual sound sources presented over headphones [12–14].

In order to compensate for the head movements of the listener, auralizations should be dynamic, meaning that the presentation of the virtual source should be updated in real-time and in response to the listener's head movements. This can be done by tracking the head movements to assess the new position of the sound sources relative to the head. The HRIRs or BRIRs of the direction corresponding to this new source position are then convolved with the signal. Such a dynamic auralization necessitates access to HRIRs or BRIRs for different head orientations. Especially in case of measurement-based room auralizations, the acquisition of BRIRs for different head orientations can get very tedious. In contrast to time-consuming BRIR measurements for different head orientations, individualized BRIRs for different head orientations can be synthesized using Room Impulse Responses (RIRs) measured with a Virtual Artificial Head (VAH).

In this chapter, the general methodology for dynamic auralizations with the VAH approach in this thesis is presented. The chapter starts in section 4.2 with the virtual rotation of the VAH for the calculation of the spectral weights for a given head orientation. In section 4.3, it is explained how the calculated spectral weights and the measured RIRs are used to synthesize the individual BRIRs for a given head orientation. In section 4.4, the technical implementations for presenting the binaural signals dynamically are reviewed, including the head tracker device and

the employed algorithms for the rea-time head-tracked signal playback. These algorithms and implementations are used for dynamic auralizations in chapters 5 and 6.

4.2 VAH spectral weights for different head orientations

An important feature of a VAH is the possibility of calculating the spectral weights for different head orientations without requiring individual HRTFs for these head orientations. This enables head rotations to be easily taken into account via head tracking during signal playback. For a given head orientation $\Theta_h = (\theta_h, \phi_h)$, with θ_h and ϕ_h denoting the horizontal and vertical angles of the head orientation, respectively, spectral weights can be calculated by considering the desired directivity pattern $\mathrm{D}(f, \Theta_k)$, $k = 1, 2, ..., P$, together with spatially shifted steering vector $\mathbf{d}(f, \Theta_s)$ with $\Theta_s = (\theta_s = \theta_k + \theta_h, \phi_s = \phi_k + \phi_h)$ into the calculation of the spectral weights using the methods described in chapter 3. This can be interpreted as a virtual rotation of the VAH to the head orientation Θ_h.

For the dynamic auralizations in this thesis, individual spectral weights were calculated for 37×5=185 head orientations (37 horizontal directions $-90° \leq \theta_h \leq 90°$ in 5° steps and 5 vertical directions $-15° \leq \phi_h \leq 15°$ in 7.5° steps), with positive horizontal head orientations corresponding to the head rotated to the left and positive vertical head orientations corresponding to the head rotated upwards. The spatial resolution of 5° for the horizontal head orientations was chosen based on the measurement setup described in Appendix A and in accordance with the resolution reported to be sufficient for non-critical signals such as music [99]. It should be mentioned that the measured steering vectors were not available for all shifted vertical directions ϕ_s. For such cases, the steering vectors were shifted to the nearest available vertical direction. For example, consider the case that $\phi_k = 30°$ (elevation of the k^{th} calibration direction) and the spectral weights were supposed to be calculated for the vertical head orientation $\phi_h = 7.5°$, such that $\phi_s = 30° + 7.5° = 37.5°$. Since the measurement setup described in Appendix A did not include the vertical direction at 37.5°, steering vectors measured at elevation 45° were used instead to calculate the spectral weights for this vertical head orientation.

4.3 Synthesizing individual BRIRs for different head orientation with a VAH

For dynamic auralizations with a VAH in this thesis, individual BRIRs for different head orientations were synthesized as follows (see also Figure 4.1). For the source at a given synthesis direction Θ', RIRs were measured between the source and the N microphones of the VAH placed at a defined position in the room. The N individually calculated left and right spectral weights for a given head orientation Θ_h were transformed to filter coefficients of FIR filters using the inverse Fourier transform. The RIRs measured with each of the N microphones were convolved

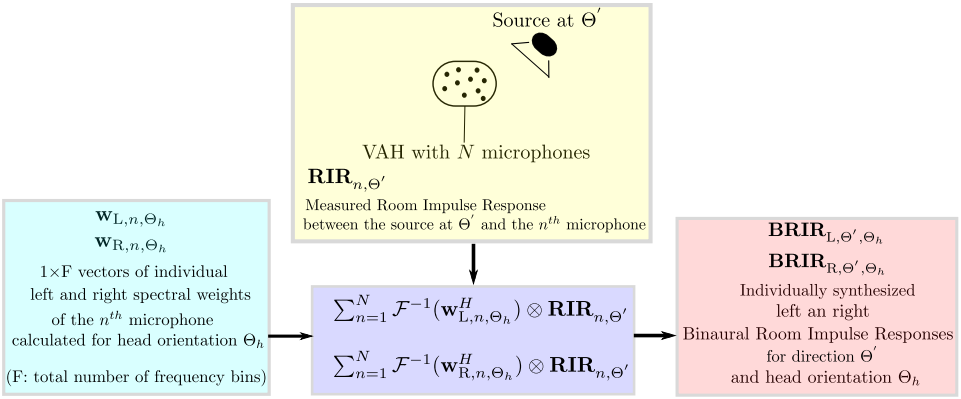

Fig. 4.1: Synthesizing individual BRIRs for direction Θ' and head orientation Θ_h from Room Impulse Responses (RIRs) measured with a VAH.

with the associated FIR filter. The sum of the filtered RIRs over the N channels resulted in synthesized left and right BRIRs for the source at direction Θ' and head orientation Θ_h.

The orientation of the VAH during the acquisition of RIRs remained to the same fixed frontal direction $\theta_h = 0°$ and $\phi_h = 0°$ as during the measurement of steering vectors as described in Appendix A.

4.4 Technical implementations

For capturing head movements during signal playback, a custom-made head tracker was used which was developed at the *Institut für Hörtechnik und Audiologie* at the Jade University of Applied Sciences. The head tracker consisted of a Bosch Sensortec Lagesensor (BNO055 9-Axis Absolute Orientation Sensor) in combination with a Teensyduino microcontroller, which could be connected to the PC or laptop via USB cable. The tracker data was updated each 10 ms. The head tracker was mounted on the top of the used headphones (Sennheiser HD800) as shown in Figure 4.2. In addition, a push button was mounted on the upper right corner of the headphones which was used during perceptual evaluations in chapter 5 to enable the listener to switch back and forth between headphone and loudspeaker presentations (as also implemented in [92]).

The real-time head-tracked binaural signal playback was performed using an algorithm written in C++ [179,180]. Upon starting the program, the BRIRs for all considered head orientations were loaded, partitioned in blocks of 256 samples, and prepared as Fourier-transformed filters. The test signal was similarly partitioned in 50% overlapping blocks of 256 samples using a Hann window. According to the data supplied by the head tracker, a search algorithm chose the Fourier-transformed

Fig. 4.2: Custom-made head tracker mounted on the headphones (Sennheiser HD800), used for perceptual evaluations in chapters 5 and 6. The push button installed on the upper right corner of the headphones enabled the listener to switch between headphone and loudspeaker presentations (see section 5.3.3).

BRIR of the head orientation with the smallest Euclidean distance to the listener's current head orientation. Associated partitions of the Fourier-transformed test signal and the chosen filters were multiplied and transformed into time domain via inverse Fourier transform. The latency caused by this algorithm was equal to one block, i.e. 5.8 ms at f_s=44100 Hz, and therefore small enough compared to reported latency thresholds [100, 101, 181] to consider the real-time dynamic presentation artifact-free.

The BRIRs were convolved with individual inverse Headphone Impulse Responses (HPIRs) and saved in the SOFA format (Spatially Oriented Format for Acoustics) [182], prior to being loaded for the real-time dynamic presentation.

4.5 Summary

In this chapter, the general methodology employed for dynamic auralizations in this thesis were presented. The calculation of the Virtual Artificial Head (VAH) spectral weights for a given head orientation and the BRIR synthesis using these spectral weights and RIRs measured with a VAH were explained. Finally, the technical implementations used for real-time dynamic auralizations were reviewed. In the next two chapters, the VAH approach is investigated in dynamic auralizations using the described algorithms and implementations in this chapter.

5

DYNAMIC AURALIZATION OF ANECHOIC AND REVERBERANT ENVIRONMENTS USING A VAH

5.1 Introduction

An important application of binaural technology is auralization of acoustical environments, where the source signal is convolved with BRIRs and presented over headphones [77]. For simulation-based auralizations, the room can be simulated via geometrical acoustical models representing the direct and reflected sound propagation from the source to the listener [78, 80], whereas for measurement-based auralizations, the BRIRs are measured, either with individual listeners or more commonly with artificial heads. Dynamic head-tracked presentation of the auralized environment can greatly enhance the realism of the playback by reducing localization ambiguities and improving the externalization [12–14]. To enable a dynamic measurement-based auralization, the BRIRs need to be measured for different head orientations [93, 94]. However, this is a very time-consuming task, especially if BRIRs for different head orientations need to be measured individually and in different environments.

This chapter is based on:
[151] M. Fallahi, M. Hansen, S. Doclo, S. van de Par, D. Püschel, and M. Blau, "Evaluation of head-tracked binaural auralizations of speech signals generated with a virtual artificial head in anechoic and classroom environments", *Acta Acustica*, vol. 5, no. 30, 2021.

[152] M. Fallahi, M. Hansen, S. Doclo, S. van de Par, D. Püschel, and M. Blau, "Binaural Reproduction of Signals captured in a reverberant Room with a Virtual Artificial Head", *Proc. of Fortschritte der Akustik - DAGA*, Rostock, Germany, pp. 619-622, 2019.

[153] M. Fallahi, M. Hansen, S. Doclo, S. van de Par, D. Püschel, and M. Blau, "Individualized dynamic binaural auralization of classroom acoustics using a virtual artificial head", *Proc. 23rd International Congress on Acoustics - ICA*, Aachen, Germany, pp. 731-738, 2019.

As an alternative, and as discussed in chapter 4, individual BRIRs for different head orientations can be synthesized with a Virtual Artificial Head (VAH) by applying the spectral weights, calculated for different head orientations, to Room Impulse Responses (RIRs) measured with a single orientation of the VAH. The present chapter deals with the VAH approach in measurement-based auralizations, with spectral weights calculated using the constrained optimization method proposed in chapter 3. The proposed constrained optimization method was evaluated previously in a simulation-based dynamic auralization and with the VAH as a simulated array with 32 microphones [92]. The synthesized HRTFs with the VAH were used to simulate BRIRs in a lecture room for different head orientations. The room was simulated based on recorded reverberation times and coarse geometries using the RAZR room simulation package [80]. The synthesized BRIRs with the VAH were rated slightly, but not significantly, lower than original BRIRs. The study offered a starting point for evaluating the VAH approach in dynamic auralizations. With simulations done in [92], however, optimized solutions for the VAH could not be evaluated with respect to robustness. In practice, the robustness of the microphone array is important, because certain solutions will be highly sensitive to small errors in microphone positions and characteristics, as well as to microphone self-noise.

In this chapter, instead of simulations, real measurements are performed with the VAH of Rasumow et al. [15] (planar microphone array with 24 microphones) using the constrained optimization method proposed in chapter 3. Individual BRIRs are synthesized with the VAH, each for 185 head orientations, as described in chapter 4. Different constraint parameters are considered when using the constrained optimization method, namely the discrete source directions included in the calculation of the spectral weights and the minimum desired value of $\mathrm{WNG_m}$ (i.e. β). The individually synthesized BRIRs as well as measured non-individual BRIRs of a conventional artificial head and a rigid sphere are used for dynamic auralizations of an anechoic and a reverberant environment, followed by perceptual evaluations with direct comparison to real loudspeaker signals. The specific research questions to be addressed are: (1) How well does the VAH perform in head-tracked auralizations? (2) Which constraint parameters lead to the best performance of the VAH for auralizing the considered environments? (3) What is the influence of a reverberant environment compared to an anechoic environment on the performance of the VAH? and (4) How well do individually synthesized BRIRs with the VAH perform compared to non-individual BRIRs of an artificial head?

The chapter continues with a review of the methods and parameters, which were used to calculate the individual spectral weights in section 5.2. Section 5.3 describes the methods used for signal preparation as well as for the perceptual evaluations. Perceptual results for the experiment in the reverberant and anechoic environments are presented in sections 5.4 and 5.5, respectively, and the results are discussed in section 5.6.

5.2 VAH methods and parameters

In this section, first, a short review of the VAH as a filter-and-sum beamformer and the optimization methods for the calculation of the spectral weights is given. Detailed information can be found in chapter 3. Then, the VAH and the parameters used to calculate the spectral weights are introduced.

5.2.1 *Calculation of spectral weights*

Virtual Artificial Head (VAH) as a filter-and-sum beamformer consists of N spatially distributed microphones. The directivity pattern of the VAH at frequency f and direction $\Theta = (\theta, \phi)$, with θ the azimuth angle and ϕ the elevation angle, is given by

$$\mathrm{H}(f, \Theta) = \mathbf{w}^H(f)\mathbf{d}(f, \Theta), \tag{5.1}$$

with $(.)^H$ denoting the Hermitian transpose. The $N \times 1$ steering vector $\mathbf{d}(f, \Theta)$ describes the free-field acoustical transfer function at frequency f between a source at direction Θ and the N microphones, and the $N \times 1$ vector $\mathbf{w}(f)$ contains the complex-valued spectral weights for each of the N microphones. Here, the aim was to synthesize the desired directivity pattern $\mathrm{D}_\zeta(f, \Theta_k)$, $\zeta \in \{\mathrm{L,R}\}$, of the left or right HRTFs at P discrete directions $\Theta_k, k = 1, 2, ..., P$. The spectral weights $\mathbf{w}_\mathrm{L}(f)$ and $\mathbf{w}_\mathrm{R}(f)$ were calculated by minimizing a narrow-band least-squares cost function, which is defined as the sum of the squared absolute differences between desired and synthesized directivity patterns over P discrete directions, i.e.

$$\mathrm{J}_\mathrm{LS}(\mathbf{w}_\zeta(f)) = \sum_{k=1}^{P} |\mathrm{H}_\zeta(f, \Theta_k) - \mathrm{D}_\zeta(f, \Theta_k)|^2, \tag{5.2}$$

where $\mathrm{H}_\zeta(f, \Theta_k)$ indicates the synthesized HRTFs for the left and right ears as defined in Eq. (5.1). The cost function J_LS was minimized separately for the left and the right ears. Aiming at achieving a small synthesis error at all P directions, as proposed in chapter 3, constraints were imposed onto the Spectral Distortion (SD), defined as

$$\mathrm{SD}_\zeta(f, \Theta_k) = 10 \log_{10} \frac{|\mathbf{w}_\zeta^H(f)\mathbf{d}(f, \Theta_k)|^2}{|\mathrm{D}_\zeta(f, \Theta_k)|^2} \mathrm{dB}. \tag{5.3}$$

Constraints were imposed on the SD such that at each direction Θ_k

$$\mathrm{L}_\mathrm{Low} \leq \mathrm{SD}_\zeta(f, \Theta_k) \leq \mathrm{L}_\mathrm{Up}, \quad k = 1, 2, ..., P, \tag{5.4}$$

where L_Up and L_Low denote the upper and lower boundary, respectively. An additional constraint was imposed onto the *mean* white Noise Gain ($\mathrm{WNG_m}$) in dB [17] (see Eq. (3.3)), in order to increase the robustness of the VAH against small devi-

ations in microphone positions and characteristics and limit microphone self-noise amplification

$$10\log_{10}(\frac{1}{P}\sum_{k=1}^{P}\frac{|\mathbf{w}_\zeta^H(f)\mathbf{d}(f,\Theta_k)|^2}{\mathbf{w}_\zeta^H(f)\mathbf{w}_\zeta(f)})\text{dB} \geq \beta. \tag{5.5}$$

To solve the constrained optimization problem of minimizing J_{LS} in Eq. (5.2) subject to P+1 constraints defined in Eqs. (5.4) and (5.5), an iterative Interior-Point algorithm, as implemented in function fmincon in the MATLAB optimization toolbox (ver. R2018b) was used.

5.2.2 *Microphone array and constraint parameters*

In this chapter, the planar microphone array with 24 microphones of Rasumow et al. [15] shown in Figure 1.4 was used. This microphone array was previously evaluated perceptually in [15] for a static scenario (i.e. without head tracking).

The upper and lower boundaries L_{Up} and L_{Low} for the SD constraints were chosen as 0.5 dB and −1.5 dB, respectively. As discussed in chapter 3, satisfying the SD constraints with these values of L_{Up} and L_{Low} results in a maximum deviation of 2 dB in the resulting Interaural Level Differences (ILDs) at all P directions. A deviation of 2 dB was considered reasonable based on the reported Just Noticeable Differences in ILD deviations [5]. For the minimum desired value of WNG_m, i.e. β in Eq. (5.5), two values of 0 dB and −10 dB were considered, labeled as $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_{-10}}$ in the remaining discussion. The choice of β=0 dB was based on the results in [142], while β=−10 dB was chosen to investigate the effect of a lower resulting WNG_m and reduced robustness.

It should be noted that the P directions considered in the calculation of the spectral weights, i.e. both in the cost function in Eq. (5.2) as well as in the SD constraints in Eq. (5.4), have a major influence on the resulting synthesized HRTFs, spectral distortion and WNG_m. It is therefore interesting to investigate the extent to which it is necessary to include directions other than horizontal directions into the calculation of the spectral weights, in order to account for non-horizontal source positions as well as room reflections. Three cases for P were considered in this chapter: (1) P=72 horizontal directions (5° azimuthal resolution), (2) P=3×72=216 directions from elevations −15°, 0° and +15° and (3) P=3×72=216 directions from elevations −30°, 0° and +30°, labeled as **V0, V0±15** and **V0±30**, respectively, in the remaining discussion. Table 5.1 summarizes the constraint parameters P and β used for the calculation of the spectral weights in this chapter.

As an example, spectral weights were calculated with a set of measured steering vectors and the individual HRTFs of one of the subjects in this chapter (subject 1) for different values of P and β. The calculated spectral weights were then applied to the same measured steering vectors using Eq. (5.1) to result in the synthesized

Table 5.1: Overview of values chosen for the parameters P and β, resulting in six sets of spectral weights. Each set of spectral weights was calculated for 185 head orientations.

Label	Constraint parameter P and β
V0/β_0	P=72 (Elevation: 0°), β=0 dB
V0±15/β_0	P=3×72=216 (Elevations: −15°, 0°, 15°), β=0 dB
V0±30/β_0	P=3×72=216 (Elevations: −30°, 0°, 30°), β=0 dB
V0/β_{-10}	P=72 (Elevation: 0°), β=−10 dB
V0±15/β_{-10}	P=3×72=216 (Elevations: −15°, 0°, 15°), β=−10 dB
V0±30/β_{-10}	P=3×72=216 (Elevations: −30°, 0°, 30°), β=−10 dB

HRTFs for subject 1 for the frontal head orientation. Figures 5.1a and 5.1b show the resulting SD for the synthesized left HRTFs at elevations 0°, 15°, and 22.5°, as well as the resulting $\mathrm{WNG_m}$ for V0/β_0 and V0/β_{-10}, respectively. At elevation 0°, it can be observed that up to about 5 kHz, the SD, as well as $\mathrm{WNG_m}$ constraints could be satisfied. However, at frequencies above 5 kHz, the SD constraints could not always be satisfied. At elevations 15° and 22.5°, the resulting SD clearly increased compared to the resulting SD at elevation 0°, since these non-horizontal directions were not included in the calculation of the spectral weights. Also, the resulting Temporal Distortion (TD, see Eq. (3.7)) at elevations 15° and 22.5° clearly increased compared to the resulting TD at elevation 0°.

Figures 5.2a and 5.2b show the results for V0±15/β_0 and V0±15/β_{-10}, respectively. Compared to the results shown in Figures 5.1a and 5.1b, for frequencies up to about 4 kHz, the resulting SD at elevation 15° clearly improved. The inclusion of non-horizontal directions also slightly improved the resulting SD at elevation 22.5°, although directions from this elevation were not included in the constrained optimization. At the same time, the resulting SD at elevation 0° deteriorated. In addition, the $\mathrm{WNG_m}$ constraint could, with a few exceptions, not be satisfied for frequencies below 4 kHz. Moreover, the resulting TD at all three elevations should be noted, which degraded extremely for the synthesis with horizontal and non-horizontal directions included. With β=−10 dB, the resulting TD was slightly better than with β=0 dB, however much worse than the TD shown in Figure 5.1 for the cases with only horizontal directions included.

5.2.3 *Spectral weights for the dynamic auralization*

To enable the head-tracked dynamic binaural signal playback with the VAH in this chapter, spectral weights were a priori calculated for each of the six parameters listed in Table 5.1 for 37×5=185 head orientations, corresponding to 37 azimuth

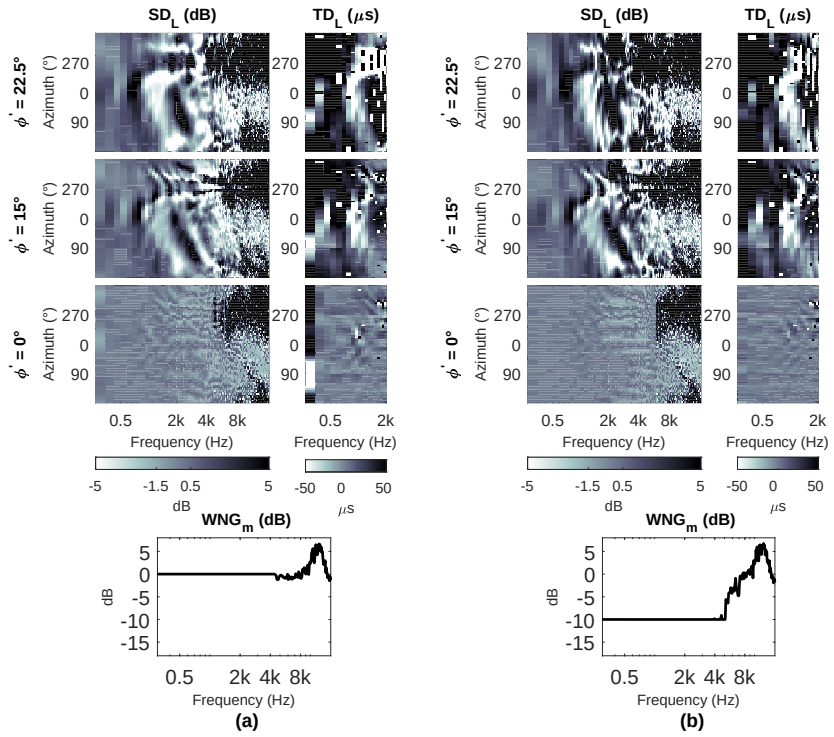

Fig. 5.1: The resulting SD and TD at elevations 0°, 15° and 22.5° and the resulting $\mathrm{WNG_m}$. The spectral weights were calculated with **(a):** $\mathrm{V0}/\beta_0$ and **(b):** $\mathrm{V0}/\beta_{-10}$, using the steering vectors measured with the planar microphone array with 24 microphones shown in Figure 1.4. Results are shown for the left ear of subject 1 in this chapter.

angles θ_h of $-90°$ to $+90°$ in $5°$ steps and 5 elevation angles ϕ_h of $-15°$ to $+15°$ in $7.5°$ steps (c.f. section 4.2).

5.3 Methods

The study in this chapter consisted of two experiments with measurements and perceptual evaluations in two different acoustical environments: a reverberant lecture room (Experiment 1) and an anechoic room (Experiment 2). Experiment 2 was performed after completing Experiment 1 and was motivated by questions arising from the results of Experiment 1. The methods and technical implementations were to a large extent the same for both experiments. The information provided in this section applies to both experiments. Specific information on each experiment (room characteristics, source and listener positions) is provided in more detail in sections 5.4 and 5.5 as well as in Figure 5.3. The description of the applied methods

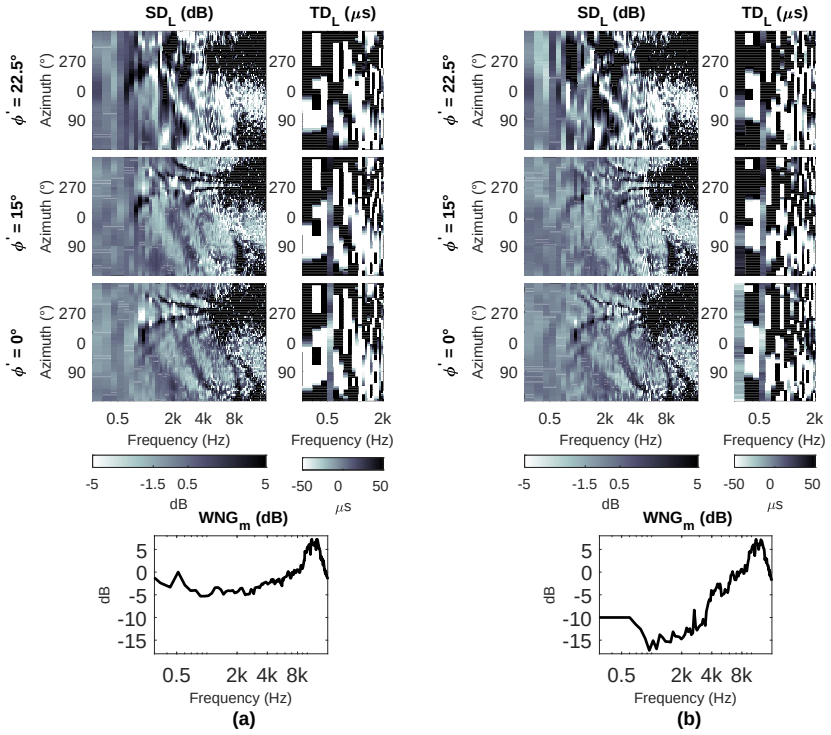

Fig. 5.2: The resulting SD and TD at elevations 0°, 15° and 22.5° and the resulting $\mathrm{WNG_m}$. The spectral weights were calculated with **(a):** V0±15/β_0 and **(b):** V0±15/β_{-10}, using the steering vectors measured with the planar microphone array with 24 microphones shown in Figure 1.4. Results are shown for the left ear of subject 1 in this chapter.

starts with introducing the preparatory measurements in section 5.3.1, followed by the methods applied for the acquisition of BRIRs in section 5.3.2. Technical implementation for listening tests and the criterion to exclude non-consistent ratings are discussed in sections 5.3.3 and 5.3.4, respectively.

5.3.1 *Preparatory measurements*

Individual Head Related Impulse Responses (HRIRs) and VAH steering vectors at 864 directions, including the ones listed in Table 5.1, as well as individual Headphone Impulse Responses (HPIRs) for Sennheiser HD800 headphones were measured with the measurement setup and methods described in Appendix A. After transferring the measured HRIRs into the frequency domain, the HRTFs were spectro-spatially smoothed according to [16]. The smoothed desired directivity patterns D and the

measured, Fourier-transformed steering vectors were used to calculate the individual spectral weights with L_{Up}=0.5 dB, L_{Low}=−1.5 dB, and the listed values for the parameters P and β in Table 5.1. The individual spectral weights were calculated for 185 head orientations, as described in section 5.2.3 as well as in section 4.2.

5.3.2 *BRIR acquisition in the two auralized environments*

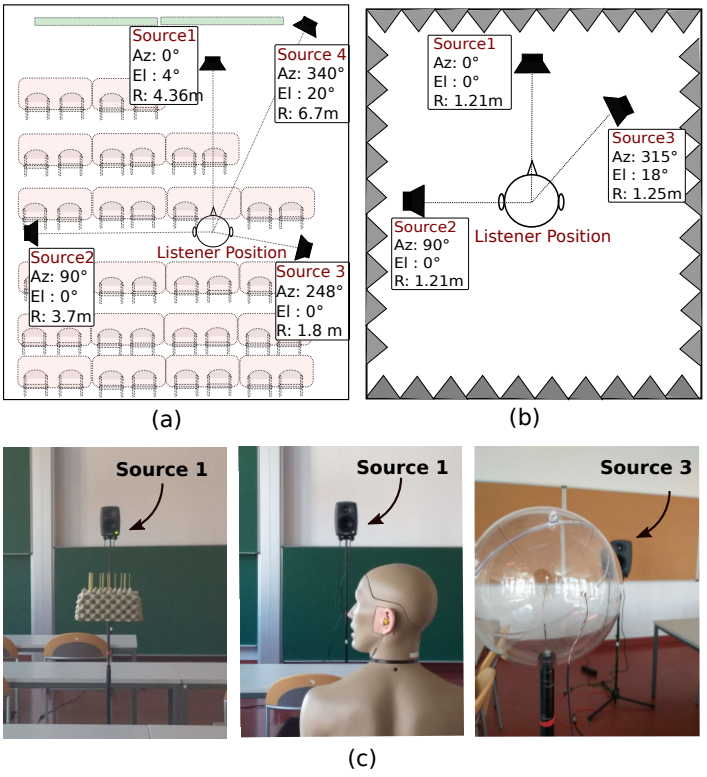

Fig. 5.3: Listener and source positions (Az: azimuth, El: elevation, R: distance to listener) in **(a):** lecture room (Experiment 1) and **(b):** anechoic room (Experiment 2). **(c):** (from left to right) VAH, KEMAR artificial head and the rigid sphere in the lecture room.

In both environments, one listener position and different source positions were defined, which are shown in Figures 5.3a and 5.3b. The VAH was placed at the listener position and RIRs were measured between the different source positions and the 24 microphones of the VAH. These RIRs were filtered with the FIR filters corresponding to the individually calculated left and right spectral weights (for each of the 185 head orientations) and added up over the N channels into the left and right BRIRs (see also section 4.3). These BRIRs are referred to as **VAH**

BRIRs.

In both environments, BRIRs were also measured with a commercial artificial head (KEMAR type 45BB, GRAS Sound & Vibration A/S, Holte, Denmark) as well as with a head-sized rigid sphere (radius = 8.5 cm) with two MEMS Knowles PSV0840LR5H microphones positioned at ±100° on the equator (see Figure 5.3c). In order to enable a head-tracked signal presentation with the BRIRs captured with the KEMAR artificial head or the rigid sphere at least for horizontal head orientations, the BRIR measurement was repeated 37 times for 37 horizontal orientations of the artificial head and rigid sphere (−90° to 90° in 5° steps). Note that this scenario of a moving artificial head is obviously non-realistic and cannot be employed in standard applications. It was considered in this study nonetheless since the usual static scenario for the KEMAR artificial head and the rigid sphere would have been too easy to discriminate from the head-tracked VAH BRIRs in listening tests. The BRIRs measured for different orientations of the KEMAR artificial head or the rigid sphere are referred to as **HTK BRIRs** (**H**ead-**T**racked **K**EMAR) and **HTS BRIRs** (**H**ead-**T**acked **S**phere), respectively.

All BRIRs were measured at f_s=44100 Hz, using the Multiple Exponential Sweep Method (MESM) [66], with modification as proposed in [183], with sweeps of 20 s duration, from 20 Hz to $f_s/2$ with 4 s shift between subsequent excitations. In the lecture room, the measured impulse responses were truncated to a length of 18000 samples using a 50-point half-Hann window. After convolution with the individual inverse HPIRs, the VAH, HTK, and HTS BRIRs were truncated again to a final length of 18000 samples, corresponding to 408 ms at f_s=44100 Hz and a decay of over 40 dB, which enabled to cover the usable dynamic range in the room (see section 5.4). In the anechoic room, the measured impulse responses were truncated to a length of 1024 samples using a 50-point half-Hann window. After convolution with the individual inverse HPIRs, the VAH, HTK, and HTS BRIRs had a final length of 3071 samples.

It should be noted that although the anechoic room could be considered as a free-field environment, the measured or synthesized binaural impulse responses in Experiment 2 are denoted as BRIRs (instead of HRIRs) to reflect the influence of the experimental apparatus in the room.

5.3.3 *Listening test - Technical implementation*

To evaluate the quality of the (individually synthesized) VAH BRIRs as well as the (non-individual) HTK and HTS BRIRs, two listening tests (Experiment 1 and Experiment 2) were performed. In both tests, head tracking was employed. During the listening tests, subjects sat at the same listener position as defined for the BRIR measurements. They were asked to rate different auralizations, generated either with VAH BRIRs for different parameter sets or with HTK and HTS BRIRs, in comparison to a reference signal (real loudspeaker playback in the

room). Subjects could decide by themselves when to listen to headphone or the reference signal and were asked to take off the headphones when listening to the loudspeaker. Playback was conveniently switched between the loudspeaker and headphone presentation via a push button on the headphones, as implemented in [92] (see also Figure 4.2 and section 4.4). Subjects had no information about the BRIR condition which was presented at any time. Loudspeakers and their positions, as well as all other features such as visual cues or the arrangement of the objects in the room, remained the same as during the BRIR measurements.

A custom-made head tracker (see section 4.4) was mounted on the top of the Sennheiser HD800 headphones (the same headphones as used for measuring the individual HPIRs) and methods described in section 4.4 were used for the real-time head-tracked binaural playback. In both experiments, the signals were played back over an external audio interface (RME Fireface UC). For the headphone signals, a headphone amplifier (Lake People Phone-Amp G103) was used. The loudness of the real sources was adjusted manually by the experimenters to have the same loudness impression as the headphone signals.

The different BRIRs were compared with the reference signal in terms of perceptual attributes (the same as used in [92]) "Halligkeit" (Reverberance), "Quellbreite" (Source Width), "Quelldistanz" (Source Distance), "Schallquellenrichtung" (Source Direction) and "Gesamtqualität" (Overall Quality). The perceptual attributes were presented always in the same order as given above, i.e. Experiment 1 started for all subjects with evaluating the attribute Reverberance, continuing to Source Width and so forth. The attribute Reverberance was not evaluated in Experiment 2. Therefore, Experiment 2 started with the attribute Source Width and so forth. In order to limit the number of evaluations, the perceptual attribute Spectral Coloration was not explicitly evaluated but was assumed to be included in the perceptual attribute Overall Quality. To give their ratings with respect to the perceptual attribute Overall Quality, subjects were instructed to exclude all aspects related to the previous attributes and to focus on everything not included yet. Subjects rated the attributes on a 9-point scale with five German labels "schlecht" (bad), "dürftig" (poor), "ordentlich" (fair), "gut" (good) and "ausgezeichnet" (excellent) and four unlabeled intermediate points (the scale point names and their English translations were taken from [175]). To obtain the ratings, a Graphical User Interface (GUI) was presented to the subjects, with sliders which could be moved with a mouse. Before starting the experiment, subjects could get familiar with the environment, with the GUI as well as with the equipment. The main experiment began after this familiarization by explaining the first perceptual attribute. After completing the ratings for one perceptual attribute and before continuing to the next one, subjects were provided with the explanation of the next perceptual attribute. Perceptual attributes were explained with a short description in German language.

Each of the source positions shown in Figure 5.3 appeared three times during the evaluation in a randomized order. Subjects were allowed to switch freely between different headphone signals and between headphone and loudspeaker presentations.

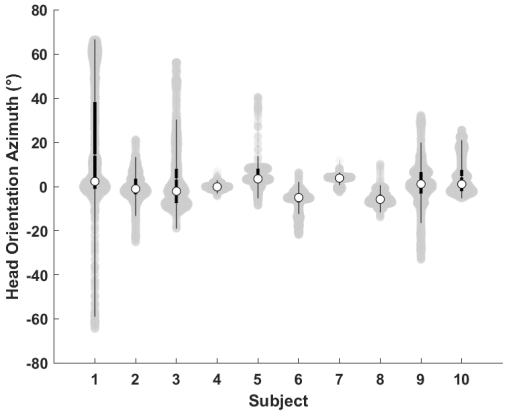

Fig. 5.4: Violin plots showing subjects' horizonztal head orientations when listening to headphone presentations of Source 2 and evaluating the perceptual attribute Overall Quality in Expriment 1.

They were informed that head rotations were permitted in the horizontal and vertical range of $\pm 90°$ and $\pm 15°$, respectively. However, no explicit instruction was given to the subjects to rotate their heads while listening to the signals. Subjects were asked to reset the head tracker by keeping the head to the front and clicking a "Reset" button on the GUI, before evaluating a given source position and perceptual attribute.

The stimulus was a dry recorded speech utterance of 15 s duration ("Nordwind und Sonne", text version from the IPA Handbook [184], first sentence), spoken by a female speaker in German. This audio sample was repeated to a total length of about three minutes to provide the subjects with enough time to compare and rate the different presentations. In case that subjects were not finished by the end of the 3-minute long signal playback, they could easily repeat the playback from the beginning. For a given source position and perceptual attribute, it took the subjects on average 2.5 minutes to complete the comparison between different headphone presentations and the reference signal.

Ten normal-hearing subjects (six male, four female, aged 20 to 52 years old, all having a hearing threshold of 15 dB HL or better verified by a pure tone audiometry between 125 Hz and 8 kHz) participated in the experiments. Eight subjects reported to have extensive experience with perceptual listening tests, while two subjects reported to not have much prior experience. For all subjects, individually measured HRIRs and HPIRs as well as individually calculated spectral weights for parameter sets listed in Table 5.1 for 185 head orientations were prepared.

It should be mentioned that although no explicit head movement instructions were given, all subjects moved the head during headphone signal presentation. The ampli-

tude and trajectory of head movements varied among the subjects. The intra-subject amplitude of head movements, on the other hand, remained stable across the perceptual attributes. Figure 5.4 shows exemplary horizontal head orientations of the subjects, collected by the tracker device when listening to headphone presentations of Source 2 and evaluating the perceptual attribute Overall Quality in Expriment 1.

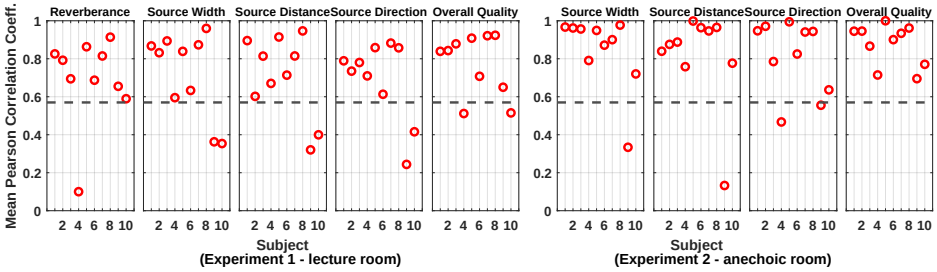

Fig. 5.5: Mean Pearson correlation coefficients $\overline{r}$ for the three presentation pairs (1-2, 1-3, 2-3) as a measure for consistent ratings. The dashed horizontal line indicates the chosen lower threshold for $\overline{r}$. Subjects with a $\overline{r}$ below this threshold were excluded from the evaluations. To calculate the correlations coefficients, 32 combinations (4 loudspeaker positions × 8 BRIR sets) for Experiment 1 and 18 combinations (3 loudspeaker positions × 6 BRIR sets) for Experiment 2 were considered for each of the three presentation pairs.

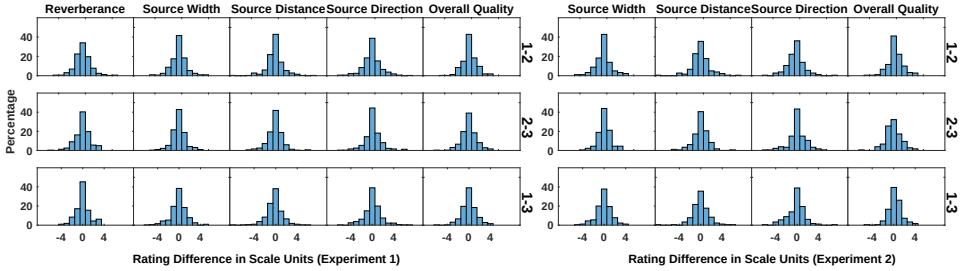

Fig. 5.6: Histogram of rating differences between the three presentation pairs (1-2, 1-3, 2-3) after excluding the non-consistent ratings. Two scale units correspond to the difference between adjacent labeled scale points.

5.3.4 *Exclusion of non-consistent ratings*

As already mentioned in section 5.3.3, for each perceptual attribute each source position in the room was presented and evaluated three times. To assess the consistency of the ratings over the three repetitions, the similar method as applied in section 3.6.2 and as in [92] was used: the Pearson correlation coefficients between the three presentation pairs (1-2, 1-3, 2-3) were calculated separately for each attribute and each subject. As a measure of repeatability, the mean correlation coef-

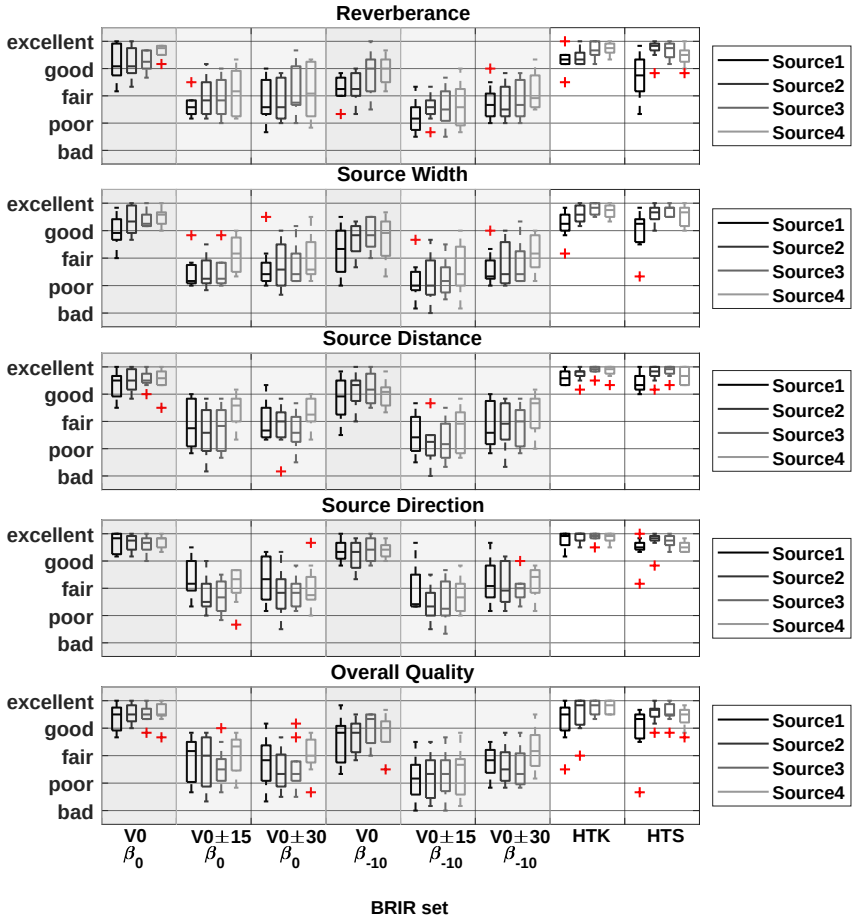

Fig. 5.7: (Experiment 1) Perceptual evaluations averaged over three repetitions, for five perceptual attributes, four source positions, and eight different BRIR sets.

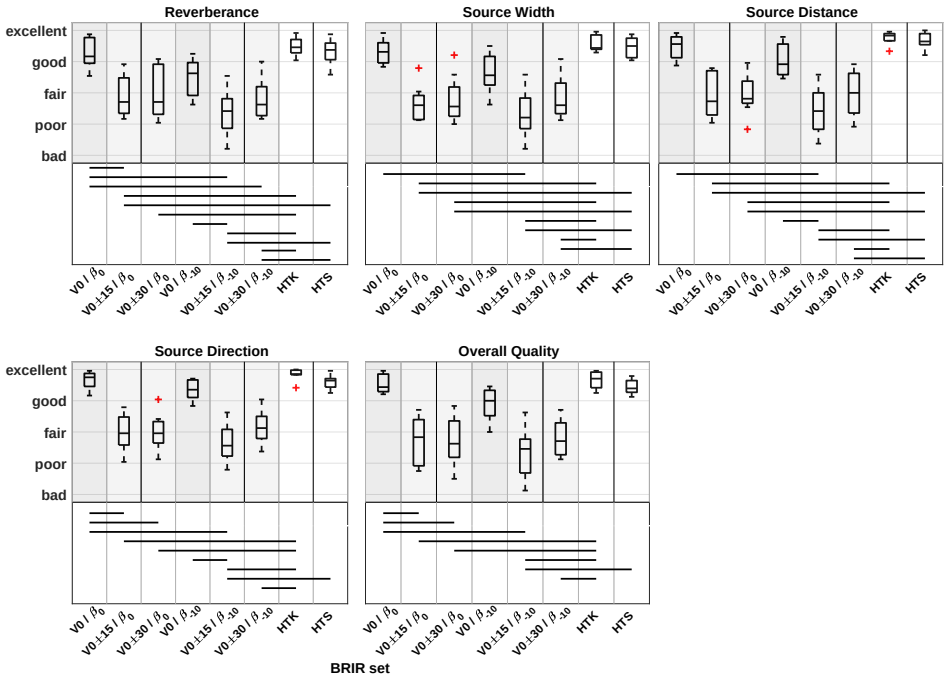

Fig. 5.8: (Experiment 1) Averaged ratings over four source positions for different BRIR sets and perceptual attributes. Significant different ratings are marked with horizontal lines ($p<0.05$).

ficient $\overline{r}$ was evaluated in relation to Cronbach's standardized coefficient [177] as in Eq. (3.15). With three repetitions and with $\alpha > 0.8$ as *good*, ratings with $\overline{r}$ >0.57 were considered as consistent and repeatable. If not, the ratings of this subject for the investigated perceptual attribute were excluded. The mean Pearson coefficients $\overline{r}$ for all subjects and for all perceptual attributes in both experiments are shown in Figure 5.5. For subjects fulfilling the repeatability criterion, the averaged ratings over three repetitions were considered for further analysis.

5.4 Experiment 1 - Auralization of the reverberant environment

The first experiment was performed in a lecture room (7.12 m × 11.94 m × 2.98 m) with an average reverberation time of 0.58 s and with six rows of tables and chairs (see Figure 5.3a). The listener position was chosen in the third row in the middle slightly shifted to the right, at 1.30 m height, which was assumed to be the height of the ear axis for subjects sitting at the listener position. Four sources were considered in the room: Source 1 (Genelec type 8030c) was located ahead of the listener at a slightly higher position than the ears. Two other sources, Source 2 and Source 3 (Genelec type 8030b), were located at the left and behind the listener at the right side, both at the same height as the ears. Source 4 (Event active studio monitor 20/20 bas V3), was located at the frontal upper right corner of the room at an elevation of about 20°. The sound pressure level at listener position was 60 dBA with signals played back from Source 1 and the background noise in the room was measured at around 20 to 25 dBA, depending on outdoor conditions.

For each source position, the six individually calculated VAH BRIRs, synthesized using the parameter sets listed in Table 5.1, as well as the non-individual HTK and HTS BRIRs, measured for the KEMAR artificial head and the rigid sphere, respectively, were evaluated.

5.4.1 *Experiment 1: results*

By excluding the non-consistent ratings as described in section 5.3.4, the number of subjects was reduced to eight for the perceptual attributes Source Width, Source Distance, Source Direction and, Overall Quality (Figure 5.5). For the perceptual attribute Reverberance, one subject was excluded. It should be noted that seven of the nine exclusions pertained to the two subjects with less experience (subject 9 and subject 10).

Figure 5.6 shows the histogram of rating differences between the three presentation pairs after excluding the non-consistent ratings. Between 35% and 48% of the ratings were identical (i.e. the difference was zero), and between 74% and 85% were within ±1 scale units. The symmetrical distribution of differences with respect to zero difference, similarly for all presentation pairs and attributes, indicates that there were no substantial learning effects over time.

Figure 5.7 shows the perceptual evaluations for the five perceptual attributes, four source positions, and eight different BRIR sets. For almost all perceptual attributes and source positions, the VAH BRIRs with V0/β_0 and the HTK and HTS BRIRs were rated similarly high, with median values between good and excellent. In comparison, the VAH BRIRs including non-horizontal directions (V0±15 and V0±30) were rated lower, regardless of the parameter β. Even for Source 4, which was located markedly out of the horizontal plane, the VAH BRIRs with V0±15 and V0±30 were rated lower than the VAH BRIRs optimized using only horizontal directions (V0/β_0 and V0/β_{-10}). For all source positions and perceptual attributes, the VAH BRIRs with V0/β_{-10} were rated lower than the VAH BRIRs with V0/β_0, but higher than the VAH BRIRs with V0±15 and V0±30, regardless of the parameter β.

Table 5.2: p-values (Friedman test) for investigating the effect of source position on the ratings given to different BRIR sets for each perceptual attribute in Experiment 1 (Exp.1) and Experiment 2 (Exp.2). p-values indicating significant different ratings ($p<0.05$) are depicted as bold numbers.

BRIR set	Reverberance		Source Width		Source Distance		Source Direction		Overall Quality	
	Exp.1	Exp.2	Exp.1	Exp.2	Exp.1	Exp.2	Exp.1	Exp.2	Exp.1	Exp.2
V0/β_0	0.058	–	**0.034**	**0.015**	0.526	0.748	0.666	0.145	0.231	0.581
V0±15/β_0	0.228	–	**0.001**	0.670	**0.010**	**0.020**	**0.015**	**0.011**	0.637	0.575
V0±30/β_0	**0.032**	–	0.286	–	**0.017**	–	0.409	–	0.050	–
V0/β_{-10}	**0.004**	–	0.073	0.428	0.365	0.176	0.436	0.067	0.057	0.478
V0±15/β_{-10}	**0.022**	–	0.190	**0.023**	**0.038**	0.648	0.075	0.115	0.078	**0.011**
V0±30/β_{-10}	**0.006**	–	**0.013**	–	**0.034**	–	0.830	–	**0.011**	–
HTK	**0.012**	–	**0.000**	0.434	**0.021**	0.314	0.625	0.050	0.138	0.335
HTS	**0.002**	–	**0.002**	0.393	**0.003**	0.661	0.060	0.184	**0.020**	0.823

According to the Shapiro-Wilk test of normality, applied to the ratings for each combination of source position, BRIR set, and perceptual attribute, ratings could not be assumed to be normally distributed for all cases ($p<0.05$). Therefore, a non-parametric method (Friedman test) was used to statistically analyze the ratings. According to the Friedman test, for 20 out of 40 combinations of BRIR sets and perceptual attributes, a significant effect of the source position could be observed. p-values are shown in Table 5.2, with bold cases indicating $p<0.05$. The effect of source position was often significant for the three perceptual attributes Reverberance, Source Width and Source Distance. However, since for each of the evaluated source positions the experiment design focused on the comparison of different BRIR sets, the ratings were averaged over the four source positions in order to statistically analyze the effect of the BRIR sets. The averaged ratings are shown in Figure 5.8. According to the Shapiro-Wilk test of normality, also the ratings averaged over the source positions could not be assumed for all BRIRs to be normally distributed. Therefore, the Friedman test was applied which revealed for all attributes a significant effect of BRIR set ($p<10^{-4}$). As indicated by

multiple comparisons after Friedman test (function friedmanmc in the statistical software R [178]), significantly lower ratings were given to VAH BRIRs with V0±15/$\beta_{0,-10}$ and V0±30/$\beta_{0,-10}$. For all perceptual attributes, there were no significant differences between the VAH BRIRs with V0/$\beta_{0,-10}$ and the HTK or HTS BRIRs. There were also no significant differences between V0/β_0 and V0/β_{-10}.

5.5 Experiment 2 - Auralization of the anechoic environment

The results in Experiment 1 revealed a perceptually successful performance of the VAH BRIRs as well as HTK and HTS BRIRs. The extent to which the room effects might have had an impact on the perception of different BRIRs was however not clear. Reverberation is expected to reduce source localization accuracy by itself, which may interact with the ratings of the subjects. It was interesting to see whether a similar performance with the tested BRIRs can also be achieved in the absence of room effects. Therefore, a similar experiment (Experiment 2) was performed in an anechoic environment. Since in Experiment 1, the ratings for the VAH BRIRs including non-horizontal directions (V0±15/$\beta_{0,-10}$ and V0±30/$\beta_{0,-10}$) were similarly low, the VAH BRIRs with V0±30/β_0 and V0±30/β_{-10} were excluded in Experiment 2.

Experiment 2 was performed in the anechoic room at the Jade University of Applied Sciences in Oldenburg (3.1m × 3.4m × 2m, cutoff frequency 200 Hz). The listener position was chosen in the middle of the room (see Figure 5.3b). Three sources (Fostex 6301B) were positioned in the room. Source 1 and Source 2 were located in front and at the left of the listener, respectively, both at the same height as the ears. Source 3 was located at 45° at the right side at an elevation of about 18°. Source 1 and Source 2 in Experiment 2 were considered equivalent to Source 1 and Source 2 in Experiment 1. However, due to practical reasons, Source 3 in Experiment 2, which was chosen to represent the sound source outside the horizontal plane, had a different position than its equivalent (Source 4) in Experiment 1. Nevertheless, the two non-horizontal sources had similar elevations (20° in Experiment 1 and 18° in Experiment 2). In addition, the azimuthal position of the non-horizontal sources, both in front of the listener on the right side, coincided with one of the azimuthal directions included in the calculation of the spectral weights (5° azimuthal resolution). Consequently, the impact of the constraint parameters chosen for the calculation of the VAH spectral weights on the perceived quality of the non-horizontal sources was considered comparable in both experiments.

For each source position, the four individually synthesized VAH BRIRs (with V0/β_0, V0/β_{-10}, V0±15/β_0 and V0±15/β_{-10}) as well as the (non-individual) HTK and HTS BRIRs were evaluated for four perceptual attributes Source Width, Source Distance, Source Direction, and Overall Quality. The perceptual attribute Reverberance was not considered due to the absence of this attribute for this environment.

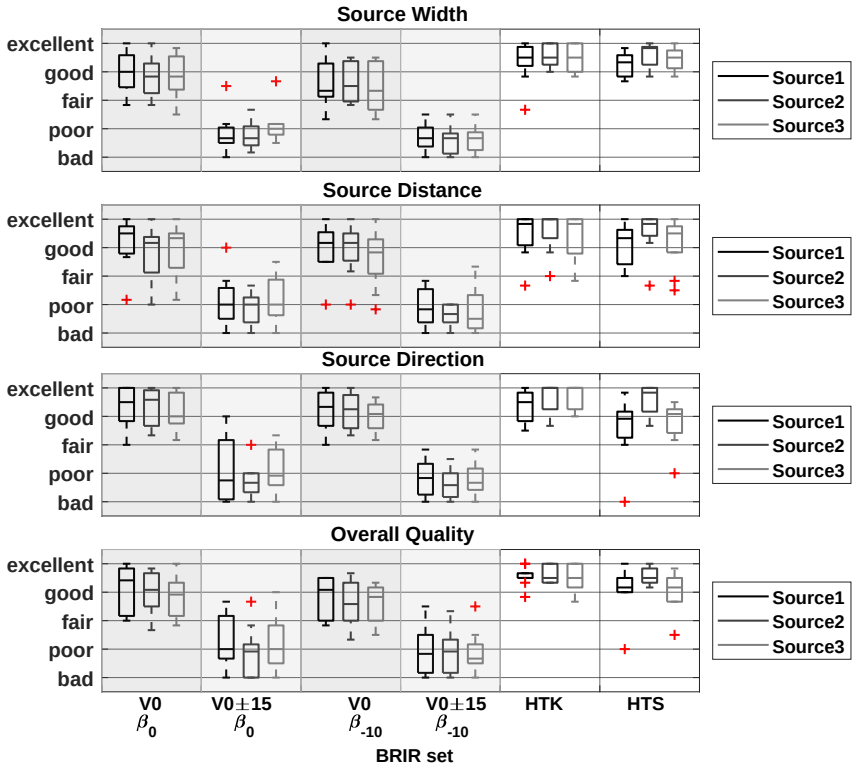

Fig. 5.9: (Experiment 2) Perceptual ratings averaged over three repetitions, for four perceptual attributes, three source positions, and six different BRIR sets.

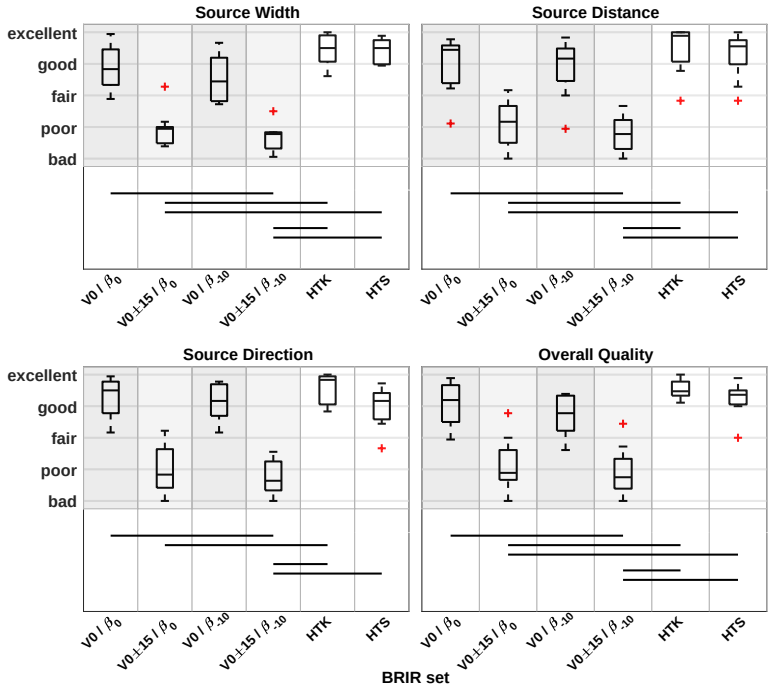

Fig. 5.10: (Experiment 2) Averaged ratings over three source positions for different BRIR sets and perceptual attributes. Significant different ratings are marked with horizontal lines ($p<0.05$).

5.5.1 *Experiment 2: results*

The consistency test described in section 5.3.4 was also used in the present experiment and led to the exclusion of one subject for the perceptual attributes Source Width and Source Distance and two subjects for the perceptual attribute Source Direction. For the perceptual attribute Overall Quality, no subjects were excluded (Figure 5.5). It should be noted that three of the four exclusions pertained to one of the subjects with less experience. Figure 5.6 shows the histograms of differences between the three repetitions. Between 32% and 45% of ratings were identical and between 71% and 84% were within ±1 scale units. A symmetrical distribution of differences with respect to zero difference can be observed for all presentation pairs and attributes.

Figure 5.9 shows the perceptual evaluations for the four perceptual attributes, three source positions, and six BRIR sets. Compared to Experiment 1, the VAH BRIRs with V0/β_0 were rated slightly lower, but still comparable with the HTK and HTS BRIRs, with median values between good and excellent in most cases. Similar to Experiment 1, the VAH BRIRs including non-horizontal directions (V0±15/β_0, V0±15/β_{-10}) were rated lower than the VAH BRIRs calculated with only horizontal directions (V0/β_0, V0/β_{-10}) and the HTK and HTS BRIRs. The median values dropped however from between fair and poor in Experiment 1 to around poor and bad. Also, similar to Experiment 1, the VAH BRIRs with V0/β_{-10} were rated slightly lower than the VAH BRIRs with V0/β_0.

According to the Shapiro-Wilk test of normality, ratings in Experiment 2 could not be assumed to be normally distributed for all cases. Therefore, the same non-parametric methods as used in Experiment 1 were applied to the ratings in Experiment 2. A significant effect of the source position was indicated by the Friedman test only for five out of 24 combinations of BRIR sets and perceptual attributes (p-values are shown in Table 5.2). Therefore, the ratings were again averaged over the three source positions. The averaged ratings are shown in Figure 5.10. The Friedman test revealed for all attributes a significant effect of the BRIR set ($p<10^{-4}$). Significantly different BRIR sets (according to the multiple comparisons after Friedman test) are indicated with horizontal lines in Figure 5.10. Significantly lower ratings were given only to VAH BRIRs with V0±15/β_0 and V0±15/β_{-10}. Similar as in Experiment 1, there were no significant differences between the VAH BRIRs with V0/$\beta_{0,-10}$ and HTK or HTS BRIRs. Also, there were no significant differences between V0/β_0 and V0/β_{-10}.

5.6 Discussion

5.6.1 *Comparison between auralization and reality*

For all attributes and for both environments, there were BRIRs for which the median values of the ratings were between good and excellent, i.e. at least 7 and more on

the 9-point scale used. As also discussed in [92], if the reference signal is known, subjects tend to avoid the highest point of the scale. Therefore, ratings between good and excellent were considered perceptually close to reality. The results suggest that it is possible to have dynamic auralizations which are perceived nearly the same as the original auditory scene, confirming the results that were obtained with simulated BRIRs in [92].

5.6.2 *Low ratings for VAH BRIRs with V0*±15 *and V0*±30

In both experiments, the ratings for VAH BRIRs calculated with horizontal and non-horizontal directions (V0±15/β_0, V0±30/β_0, V0±15/β_{-10} and V0±30/β_{-10}) were lower than the ratings for VAH BRIRs with V0/β_0 or V0/β_{-10}. This applied to all source positions. For sources in the horizontal plane (e.g., Source 2 in both experiments), one could explain this by the higher resulting SDs at horizontal directions for the case where horizontal and non-horizontal directions were included (compare the lower row in Figures 5.1a and 5.1b to the lower row in Figures 5.2a and 5.2b). However, the lower ratings of VAH BRIRs with V0±15 or V0±30 applied also to sources out of the horizontal plane (Source 4 in Experiment 1 or Source 3 in Experiment 2). These ratings cannot be explained by the SDs at non-horizontal directions, which would predict a better performance of the VAH BRIRs with V0±15 or V0±30. Instead, the ratings seem to be related to the resulting Temporal Distortion (TD). The higher TDs are suspected to have led to the lower ratings of the case with horizontal and non-horizontal directions included (see the higher resulting TDs in Figure 5.2 compared to those in Figure 5.1). Apparently, the constrained optimization algorithm sacrificed the phase accuracy to serve the large amount of constraints (216 + 1 in case of V0±15 and V0±30) which were applied to the spectral distortion (magnitude error) and mean WNG. Errors in the resulting phase (or TD) will then lead to deviations in the ITDs, which will have impeded the Source Direction ratings. In addition, the ITDs were only implicitly controlled for in the minimization of the cost function while the ILDs were explicitly controlled for as a direct consequence of constraints applied to the spectral distortion. As a result, non-matching ITDs and ILDs might have led to a spatial split or a diffuseness of the auditory event [185] or insufficient externalization, which will have impacted the Source Width and Source Distance ratings.

In case of the reverberant environment in Experiment 1, it is also of interest to consider the modified $\mathrm{RL}_{\mathrm{E}}^{'}$ (Room Level (Early)), which has been shown to correlate with the perceived Apparent Source Width (ASW) for music [186]. According to this measure, a higher $\mathrm{RL}_{\mathrm{E}}^{'}$ corresponds to a larger perceived ASW. Figure 5.11 shows the $\mathrm{RL}_{\mathrm{E}}^{'}$, calculated for the VAH BRIRs of subject 1 and the HTK and HTS using the method described in [186]. The $\mathrm{RL}_{\mathrm{E}}^{'}$s in Figure 5.11 were calculated for the frontal sound source in the lecture room and for horizontal head orientations θ_h between $-90°$ and $+90°$. The results show higher $\mathrm{RL}_{\mathrm{E}}^{'}$ of VAH BRIRs with V0±15/$\beta_{0,-10}$ and V0±30/$\beta_{0,-10}$ compared to VAH BRIRs with V0/$\beta_{0,-10}$ or HTK and HTS BRIRs, which implies that the virtual sources

generated with VAH BRIRs with V0±15 or V0±30 were difficult to be perceived at a focused position.

In general, the similarity of the results across perceptual attributes indicated that the synthesis artifacts in VAH BRIRs with V0±15 or V0±30 impacted similarly the quality of the headphone signals with respect to all of the evaluated perceptual attributes.

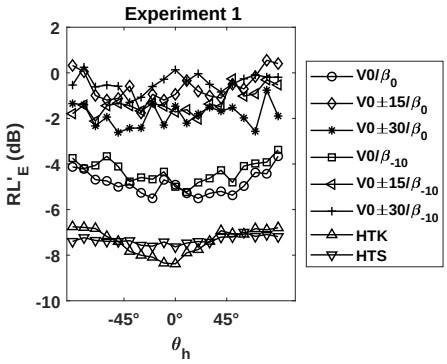

Fig. 5.11: $\mathrm{RL}_{\mathrm{E}}^{'}$ calculated for VAH BRIRs of subject 1 and the HTK and HTS BRIRs. RL_{E} was calculated for the frontal source in the lecture room.

5.6.3 *The choice of the P discrete source directions depending on the application case*

In both environments investigated in this study, the VAH BRIRs with V0/β_0 resulted in median ratings between good and excellent for most of the tested source positions and perceptual attributes. Although the resulting SDs at non-horizontal source positions were higher for these VAH BRIRs than with V0±15/$\beta_{0,-10}$ or V0±30/$\beta_{0,-10}$, it seemed that the increasing SDs towards higher frequencies for the BRIRs with V0/β_0 were not very crucial. In addition, the low-frequency TDs were lower with VAH BRIRs with only horizontal directions included. The results imply that it is advantageous to apply the constraints to horizontal directions only.

It must be noted that the advantage of calculating the spectral weights with horizontal source directions is valid for speech signals only, because the SDs of VAH BRIRs with V0 at non-horizontal directions only stay within an acceptable range in the frequency range important for speech. In case of applications using signals with a more pronounced high-frequency spectral content, additional audible artifacts are expected to occur at non-horizontal directions.

5.6.4 *The effect of the minimum desired* $\mathrm{WNG_m}$

In both experiments, for all source positions and perceptual attributes, the VAH BRIRs with V0/β_{-10} were rated slightly lower than the VAH BRIRs with V0/β_0. Although the effect of microphone self-noise was not evaluated in the same manner using the synthesized or measured BRIRs as it would be using real recordings, a possible mismatch between the measured steering vectors (see section 5.3.1) and the measured impulse responses (see section 5.3.2) was present. Since at least four months passed between measuring the steering vectors and measuring the impulse responses in the lecture room and in the anechoic room, it is possible that small deviations in microphone characteristics or positions occurred during this time period. With a lower value of parameter β, the susceptibility of the VAH synthesis to deviations in microphone characteristics increases, which possibly explains the lower ratings given to VAH BRIRs with V0/β_{-10} compared to V0/β_0. The case of β=0 dB was perceptually evaluated previously to be a proper choice for the used microphone array in this study [142] and the results in the present study confirmed it. The effect of the parameter β could also be observed for the VAH BRIRs including non-horizontal directions (V0±15/$\beta_{0,-10}$ and V0±30/$\beta_{0,-10}$), although the lower ratings for these VAH BRIRs were dominated by other factors, as discussed in section 5.6.2.

5.6.5 *The positive effect of reverberation*

The inclusion of reverberation in the binaural signals, when congruent with the reverberation of the real room (see section 5.6.6), can contribute to a better externalization even for the case that non-individual HRTFs of artificial heads are used [92, 187, 188] and help smooth out the deviations to individual HRTFs [94]. The generally higher ratings for VAH BRIRs in Experiment 1 compared to Experiment 2 implied that the synthesis errors of the VAH BRIRs were less audible in the reverberant environment.

The increase of apparent source width in the reverberant environment of Experiment 1 seems to have been particularly in favor of the ratings for Source 3. This source was located at the azimuthal position 248°, which did not match any of the azimuthal directions considered, regarding the 5° azimuthal resolution of the measured HRTFs and steering vectors. The synthesis at directions other than the ones included in the calculation of the spectral weights can be subject to audible artifacts. Although with the relatively close distance of Source 3 to the listener the direct part of the RIR had more energy than the reverberant part, the small ratio of the reverberation included in the measured RIR for Source 3 was enough to cover up the potential audible artifacts. Such artifacts would probably have been audible in the anechoic environment if sources at positions not matching the considered directions had been evaluated in Experiment 2.

The presence of reverberation and reflections were also helpful against the non-individual cues of KEMAR artificial head or the rigid sphere. However, the comparably high ratings given to HTK and HTS BRIRs in the anechoic environment suggested that also other factors promoted the high ratings for non-individual BRIRs, which are discussed in section 5.6.6.

5.6.6 *The positive effect of head tracking, the compatibility of the auralized and listening rooms, and the presence of visual cues*

The similarly high ratings given to HTK and HTS BRIRs and VAH BRIR with V0/β_0 in Experiment 2 are not in accordance with the results of Rasumow et al. [15], where individual binaural presentations generated with the VAH (the same microphone array as used in the present study) in the anechoic environment outperformed the presentations generated with a conventional artificial head. The major differences between the study in [15] and the study here were the stimulus and the presentation method. Rasumow et al. evaluated the VAH and artificial head signals with noise bursts in a static scenario, i.e. without head tracking. Broadband test signals appeal a different challenge on the spectral accuracy compared to speech signals. Furthermore, although the advantages of using individual HRTFs are known in a static signal presentation without head tracking (lack of externalization or localization ambiguities [7,98]), it has been shown that the incorporation of head tracking can significantly reduce the localization ambiguities such as front-back reversals [12] and that the effect of head-tracking is larger than the effect of using individual HRTFs [98,119].

In addition, other features promoted the quality of the signals generated with VAH, HTK, and HTS BRIRs in this chapter. For both experiments, the listening test was performed in the same environment that was also auralized, with all perceptual cues preserved as they were during the impulse response measurements. A discrepancy between the auralized room and the listening room can impact the externalization or the perceived distance of the sound source negatively [89,90].

Another relevant feature was the visual information about the sources and their positions in the room. The knowledge of the source position can help suppress front-back reversals and improve the externalization. In addition, the presence of visual information can draw the acoustically perceived source position to the visual one [58].

At any time in everyday life, the surrounding environment is being perceived and evaluated based on the information available from different modalities in accordance with each other. The present study also offered a high consistency between the acoustical and visual features. Regarding the high perceptual ratings given to non-individual BRIRs of HTK or HTS, one can question the need for individualizing the binaural signals, if the head-tracked binaural presentation, applied to less critical

signals such as speech, can maintain such a consistency, especially for cases where no external reference is provided.

5.6.7 *VAH vs. conventional artificial head*

As discussed in section 5.6.6, the possibility of applying head-tracking to the non-individual binaural signals of the conventional artificial head can be expected to improve the perceptual quality. Regarding the fact that the conventional artificial heads do not normally offer the possibility of dynamic presentation in their standard applications, the incorporation of head tracking constitutes the great advantage of the VAH approach against these conventional artificial heads. Although preparing the spectral weights for a large number of head orientations requires a large number of calculations, these spectral weights are calculated only once and can then be applied to any recording. The comparable perceptual ratings given to VAH BRIRs with V0/β_0 and HTK or HTS together with the provided ability of the VAH to allow dynamic auralizations confirmed that the VAH is the more promising alternative for head-tracked auralizations of different environments with a realistic signal such as speech.

5.7 Summary

In this chapter, the Virtual Artificial Head (VAH) of Rasumow et al. [15] (planar microphone array with 24 microphones) was used to synthesize individual Binaural Room Impulse Responses (BRIRs) in two acoustically different environments (lecture room and anechoic room). VAH spectral weights were calculated for 185 head orientations (37 horizontal $\times$ 5 vertical), individually for each listener, using different sets of parameters. Individual BRIRs were synthesized by filtering the room impulse responses measured with the VAH with the FIR filters corresponding to the inverse Fourier transform of the spectral weights.

The results of the perceptual evaluations suggest that realistically (i.e. perceptually close to the original scenario) sounding head-tracked auralizations of speech can be realized using the VAH approach. This was shown for two different acoustical environments and for sources in and out of the horizontal plane. The choice of the discrete source directions included in the calculation of the spectral weights is critical for the quality of the synthesis. According to the perceptual results, it was advantageous to include directions from the horizontal plane only. A total of 72 horizontal directions together with the 5° resolution for the horizontal head orientations was sufficient to achieve good perceptual results with the VAH. The slightly higher perceptual results for the reverberant environment indicate the positive effect of reverberation in masking the synthesis errors and thus improving the perceptual quality of the synthesis with the VAH.

The results also showed that the resulting mean White Noise Gain (WNG_m), as a measure for robustness can as well impact the quality of the binaural signals generated with the VAH. In general, it is advisable to avoid low resulting WNG_m in order to increase the robustness of the microphone array against e.g. changes in microphone positions or microphone self-noise.

Non-individual BRIRs measured with a conventional artificial head or a simple rigid sphere can also result in highly realistic auralizations of speech, provided that head tracking with sufficiently many head orientations is employed. This means that different head orientations have to be accounted for by repeating the BRIR measurements. This will only rarely be an option in BRIR measurements and not be possible in live recordings. It is still interesting to note that individual BRIRs are not necessarily required for the case that the binaural signals can be presented dynamically.

The success of the VAH by including only horizontal source directions, as reported in this chapter, applies to the tested speech signal or signals with comparable spectral content only. When listening to broadband signals, the inclusion of non-horizontal source directions is expected to be more critical for preserving the synthesis accuracy at positions outside the horizontal plane. In addition, when using other test signals, the appropriateness of the spatial resolution for different head orientations in this study (5°) should be verified as well. More accurate statements with this regard require further perceptual evaluations.

It would also be interesting to investigate the extent to which the effect of the head tracking and visual cues contributed to the results, as will be thoroughly discussed in the next chapter.

6

LOCALIZATION PERFORMANCE WITH DYNAMIC BINAURAL SIGNALS GENERATED WITH TWO VARIANTS OF THE VAH

6.1 Introduction

The aim of binaural technology is to provide the listeners with the same spatial impression of a sound field as they would have experienced if they were present in the actual sound field. To preserve spatial information, binaural signals are often recorded with artificial heads or are pre-processed with Head Related Transfer Functions (HRTFs), and then presented over headphones.

Previous studies have shown the advantage of using individual over non-individual HRTFs in reducing localization ambiguities and improving externalization [7,95,98]. However, Oberem et al. [98] showed that when incorporating head tracking during signal playback, individualization is not very important. Dynamic binaural presentation, i.e. with head tracking, was shown to greatly enhance the quality with respect to externalization and reduction of front-back reversals, regardless of using individual or non-individual HRTFs [12,13]. If sufficient care is taken regarding the system latency for head tracking, the spatial resolution of head orientations, and

This chapter is based on:
[154] M. Fallahi, M. Hansen, S. Doclo, S. van de Par, D. Püschel, and M. Blau, "Dynamic Binaural Rendering: The Advantage of Virtual Artificial Heads Over Conventional Ones For Localization With Speech Signals", *Applied Sciences*, vol. 11, pp. 6793, 2021.

[155] M. Fallahi, M. Hansen, S. van de Par, S. Doclo, D. Püschel, and M. Blau, "Localization Performance in the Absence of Visual Cues for Binaural Renderings generated with a Virtual Artificial Head", *Proc. Fortschritte der Akustik - DAGA*, Hanover, Germany, pp. 106-109, 2020.

[156] M. Fallahi, M. Hansen, S. van de Par, S. Doclo, D. Püschel, and M. Blau, "Localization Performance For Binaural Signals Generated with a Virtual Artificial Head in the Absence of Visual Cues", *Proc. e-Forum Acusticum*, Lyon, France, pp. 1937-1944, 2020.

the correct compensation of headphone transfer functions, it is essentially possible to achieve a similar localization performance with dynamically presented virtual sources as with real sources [96, 189, 190].

Compared to conventional artificial heads, a VAH not only offers the possibility to adjust to individual HRTFs by using individually optimized spectral weights, but also to adjust to different head orientations during playback. This can be achieved by calculating spectral weights for different head orientations, such that the same recording, captured for a single orientation of the VAH during the recording, can be presented dynamically during playback, i.e. with head tracking. For a binaural recording with a conventional artificial head on the other hand, the recording can be presented only for a fixed head orientation of the listener, namely the orientation of the artificial head during the recording.

The evaluations in chapter 5 showed a good performance of the VAH of Rasumow et al. [15] (planar microphone array with 24 microphones) in dynamic auralizations with speech signals with respect to different perceptual attributes, including the perceived source position. Non-individual binaural signals generated with Binaural Room Impulse Responses (BRIRs) of a conventional artificial head and a rigid sphere were evaluated as well. Dynamic presentation was artificially enabled by BRIRs measured for different head-above-torso orientations of these artificial heads, which is quite unrealistic and different from the typical application of an artificial head in practice. The signals generated with such non-individual BRIRs showed also a good perceptual performance. Since in experiments in chapter 5 subjects could see the real sources in the room, the first open question was to which extent the visual information about the sound sources contributed to the successful performance of the VAH and the non-individual binaural signals, especially with respect to the perceived source position. The second open question concerned the possibly positive impact of head tracking on the perceptual results. This chapter aims at answering both questions.

This chapter consists of two parts. In part I, the performance of localizing virtual sources is assessed in the absence of visual cues. A localization experiment is performed with dynamically presented virtual sources generated with two VAHs (the same planar microphone array with 24 microphones as used in chapter 5 and a three-dimensional array with 31 microphones) as well as with a conventional artificial head (the same artificial head as evaluated in chapter 5). Subjects are asked to localize the virtual sources, while listening to dynamic headphone signals in darkness, i.e. without being supplied with any visual information about the sources. Subjects map the perceived source position on a Graphical User Interface (GUI). The same localization test is also performed with hidden real sound sources in order to verify the employed GUI. In part II, the impact of dynamic presentations on the localization performance with virtual sources generated with the VAHs and the conventional artificial head is assessed by two separate localization experiments, one with and one without head tracking.

The chapter continues in section 6.2 with a review of the methods and parameters used to calculate the individual spectral weights. Sections 6.3 and 6.4 present the methods, the results, and the discussion of the results for both localization experiments. Finally, some general discussion is offered in section 6.5.

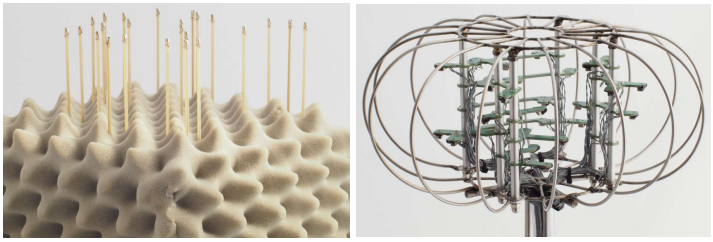

Fig. 6.1: VAHs used in this chapter. VAH1: planar microphone array with 24 microphones (20 cm×20 cm) [15]. VAH2: three-dimensional microphone array with 31 microphones (11 cm (Width)× 11 cm (Length)× 6 cm (Height))[1].

6.2 Microphone arrays and constraint parameters

For assessing the localization performance in this chapter, two different microphone arrays, referred to as VAH1 and VAH2, as shown in Figure 6.1 were used. VAH1 was the planar microphone array with 24 microphones, developed by Rasumow et al. [15], which was also used in dynamic auralizations in chapter 5 (also shown in Figure 1.4). VAH2 was a three-dimensional microphone array, 11 cm (Width)× 11 cm (Length)× 6 cm (Height), consisting of 31 microphones (TDK InvenSense ICS-40730). In both VAHs, the microphones were spatially distributed such that the inter-microphone distances were as different as possible in all possible directions. This was achieved by placing the microphones based on the Golomb ruler [144]. Individual spectral weights were calculated for both microphone arrays with measured steering vectors as described in Appendix A and by imposing constraints on Spectral Distortion (SD) and mean White Noise Gain ($\mathrm{WNG_m}$), as described in chapter 3. The SD constraint parameters $\mathrm{L_{Up}}$ and $\mathrm{L_{Low}}$ were chosen as 0.5 dB and −1.5 dB, respectively, as in chapter 5. The minimum desired $\mathrm{WNG_m}$ was chosen as β=0 dB. Although this chosen value for β was evaluated only for VAH1 in [142] as well as in chapter 5, the same value was also used for VAH 2. Two sets of directions were considered in the optimization: P=72 directions, equally spaced in the horizontal plane (i.e. 5° resolution), and P=3×72=216 directions, i.e. the 72 directions from the horizontal plane as well as the same 72 azimuths at two elevations ±15°. A summary of the constraint parameter P and the VAHs used in this chapter is given in Table 6.1. For both VAHs in this chapter, individual

1 Photos taken by Dr. rer. nat. Ralph Nolte-Holube.

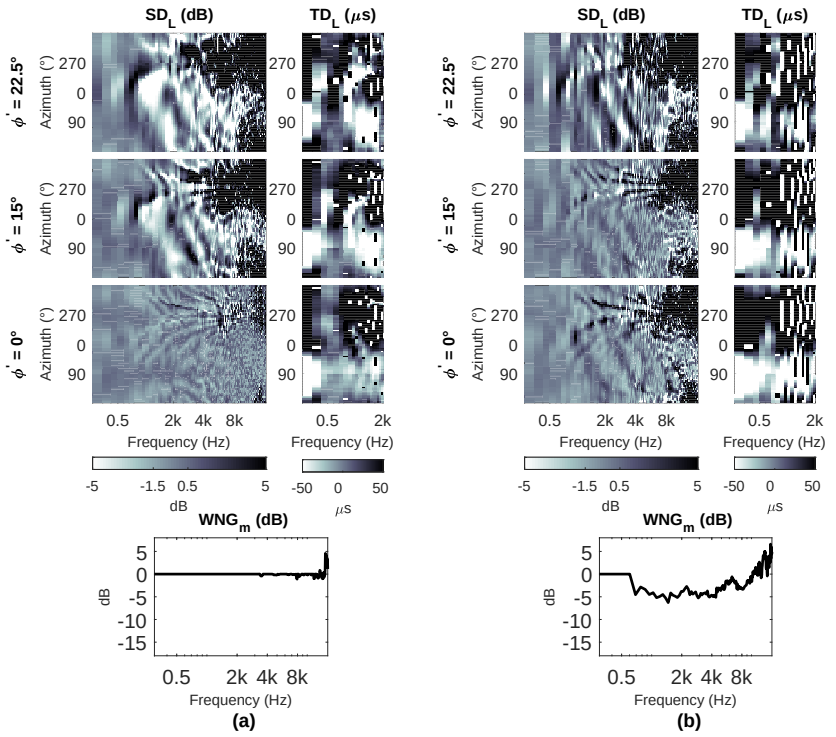

Fig. 6.2: The resulting SD and TD at elevations 0°, 15° and 22.5° and the resulting $\mathrm{WNG_m}$. The spectral weights were calculated with **(a):** P=72 horizontal directions and **(b):** P=3×72=216 directions from elevations −15°, 0° and +15°, both with β=0 dB and using the steering vectors measured with VAH2 (microphone array with 31 microphones shown in Figure 6.1). Results are shown for the left ear of subject 1 in chapter 5.

spectral weights labeled **V11**, **V13**, **V21** and **V23** (see Table 6.1) were each calculated for 37×5=185 head orientations, corresponding to 37 azimuth angles θ_h of −90° to +90° in 5° steps and 5 elevation angles ϕ_h of −15° to +15° in 7.5° steps (c.f. section 4.2). The chosen spatial resolutions for head orientations were assumed to be sufficient for speech signals, motivated by findings in chapter 5.

As already discussed in chapter 5, the directions included in the calculation of the spectral weights influence the synthesis performance at these and other directions. On the one hand, the spectral distortion is typically lower at the directions included in the calculation of the spectral weights than at other directions. On the other hand, the more directions are included, the more difficult it becomes to satisfy the increased number of SD constraints and the $\mathrm{WNG_m}$ constraint. It was shown in chapter 5 that this leads to a deterioration of the phase accuracy, referred to as

Temporal Distortion (TD). For VAH1, exemplary resulting SD, TD and $\mathrm{WNG_m}$ when synthesizing HRTFs at elevations 0°, 15° and 22.5° with VAH1 with P=72 horizontal directions (V11) and P=216 directions, i.e. the 72 directions from the horizontal plane as well as the same 72 azimuths at two elevations ±15° (V13) are shown in Figures 5.1a and 5.2a in chapter 5. For VAH2, exemplary resulting SD, TD and $\mathrm{WNG_m}$ when synthesizing HRTFs at elevations 0°, 15° and 22.5° with P=72 horizontal directions (V21) and P=216 directions (the 72 directions from the horizontal plane as well as the same 72 azimuths at two elevations ±15°) (V23) are shown in Figures 6.2a and 6.2b, respectively.

For the parameter sets considered in this chapter, this means that V11 and V21 provide a more accurate synthesis (smaller SD and TD) at horizontal directions and a less accurate synthesis (higher SD and TD) at non-horizontal directions, because the non-horizontal directions were not included in the calculation of the spectral weights. In contrast, when including both horizontal as well as non-horizontal directions in V13 and V23, the overall accuracy is distributed over a large number of directions. As a result, the spectral synthesis accuracy improves at non-horizontal directions compared to V11 and V21, while it degrades at horizontal directions. However, the large number of directions in V13 and V23 leads to higher TDs at all directions compared to V11 and V21. The localization performance is therefore expected to be less accurate with V13 and V23 compared to V11 and V21, at least with respect to azimuth accuracy and externalization. In perceptual evaluations in chapter 5, for sound sources in and outside the horizontal plane, speech signals synthesized with spectral weights including only 72 horizontal directions perceptually outperformed the signals synthesized with spectral weights including 216 directions from horizontal and non-horizontal directions.

Table 6.1: Overview of parameter P and the VAHs used to calculate the spectral weights.

Label	Constraint parameter P and the used VAH
V11	VAH1 - P=72 (Elevation: 0°)
V13	VAH1 - P=3×72=216 (Elevations: −15°, 0°, 15°)
V21	VAH2 - P=72 (Elevation: 0°)
V23	VAH2 - P=3×72=216 (Elevations: −15°, 0°, 15°)

6.3 Part I: Localization of real and virtual sources in the absence of visual cues

The localization study in part I consisted of two listening tests. The first listening test, referred to as **TestVR**, was performed to assess the localization performance when listening to individual binaural signals generated with both VAHs as well as with non-individual binaural signals of a conventional artificial head, for which the BRIRs were measured for different head orientations. During TestVR, the virtual

sound source was presented at different target source positions dynamically, i.e. with head tracking, over headphones. Subjects sat in a darkened room with very limited visual information about the surroundings and were asked to indicate the perceived source position using a Graphical User Interface (GUI). In a second listening test, referred to as **TestReal**, subjects listened to signals played back in the same darkened room from real (hidden) sound sources and were asked to indicate the perceived source position using the same GUI as in TestVR. Both listening tests, as well as the measurements which were done to prepare the binaural signals with the VAHs (Room Impulse Response (RIR) measurements with both VAHs as well as the BRIR measurements for the conventional artificial head), were performed in the anechoic room (3.1m×3.4m×2m, cutoff frequency=200 Hz) at the Jade University of Applied Sciences in Oldenburg. For each test, a different set of 15 target positions, as shown in Figure 6.3a and Figure 6.3b was considered. The azimuthal target positions were chosen randomly at multiples of 5° between 0° and 355°. Six different elevations (0°, ±10°, ±20° and +25°) were assigned to the 15 target positions such that a balance between the number of positive, negative and zero elevations in front and in back could be maintained. Target positions were distributed non-uniformly across both listening tests to prevent subjects from guessing the presentation grid. Individual Head Related Impulse Responses (HRIRs) and steering vectors for both VAHs, as well as individual Headphone Impulse Responses (HPIRs) for Sennheiser HD800 headphones were measured with the measurement setup and methods described in Appendix A. All procedures were approved by the ethics committee of the Carl von Ossietzky University of Oldenburg.

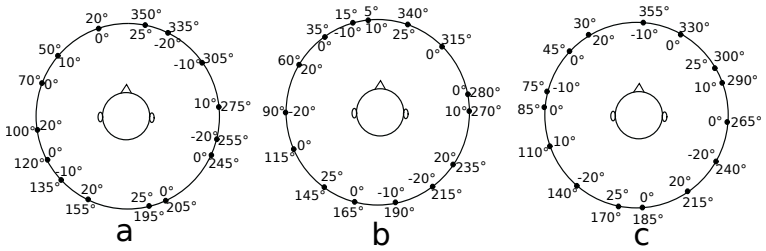

Fig. 6.3: Target positions when localizing with **(a):** real sources in TestReal, **(b):** virtual sources in TestVR, and **(c):** virtual sources in TestDyamic and TestStatic. Numbers outside and inside the circle indicate the azimuth and the elevation of the target sources, respectively.

6.3.1 *Experiment design*

6.3.1.1 *Target source positions in the room*

A loudspeaker arc of 1.2 m radius, hanging vertically from a turntable installed in the ceiling of the anechoic room was used to represent the target source positions

in TestReal as well as to prepare the virtual sources in TestVR, as described later in section 6.3.1.3 (RIR measurements with both VAHs as well as the BRIR measurements for the conventional artificial head). The center of the loudspeaker arc at 1.24 m height was defined as the listener position. Six loudspeakers (SPEEDLINK XILU SL-8900-GY) were mounted in the arc at elevations 0°, ±10°, ±20° and +25°. The loudspeaker arc could be rotated by the turntable to any azimuth. Signals were played back through loudspeakers over an ADI-8DS RME audio interface. Loudspeakers were individually equalized (in amplitude and phase) using 256-tap FIR filters, calculated as the regularized inverse [191] of transfer functions measured with a calibration microphone (GRAS 40AF), using a regularization parameter of $\beta_{inversion}$=0.3 times the square value of the average of the impulse responses measured with the calibration microphone.

6.3.1.2 *Localization of real sound sources (TestReal)*

During TestReal, subjects sat with their interaural center at the listener position in the anechoic room. In order to eliminate any visual information about the source positions, subjects sat inside an acoustically transparent tent (see Figures 6.4a and 6.4b) and the room was darkened. The only source of light was a tablet monitor, installed in front of the subjects and used by them to conduct the experiment and to give their responses. The loudspeaker arc was rotated to one of the 15 azimuthal target positions shown in Figure 6.3a. The test signal was played back from the loudspeaker channel corresponding to the target elevation. Subjects were encouraged to rotate their heads when listening to the signals within an allowable range of ±90° in horizontal and ±15° in vertical directions and not to exceed this range even if they perceived the sound source behind them. Each of the 15 target source positions was presented once and the presentation order was randomized. Five target positions were chosen randomly to be presented at the beginning for familiarization. Responses given to these five target positions were discarded from the evaluations. No feedback was given to the subjects during the familiarization as well as during the listening test.

6.3.1.3 *Localization of virtual sources (TestVR)*

To generate the binaural signals for TestVR, both VAHs were positioned at the listener position in the anechoic room. The same loudspeaker arc in combination with the turntable as used in section 6.3.1.2 was used to measure the RIRs between the microphones of VAH1 and VAH2 and each of the target positions shown in Figure 6.3b. In order to keep the acoustical conditions comparable to TestReal, the VAHs were placed inside the acoustically transparent tent during the measurement of RIRs (see Figure 6.4c). These RIRs were filtered with the FIR filters corresponding to the individually calculated left and right spectral weights (for each of the 37×5=185 head orientations and with the parameters listed in Table 6.1) and added up over the N channels into the left and right BRIRs (see also section 4.3). This resulted in four sets of individually synthesized VAH BRIRs,

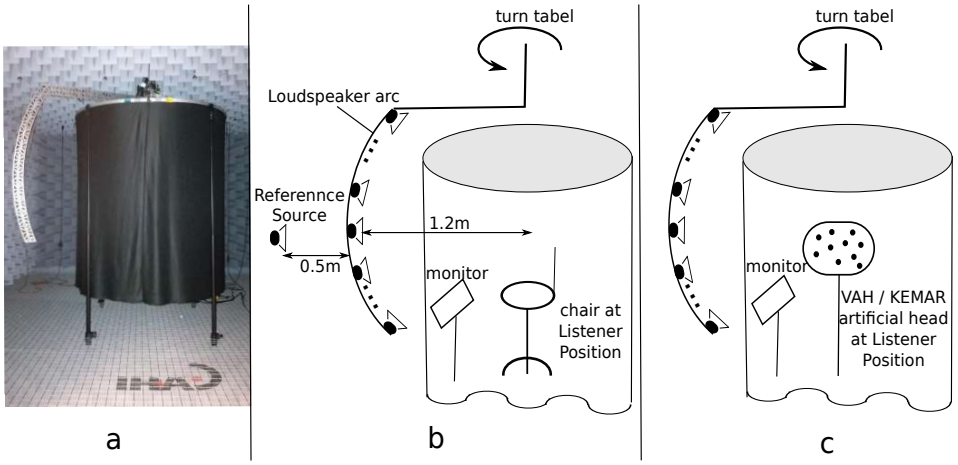

Fig. 6.4: **(a):** Acoustically transparent tent and the loudspeaker arc in the anechoic room. **(b):** Experiment setup during TestReal and TestVR (During TestReal, the loudspeaker arc was used to represent the target sources. During TestVR, virtual target sources were presented over headphones). **(c):** Setup for room impulse response measurements with the VAHs or BRIR measurements for the KEMAR artificial head inside the acoustically transparent tent.

synthesized with V11, V13, V21 and V23, each for 185 head orientations.

In addition, for each of the target source positions shown in Figure 6.3b, BRIRs were acquired with the two ears of a KEMAR artificial head (KEMAR type 45BB, GRAS Sound & Vibration A/S, Holte, Denmark) placed at the listener position inside the acoustically transparent tent. In order to enable a dynamic presentation with these BRIRs and similar to chapter 5, for each of the 15 target source positions shown in Figure 6.3b, the BRIR measurement was repeated for 37 head-above-torso orientations of the KEMAR artificial ($-90°$ to $+90°$ in $5°$ steps), resulting in $37\times15=555$ measurements. It should be mentioned again that for signals recorded with conventional artificial heads, dynamic presentation of the recordings is not possible in practice. The additional effort to measure BRIRs for different head-above-torso orientations of the KEMAR artificial head was only accepted in the present study because otherwise, signals generated with KEMAR BRIRs would clearly lose out against signals generated with the VAH BRIRs during the localization experiments. The BRIRs measured for 37 head-above-torso orientations of the KEMAR artificial head are referred to as **HTK** (**H**ead-**T**racked **K**EMAR).

It should be noted that although the anechoic room can be considered as a free-field environment, the synthesized or measured binaural impulse responses were denoted as BRIRs rather than HRIRs to reflect the influence of measurement equipment in

the room.

During TestVR, subjects sat with their interaural center at the listener position inside the acoustically transparent tent and the room was darkened. Subjects wore the headphones (the same as used to measure the individual HPIRs) with a custom-made head tracker mounted on top (see Figure 4.2. The push button was not used for the localization experiments in this chapter). The real-time head-tracked binaural playback was generated with the methods described in section 4.4. Audio signals were presented over an RME Fireface UC sound card and a Lake People Phone-Amp G103 headphone amplifier. Subjects were instructed to reset the head tracker before listening to the virtual source by keeping their head oriented to the frontal direction, indicated with a mark on the top of the tablet monitor in front of them, and press the "Reset" button on the GUI. Subjects were encouraged to make use of the possibility to rotate their heads within an allowable range of $\pm 90°$ in horizontal and $\pm 15°$ in vertical directions. Each of the 15 target source positions was presented five times, i.e. once with each of the five BRIRs (V11, V13, V21, V23, and HTK), which resulted in 75 virtual sources, presented in a randomized order. Five of these 75 virtual sources were chosen randomly to be presented at the beginning for familiarization. Responses given to these five target positions were discarded from the evaluations. No feedback was given to the subjects during the familiarization as well as during the listening test.

6.3.1.4 *Response method*

The localization task in this chapter consisted of providing information about the perceived azimuth, elevation and distance of the real or virtual sources. The GUI shown in Figure 6.5 was used to collect the responses. This GUI was presented on the tablet monitor positioned in front of the subject. For collecting the azimuth responses, the GUI showed the head as seen from above, with a circle around it. To enter the perceived source azimuth, subjects could click on any point on this circle. For collecting the elevation responses, the GUI showed an equivalent depiction of the head as seen from the side. The frontal direction of azimuth=0° and elevation=0°, corresponding to the frontal head orientation, was marked with a colored point on the GUI. To give the perceived source distance, subjects were supplied with a real reference sound source (the same loudspeaker type as mounted in the loudspeaker arc), which was positioned at a fixed position in front of the subject outside the acoustically transparent tent. By clicking the "Play" and "Reference" buttons, subjects could switch between the (real or virtual) target source and the reference source, respectively. In TestVR, subjects were asked to take off the headphones while listening to the reference source. Subjects had to judge the perceived distance either with respect to their own body or compared to the reference source using an ordinal scale from 0 to 4, corresponding to perception (0) in the head, (1) outside but near the head, (2) outside the head and closer than the reference, (3) outside the head and at the reference distance, or (4) outside the head and at a further distance than the reference. This scale was inspired by similar scales commonly used in studies investigating externalization of virtual sources [14, 83].

The reference source was first positioned at the same distance as the target sources and the target and reference sources were adjusted to have the same level (55 dB SPL) at the listener position. For both tests (TestReal and TestVR), the reference source was then displaced back by 50 cm in order not to interrupt the target source presentation during TestReal using the rotating loudspeaker arc (see Figure 6.4b). Subjects had no information about the exact position of the reference source and were instructed not to consider this source as a reference for azimuth and elevation, but only for the perceived distance. The "Reset" button was used during TestVR to reset the head tracker. For TestReal, this button was omitted from the GUI.

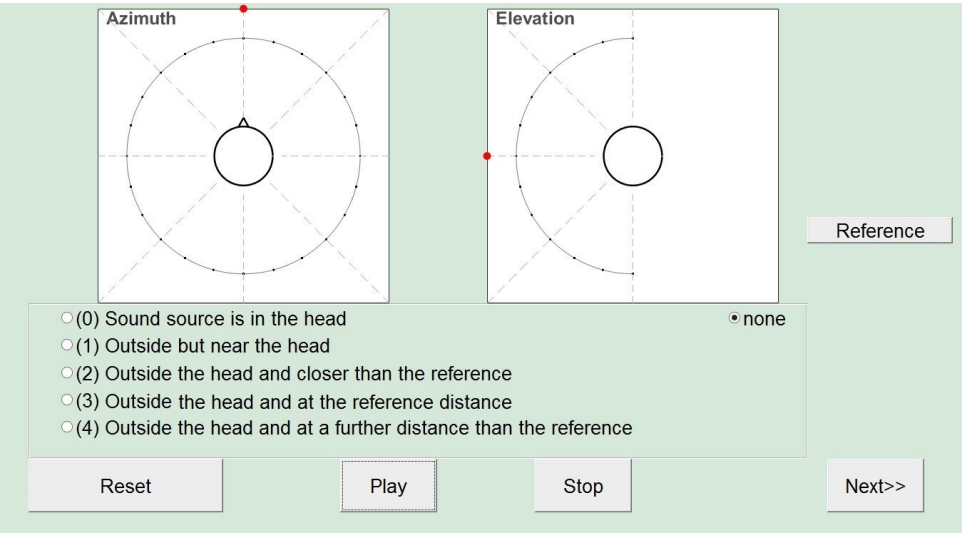

Fig. 6.5: GUI for collecting the responses on source azimuth, elevation and distance. By clicking the "Reference" button, subjects could switch to the signal coming from the reference source in the room, to give the perceived source distance between 0 (inside the head) and 4 (outside the head and at a further distance than the reference). The "Reset" button was active only during TestVR and was used to reset the head tracker.

6.3.1.5 *Subjects and test signal*

A total of 14 (self-reported) normal-hearing subjects took part in TestReal and TestVR. For all of them, individual HRTFs and HPIRs were measured, and the VAH BRIRs V11, V13, V21, and V23, each for 185 head orientations, were prepared as described in section 6.3.1.3. Seven subjects started with TestReal whereas the other seven subjects started with TestVR. For each subject, there was at least a pause of one day between the two tests.

The test signal was a dry recorded speech utterance of 15 s duration, spoken by a female speaker (the same signal as used in chapter 5). This utterance was repeated to a total length of about three minutes to provide the subjects with enough time to give their responses. For TestReal, the test signal was played back from real

sound sources. For TestVR, the test signal was convolved with different BRIRs and presented dynamically over headphones.

6.3.2 *Results*

Figure 6.6 shows response vs. target azimuths and elevations in TestReal (Real Source) and TestVR (V11, V13, V21, V23, HTK). Each circle represents an individual response of a subject. If a pair of target and response azimuths were at the two difference sides of the interaural axis, a front-back reversal was suspected. Responses classified as front-back reversals are indicated with a $\times$ in the upper row of Figure 6.6, where target and response pairs within $\pm 7.5°$ off the interaural axis were not checked for reversals. For one subject, some target and response azimuths were swapped in the front-back as well as in the left-right directions. These cases, indicated with a $\diamondsuit$ in Figure 6.6, were suspected to be caused either by wearing the headphones inversely (left and right ears switched) or not resetting the tracker, and were therefore classified as invalid and discarded from further analysis.

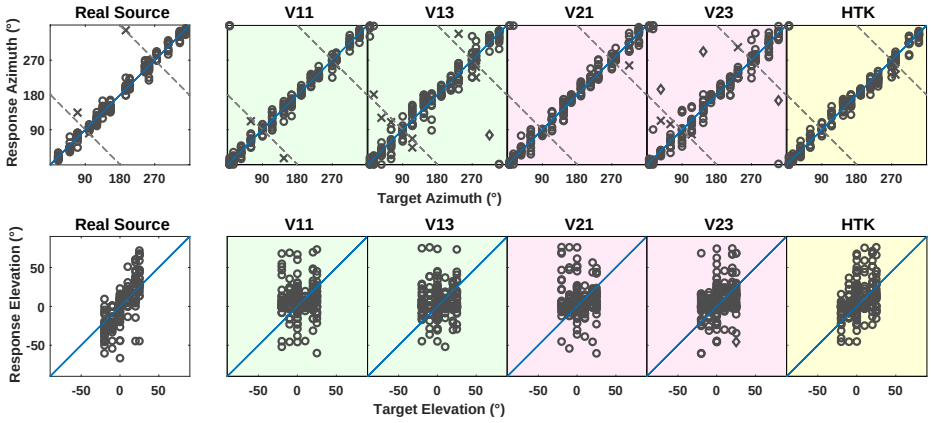

Fig. 6.6: **Top:** Response azimuth (ordinate) vs. target azimuth (abscissa), **Bottom**: Response elevation (ordinate) vs. target elevation (abscissa), when listening to real sources in TestReal (Real Source) as well as to virtual sources in TestVR (generated with V11, V13, V21, V23 and HTK). The circles represent the responses of each of the 14 subjects. Responses marked with a $\diamondsuit$ indicate invalid localizations. Responses classified as front-back reversals are marked with a $\times$ in the top row. Dashed lines represent possible subject responses in case of a perfect front-back confusion.

Azimuth: The azimuth error was calculated as the absolute difference between target and response azimuths. Front-back reversals were excluded from the error calculation. Figure 6.7a shows the azimuth error averaged over 14 subjects and 15 target positions, for different conditions of listening to real and virtual sources. With real sources in TestReal, the average absolute error was 8°. In TestVR, a

comparable average azimuth error as with real sources was achieved with V11 (8.3°), while larger average azimuth errors occurred with HTK (9.9°) and with V21 (10.1°), followed by V23 (11.7°) and V13 (13.3°). VAH syntheses including horizontal and non-horizontal directions in the calculation of the spectral weights (V13 and V23) led to larger average azimuth errors compared to real sources and VAH syntheses including only horizontal directions (V11 and V21). According to the Shapiro-Wilk test of normality, the azimuth error could be assumed to be normally distributed. Accordingly, a one-way repeated-measures ANOVA was performed, which revealed a significant difference in the azimuth error when localizing real or different virtual sources ($F(5,65)=10.1$, $p<0.001$). The post-hoc multiple comparisons with Bonferroni correction ($p<0.05$) indicated significantly higher average azimuth errors for V13 compared to V11 and real sources.

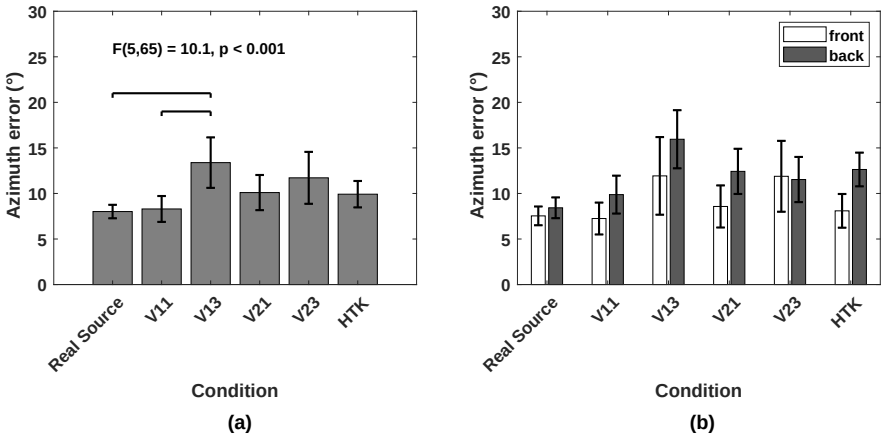

Fig. 6.7: **(a):** Azimuth error, averaged over 14 subjects and all target sources, when localizing real sources in TestReal (Real Source) and virtual sources in TestVR (generated with V11, V13, V21, V23 and HTK). Horizontal bars indicate significant differences (post-hoc multiple comparisons with Bonferroni correction, $p<0.05$). **(b):** Azimuth error, averaged over 14 subjects for target sources grouped into front and back. All error bars indicate 95% confidence intervals.

In Figure 6.7b, the average azimuth error over 14 subjects is shown separately for target positions grouped into front and back. Remember that virtual sources in TestVR were rendered for horizontal head orientations restricted to the frontal range ($-90° \leq \theta_h \leq +90°$) and in TestReal, subjects were asked not to move their heads beyond this range. With the exception of V23, azimuth errors were lower for sources in front than in back. For target sources in the frontal hemisphere, subjects could rotate their heads towards the target source to a region, where the interaural differences and the minimum audible angle were the smallest. When facing the sound source directly, subjects could give a more accurate azimuth response than

target sources in the back.

Elevation: The elevation error was calculated as the absolute difference between target and response elevations, separately for negative (<0°, **N**), zero (=0°, **Z**) and positive (>0°, **P**) target elevations. The lower part of Figure 6.8 shows the elevation error averaged over 14 subjects for different conditions of listening to real and virtual sources. According to the Shapiro-Wilk test of normality, the calculated elevation errors could not be assumed to be normally distributed. Therefore, the Friedman test was applied, separately to each of the negative, zero and positive target elevations. According to the Friedman test, there were significant differences in the average elevation error when localizing real or different virtual sources for negative ($p<10^{-4}$) and zero (p=0.01) target elevations. The multiple comparisons after Friedman test (function friedmanmc in the statistical software R [178]) revealed significantly different average elevation errors between Real Source and V11 as well as between Real Source and V21 (for negative target elevations), and between Real Source and V23 (for zero target elevations). For positive target elevations, the difference between different conditions (real source or different virtual sources) was not significant.

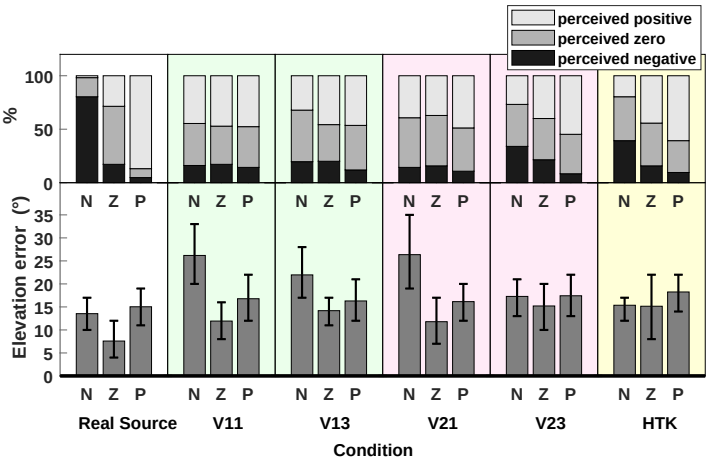

Fig. 6.8: Average elevation error, when localizing real sources in TestReal (Real Source) and virtual sources in TestVR (generated with V11,V13, V21, V23 and HTK). **Bottom:** Absolute error, averaged over 14 subjects for negative (N), zero (Z) and positive (P) target elevations. Error bars indicate 95% confidence intervals. **Top:** Percentage of response elevations, which were perceived negative (below −5°), zero (between −5° and +5°) or positive (above +5°).

To offer some information about the sign of the response elevations, the upper part of Figure 6.8 shows, for each average elevation error split in N, Z and P, the percentage of response elevations which were positive, zero or negative. To calculate these percentages, response elevations were classified as positive or negative, if they

were above $+5°$ or below $-5°$, respectively, and they were classified as zero, if they were between $-5°$ and $+5°$.

With real sources, the sign of response elevations was in good agreement with the sign of target elevations, i.e. the majority of negative, zero and positive real target sources were perceived correctly as negative, zero and positive, respectively. Nevertheless, it should be noted that, as shown in the lower part of Figure 6.6, the response elevations for TestReal extended from below $-50°$ to above $+70°$, although the target elevations varied between $-20°$ and $+25°$ only. Indeed, subjects tended to underestimate negative elevations and overestimate positive elevations. This was presumably caused by response mapping to the GUI, which included an additional step of translating the perceived elevation into the vertical angle difference to the horizontal plane, which apparently could not be performed correctly by the subjects.

With virtual sources on the other hand, the accordance of the signs between target and response elevations could be observed only in a weak form for V23 and HTK; for V11, V13 and V21, target elevations were perceived as zero or positive most of the time, regardless of the sign of the target elevation. With V11 and V21, (i.e. when only horizontal directions were included in the calculation of the spectral weights), one could expect at least the zero target elevations to be perceived correctly. However, even with V11 and V21, the zero target elevations were perceived positive as often as zero. Also with HTK, a large percentage of zero target elevations were perceived positive.

Externalization rate: The responses given for the source distance were divided into two groups: "not externalized" (scores 0 and 1) and "externalized" (scores 2, 3, and 4). Figure 6.9 shows the externalization rate, defined as the percentage of responses classified as externalized, over target azimuths.

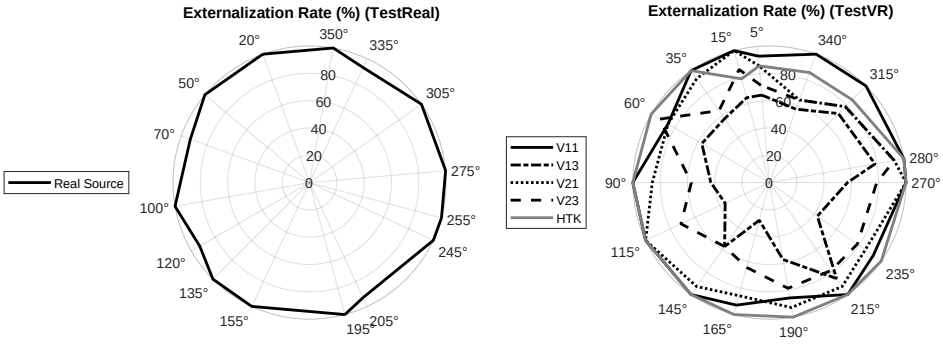

Fig. 6.9: Externalization rate, defined as the percentage of responses classified as externalized, when localizing real sources in TestReal (Real Source) and virtual sources in TestVR (V11, V13, V21, V23 and HTK).

As the polar diagrams in Figure 6.9 show, real sources were almost always externalized, which is an expected result. Only one subject perceived the real target sources at 70°, 120°, 205° and 335° azimuth outside but near the head, i.e. not enough externalized, which could be due to the extraordinary listening situation (darkened room and missing visual information). Externalization rates for virtual sources generated with V11, V21 and HTK were comparable to externalization rates for real sources, whereas for virtual sources generated with V13 and V23, externalization rates were markedly lower. Figure 6.10 shows the externalization rates averaged over target positions. According to the Shapiro-Wilk test of normality, the average externalization rates could not be assumed to be normally distributed. Therefore, the Friedman test was applied, which revealed significant differences in the average externalization rates when listening to real sources or different virtual sources ($p<10^{-4}$). According to the multiple comparisons after Friedman test, virtual sources generated with V13 and V23 were significantly less externalized than other sources. There were no significant differences between externalization rates of virtual sources generated with V11, V21, HTK and the real source.

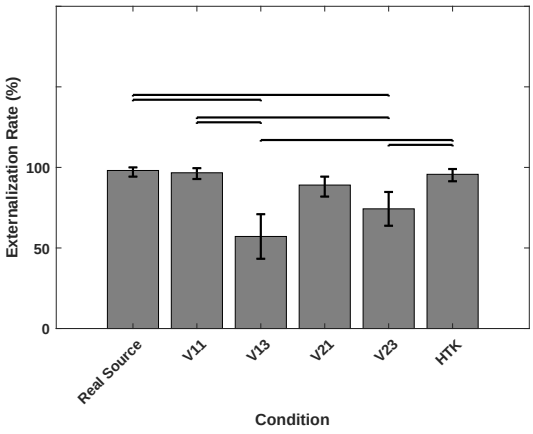

Fig. 6.10: Externalization rate averaged over target positions, when listening to real sources in TestReal (Real Source) and to virtual sources in TestVR (generated with V11, V13, V21, V23 and HTK). Horizontal bars indicate significant differences (multiple comparison after Friedman test, $p<0.05$). Error bars indicate 95% confidence intervals

Reversal rates: Table 6.2 shows the reversal rates, defined as the percentage of responses classified as front-back reversals, for different conditions of listening to real and virtual sources. With real sources, this rate was equal to 1.4%. With virtual sources generated with V11, V21 and HTK, reversal rates were smaller or comparable to real sources (between 0.47% and 1.4%), while reversal rates were slightly higher with V23 (1.9%) and much higher with V13 (4.2%). Reversals are known

to occur very often when listening to virtual sources which are presented statically, i.e. without head tracking [96, 192]. However, when dynamically presenting signals as in TestVR, head movements lead to changes in the interaural differences, which provide important cues to distinguish between sources in front and back. Especially, the low-frequency ITDs provide important dynamic cues when listening to speech signals [9, 193].

Table 6.2: Reversal rate when localizing real sources in TestReal (Real Source) and virtual sources in TestVR (generated with V11, V13, V21, V23 and HTK).

Presented source (real or virtual)	Real Source	V11	V13	V21	V23	HTK
Reversal rate	1.4%	1.4%	4.2%	0.47%	1.9%	0.95%

6.3.3 *Discussion*

According to the results, average azimuth errors and externalization rates were similar with real sources and with virtual sources generated with V11 and V21. In chapter 5 as well as in [92], where a good perceptual performance of the VAH could be observed, not only could subjects see the sound source, but also the experiment task was different. In chapter 5 and in [92], a subjective rating was given for the match between the position of real and virtual sources rather than giving direct judgements about the perceived azimuth and elevation. Nevertheless, the results in this chapter showed that also in the absence of visual cues and for a more challenging localization task, it is possible to generate virtual sources with a VAH with comparable externalization and azimuth localization performance as real sound sources. In line with the results in chapter 5, VAH BRIRs synthesized with spectral weights with only horizontal directions included (V11 and V21) performed better than the syntheses with horizontal and non-horizontal directions included (V13 and V23). As already discussed in section 6.2, the inclusion of horizontal and non-horizontal directions introduced increased spectral and temporal distortions, which impacted the localization performance negatively. In addition, according to [83], both spectral as well as binaural cues should be preserved in the binaural synthesis in order to externalize the headphone signals. The degradation in binaural cues, due to temporal and spectral distortions, led to the lack of externalization with V13 and V23. In addition, the deteriorated low-frequency ITDs could also explain the higher reversal rates with V13 and V23, despite a dynamic presentation with head tracking.

According to the results, azimuth accuracy, externalization and reversal rates were very convincing for virtual sources generated with BRIRs of the KEMAR artificial head (HTK), despite the non-individuality of these BRIRs. Listening to non-individual recordings can degrade the externalization in static scenarios without head tracking [7, 98], whereas it has been shown in previous studies that when head movements are enabled, externalization is improved when listening to

non-individual binaural signals [98, 119]. Similarly, front-back reversals have been observed to occur more frequently with non-individual binaural signals, however in static scenarios [7, 95, 98]. In contrast, it has been shown that with dynamic signal presentations, reversal rates can be reduced, regardless of listening to individual or non-individual binaural signals [12, 98]. The fact that the non-individual signals of HTK could be presented dynamically was advantageous in improving externalization and reducing the reversals. In general, the results confirm the results in chapter 5 as well as the results in [92], indicating that for speech signals and with head tracking, individual binaural signals do not constitute a major advantage over non-individual signals of a conventional artificial head. However, the effort to acquire the non-individual BRIRs of the artificial head for different head orientations should not be neglected.

Elevation perception with virtual sources was in general not very convincing in comparison to real sources. When listening to virtual sources, subjects seemed to have had difficulty to deal with the task of vertical localization. This will be discussed more thoroughly in section 6.5.

6.4 Part II: The impact of head tracking on the localization performance

Although in the first part head tracking was included, in the second part we wanted to explicitly assess the impact of head tracking by comparing the localization performance of virtual sources with two listening tests, either with head tracking (referred to as **TestDynamic**) or without head tracking (referred to as **TestStatic**). The same VAHs as in part I (shown in Figure 6.1) were used to synthesize individual BRIRs with spectral weights calculated with the parameters listed in Table 6.1. The same 14 subjects, who participated in TestReal and TestVR in part I, took part in the new listening tests. Therefore, the individually calculated spectral weights for 185 head orientations of all subjects were already available. Since the BRIRs synthesized with V13 in part I performed worse than other VAH BRIRs with respect to azimuth accuracy and externalization rates, V13 was not considered in the experiments in part II.

For both listening tests, individually synthesized BRIRs with V11, V21 and V23, each for 185 head orientations, as well as KEMAR BRIRs acquired for 37 horizontal head orientations, were considered. Measurements were performed inside the acoustically transparent tent in the anechoic room, in the same way as for TestVR described in section 6.3.1.3, however, for another set of 15 target source positions (see Figure 6.3c). The BRIRs were the same in TestDynamic and TestStatic; the difference between both tests concerned solely the presentation method, i.e. with or without head tracking. In TestDynamic, the virtual sources were presented according to the listener's head movements using the BRIRs corresponding to different head orientations, whereas in TestStatic, the virtual sources were presented only

using the BRIRs corresponding to the frontal head orientation, regardless of head movements. In order to distinguish between the BRIRs in both tests, for TestStatic, BRIRs were assigned with the subscript $\{\}_s$ ($\mathbf{V11_s}$, $\mathbf{V21_s}$, $\mathbf{V23_s}$, $\mathbf{HTK_s}$).

6.4.1 *Experiment design*

6.4.1.1 *Response method: GUI*

The localization task in TestDynamic and TestStatic consisted of reporting the perceived azimuth, elevation and distance of the virtual sources. The GUI used in Part I, shown in Figure 6.5, was used with a few modifications. Since the listening tests were designed to have only virtual presentations, no reference signal was provided for the perceived source distance. Therefore, the button "Reference" was omitted from the GUI. Subjects had to judge the perceived distance solely with respect to their own body as a reference, using an ordinal scale ranging from 0 to 3, corresponding to (0) in the head, (1) outside but near the head, (2) outside the head and within reach, or (3) outside the head and further away. The frontal direction of azimuth=0° and elevation=0°, corresponding to the frontal head orientation, was marked in front of the subjects in the room. In TestDynamic, subjects were asked to orient their head towards this point in the room and click the "Reset" button on the GUI to reset the head tracker before listening to the virtual sources. In TestStatic, the "Reset" button was omitted from the GUI.

6.4.1.2 *Experiment setup, subjects and test signal*

Both listening tests, TestDynamic and TestStatic, were performed in the control room of the recording studio at the Jade University of Applied Sciences (4.7m×5.1m×3m, average reverberation time: 0.35s). The choice of different recording and listening environments was in line with practical applications, since in general it is not always straightforward to listen to binaural signals in the same environment as where they were captured, e.g., signals recorded in a moving vehicle or during a concert. During TestDynamic and TestStatic, the room was normally illuminated and subjects could see the room and the objects in it (desk, mixing console, loudspeakers, etc.). Subjects were informed that there was no association between the virtual target sources and any objects in the room. Subjects were in addition informed that the signals had been captured in an anechoic environment, and that they should imagine themselves in an anechoic room to judge the distance of the presented sources. The GUI was presented to the subjects via a laptop, connected to the audio interface and headphone amplifier. In each test, each of the 15 target positions was presented four times, i.e. once with each of the four BRIRs (V11, V21, V23 and HTK in TestDynamic and $\mathrm{V11_s}$, $\mathrm{V21_s}$, $\mathrm{V23_s}$ and $\mathrm{HTK_s}$ in TestStatic). This resulted in a total of 60 virtual sources for each test, presented in a randomized order. Prior to each test, five of these 60 virtual sources were chosen randomly to be presented for familiarization and were discarded from the evaluations. No feedback was given to the subjects during the familiarization as

well as during the listening test.

The same 14 subjects, who participated in TestReal and TestVR, took part. Seven subjects started with TestDynamic whereas the other seven subjects started with TestStatic. For each subject, there was a pause of at least one day between both tests. During TestDynamic, subjects were encouraged to move their heads within the allowable range of $\pm 90°$ in horizontal and $\pm 15°$ in vertical directions. During TestStatic, subjects were made aware that head tracking was switched off. The test signal was the same speech signal spoken by a female speaker as in TestReal and TestVR in part I.

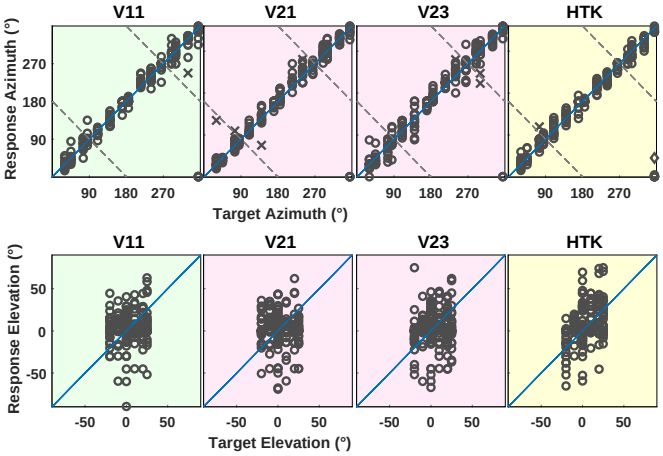

Fig. 6.11: **Top:** Response azimuth (ordinate) vs. target azimuth (abscissa), **Bottom:** Response elevation (ordinate) vs. target elevation (abscissa) when listening to virtual sources generated with V11, V21, V23 and HTK in TestDynamic. Responses marked with a $\times$ indicate front-back reversals and the response marked with a $\diamondsuit$ indicates an invalid localization. Dashed lines represent possible subject responses in case of a perfect front-back confusion.

6.4.2 *Results*

Figures 6.11 and 6.12 show response vs. target azimuths and elevations of 14 subjects in TestDynamic and Teststatic, respectively. Responses marked with a $\times$ indicate front-back reversals and the response marked with a $\diamondsuit$ for HTK indicates an invalid localization (for the same subject as in part I) and was excluded from further analysis.

Azimuth: Figure 6.13a shows the azimuth error averaged over 14 subjects and 15 target source positions for different BRIRs in TestDynamic and TestStatic. Front-back reversals were excluded from the error calculation. According to

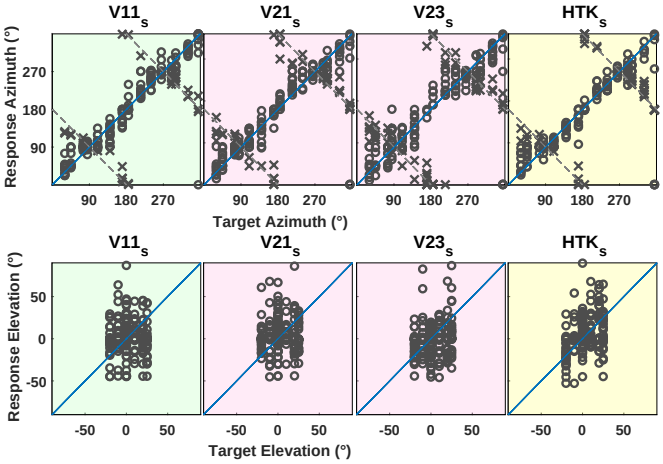

Fig. 6.12: **Top:** Response azimuth (ordinate) vs. target azimuth (abscissa), **Bottom:** Response elevation (ordinate) vs. target elevation (abscissa) when listening to virtual sources generated with $V11_s$, $V21_s$, $V23_s$ and HTK_s in TestStatic. Responses marked with a $\times$ indicate front-back reversals. Dashed lines represent possible subject responses in case of a perfect front-back confusion.

the Shapiro-Wilk test of normality, the azimuth error could be assumed to be normally distributed. Therefore, a paired t-test was applied, which revealed that for all BRIRs, the average azimuth error was significantly higher in TestStatic compared to TestDynamic (t-values shown in Figure 6.13a). In Figure 6.13b, the average azimuth error is shown over 14 subjects separately for target positions grouped into front and back. For target sources both in front and in back, the average azimuth error was higher in TestStatic compared to TestDynamic. Although in TestDynamic head movements were limited only to orientations within the frontal hemisphere, head movements in this limited range were advantageous for localization accuracy of virtual sources both in front and in back. Since front-back reversals were excluded from the azimuth error calculation, the results indicate that also for cases where no front-back reversals occurred, the azimuth error was smaller with head movements than without head movements.

A significant effect of BRIRs on the average azimuth error was observed only in TestStatic ($F(3,39)=9.5$, $p<0.001$), with significantly higher azimuth errors for $V23_s$ compared to $V11_s$, $V21_s$ and HTK_s ($p<0.05$, Bonferroni correction). In TestDynamic, there were no significant differences between average azimuth errors of VAH or HTK BRIRs. While the syntheses with V23 and $V23_s$ were both subject to large spectral and temporal distortions due to the inclusion of horizontal and non-horizontal directions in the calculation of the spectral weights, these distortions seem to be less critical in TestDynamic compared to TestStatic.

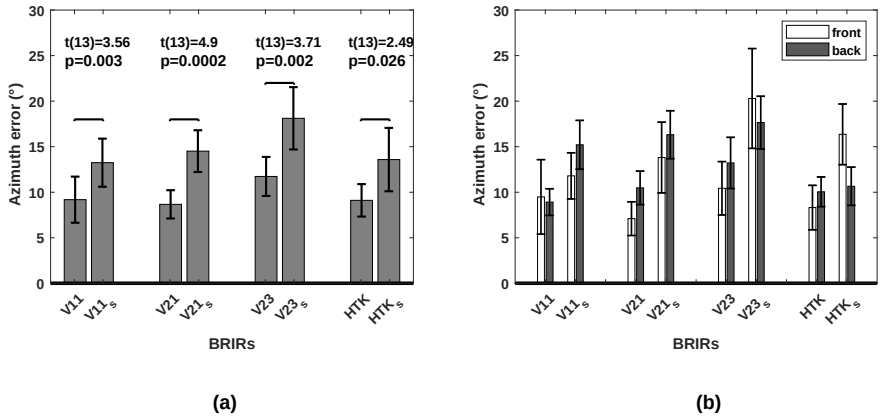

Fig. 6.13: **(a):** Azimuth error, averaged over 14 subjects and all target sources when localizing virtual sources generated with V11, V21, V23 and HTK in TestDynamic and with $V11_s$, $V21_s$, $V23_s$ and HTK_s in TestStatic. Horizontal bars indicate significant differences (according to paired t-test). **(b):** Azimuth error in TestDynamic and TestStatic, averaged over 14 subjects for target sources grouped into front and back. All error bars indicate 95% confidence intervals.

Elevation: The lower part of Figure 6.14 shows the elevation error averaged over 14 subjects in both tests, separately for negative, zero and positive target elevations. The upper part of Figure 6.14 shows the percentage of response elevations, which were positive, zero or negative, calculated in a similar way as in section 6.3.2. According to the Shapiro-Wilk test of normality, the average elevation errors could not be assumed to be normally distributed. Therefore, the Wilcoxon signed-rank test was applied separately to the negative, zero and positive target elevations, which revealed no significant differences in the average elevation error in TestDynamic compared to TestStatic for any groups of positive, zero and negative target elevations and for any of the BRIRs. This means that there was apparently no effect of head movements on the vertical localization accuracy. The accordance of the signs between target and response elevations could be observed, though in a weak form, only for HTK and HTK_s and for V23 and $V23_s$. To compare the average elevation error within TestDynamic and TestStatic, the Friedman test was applied, which revealed no significant effect of BRIRs.

Externalization rate: Responses given for the source distance were divided into two groups: "not externalized" (scores 0 and 1) and "externalized" (scores 2 and 3). Figure 6.15 shows the externalization rate over the target azimuths.

The polar diagrams in Figure 6.15 show that externalization rates were in general lower in TestStatic than in TestDynamic. Externalization rates in TestStatic were considerably lower for sources close to the median plane than for lateral sources,

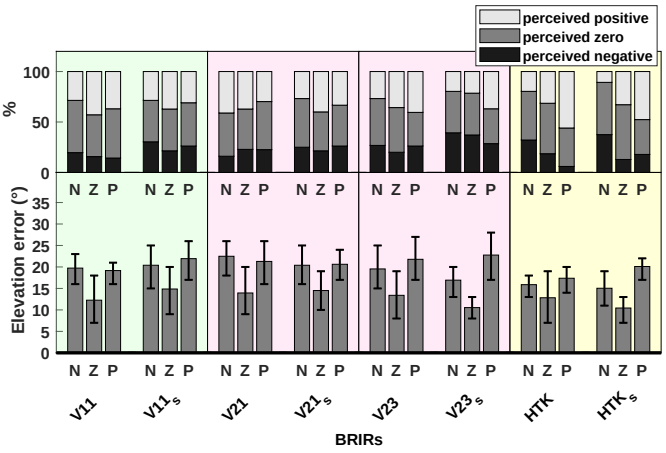

Fig. 6.14: Average elevation error, when localizing virtual sources generated with V11, V21, V23 and HTK in TestDynamic and with $V11_s$, $V21_s$, $V23_s$ and HTK_s in TestStatic. **Bottom:** Absolute error, averaged over 14 subjects for negative (N), zero (Z) and positive (P) target elevations. Error bars indicate 95% confidence intervals. **Top:** The percentage of response elevations, which were perceived negative (below $-5°$), zero (between $-5°$ and $+5°$) or positive (above $+5°$).

which is a known phenomenon also reported in previous studies [13,87]. Figure 6.16 shows the externalization rates averaged over the target positions. According to the Shapiro-Wilk test of normality, the average externalization rates could not be assumed to be normally distributed. Therefore, the Wilcoxon signed-rank test was applied, which revealed that for all BRIRs, average externalization rates were higher in TestDynamic than in TestStatic (p-values shown in Figure 6.16). The results confirm the positive impact of head tracking on the externalization of virtual sources presented over headphones [13, 14, 98]. To compare the average externalization rates within TestDynamic and TestStatic, the Friedman test was applied. In TestStatic, there was no significant effect of the BRIRs. In TestDynamic, there was a significant effect of the BRIRs ($p=0.001$), with a significant difference between the externalization rates of VAH BRIRs V11 and V23 (multiple comparisons after Friedman test, $p<0.05$).

Reversal rates: Table 6.3 shows the reversal rates in both tests. When localizing virtual sources generated with $V11_s$, $V21_s$, $V23_s$ and HTK_s in TestStatic, reversal rates were considerably larger compared to localizing virtual sources generated with V11, V21, V23 and HTK in TestDynamic. This result clearly confirmed the positive impact of head movements on the reduction of front-back reversals [12, 98, 194].

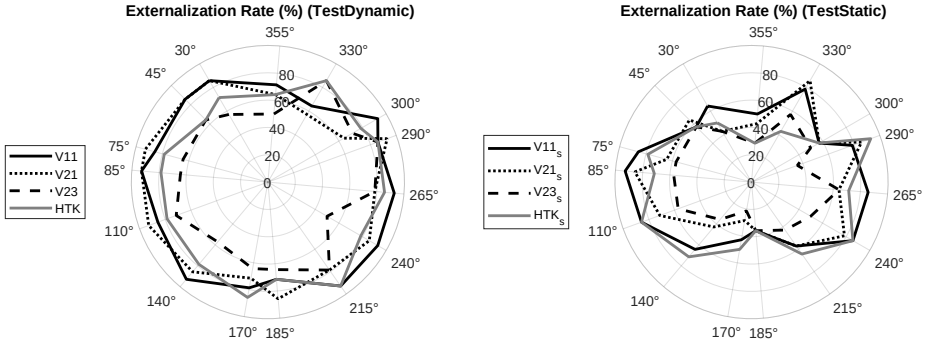

Fig. 6.15: Externalization rate shown over target azimuths, when localizing virtual sources generated with V11, V21, V23 and HTK in TestDynamic (left) and with $V11_s$, $V21_s$, $V23_s$ and HTK_s in TestStatic (right).

Table 6.3: Reversal rate when localizing virtual sources generated with V11, V21, V23 and HTK in TestDynamic and with $V11_s$, $V21_s$, $V23_s$ and HTK_s in TestStatic.

Presented virtual source	V11	V21	V23	HTK		$V11_s$	$V21_s$	$V23_s$	HTK_s
Reversal rate	0.47%	1.90%	1.42%	0.47%		15.24%	23.33%	30%	20%

6.4.3 *Discussion*

In line with previous studies [12–14, 98, 194], the dynamic presentation in TestDynamic improved the localization performance of virtual sources with respect to azimuth, externalization and the reduction of front-back reversals compared to TestStatic. Horizontal head movements limited to the frontal hemisphere were advantageous also for the localization accuracy of virtual sources located in the back. Dynamic presentation was also advantageous for VAH BRIRs with relatively large synthesis errors; while in TestStatic the average azimuth error with $V23_s$ was significantly higher than with other BRIRs, in TestDynamic (and similarly in TestVR in Part I) the difference between V23 and other VAH BRIRs was not significant.

No effect of head movements on the vertical localization accuracy could be observed. Some previous studies reported the positive impact of horizontal head movements on vertical localization. Such an advantage, however, was only found for sources limited to the median plane or the left lateral plane [9, 11] or only for low-frequency noise signals [10]. In the current study, target sources were distributed all around the listener and a more broadband signal (speech) was used. Therefore, the results were more in line with localization studies using speech signals [12] or a wider range of source positions [194], which also reported no specific effect of head movements

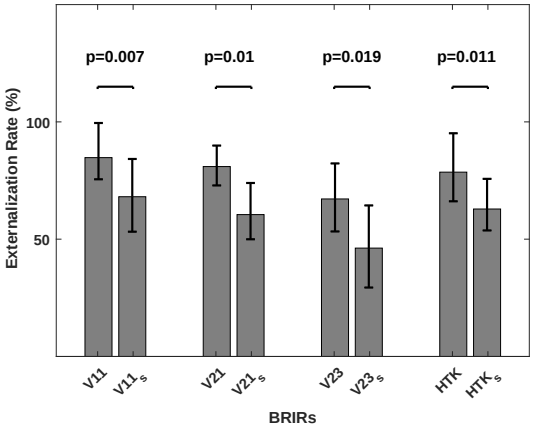

Fig. 6.16: Externalization rate, averaged over target positions, when listening to virtual sources generated with V11, V21, V23 and HTK in TestDynamic and with $V11_s$, $V21_s$, $V23_s$ and HTK_s in TestStatic. Horizontal bars indicate significant differences (according to the Wilcoxon signed-rank test). Error bars indicate 95% confidence intervals.

on the elevation error. The unconvincing performance in vertical localization in TestDynamic and TestStatic was similar to TestVR in part I and suggested that independent of the presentation method, subjects had difficulty to deal with the task of vertical localization.

Apart from generally poor elevation results, the localization performance of virtual sources generated with the non-individual BRIRs (HTK or HTK_s) was comparable to the localization performance of virtual sources generated with the individually synthesized VAH BRIRs (V11 and V21 or $V11_s$ and $V21_s$). Analyzing the results of TestDynamic and TestStatic separately, the individually synthesized VAH BRIRs offered no special advantage over the non-individual KEMAR BRIRs. It should however be kept in mind that the dynamic presentation of KEMAR BRIRs was artificially enabled in this study. In standard applications of conventional artificial heads, dynamic presentation of binaural recordings is not possible; in contrast, a VAH straightforwardly allows to dynamically present recordings. Therefore, from a practical point of view, one should compare the localization performance of VAH BRIRs in TestDynamic with KEMAR BRIRs in TestStatic. Not only the reversal rate was higher for HTK_s in TestStatic (20%) compared to VAH BRIRs in TestDynamic (below 2%), but also the average azimuth error for HTK_s was significantly higher than for V11 ($t(13)=3.21$, $p=0.006$) and V21 ($t(13)=2.94$, $p=0.014$). Furthermore, according to the Wilcoxon signed-rank test, the average externalization rate for HTK_s was significantly lower compared to V11 ($p=0.002$) and V21 ($p=0.034$). This analysis indicated that besides individualization, an

important advantage of the VAH approach over conventional artificial heads is the dynamic presentation of recorded signals; a feature shown in this study to largely improve the localization performance of virtual sources.

6.5 Discussions concerning part I and part II

This section provides some general discussion which applies to both parts of the study.

It has been shown in [195] that the response method used for localization experiments can impact the accuracy of the results. In this chapter, using the same GUI for both listening tests was a suitable method to verify the used response method, which provided a means to identify differences between real and virtual sources or between different presentation methods of virtual sources.

In both tests in which head tracking was applied, i.e. TestVR and TestDynamic, the average azimuth and elevation errors and the reversal rates were in general in a comparable range. However, average externalization rates dropped consistently for all BRIRs in TestDynamic compared to TestVR, with mean values changing from 96.6% to 84.7% (V11), from 89% to 80.9% (V21), from 74.3% to 67.1% (V23) and from 95.7% to 78.5% (HTK). In both tests, the sound source was classified as not externalized, if it was perceived in the head or outside but near the head, independent of the reference source which was absent in TestDynamic. The difference in the externalization rates of both tests was therefore suspected to be due to the presentation rooms. For both tests, BRIRs were acquired in the anechoic room, however, the presentation rooms were different (anechoic room in TestVR vs. control room in TestDynamic). The reduction of externalization rates in TestDynamic was suspected to be due to the acoustical incongruency between recording and presentation rooms, which is known to reduce the externalization of virtual sources [89, 90].

With respect to the correct perception of the target elevation sign, the performance for virtual sources was poorer than for real sources. On average, virtual target sources both in and outside the horizontal plane were often perceived higher than they were. The tendency to perceive virtual sources with a positive bias was reported in some previous studies, however using non-individual HRTFs [97, 196, 197]. The missing individual features in BRIRs could not have been the only reason for the elevation bias in the present study, since otherwise this phenomenon should not have occurred for horizontal virtual sources generated with the VAHs including only horizontal directions in the calculation of the spectral weights (V11 and V21), because the monaural spectral and temporal features of individual horizontal BRIRs could be preserved well with these syntheses.

However, the elevation bias observed in the present study was an average tendency and did not apply to all subjects. One subject for instance perceived over 60% of all virtual target elevations negative, whereas two other subjects mostly gave elevation responses zero, independent of target elevation or the BRIR. The pattern of elevation perception was not consistent among all subjects. The clear trend is that the vertical localization was more difficult with virtual than with real sources. One can argue that when listening to speech signals, the main elevation cues related to the pinna reflections (above about 5 kHz [24, 72]) cannot be presented well. However, the vertical localization performed quite convincingly with the same speech signal when listening to real sources. It is not clear whether the poor vertical localization with virtual sources was solely related to the synthesis inaccuracy of VAH BRIRs and the non-individuality of KEMAR BRIRs, or whether also factors related to the rendering process played a role. The localization experiment should be repeated with individually measured BRIRs in order to provide more details with this respect.

Despite the generally non-convincing elevation perception with the VAH BRIRs, the target elevation sign could be better perceived with VAH2 with horizontal and non-horizontal directions included (V23 in TestVR) compared to other VAH BRIRs. This was however not the case for VAH1 with horizontal and non-horizontal directions included (V13 in TestVR). The three-dimensional topology of VAH2 was suspected to have contributed to the better elevation perception compared to VAH1. It seemed that the sound incidence from different elevations could be better recognized with microphones having different positions along the z-axis than with microphones distributed at the same height. The additional seven microphones of VAH2 compared to VAH1 could also have been advantageous. Further investigations with VAH1 and VAH2 or other array topologies are required to confirm this statement.

An important point to be mentioned is that the results achieved in both parts of the study in this chapter were obtained using speech signals. In general, the synthesis accuracy of the VAHs decreases towards higher frequencies (see section 5.2.2 for VAH1 and section 6.2 for VAH2). In addition, in case of measured BRIRs of a conventional artificial head, the non-individuality gets more prominent at higher frequencies. In applications with signals with a more pronounced high-frequency spectral content, the localization performance with synthesized VAH BRIRs or non-individual BRIRs of a conventional artificial head could be different from the current study.

6.6 Summary

In this chapter, the localization performance when dynamically presenting virtual sources generated with a Virtual Artificial Head (VAH) was evaluated, where both the difference with real sources and the impact of head tracking were investigated. Two different VAHs were used to capture the room impulse responses for different

three-dimensional source positions in an anechoic room. Spectral weights calculated for 185 head orientations of individual listeners were applied to the measured room impulse responses to result in individually synthesized BRIRs, which were convolved with the test signal (speech) and presented over headphones.

In comparison to the localization performance of real sound sources, similar localization accuracies could be achieved with the VAHs in azimuth, externalization and number of front-back reversals. The vertical localization performance of virtual sources generated with the VAHs was not as convincing as the vertical localization performance of real sources. The localization experiments with and without head tracking confirmed the importance of dynamic presentation for the localization accuracy when listening to virtual sources generated with the VAHs.

In both parts of the study in this chapter, signals generated with non-individual BRIRs measured for a conventional artificial head were also evaluated, for which the dynamic presentation was artificially enabled by laborious measurement of BRIRs for different head orientations. When dynamically presented, similar localization performances could be achieved with the non-individual BRIRs of the artificial head as with individually synthesized VAH BRIRs. However, in practical applications, binaural recordings with conventional artificial heads cannot be presented dynamically. Hence, it can be concluded that the possibility of presenting binaural signals recorded with a VAH dynamically is the major advantage of virtual over the conventional artificial heads.

7

SUMMARY, CONCLUSION AND FURTHER RESEARCH

This chapter provides a summary of the main contributions of the thesis and discusses possible directions for further research.

7.1 Summary and conclusion

The main objective of this thesis was to improve and further investigate the Virtual Artificial Head (VAH) approach as developed by Rasumow et al. [15–17]. A VAH is a microphone array which employs filter-and-sum beamforming to synthesize individual HRTF directivity patterns. This is done by applying complex-valued spectral weights to the microphone signals and adding them. In comparison to conventional artificial heads, the VAH approach not only offers the possibility to adjust to individual HRTFs by using individually calculated spectral weights, but also to adjust to different head orientations during playback. For binaural recordings with conventional artificial heads on the other hand, the recording can be presented only for a fixed head orientation of the listener. The general concept is based on calculating individual spectral weights by minimizing a least-squares cost function subject to a constraint on the mean White Noise Gain (WNG), which is applied to increase the robustness and to limit amplification of microphones self-noise. The least-squares cost function is defined as the sum over calibration directions of the squared absolute deviations between the desired and synthesized HRTF directivity patterns and is minimized independently for each frequency bin. The desired directivity pattern to be synthesized is defined for specific discrete directions. Although the synthesized directivity pattern of a VAH implicitly interpolates between these directions, the spatial resolution, i.e. the resolution of directions for which the synthesis can be considered acceptable depends on these discrete directions. Considering a larger number of directions for the calculation of the spectral weights improves the spatial resolution only if the number of microphones is increased as well. This thesis aimed at improving the horizontal spatial resolution of a VAH without increasing the number of microphones. In addition, the impact of array topology on the VAH performance was investigated. Moreover, the VAH approach was evaluated in acoustical scenarios which were not considered before, namely for dynamic auralizations by

including head tracking during listening, considering horizontal and non-horizontal directions and assessing the performance in anechoic and reverberant environments.

In **chapter 3** we proposed to improve the spatial resolution of the VAH by including more directions in the calculation of the spectral weights and imposing additional constraints on the monaural spectral error at these directions. The monaural spectral error, referred to as Spectral Distortion (SD) in this thesis, was defined as the spectral difference in dB between the desired and the synthesized HRTFs. As upper and lower boundaries for the SD, we used 0.5 dB and -1.5 dB, respectively, in order to keep the deviation between the Interaural Level Differences (ILDs) of the synthesized and desired HRTFs below 2 dB for all considered discrete directions. With the proposed additional SD constraints and for a simulated planar microphone array with 24 microphones, it was shown that the SD could be kept between the chosen upper and lower boundaries for 72 horizontal directions (i.e. 5° resolution) up to around 5 kHz at contralateral directions and higher than 5 kHz at ipsilateral directions. This frequency range was higher than the range for which an acceptable synthesis could be achieved without applying the SD constraints. In addition, synthesized HRTFs applying the SD constraints outperformed the synthesized HRTFs without applying SD constraints in perceptual evaluations with respect to coloration, localization and overall quality. The method proposed in chapter 3 was used to calculate individual spectral weights for dynamic auralizations in **chapter 5** and **chapter 6**.

As a next step, the impact of array topology on the VAH performance using the proposed optimization method was investigated. In general, smaller inter-microphone distances help satisfy the SD constraints for an extended frequency range by shifting the spatial aliasing effects to higher frequencies. However, with smaller inter-microphone distances the minimization of the cost function becomes more ill-conditioned, especially at low frequencies. This may make satisfying the mean WNG constraint more difficult. When aiming at higher robustness by setting the minimum desired mean WNG to a higher value, the results showed that the phase accuracy, referred to as Temporal Distortion (TD) in this thesis, increased drastically for arrays with small inter-microphone distances. In order to maintain a good synthesis accuracy at higher frequencies and a good phase accuracy at lower frequencies, it was suggested to use a combination of dense and sparse inter-microphone distances. Objective results based on simulations confirmed that such an array topology enabled to preserve the synthesis accuracy at high and low frequencies while at the same time achieving an appropriate robustness. In addition, perceptual evaluations indicated that the binaural signals generated using the proposed mixed array topology resulted in the best perceptual ratings compared to the other tested microphone arrays, which resulted in either more high-frequency SD or more low-frequency TD.

The next chapters of the thesis dealt with the evaluation of dynamic auralizations using the VAH approach. Spectral weights for different head orientations of the listener were calculated by virtually rotating the VAH prior to calculating

the spectral weights. In **chapter 4**, the overall methods applied for dynamic auralizations including the head tracker device and the employed algorithms for the real-time head-tracked signal playback were presented. In this thesis, we considered 37×5=185 head orientations in the frontal hemisphere corresponding to 37 azimuth angles of −90° to +90° in 5° steps and 5 elevation angles of −15° to +15° in 7.5° steps. Individual BRIRs for these 185 head orientations were synthesized with two VAHs, namely a planar array with 24 microphones and a three-dimensional array with 31 microphones.

In **chapter 5**, the VAH consisting of 24 microphones with a planar array topology as developed by Rasumow et al. [15] was used to auralize a reverberant and an anechoic room, with sound sources both in and outside the horizontal plane. Whereas the upper and lower boundaries for the SD were the same as in chapter 3, two different values for the minimum desired mean WNG were investigated. In addition, the directions included in the calculation of the spectral weights varied, either including 72 horizontal directions as in chapter 3 or 3×72=216 directions from horizontal as well as two non-horizontal planes. Dynamic auralizations with the synthesized BRIRs were evaluated in comparison to real (visible) sound source presentations. Perceptual evaluations with speech signals indicated close-to-reality auralizations with the VAH in both environments. Despite the fact that the virtual sources could be both inside and outside the horizontal plane in the auralizations, using spectral weights including 72 directions from the horizontal plane resulted in better perceptual results than including horizontal and non-horizontal directions. This could be mainly explained by the increased TDs in the synthesis including horizontal and non-horizontal directions, which were more important than the reduced SDs at non-horizontal directions. In addition, by decreasing the minimum desired mean WNG, perceptual ratings degraded, indicating that it is advisable to avoid low resulting mean WNG values. Moreover, comparing the perceptual evaluations in reverberant and anechoic environments showed that the VAH synthesis error was less audible in the presence of reverberation. Interestingly, non-individual dynamic auralizations realized with a conventional artificial head and a simple rigid sphere with two microphones led to a realistic impression of the sources in the auralized environments as well. Dynamic auralization was artificially enabled by BRIRs measured for different head-above-torso orientations of the conventional artificial head and the rigid sphere, which was quite unrealistic and different from the typical applications of an artificial head. The results indicated that if dynamic auralizations are enabled, they do not need to be individualized to be perceptually convincing. The 5° × 7.5° resolution for head orientations was evaluated as sufficient for speech signals and was therefore used in the next chapter again.

In **chapter 6**, the VAH approach was evaluated in localization experiments for dynamic auralizations in an anechoic environment. The two main motivations for this study were to investigate whether the successful performance of the VAH assessed in auralizations in chapter 5 was enhanced by the presence of visual cues during the evaluations and whether head tracking had a possibly positive impact on the perceptual results. Two VAHs, namely the planar array with 24 microphones

of Rasumow et al. [15] and a three-dimensional array with 31 microphones, were used to auralize a virtual source at different positions. In the first part of the study, localization performance when listening to virtual sources generated with the VAHs was compared to localization performance when listening to real sound sources in the same anechoic environment. The localization experiments in the first part of the study in chapter 6 took place in a darkened room, where no visual information was offered to the subjects. Localization experiments with 14 subjects indicated that even in the absence of visual cues, virtual sources generated with the VAHs were localized with a similar accuracy with respect to azimuth, externalization and the occurrence of front-back reversals as real sources. In line with the results of chapter 5, including only horizontal directions in the calculation of the spectral weights led to a better localization performance compared to including horizontal and non-horizontal directions in the calculation of the spectral weights. In the second part of the study, the localization performance when listening to virtual sources was assessed in two separate listening tests, one with and one without head tracking. The results confirmed the importance of dynamic auralizations for the localization accuracy of virtual sources with respect to azimuth, externalization and the occurrence of front-back reversals. This applied to virtual sources generated both with the BRIRs synthesized with the VAHs as well as with BRIRs measured with a conventional artificial head, for which BRIRs had to be measured repeatedly for different head orientations.

The studies in chapters 5 and 6 focused on auralizations with synthesized or measured BRIRs. It should be noted again that spectral weights for different head orientations can be applied to any signal recorded with a VAH. A fine grid of head orientations requires steering vectors measured at a high number of directions and the calculation of the spectral weights for many different head orientations, which is associated with a high number of calculations. However, these spectral weights are calculated only once and can then be applied to any recording. In contrast, in practical applications of conventional artificial heads outside the laboratory, it is almost impossible to present the recordings dynamically. Although individualization is an important capability of the VAH approach, the studies in chapters 5 and 6 showed that the possibility of presenting the recorded signals dynamically is the main advantage of the VAH approach over conventional artificial heads.

Without increasing the number of microphones, it is hardly possible to achieve a good synthesis accuracy at a fine grid of both horizontal and non-horizontal directions. The results in chapters 5 and 6 indicated that when listening to speech signals, and especially if visual cues are present, the synthesis error at non-horizontal directions caused by including only horizontal directions in the calculation of the spectral weights is imperceptible. In line with the fact that in many applications the sound sources are in or close to the horizontal plane, it is practically relevant to focus on achieving synthesis accuracy at horizontal directions with a high spatial resolution. In conclusion, the thesis showed that for practical applications using speech signals and with the supportive role of visual

cues, the VAH consisting of 24 or 31 microphones can be used as a suitable system for spatial sound reproduction.

7.2 Suggestions for further research

The perceptual evaluation of dynamic auralizations in this thesis was performed only with speech signals. In general, the dynamic auralization with the VAHs should be perceptually verified with other signals, especially signals containing more high-frequency energy. Using such signals, it should be investigated, whether the $5^\circ \times 7.5^\circ$ resolution of head orientations should be increased to avoid audible artifacts during head movements [99]. Furthermore, with signals containing more high-frequency energy, it should be assessed to what extent the synthesis error at non-horizontal directions would be audible, in case that they are not included in the calculation of the spectral weights. If this error is audible, one possible solution towards a better synthesis accuracy at non-horizontal directions is to include them in the optimization process, and in parallel consider a modification of the constraint parameters. Whereas throughout this thesis the boundaries for spectral distortion were direction-independent, these constraint parameters could be modified to assign more importance to certain directions and allow more synthesis error at perceptually less relevant, e.g. contralateral elevated directions off the median plane (see Appendix B for preliminary results). Another suggestion is to further improve the array topology. In this thesis, the impact of array topology on the VAH performance was studied only for synthesizing horizontal HRTF directivity patterns. Similar investigations can be performed with simulated microphone arrays, preferably with three-dimensional topologies, by testing the impact of array extension also in the third dimension (along the z-axis) and the distribution of the microphones. Furthermore, it would be interesting to investigate whether non-horizontal HRTF directivity patterns could be even more smoothed in the spatial domain than done in this thesis since this would further facilitate the synthesis with the VAH, when including both horizontal and non-horizontal HRTFs in the optimization.

An important issue that should be considered is the impact of microphone self-noise. In real-world applications, the recorded signals with a VAH contain some noise induced by the microphones. With synthesized BRIRs it could not be assessed to what extent microphone self-noise is perceived for different resulting mean WNG values over frequencies. In real recordings with the VAHs, microphone self-noise amplification can be audible depending on the resulting mean WNG. Particularly the different levels of self-noise among the microphones (mismatch between the microphones) could lead to audible changes in the perceived microphone self-noise depending on the orientation of the head (see e.g. [198]). The extent of such effects and their audibility for different resulting mean WNG should ideally be assessed in recordings with the VAHs in future investigations.

In addition, whereas the sound sources in this thesis were all stationary (i.e. at a fixed position), in further studies the spatial resolution of the VAH synthesis could be assessed for moving sources as well. The required spatial resolution is predicted to depend on signal bandwidth, the position of the source and trajectory and velocity of its movement [199], as well as on head movements [200], and the recording environment [201].

Another case of investigation concerns the challenging situation of nearby sound sources. In violation of the far-field assumption, the steering vectors and the HRTFs do not only depend on direction, but also on distance [31, 138, 202]. For near-field HRTFs, pinna cues are observed to vary depending on the distance especially for the sources near the interaural axis [202], while ILDs increase with decreasing distance [31]. In this case, distance should also be considered in the calculation of the spectral weights. Firstly, it should be clarified how many and which directions and distances should be included and how to handle the synthesis accuracy at all of them. Secondly, in addition to new steering vector measurements, individual distance- and direction-dependent near-field HRTFs should be acquired e.g. by measurements [203], directional equalization of far-field HRTFs [204] or numerical calculations [205]. Thirdly, the acoustical parallax effect (substantial difference between the source angle relative to the head and to the ears) [206] imposes a challenge to dynamic presentations as well as synthesizing moving sources, which does not concern only the VAH but applies to other headphone-based approaches in binaural technology as well. With decreasing distance to the head, however, the increased shadowing effects of the head introduces a low-pass filtering effect into the HRTFs [31], such that the directivity patterns could probably be further simplified at contralateral directions and higher frequencies, thus promoting the VAH synthesis. Despite the challenges involved, studying the VAH approach for nearby sources would be very worthwhile for small enclosures such as inside a vehicle.

Finally, it would be interesting to perceptually evaluate the VAH in comparison to other microphone array-based approaches [104–114]. Whereas separately performed studies enable only indirect comparisons, evaluating the VAH with other approaches within one study offers new insights, both for limitations and potentials, as well as new directions to further improve the VAH performance.

A

MEASUREMENT SETUP FOR ACQUIRING INDIVIDUAL HRIRS AND ARRAY STEERING VECTORS

The measurement setup for acquiring individual HRIRs and steering vectors consisted of a vertical circular loudspeaker arc of 1.25 m radius, with 24 small active Speedlink SL-8902-GY Xilu loudspeakers. The loudspeakers were distributed symmetrically on the two halves of the vertical arc, covering 12 elevations $-30°$, $-22.5°$, $-15°$, $-7.5°$, $0°$, $7.5°$, $15°$, $22.5°$, $30°$, $45°$, $60°$, and $75°$. The batteries of the small active loudspeakers were removed and the loudspeakers were powered by a power supply. The loudspeaker arc hang from a turntable (Outline ET250-3D), installed in the ceiling, in the middle of an acoustical laboratory (10m×7.75m×3m, reverberation time: 0.46s). For measuring individual HRIRs, subjects were seated with their interaural center positioned in the center of the arc (Figure A.1a). Similarly, for measuring steering vectors, the center of the Virtual Artificial Head (VAH) was positioned in the center of the arc (Figure A.1b). The loudspeaker arc was rotated by the turntable around the subject or the VAH in $5°$ steps. At each azimuthal position of the loudspeaker arc, impulse responses were measured at f_s=44100 Hz sampling frequency using the Multiple Exponential Sweep Method (MESM) [66] with modification as proposed in [183] over a 32-channel audio interface (Antelope Orion). The excitation was done with exponential sweeps of 17 s duration between 100 Hz and $f_s/2$ with 0.35 s shift between subsequent excitations. In order to eliminate the room reflections, the floor and the ceiling were covered with absorbent foams and the measured impulse responses were truncated to 256 samples using a 50-point half-Hann window [64].

The symmetrical distribution of loudspeakers in the left and right halves of the loudspeaker arc enabled to measure the impulse responses for the current azimuthal position and its mirrored position across the interaural axis simultaneously, such that the complete azimuthal range $0°$ to $360°$ could be covered within only $180°$ rotation of the arc. Thus, the impulse responses for 72 azimuthal ($5°$ resolution) and 12 vertical directions (in total 864 directions) could be captured in less than 40 minutes.

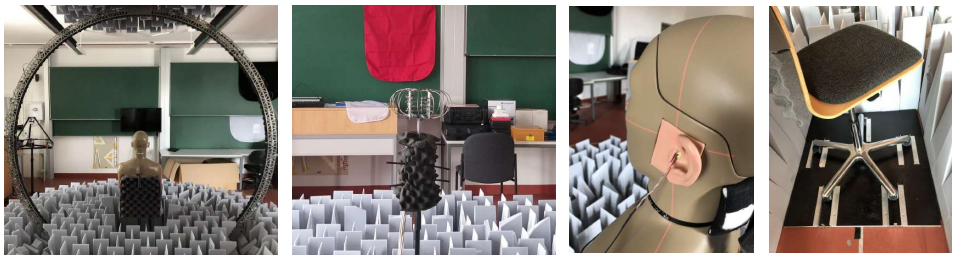

Fig. A.1: Measurement setup in the acoustic laboratory: loudspeaker arc hanging from the turntable installed in the ceiling with **(a)**: subject or **(b)**: virtual artificial head, positioned at the center. **(c)**: Exact positioning of the subject's interaural center using the laser pointers and the microphone positioning at the blocked ear canal. **(d)**: Adjustable chair for measuring individual HRIRs[1].

During the HRIR measurement, subjects sat on a chair at the center of the loudspeaker arc. The exact position of the head could be controlled with two laser pointers attached to the two halves of the loudspeaker arc at 0° elevation (Figure A.1c). The chair was adjustable in the height and could be moved slightly back- and forward, while the position of the chair in the other degree of freedom to the left or right was already adjusted to a fixed point (Figure A.1d). To help the subjects maintain the correct head position during the measurement, a head rest was used. The HRIRs were measured with two MEMS microphones (Knowles SPV0840LR5H), each attached to a small holder and then mounted into foam ear plugs (see also [207]). The ear plugs were placed inside subject's ear canals such that the microphones were positioned at the entrance of the blocked ears (Figure A.1c).

To equalize the frequency response of the loudspeakers as well as the effect of the measurement system, the two MEMS microphones used for the HRIR measurement were positioned close together (separated by a few millimeters) at the center of the loudspeaker arc and the impulse responses were measured for all 24 loudspeakers. The measured transfer functions were inverted using the regularized inversion method described in [191] with a regularization parameter of $\beta_{inversion}$=0.3 times the mean square value of the average of the impulse responses measured with the two MEMS microphones. The Loudspeaker Equalization (LSEQ) filters were calculated and saved prior to each HRIR measurement. The exponential sweep was filtered with the LSEQ filters before being emitted to the loudspeakers for the HRIR measurement. The delay for sound propagation between loudspeaker and the head was removed such that the obtained impulse responses had minimal initial delays. Subsequent to HRIR measurements, individual Headphone Impulse Responses (HPIRs) were measured with the microphones still left in the ear canals. The method as also described in [15] was applied. The HPIR measurement was

1 Pictures (a), (c) and (d) taken by Armin Budnik [208].

repeated up to nine times, each after repositioning the headphones. HPIRs resulting in the smallest dips in the frequency domain between 8 kHz and 12 kHz were chosen for the calculation of the individual inverse HPIRs using the regularized inversion method [191] with a regularization parameter of $\beta_{inversion}$=10 times the mean square value of the headphone impulse response. Measured HRIRs and inverse HPIRs were finally saved in the SOFA format (Spatially Oriented Format for Acoustics) [182].

For the measured steering vectors, the calculated LSEQ filters were applied after the impulse response measurement, prior to calculating the individual spectral weights, by using the LSEQs from the measurement of the individual HRIR. Figure A.2 shows the measured steering vectors with the two VAHs in this thesis (VAH1 and VAH2, as shown in Figure 6.1), exemplary for three microphones and for the source at $(\theta = 90°, \phi = 0°)$.

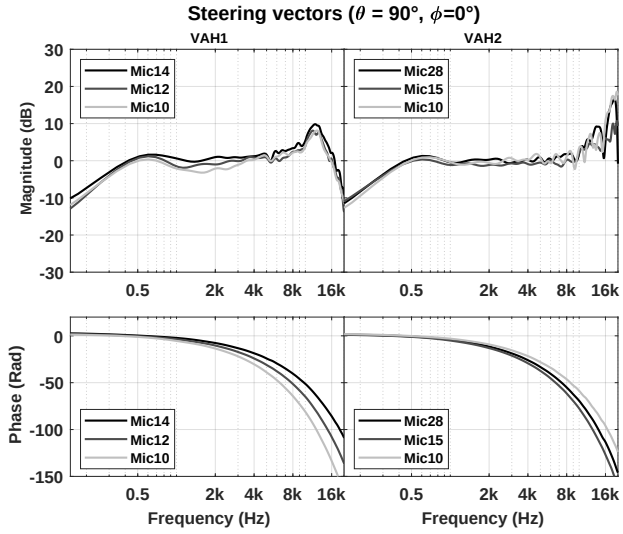

Fig. A.2: Measured steering vectors (magnitude and phase) with the two VAHs used in this thesis (VAH1 and VAH2), for the source at azimuth $\theta = 90°$ and elevation $\phi = 0°$ and three exemplary microphones.

B

THE IMPACT OF CONSTRAINT RELAXATION ON THE VAH PERFORMANCE

The study of constraint relaxation in this Appendix was partly published in [209, 210].

In order to evaluate the impact of constraint relaxation on the performance of a VAH, four constraint cases were considered as follows:

- Fixed values L_{Up}=0.5 dB, $L_{Low} = -1.5$ dB and β=0 dB were considered, referred to as **Fixed**.
- In the second case, referred to as **Relaxed L_{Low}**, the lower boundary L_{Low} for the allowable SD at *contralateral* directions was relaxed. This was motivated by the fact that the auditory system does not seem to need all spectral information from the contralateral directions. Different previous studies have discussed and utilized the decreasing importance of the contralateral ear with increasing lateralization [16, 26, 211–213]. The contralateral directions θ_{cl} were defined as 20° off the mid-line, i.e. 200° $\leq \theta_{cl} \leq$340° for the left and 20° $\leq \theta_{cl} \leq$160° for the right ear. At each contralateral direction θ_{cl}, L_{Low} was decreased as a function of the difference between the amplitude of the desired directivity pattern D at that direction and the maximum amplitude of the desired directivity pattern ($|D|_{max}$), i.e.

$$L_{Low}(\theta_{cl}) = -1.5\text{dB} - \alpha_R\{20\log_{10}(|D|_{max}) - 20\log_{10}(|D(\theta_{cl})|)\}. \quad \text{(B.1)}$$

  The factor α_R determined individually for each frequency, how much L_{Low} was reduced. The optimization started with α_R=0 (i.e. $L_{Low} = -1.5$ dB). If the WNG_m constraint or the SD constraints could not be satisfied for all directions, then α_R was increased in steps of 0.1 and the constrained optimization was repeated. An upper limit of 0.6 was chosen for α_R to limit the computation time. The two other constraint parameters remained unchanged, i.e. $L_{Up} = 0.5$ dB and β=0 dB.
- In the third constraint case, referred to as **Relaxed WNG**, the minimum desired value for WNG_m was relaxed. The optimization started with β=0 dB. If not all constraints could be satisfied, β was decreased in 1 dB steps until

all constraints could be satisfied or until β reached -13 dB as stop criterion. The two other parameters remained unchanged, i.e. $L_{Up} = 0.5$ dB and $L_{Low} = -1.5$ dB.

- In the fourth constraint case, the relaxation of β as described above was combined with a reduction of the number P of the calibration directions to $P/2$, by excluding every second directions from the calculation of the spectral weights. This case was referred to as **Relaxed WNG+Res.** This last case was obviously contrary to the objective of increasing the spatial resolution. However, it was considered to investigate the effect of the number of constraints on the ability to satisfy them.

To evaluate the impact of constraint relaxation on the success of constrained optimization, a success rate was computed, which was defined as the percentage of the frequencies in the range 170 Hz$\leq f \leq$16 kHz for which all of the defined constraints could be satisfied.

The four constraint cases were applied to A-100% (see Figure 3.4) as well as the two simulated microphone arrays, A-37.5% and A-Rand32, shown in Figure B.1. A-37.5% was the 37.5% down-scaled copy of A-100%. A-Rand32 was an array of almost the same extensions of A-100%, with 32 microphones, consisting of 24 outer microphones and 8 microphones close to the center of the array. To depict the impact of constraint relaxation, spectral weights were calculated for synthesizing P=72 horizontal HRTFs with 5° resolution of one exemplary subject.

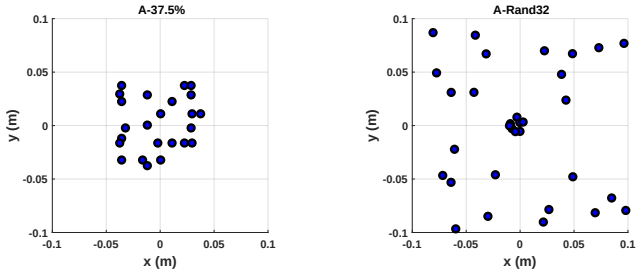

Fig. B.1: Microphone positions for the two simulated microphone arrays. **Left**: A-37.5%, down-scaled version of A-100% shown in Figure 3.4 to 37.5% of the original size. **Right**: A-Rand32, array consisting of 32 microphones, with 24 outer microphones and 8 microphones close to the center of the array.

Figure B.2a shows the resulting SD, TD and WNG_m, for the left ear and for the synthesis with A-100%. Also, the desired minimum value for WNG_m (i.e. β) is shown as a grey curve. Frequencies, for which the resulting WNG_m was less than β, indicate a failure of satisfying the WNG_m constraint. This was the case for frequencies above about 5 kHz for the Fixed and Relaxed L_{Low} cases. The resulting L_{Low} and the corresponding factor α_R for the Relaxed L_{Low} case, as well as the success rate for the four constraint cases are shown in Figure B.2b and Figure B.2c,

respectively. While for frequencies up to about 5 kHz, there was no need to reduce the L_{Low} (α_R=0), for frequencies above this range, the reduction of L_{Low} even to the maximum allowable limit (i.e. α_R=0.6) could not help satisfy the constraints. In other words, neither the resulting SD and WNG_m nor the success rate showed any remarkable improvement for the Relaxed L_{Low} case compared to the Fixed case. The reduction of β for Relaxed WNG case was helpful in satisfying the constraints only for frequencies between approx. 5 kHz and 6 kHz. For frequencies above that, although a maximum relaxation of β down to -13 dB was not necessary to satisfy the WNG_m constraint, it was no help against the positive contralateral SDs. Compared to the Fixed case, there was only a slight increase in success rate for the Relaxed WNG case. A prominent increase in success rate compared to the Fixed case could be observed for the Relaxed WNG+Res. case. At the directions which were included in the calibration, for most frequencies, the resulting SD could be kept in the desired range and the WNG_m constraint could be satisfied. At the same time, the resulting SD was too high at the excluded directions, as can be seen in the lower row of Figure B.2a. For the three new constraint cases (Relaxed L_{Low}, Relaxed WNG and Relaxed WNG+Res.), the effect of the constraint relaxation appeared at frequencies above about 2 kHz. As a result, the resulting TD was almost the same for the different constraint relaxation cases.

Figure B.3 shows the resulting SD, TD and WNG_m for the left ear and for the synthesis with A-37.5%. For this array, the success rate for the Fixed case was smaller than the Fixed case of A-100% (see Figure B.2c). Although the small inter-microphone distances of A-37.5% were helpful for satisfying the SD constraints at higher frequencies, the WNG_m constraints could not be satisfied at above 2 kHz due to the smaller extension of A-37.5% compared to A-100%. As expected, satisfying the SD and WNG_m constraints at frequencies below 2 kHz impacted the resulting TD negatively. In contrast to A-100%, for A-37.5% the Relaxed L_{Low} case led to an increased success rate compared to the Fixed case. This constraint relaxation was effective at frequencies above about 1.5 kHz, where the relaxation of L_{Low} helped satisfy the SD and WNG_m constraints. At the same time, this improved success rate was subject to increased negative contralateral SDs down to -15 dB or less. Constraint relaxations as in Relaxed WNG and Relaxed WNG+Res. led as well to an increased success rate compared to the Fixed case for A.37.5%. However, this improvement was at the cost of less resulting WNG_m or expanded positive and negative SDs at directions which were excluded from the constrained optimization, as shown in Figure B.3a for the Relaxed WNG+Res. case. Since the effect of different constraint relaxations appeared at frequencies above 1.5 KHz, the resulting TDs were similar for the four constraint cases.

Figure B.4 shows the results for the synthesis with A-Rand32. For this array, the constraints could be better satisfied compared to A-100% and A-37.5%, even for the Fixed case. This can be explained by the advantageous topology of A-Rand32: the combination of dense and sparse microphone distances helped satisfying the SD constraint up to higher frequencies and the WNG_m at low and mid-frequency range. For A-Rand32, constraints could be satisfied up to about 6 kHz without the

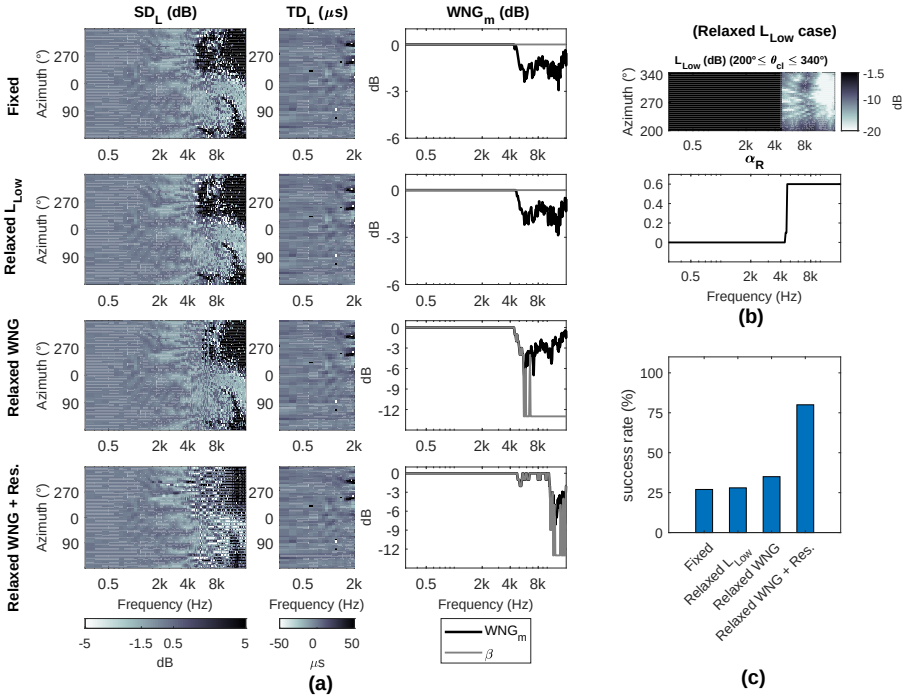

Fig. B.2: Synthesizing 72 horizontal left-ear HRTFs with A-100%, when applying the four different constraint cases. **(a)**: Resulting SD, TD, $\mathrm{WNG_m}$ as well as the desired value of β. **(b)**: $\mathrm{L_{Low}}$ at the contralateral directions ($200^\circ \leq \theta_{cl} \leq 340^\circ$ for the left ear) for the Relaxed $\mathrm{L_{Low}}$ case and the reduction factor α_R as implied in Eq. (B.1). **(c)** Success rate, defined as the percentage of narrow-band constrained optimizations at 170 Hz$\leq f \leq$ 16 kHz, for which all constraints could be satisfied.

need for relaxing $\mathrm{L_{Low}}$ (see Figure B.4b). With the relaxation of $\mathrm{L_{Low}}$, the $\mathrm{WNG_m}$ constraint could be satisfied up to about 7 kHz. At frequencies above this range, the relaxation of $\mathrm{L_{Low}}$ was no help for satisfying the constraints. The relaxation of β as in the Relaxed WNG case, and especially in combination with a reduced number of constraints as in Relaxed WNG+Res. case could improve the success rate, however either at the cost of lower resulting $\mathrm{WNG_m}$ for the Relaxed WNG case or degraded SDs at the contralateral directions for the Relaxed WNG+Res. case.

B.1 Perceptual evaluation

The perceptual quality of the synthesis with A-37.5% and A-Rand32 and the four constraint cases were investigated in a listening test. The simulated microphone arrays were considered as perfectly robust, i.e. neither microphone self-noise nor deviations in microphone characteristics and positions were considered. It should

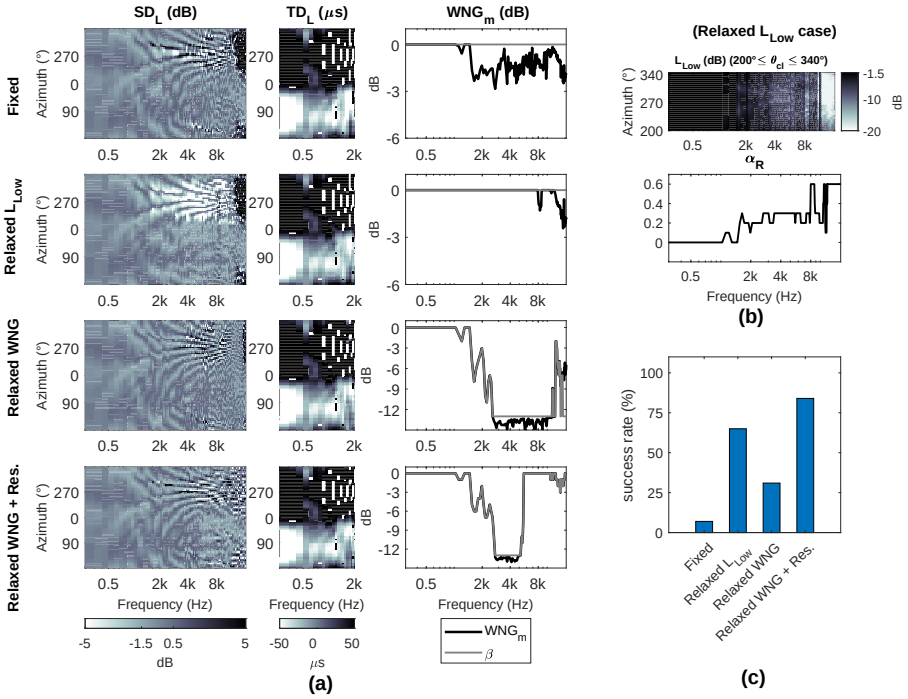

Fig. B.3: Synthesizing 72 horizontal left-ear HRTFs with A-37.5%, when applying the four different constraint cases. **(a)**: Resulting SD, TD, $\mathrm{WNG_m}$ as well as the desired value of β. **(b)**: $\mathrm{L_{Low}}$ at the contralateral directions ($200° \leq \theta_{cl} \leq 340°$ for the left ear) for the Relaxed $\mathrm{L_{Low}}$ case and the reduction factor α_R as implied in Eq. (B.1). **(c)** Success rate, defined as the percentage of narrow-band constrained optimizations at 170 Hz$\leq f \leq$ 16 kHz, for which all constraints could be satisfied.

be mentioned that besides the improvement of the resulting spectral accuracy, the relaxation of β can also lead to a loss of robustness. In the evaluations here however, the robustness was not investigated for different resulting $\mathrm{WNG_m}$ values. As a result, the effect of the relaxation of the $\mathrm{WNG_m}$ constraint as well as other constraint relaxation cases was only evaluated based on the resulting spectral accuracy.

Individual horizontal HRTFs were measured with 7.5° azimuthal resolution, i.e. for 48 source directions inside an anechoic room, as described in detail in [214]. A total of ten (self-reported) normal hearing subjects (five male, five female, aged between 28 to 52 years) took part in the listening test. All subjects had long-lasting experience with binaural psychoacoustic listening tests. All procedures were approved by the ethics committee of the Carl von Ossietzky University of Oldenburg.

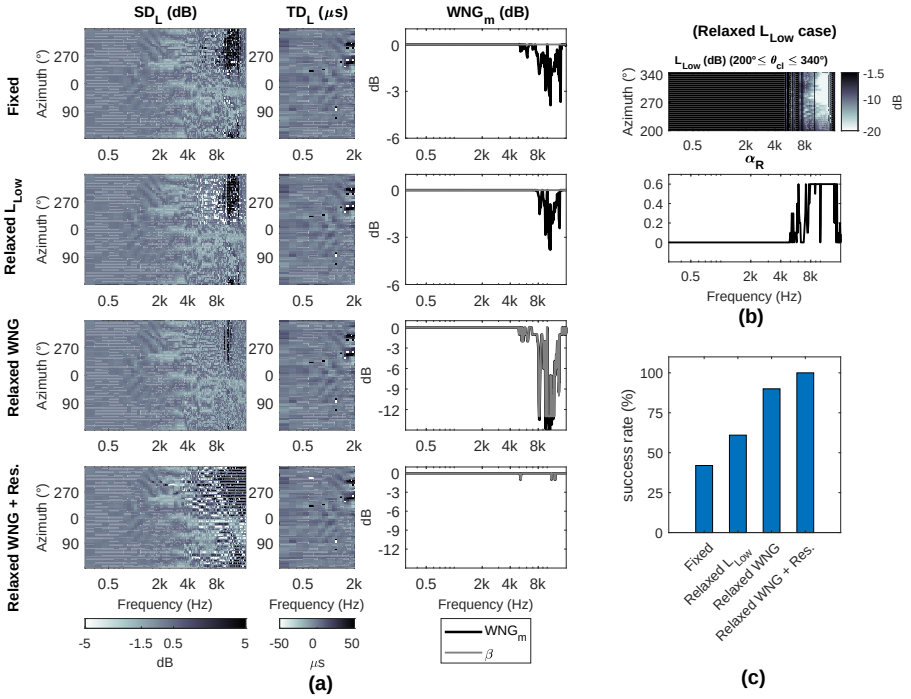

Fig. B.4: Synthesizing 72 horizontal left-ear HRTFs with A-Rand32, when applying the four different constraint cases. **(a)**: Resulting SD, TD, $\mathrm{WNG_m}$ as well as the desired value of β. **(b)**: $\mathrm{L_{Low}}$ at the contralateral directions ($200^\circ \leq \theta_{cl} \leq 340^\circ$ for the left ear) for the Relaxed $\mathrm{L_{Low}}$ case and the reduction factor α_R as implied in Eq. (B.1). **(c)** Success rate, defined as the percentage of narrow-band constrained optimizations at 170 Hz$\leq f \leq$ 16 kHz, for which all constraints could be satisfied.

The measured HRTFs were synthesized with A-37% and A-Rand32 separately, applying each of the four constraint cases Fixed, Relaxed $\mathrm{L_{Low}}$, Relaxed WNG and Relaxed WNG+Res. The test signal was filtered either with the individually measured original HRTFs or with individually synthesized HRTFs, and subsequently with the inverse individual headphone transfer functions prior to headphone presentation. Binaural signals generated with HRTFs measured for the KEMAR artificial head were also presented as a hidden anchor signal. Subjects were asked to rate the presentations generated with different synthesized HRTFs as well as the hidden anchor signal, compared to the reference signal (binaural signals generated with the individually measured HRTFs). This resulted in nine test stimuli and one reference stimulus. Three perceptual attributes were used for the ratings in the listening test: Spectral Coloration, Localization and Overall Quality. The test signal for perceptual attributes Coloration and Localization consisted of short bursts of pink noise with a spectral content of 300 Hz$\leq f \leq$ 16 kHz. For evaluating Spectral Coloration, three pink noise bursts were presented, each lasting $\frac{1}{3}$s with

1 ms onset-offset ramps, followed by $\frac{1}{6}$s silence. The frontal direction $\theta = 0°$, as shown in Figure B.5a, as well as two lateral directions $\theta = 60°$ and 247° were considered as nominal azimuthal positions for the virtual source.

For ratings regarding Localization, the same three nominal positions were considered. For each nominal position, a sound source moving over a course of seven subsequent positions in 7.5° steps was presented, either from 22.5° to −22.5° (0°±22.5°), or from 37.5° to 82.5° (60°±22.5°), or from 217.5° to 262.5° (240°±22.5°). As an example, the positions of the moving virtual source at 0°±22.5° are shown in Figure B.5b. One single pink noise burst of $\frac{1}{6}$s duration followed by $\frac{1}{6}$s silence was presented from each of the seven virtual sources.

For perceptual attribute Overall Quality, the test signal consisted of a piece of music recorded separately for male voice and two guitars. A virtual musical scene was created at $\theta = 0°$ by filtering each recorded part with the (synthesized, individually measured or KEMAR) HRTFs at the directions shown in Figure B.5c, i.e. 15° and 337.5° for the first and second guitar, respectively, and 352.5° for the vocal part. The musical scene was also generated for two other directions by rotating the musical scene shown in Figure B.5c to $\theta = 60°$ and $\theta = 240°$. A segment of 30s of the test signal was presented in a continuous loop.

The listening test took place with subjects seated in an empty anechoic room, wearing the headphones. Similar to perceptual evaluations in chapter 3.6, subjects gave their evaluation on a 9-point scale between 1 and 5 in 0.5 steps, with the five German labels "schlecht" (bad), "dürftig" (poor), "ordentlich" (fair), "gut" (good) and "ausgezeichnet" (excellent) and four unlabeled intermediate points using a Graphical User Interface (GUI). Subjects could listen to different headphone presentations or reference signals as often as they liked and could sort the test stimuli according to their current rating by clicking a sort button on the GUI, in order to ease the comparison.

For each perceptual attribute, each of the three source positions was presented three times in a randomized order, resulting in nine rounds. The experiment started with ratings on Spectral Coloration, followed by Localization and Overall Quality. Each part lasted on average 30 minutes. Subjects were encouraged to take a break between the parts.

B.2 Results

The mean success rates are shown in Figure B.6. The individual resulting left and right SD and $\mathrm{WNG_m}$ of the ten subjects, when applying the four constraint cases for A-37.5% and A-Rand32 were in general similar to the results shown in Figures B.3 and B.4 and are therefore not shown here.

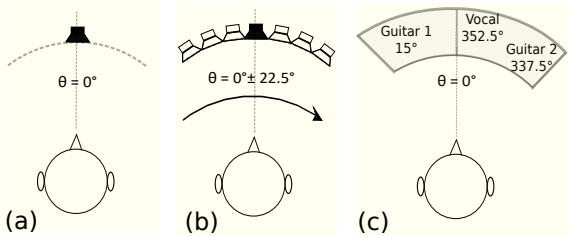

Fig. B.5: Spatial setup for the subjective listening test. **(a)**: Three noise pulses presented from the source at $\theta = 0°$ for ratings on Spectral Coloration. **(b)**: Moving virtual sound source over seven source positions $\theta = 22.5°$ to $\theta = -22.5°$ for ratings on Localization, each source presented with a single noise pulse. **(c)**: Virtual musical scene at $\theta = 0°$ for ratings on Overall Quality.

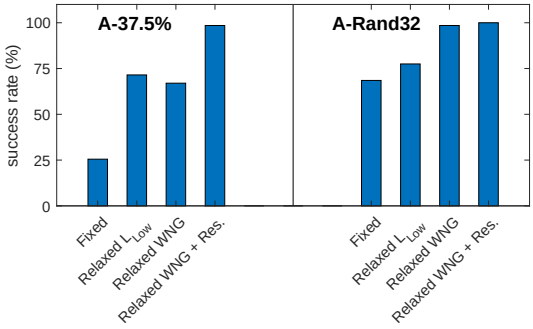

Fig. B.6: Success rate, averaged over ten subjects when listening to 48 horizontal HRTFs with A-37.5% and A-Rand32, applying different constraint cases.

Since for each perceptual attribute, the three nominal source positions were presented three times, the consistency of the ratings given to the three presentations was assessed similar to the method in section 3.6.2, according to which, Localization ratings of three subjects as well as Overall Quality ratings of one subject were excluded. All other ratings were averaged over the three repetitions and the results are shown in Figures B.7 - B.9. For statistical analysis, the Friedman test was performed, since according to the Shapiro-Wilk test of normality, for some HRTF sets, the ratings could not be assumed to be normally distributed. The Friedman test confirmed for all perceptual attributes and source positions a significant effect of HRTF set ($p<10e\text{-}4$). Significantly different cases were indicated by multiple comparisons after Friedman test and are indicated with horizontal lines.

Coloration ratings: The Coloration ratings of ten subjects, averaged over three repetitions are shown in Figure B.7. The median values for the Fixed case were between fair and good for both arrays. Ratings appeared in general higher for A-Rand32 than for A-37.5%. For almost all cases, ratings for both arrays were noticeably higher than ratings for the anchor signal (measured KEMAR HRTFs).

For the Relaxed L_{Low} case, the excess negative contralateral SD of A-37.5% in the frequency range between 1 kHz and 8 kHz (compare Figure B.3) led to slightly lower ratings at the two lateral source positions compared to the Fixed case. For A-Rand32, applying the Relaxed L_{Low} case did not lead to prominent changes in the Coloration ratings at the lateral positions in comparison to the Fixed case; Coloration ratings improved slightly for the frontal source position. The Relaxed WNG case did not noticeably change the Coloration ratings for A-37.5% compared to the Fixed case. According to Figure B.6, the Relaxed WNG improved the chance of satisfying the constraints for A-37.5% compared to the Fixed case. This improvement was mainly with respect to the WNG_m, the effect of which was not evaluated perceptually here. For the Relaxed WNG+Res. case, there was a pronounced degradation in the Coloration ratings at $\theta = 247.5°$ for A-Rand32, with ratings comparable or even worse than the anchor signal. This was not an unexpected result, since for the Relaxed WNG+Res. case the direction $\theta = 247.5°$ was excluded during the optimization. For A-37.5%, spectral errors at the excluded directions began to appear at higher frequencies compared to A-Rand32, especially at the contralateral directions. Coloration ratings for A-37.5% showed a smaller deterioration for the Relaxed WNG+Res. case compared to A-Rand32. For all cases of A-37.5% and A-Rand32, Coloration ratings were significantly higher than ratings for the anchor signal. The exception was the Relaxed WNG+Res. case of A-Rand32 at $\theta = 247.5°$.

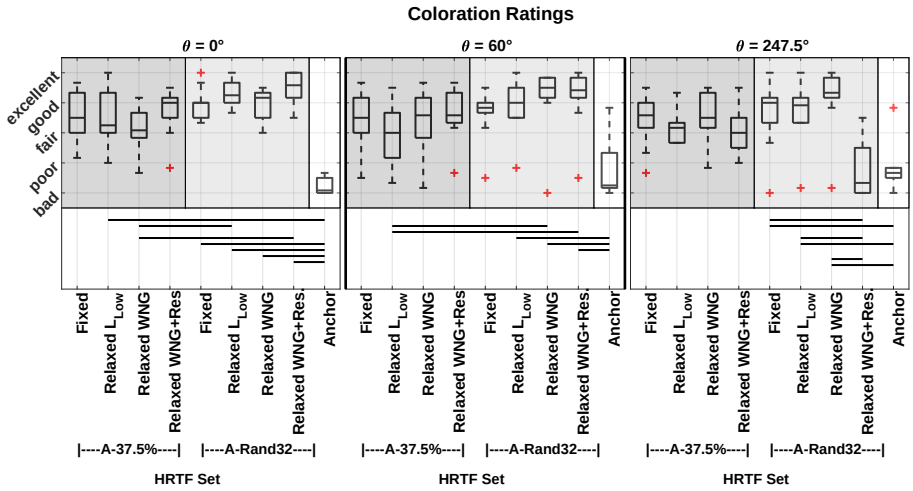

Fig. B.7: Results of perceptual evaluation with respect to Spectral Coloration for ten subjects, averaged over three repetitions, as a function of HRTF set. Cases with significant differences are marked with horizontal bars ($p<0.05$).

Localization ratings: Over three repetitions averaged perceptual results of seven subjects are shown in Figure B.8. Again, ratings were generally higher for A-Rand32 than for A.37.5%, and for both arrays higher than ratings for the anchor signal. It is interesting that the increasing negative contralateral SD of

A-37.5% for Relaxed L_{Low} case did not lead to noticeably lower Localization ratings compared to the Fixed case. The Relaxed L_{Low} case also did not change the Localization ratings for A-Rand32. An apparent degradation in the Localization ratings compared to the Fixed case was observed for the Relaxed WNG+Res. at all source positions for A-Rand32. This is explained by the fact that each second direction was excluded from the optimization. Localization ratings for A-37.5% suffered less from the synthesis with reduced azimuthal resolution.

At all directions, Localization ratings for Relaxed WNG+Res. case of A-Rand32 were significantly lower than ratings for other cases of this array and different cases of A-37.5%. In contrast, significant differences to the anchor signal were not observed, which confirms that the Relaxed WNG+Res. of A-Rand32 performed worse than the anchor signal.

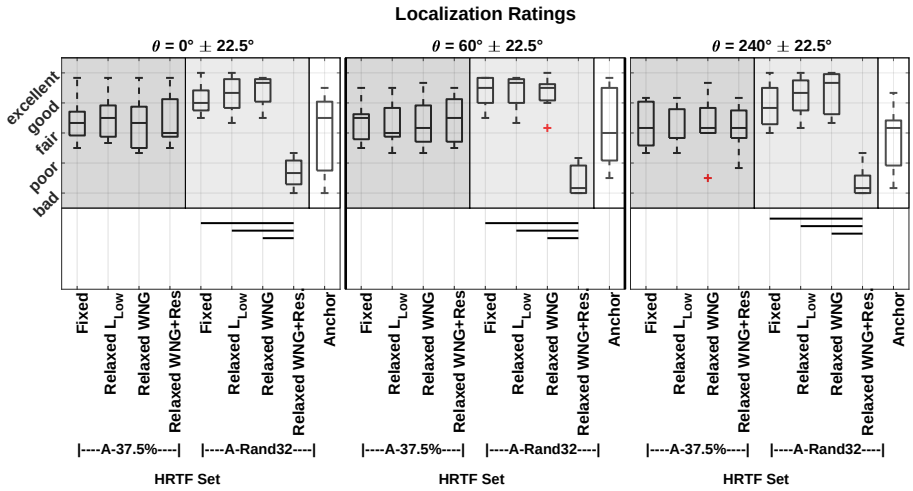

Fig. B.8: Results of perceptual evaluation with respect to Localization for seven subjects, averaged over three repetitions, as a function of HRTF set. Cases with significant differences are marked with horizontal bars ($p<0.05$).

Overall Quality ratings: Ratings on Overall Quality, averaged over the three repetitions of nine subjects are shown in Figure B.9. Ratings for A-Rand32 were in general higher than for A-37.5%. Median ratings were between fair and good for A-37.5% and between good and excellent for A-Rand32. One exception was the Relaxed WNG+Res. case for A-Rand32, were the ratings showed a clear degradation compared to the ratings for Fixed, Relaxed L_{Low} and Relaxed WNG cases, especially for the two lateral directions of the musical scene at $\theta = 60°$ and 240°. This was due to the increased error in the synthesis of vocal and second guitar parts of the stimuli, located at $\theta = 352.5°$ and 337.5°, respectively, for the frontal scene (see Figure B.5c). These two directions and equivalently the directions 52.5° and 37.5°, and the directions 232.5° and 217.5° for the two other rotated musical

scenes to 60° and 240°, were excluded from the calibration. On the other hand, the Relaxed WNG+Res. case did not show any deterioration for A-37.5%. This was due to the smaller extension of A-37.5% compared to A-Rand32 and the fact that the increasing SD at the excluded directions appear first at higher frequencies, especially at ipsilateral directions. The increasing negative contralateral SD of A-37.5% for Relaxed L_{Low} case did not lead to a major deterioration of the subjective ratings on Overall Quality.

At $\theta = 0°$, the ratings for both arrays were significantly higher than the ratings given to the anchor signal. For the two lateral directions at $\theta = 60°$ and 240°, the Fixed, Relaxed L_{Low} and Relaxed WNG cases of A-Rand32 were significantly better evaluated than the anchor signal, and also partly better than different cases of A-37.5%.

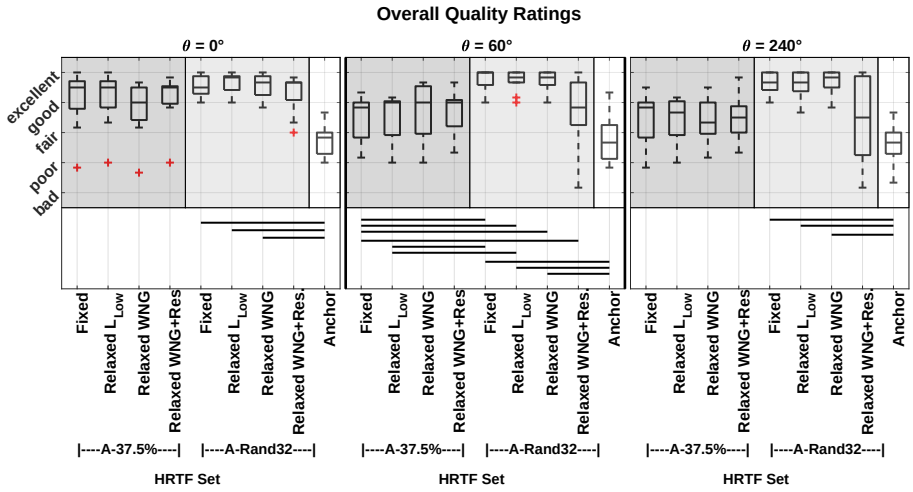

Fig. B.9: Results of perceptual evaluation with respect to Overall Quality for nine subjects, averaged over three repetitions, as a function of HRTF set. Cases with significant differences are marked with horizontal bars ($p<0.05$).

B.3 Discussion

Relaxing the constraints improved the success of satisfying the constraints for the two considered arrays. With the exception of the Relaxed WNG+Res. for A-Rand32, the results showed higher perceptual ratings for the different constraint cases of both arrays in comparison to the anchor signal (KEMAR HRTFs), with often significantly higher ratings for A-Rand32. With the exception of the Relaxed WNG+Res., with A-Rand32 more constraints could be satisfied and at the same time better perceptual ratings compared to A-37.5% could be achieved. The results indicated that with properly chosen constraints, the VAH can be a perceptually well-suited substitute for the individually measured HRTFs for a high spatial

resolution grid of horizontal directions.

The lower success rate of A-37.5% compared to A-Rand32 for the Fixed case was mostly due to the failure to satisfy the $\mathrm{WNG_m}$ constraint, the effect of which was not considered here. Whereas the Coloration ratings for the Fixed case were similar for both arrays, the Localization ratings for A-37.5% were lower, which is explained by the smaller extension of A-37.5% compared to A-Rand32 and the high TDs caused by satisfying the SD and the more challenging $\mathrm{WNG_m}$ constraints at lower frequencies.

An interesting point to discuss is the Relaxed $\mathrm{L_{Low}}$ case for A-37.5% at the two lateral source positions for all three perceptual attributes ($\theta = 60°$ and $247.5°$ for Coloration, $\theta = 60° \pm 22.5°$ and $240° \pm 22.5°$ for Localization and $\theta = 60°$ and $240°$ for Overall Quality). The increased negative contralateral SD at approximately 1 kHz$\leq f \leq$8 kHz did not lead to significantly lower ratings. This is in accordance with previous studies, confirming the limited importance of the contralateral ear at lateral directions [16, 26, 213].

The synthesis error with different cases of the applied constraint increased generally towards higher frequencies. Since noise and noise pulses, as used for the evaluation of Coloration and Localization, exhibit more high-frequency content than music as used for the evaluation of Overall Quality, the Overall Quality ratings were, on the whole, higher than the Coloration and Localization ratings. It may be concluded that the VAH approach will be suitable for more relevant practical cases using speech or music than for critical signals such as noise and noise pulses.

BIBLIOGRAPHY

[1] H. Lehnert, "Auditory Spatial Impression," *Proc. AES 12th International Conference: The Perception of Reproduced Sound*, pp. 40–46, 1993. (Cited on page 1.)

[2] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, "Spatial Sound With Loudspeakers and Its Perception: A Review of the Current State," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 1920–1938, 2013. (Cited on page 1.)

[3] W. G. Gardner, "3-D Audio Using Loudspeakers," *Ph.D. thesis, Massachusetts Institute of Technology*, 1997. (Cited on page 1.)

[4] H. Møller, "Fundamentals of Binaural Technology," *Applied Acoustics*, vol. 36, no. 3-4, pp. 171–218, 1992. (Cited on pages 1, 7, and 10.)

[5] J. Blauert, *Spatial Hearing. The psychophysics of human sound localization*, revised ed. Cambridge, Massachusetts: MIT Press, 1997. (Cited on pages 1, 3, 4, 5, 13, 39, 40, 41, and 60.)

[6] S. Paul, "Binaural Recording Technology: A Historical Review and Possible Future Developments," *Acta Acustica united with Acustica*, vol. 95, pp. 767–788, 2009. (Cited on pages 1 and 8.)

[7] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural Technique: Do We Need Individual Recordings?" *J. Audio Eng. Soc.*, vol. 44, no. 6, pp. 451–469, 1996. (Cited on pages 1, 8, 80, 83, 98, and 99.)

[8] P. Minnaar, S. K. Olesen, F. Christensen, and H. Møller, "Localization with Binaural Recordings from Artificial and Human Heads," *J. Audio Eng. Soc.*, vol. 49, no. 5, pp. 323–336, 2001. (Cited on pages 1 and 8.)

[9] H. Wallach, "The Role of Head Movements and Vestibular and Visual Cues in Sound Localization," *J. Exp. Psychol.*, vol. 27, no. 4, pp. 339–368, 1940. (Cited on pages 2, 5, 98, and 105.)

[10] W. R. Thurlow and P. S. Runge, "Effect of Induced Head Movements on Localization of Direction of Sounds," *J. Acoust. Soc. Am.*, vol. 42, no. 2, pp. 480–488, 1967. (Cited on pages 2, 5, and 105.)

[11] S. Perrett and W. Noble, "The effect of head rotations on vertical plane sound localization," *J. Acoust. Soc. Am.*, vol. 102, no. 4, pp. 2325–2332, 1997. (Cited on pages 2, 5, and 105.)

[12] 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," *J. Audio Eng. Soc.*, vol. 49, no. 10, pp. 904–916, 2001. (Cited on pages 2, 13, 53, 57, 80, 83, 99, 104, and 105.)

[13] W. O. Brimijoin, A. W. Boyd, and M. A. Akeroyd, "The Contribution of Head Movement to the Externalization and Internalization of Sounds," *PloS one*, vol. 8, no. 12, 2013, e83068. (Cited on pages 2, 13, 53, 57, 83, 104, and 105.)

[14] 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," *J. Acoust. Soc. Am.*, vol. 141, no. 3, pp. 2011–2023, 2017. (Cited on pages 2, 13, 53, 57, 91, 104, and 105.)

[15] E. Rasumow, M. Blau, S. Doclo, S. van de Par, M. Hansen, D. Püschel, and V. Mellert, "Perceptual Evaluation of Individualized Binaural Reproduction Using a Virtual Artificial Head," *J. Audio Eng. Soc.*, vol. 65, no. 6, pp. 448–459, 2017. (Cited on pages 2, 14, 16, 18, 19, 22, 28, 34, 43, 44, 46, 49, 50, 51, 58, 60, 80, 81, 84, 85, 111, 113, 114, and 118.)

[16] E. Rasumow, M. Blau, M. Hansen, S. van de Par, S. Doclo, V. Mellert, and D. Püschel, "Smoothing individual head-related transfer functions in the frequency and spatial domains," *J. Acoust. Soc. Am.*, vol. 135, no. 4, pp. 2012–2025, 2014. (Cited on pages 2, 16, 18, 25, 30, 31, 34, 37, 38, 63, 111, 121, and 132.)

[17] E. Rasumow, M. Hansen, S. van de Par, D. Püschel, V. Mellert, S. Doclo, and M. Blau, "Regularization Approaches for Synthesizing HRTF Directivity Patterns," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 215–225, 2016. (Cited on pages 2, 16, 17, 25, 28, 29, 34, 36, 44, 50, 59, and 111.)

[18] P. Zahorik, "Spatial Hearing in Rooms and Effects of Reverberation," in *Binaural Hearing. With 93 Illustrations*, R. Y. Litovsky, M. J. Goupell, R. R. Fay, and A. N. Popper, Eds. Springer, Cham, 2021, ch. 9, pp. 243–280. (Cited on page 3.)

[19] B. R. Shelton and C. L. Searle, "The influence of vision on the absolute identification of sound-source position," *Perception & Psychophysics*, vol. 28, no. 6, pp. 589–596, 1980. (Cited on pages 3 and 7.)

[20] P. Bertelson and M. Radeau, "Cross-modal bias and perceptual fusion with auditory-visual spatial discordance," *Perception & Psychophysics*, vol. 29, no. 6, pp. 578–584, 1981. (Cited on pages 3 and 7.)

[21] J. C. Middlebrooks, J. C. Makous, and D. M. Green, "Directional sensitivity of sound-pressure levels in the human ear canal," *J. Acoust. Soc. Am.*, vol. 86, no. 1, pp. 89–108, 1989. (Cited on page 4.)

[22] V. Benichoux, M. Rébillat, and R. Brette, "On the variation of interaural time differences with frequency," *J. Acoust. Soc. Am.*, vol. 139, no. 4, pp. 1810–1821, 2016. (Cited on page 4.)

[23] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp. 1648–1661, 1992. (Cited on page 4.)

[24] E. A. G. Shaw and R. Teranishi, "Sound Pressure Generated in an External-Ear Replica and Real Human Ears by a Nearby Point Source," *J. Acoust. Soc.*

Am., vol. 44, no. 1, pp. 240–249, 1968. (Cited on pages 4, 30, and 108.)

[25] M. B. Gardner and R. S. Gardner, "Problem of localization in the median plane: effect of pinnae cavity occlusion," *J. Acoust. Soc. Am.*, vol. 53, no. 2, pp. 400–408, 1973. (Cited on page 4.)

[26] M. Morimoto, "The contribution of two ears to the perception of vertical angle in sagittal planes," *J. Acoust. Soc. Am.*, vol. 109, no. 4, pp. 1596–1603, 2001. (Cited on pages 4, 121, and 132.)

[27] M. B. Gardner, "Some monaural and binaural facets of median plane localization," *J. Acoust. Soc. Am.*, vol. 54, no. 6, pp. 1489–1495, 1973. (Cited on pages 4 and 10.)

[28] V. R. Algazi, C. Avendano, and R. O. Duda, "Elevation localization and head-related transfer function analysis at low frequencies," *J. Acoust. Soc. Am.*, vol. 109, no. 3, pp. 1110–1122, 2001. (Cited on pages 4 and 10.)

[29] J. Hebrank and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," *J. Acoust. Soc. Am.*, vol. 56, no. 6, pp. 1829–1834, 1974. (Cited on page 5.)

[30] K. I. McAnally and R. L. Martin, "Sound localization with head movement: implications for 3-d audio displays," *Frontiers in neuroscience*, vol. 8, no. 210, 2014. (Cited on page 5.)

[31] D. S. Brungart and W. M. Rabinowitz, "Auditory localization of nearby sources. Head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1465–1479, 1999. (Cited on pages 5, 10, and 116.)

[32] D. S. Brungart, N. I. Durlach, and W. M. Rabinowitz, "Auditory localization of nearby sources. II. Localization of a broadband source," *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 1956–1968, 1999. (Cited on page 5.)

[33] P. D. Coleman, "Failure to Localize the Source Distance of an Unfamiliar Sound," *J. Acoust. Soc. Am.*, vol. 34, no. 3, pp. 345–346, 1962. (Cited on page 5.)

[34] S. K. Roffler and R. A. Butler, "Factors That Influence the Localization of Sound in the Vertical Plane," *J. Acoust. Soc. Am.*, vol. 43, no. 6, pp. 1255–1259, 1968. (Cited on page 5.)

[35] R. B. King and S. R. Oldfield, "The Impact of Signal Bandwidth on Auditory Localization: Implications for the Design of Three-Dimensional Audio Displays," *Human Factors*, vol. 39, no. 2, pp. 287–295, 1997. (Cited on page 5.)

[36] J. C. Middlebrooks and D. M. Green, "Sound localization by human listeners," *Annu. Rev. Psychol.*, vol. 42, pp. 135–159, 1991. (Cited on page 5.)

[37] W. A. Yost and X. Zhong, "Sound source localization identification accuracy: Bandwidth dependencies," *J. Acoust. Soc. Am.*, vol. 136, no. 5, pp. 2737–2746, 2014. (Cited on page 5.)

[38] R. A. Butler, "The bandwidth effect on monaural and binaural localization," *Hearing Research*, vol. 21, pp. 67–73, 1986. (Cited on page 5.)

[39] J. Blauert and J. Braasch, "Räumliches Hören," in *Handbuch der Audiotechnik*, S. Weinzierl, Ed. Springer, 2008, ch. 3, pp. 87–121. (Cited on page 5.)

[40] A. W. Mills, "On the Minimum Audible Angle," *J. Acoust. Soc. Am.*, vol. 30, no. 4, pp. 237–246, 1958. (Cited on page 5.)

[41] D. R. Perrott and K. Saberi, "Minimum audible angle thresholds for sources varying in both elevation and azimuth," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1728–1731, 1990. (Cited on pages 5, 6, and 31.)

[42] J. C. Makous and J. C. Middlebrooks, "Two-dimensional sound localization by human listeners," *J. Acoust. Soc. Am.*, vol. 87, no. 5, pp. 2188–2200, 1990. (Cited on page 6.)

[43] M. Mironovs and H. Lee, "On the accuracy and consistency of sound localisation at various azimuth and elevation angles," *Proc. AES 144th Convention*, 2018. (Cited on page 6.)

[44] A. Ford, "Dynamic Auditory Localization. 1. The Binaural Intensity Disparity Limen," *J. Acoust. Soc. Am.*, vol. 13, pp. 367–372, 1942. (Cited on page 6.)

[45] L. F. Elfner and D. R. Perrott, "Lateralization and Intensity Discrimination," *J. Acoust. Soc. Am.*, vol. 42, no. 2, pp. 441–445, 1967. (Cited on page 6.)

[46] H. Babkoff and S. Sutton, "Binaural Interaction of Transients: Interaural Intensity Asymmetry," *J. Acoust. Soc. Am.*, vol. 46, no. 4B, pp. 887–892, 1969. (Cited on page 6.)

[47] R. G. Klumpp and H. R. Eady, "Some Measurements of Interaural Time Difference Thresholds," *J. Acoust. Soc. Am.*, vol. 28, no. 5, pp. 859–860, 1956. (Cited on page 6.)

[48] A. Brughera, L. Dunai, and W. M. Hartmann, "Human interaural time difference thresholds for sine tones: The high-frequency limit," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. 2839–2855, 2013. (Cited on page 6.)

[49] H. Wallach, E. B. Newman, and M. R. Rosenzweig, "A Precedence Effect in Sound Localization," *J. Acoust. Soc. Am.*, vol. 21, pp. 468–468, 1949. (Cited on page 6.)

[50] B. Rakerd and W. M. Hartmann, "Localization of sound in rooms. V. Binaural coherence and human sensitivity to interaural time differences in noise," *J. Acoust. Soc. Am.*, vol. 128, no. 5, pp. 3052–3063, 2010. (Cited on page 6.)

[51] B. G. Shinn-Cunningham, N. Kopco, and T. J. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *J. Acoust. Soc. Am.*, vol. 117, no. 5, pp. 3100–3115, 2005. (Cited on page 6.)

[52] S. Klockgether and S. van de Par, "Just noticeable differences of spatial cues in echoic and anechoic acoustical environments," *J. Acoust. Soc. Am.*, vol. 140, no. 4, pp. EL352–EL357, 2016. (Cited on page 6.)

[53] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 1633–1654, 1999. (Cited on page 6.)

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

[55] A. W. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," *Nature*, vol. 397, pp. 517–520, 1999. (Cited on page 6.)

[56] M. Barron and A. H. Marshall, "Spatial Impression Due to Early Lateral Reflections in Concert Halls: The Derivation of a Physical Measure," *Journal of Sound and Vibration*, vol. 77, no. 2, pp. 211–232, 1981. (Cited on page 7.)

[57] P. Zahorik, "Estimating Sound Source Distance with and without Vision," *Optometry and Vision Science*, vol. 78, no. 5, pp. 270–275, 2001. (Cited on page 7.)

[58] C. V. Jackson, "Visual Factors in Auditory Localization," *Quarterly Journal of Experimental Psychology*, vol. 5, no. 2, pp. 52–65, 1953. (Cited on pages 7 and 80.)

[59] E. Hendrickx, M. Paquier, V. Koehl, and J. Palacino, "Ventriloquism effect with sound stimuli varying in both azimuth and elevation," *J. Acoust. Soc. Am.*, vol. 138, no. 6, pp. 3686–3697, 2015. (Cited on page 7.)

[60] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen, "Transfer Characteristics of Headphones Measured on Human Ears," *J. Audio Eng. Soc.*, vol. 43, no. 4, pp. 203–217, 1995. (Cited on page 8.)

[61] D. Pralong and S. Carlile, "The role of individualized headphone calibration for the generation of high fidelity virtual auditory space," *J. Acoust. Soc. Am.*, vol. 100, no. 6, pp. 3785–3793, 1996. (Cited on page 8.)

[62] I. Toshima, S. Aoki, and T. Hirahara, "Sound Localization Using an Acoustical Telepresence Robot: TeleHead II," *Presence*, vol. 17, no. 4, pp. 392–404, 2008. (Cited on page 8.)

[63] D. Hammershøi and H. Møller, "Sound transmission to and within the human ear canal," *J. Acoust. Soc. Am.*, vol. 100, no. 1, pp. 408–427, 1996. (Cited on page 9.)

[64] S. Köhler, M. Blau, S. van de Par, and E. Rasumow, "Simultane Messung mehrerer HRTFs in nichtreflexionsarmer Umgebung," *Proc. Fortschritte der Akustik - DAGA, Oldenburg*, pp. 202–203, 2014. (Cited on pages 10 and 117.)

[65] S. Müller and P. Massarani, "Transfer-Function Measurement with Sweeps," *J. Audio Eng. Soc.*, vol. 49, no. 6, pp. 443–471, 2001. (Cited on page 10.)

[66] P. Majdak, P. Balazs, and B. Laback, "Multiple Exponential Sweep Method for Fast Measurement of Head-Related Transfer Functions," *J. Audio Eng. Soc.*, vol. 55, no. 7-8, pp. 623–637, 2007. (Cited on pages 10, 65, and 117.)

[67] G. Enzner, "3D-Continuous-Azimuth Acquisition of Head-Related Impulse Responses using Multi-Channel Adaptive Filtering," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 18-21, New Paltz, NY*, 2009. (Cited on page 10.)

[68] S. Li and J. Peissig, "Measurement of Head-Related Transfer Functions: A Review," *Applied Sciences*, vol. 10, p. 5014, 2020. (Cited on page 10.)

[69] H. Ziegelwanger, P. Majdak, and W. Kreuzer, "Numerical calculation of listener-specific head-related transfer functions and sound localization: Microphone model and mesh discretization," *J. Acoust. Soc. Am.*, vol. 138, no. 1, pp. 208–222, 2015. (Cited on page 10.)

[70] B. F. G. Katz, "Boundary element method calculation of individual head-related transfer function," *J. Acoust. Soc. Am.*, vol. 110, no. 5, pp. 2440–2455,

2001. (Cited on page 10.)

[71] A. Kulkarni, S. K. Isabelle, and H. S. Colburn, "Sensitivity of human subjects to head-related transfer-function phase spectra," *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2821–2840, 1999. (Cited on page 10.)

[72] F. Asano, Y. Suzuki, and T. Sone, "Role of spectral cues in median plane localization," *J. Acoust. Soc. Am.*, vol. 88, no. 1, pp. 159–168, 1990. (Cited on pages 11, 30, and 108.)

[73] K. Iida, M. Itoh, A. Itagaki, and M. Morimoto, "Median plane localization using a parametric model of the head-related transfer function based on spectral cues," *Applied Acoustics*, vol. 68, pp. 835–850, 2007. (Cited on pages 11 and 30.)

[74] H. Takemoto, P. Mokhtari, H. Kato, R. Nishimura, and K. Iida, "Mechanism for generating peaks and notches of head-related transfer functions in the median plane," *J. Acoust. Soc. Am.*, vol. 132, no. 6, pp. 3832–3841, 2012. (Cited on pages 11 and 30.)

[75] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF Database," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 21-24, New Paltz, New York*, pp. 99–102, 2001. (Cited on page 11.)

[76] M. Kleiner, B. I. Dalenbäck, and P. Svensson, "Auralization - An Overview," *J. Audio Eng. Soc.*, vol. 41, no. 11, pp. 861–875, 1993. (Cited on page 11.)

[77] M. Vorländer, *Auralization. Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*, 1st ed. Berlin, Heidelberg: Springer, 2008. (Cited on pages 11 and 57.)

[78] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *J. Acoust. Soc. Am.*, vol. 138, no. 2, pp. 708–730, 2015. (Cited on pages 11 and 57.)

[79] T. Sakuma, S. Sakamoto, and T. Otsuru, *Computational Simulation in Architectural and Environmental Acoustics. Methods and Applications of Wave-Based Computation.* Japan: Springer, 2014. (Cited on page 11.)

[80] T. Wendt, S. van de Par, and S. D. Ewert, "A Computationally-Efficient and Perceptually-Plausible Algorithm for Binaural Room Impulse Response Simulation," *J. Audio Eng. Soc.*, vol. 62, no. 11, pp. 748–766, 2014. (Cited on pages 11, 57, and 58.)

[81] S. J. Schlecht and E. A. P. Habets, "Feedback Delay Networks: Echo Density and Mixing Time," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 374–383, 2017. (Cited on page 11.)

[82] N. I. Durlach, A. Rigopulos, X. D. Pang, W. S. Woods, A. Kulkarni, H. S. Colburn, and E. M. Wenzel, "On the Externalization of Auditory Images," *Presence*, vol. 1, no. 2, pp. 251–257, 1992. (Cited on page 11.)

[83] W. M. Hartmann and A. Wittenberg, "On the externalization of sound images," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3678–3688, 1996. (Cited on pages 11, 12, 33, 39, 91, and 98.)

[84] D. R. Begault, "Perceptual Effects of Synthetic Reverberation on Three-Dimensional Audio Systems," *J. Audio Eng. Soc.*, vol. 40, no. 11, pp. 895–904, 1992. (Cited on page 12.)

[85] V. Best, R. Baumgartner, M. Lavandier, P. Majdak, and N. Kopčo, "Sound Externalization: A Review of Recent Research," *Trends in Hearing*, vol. 24, 2020. (Cited on page 12.)

[86] J. Catic, S. Santurette, J. M. Buchholz, F. Gran, and T. Dau, "The effect of interaural-level-difference fluctuations on the externalization of sound," *J. Acoust. Soc. Am.*, vol. 134, no. 2, pp. 1232–1241, 2013. (Cited on page 12.)

[87] T. Leclère, M. Lavandier, and F. Perrin, "On the externalization of sound sources with headphones without reference to a real source," *J. Acoust. Soc. Am.*, vol. 146, no. 4, pp. 2309–2320, 2019. (Cited on pages 12 and 104.)

[88] 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," *J. Acoust. Soc. Am.*, vol. 139, no. 5, pp. 2992–3000, 2016. (Cited on page 12.)

[89] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, "A Summary on Acoustic Room Divergence and its Effect on Externalization of Auditory Events," *Proc. IEEE 8th International Conference on Quality of Multimedia Experience - QoMEX*, pp. 1–6, 2016. (Cited on pages 12, 80, and 107.)

[90] J. C. Gil-Carvajal, J. Cubick, S. Santurette, and T. Dau, "Spatial Hearing with Incongruent Visual or Auditory Room Cues," *Scientific Reports*, vol. 6:37342, 2016. (Cited on pages 12, 80, and 107.)

[91] C. Pörschmann, P. Stade, and J. M. Arend, "Binaural auralization of proposed room modifications based on measured omnidirectional room impulse responses," *Proc. Mtgs. Acoust.*, vol. 30, 015012, 2017. (Cited on page 12.)

[92] M. Blau, A. Budnik, M. Fallahi, H. Steffens, S. D. Ewert, and S. van de Par, "Toward realistic binaural auralizations - perceptual comparison between measurement and simulation-based auralizations and the real room for a classroom scenario," *Acta Acustica*, vol. 5, no. 8, 2021. (Cited on pages 12, 13, 47, 55, 58, 66, 68, 77, 79, 98, and 99.)

[93] A. Lindau and S. Weinzierl, "Assessing the Plausibility of Virtual Acoustic Environments," *Acta Acustica united with Acustica*, vol. 98, pp. 804–810, 2012. (Cited on pages 12, 13, and 57.)

[94] F. Brinkmann, A. Lindau, and S. Weinzierl, "On the authenticity of individual dynamic binaural synthesis," *J. Acoust. Soc. Am.*, vol. 142, no. 4, pp. 1784–1795, 2017. (Cited on pages 12, 13, 57, and 79.)

[95] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 94, no. 1, pp. 111–123, 1993. (Cited on pages 12, 33, 83, and 99.)

[96] A. W. Bronkhorst, "Localization of real and virtual sound sources," *J. Acoust. Soc. Am.*, vol. 98, no. 5, pp. 2542–2553, 1995. (Cited on pages 12, 84, and 98.)

[97] D. R. Begault and E. M. Wenzel, "Headphone Localization of Speech," *Human Factors*, vol. 35, no. 2, pp. 361–376, 1993. (Cited on pages 12 and 107.)

[98] J. Oberem, J. G. Richter, D. Setzer, J. Seibold, I. Koch, and J. Fels, "Experiments on localization accuracy with non-individual and individual HRTFs comparing static and dynamic reproduction methods," *bioRxiv*, 2020, https://doi.org/10.1101/2020.03.31.011650. (Cited on pages 12, 13, 80, 83, 98, 99, 104, and 105.)

[99] A. Lindau, H.-J. Maempel, and S. Weinzierl, "Minimum BRIR grid resolution for dynamic binaural synthesis," *Proc. Acoustics 08 Paris*, pp. 3851–3856, 2008. (Cited on pages 13, 54, and 115.)

[100] P. Stitt, E. Hendrickx, J. C. Messonnier, and B. F. G. Katz, "The Influence of Head Tracking Latency on Binaural Rendering in Simple and Complex Sound Scenes," *Proc. AES 140th Convention, Paris, France*, 2016. (Cited on pages 13 and 56.)

[101] A. Lindau, "The Perception of System Latency in Dynamic Binaural Synthesis," *Proc. Fortschritte der Akustik - DAGA, Rotterdam*, pp. 1063–1066, 2009. (Cited on pages 13 and 56.)

[102] M. Paquier and V. Koehl, "Discriminability of the placement of supra-aural and circumaural headphones," *Applied Acoustics*, vol. 93, pp. 130–139, 2015. (Cited on page 13.)

[103] J. Oberem, B. Masiero, and J. Fels, "Experiments on authenticity and plausibility of binaural reproduction via headphones employing different recording methods," *Applied Acoustics*, vol. 114, pp. 71–78, 2016. (Cited on page 13.)

[104] V. R. Algazi, R. O. Duda, and D. M. Thompson, "Motion-Tracked Binaural Sound," *J. Audio Eng. Soc.*, vol. 52, no. 11, pp. 1142–1156, 2004. (Cited on pages 14 and 116.)

[105] J. B. Melick, V. R. Algazi, R. O. Duda, and D. M. Thompson, "Customization for Personalized Rendering of Motion-Tracked Binaural Sound," *Proc. AES 117th Convention, San Francisco, CA, USA*, 2004. (Cited on pages 14 and 116.)

[106] B. Bernschütz, "Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording," *Ph.D. thesis, Technical University of Berlin*, 2016. (Cited on pages 14, 15, and 116.)

[107] J. Ahrens and C. Andersson, "Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre," *J. Acoust. Soc. Am.*, vol. 145, no. 4, pp. 2783–2794, 2019. (Cited on pages 14, 15, and 116.)

[108] J. Atkins, "Robust Beamforming and Steering of Arbitrary Beam Patterns using Spherical Arrays," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 16-19, New Paltz, NY*, pp. 237–240, 2011. (Cited on pages 14, 15, 17, 36, and 116.)

[109] D. N. Zotkin, R. Duraiswami, and N. A. Gumerov, "Regularized HRTF Fitting Using Spherical Harmonics," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 18-21, New Paltz, NY*, pp. 257–260, 2009. (Cited on pages 14, 15, and 116.)

[110] M. A. Poletti and U. P. Svensson, "Beamforming synthesis of binaural responses from computer simulations of acoustic spaces," *J. Acoust. Soc. Am.*, vol. 124, no. 1, pp. 301–315, 2008. (Cited on pages 14, 15, and 116.)

[111] C. D. S. Castaneda, S. Sakamoto, J. A. T. Lopez, J. Li, Y. Yan, and Y. Suzuki, "Accuracy of head-related transfer functions synthesized with spherical microphone arrays," *Proc. Mtgs. Acoust*, vol. 19, 055085, 2013. (Cited on pages 14, 15, and 116.)

[112] A. Avni, J. Ahrens, M. Geier, S. Spors, H. Wierstorf, and B. Rafaely, "Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. 2711–2721, 2013. (Cited on pages 14, 15, and 116.)

[113] M. Zaunschirm, M. Frank, and F. Zotter, "BRIR synthesis using first-order microphone arrays," *Proc. AES 144th Convention, Milan, Italy*, 2018. (Cited on pages 14, 15, 16, and 116.)

[114] V. Pulkki, "Spatial Sound Reproduction with Directional Audio Coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, 2007. (Cited on pages 14, 16, and 116.)

[115] J. Chen, B. D. Van Veen, and K. E. Hecox, "External ear transfer function modeling: A beamforming approach," *J. Acoust. Soc. Am.*, vol. 92, no. 4, pp. 1933–1944, 1992. (Cited on pages 14 and 16.)

[116] N. Tohtuyeva and V. Mellert, "Approximation of dummy-head recording technique by a multimicrophone arrangement," *J. Acoust. Soc. Am.*, vol. 105, no. 2, pp. 1101–1101, 1999. (Cited on pages 14 and 16.)

[117] S. Sakamoto, S. Hongo, T. Okamoto, Y. Iwaya, and Y. Suzuki, "Sound-space recording and binaural presentation system based on a 252-channel microphone array," *Acoust. Sci.& Tech.*, vol. 36, no. 6, pp. 516–526, 2015. (Cited on pages 14, 16, 18, 29, and 36.)

[118] A. Lindau and S. Roos, "Perceptual evaluation of discretization and interpolation for motion-tracked binaural (MTB) recordings," *Proc. 26. Tonmeistertagung, VDT International Convention, Leipzig, Germany*, pp. 680–701, 2010. (Cited on page 14.)

[119] D. Ackermann, F. Fiedler, F. Brinkmann, M. Schneider, and S. Weinzierl, "On the Acoustic Qualities of Dynamic Pseudobinaural Recordings," *J. Audio Eng. Soc.*, vol. 68, no. 6, pp. 418–427, 2020. (Cited on pages 14, 15, 80, and 99.)

[120] B. Rafaely, *Fundamentals of Spherical Array Processing*. Berlin, Heidelberg: Springer, 2015, vol. 8. (Cited on page 15.)

[121] H. Chen, T. D. Abhayapala, and W. Zhang, "Theory and design of compact hybrid microphone arrays on two-dimensional planes for three-dimensional soundfield analysis," *J. Acoust. Soc. Am.*, vol. 138, no. 5, pp. 3081–3092, 2015. (Cited on page 15.)

[122] P. N. Samarasinghe, H. Chen, A. Fahim, and T. D. Abhayapala, "Performance Analysis of A Planar Microphone Array for Three Dimensional Soundfield Analysis," *IEEE Workshop an Applications of Signal Processing to Audio and Acoustics, October 15-18, New Paltz, NY*, pp. 249–253, 2017. (Cited on

page 15.)

[123] B. Rafaely and A. Avni, "Interaural cross correlation in a sound field represented by spherical harmonics," *J. Acoust. Soc. Am.*, vol. 127, no. 2, pp. 823–828, 2010. (Cited on page 15.)

[124] M. J. Evans, J. A. S. Angus, and A. I. Tew, "Analyzing head-related transfer function measurements using surface spherical harmonics," *J. Acoust. Soc. Am.*, vol. 104, no. 4, pp. 2400–2411, 1998. (Cited on page 15.)

[125] G. D. Romigh, D. S. Brungart, R. M. Stern, and B. D. Simpson, "Efficient Real Spherical Harmonic Representation of Head-Related Transfer Functions," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 921–930, 2015. (Cited on page 15.)

[126] J. Sheaffer, S. Villeval, and B. Rafaely, "Rendering Binaural Room Impulse Responses From Spherical Microphone Array Recordings Using Timbre Correction," *Proc. EAA Joint Symposium on Auralization and Ambisonics, April 3-5, Berlin*, 2014. (Cited on page 15.)

[127] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, and B. Rafaely, "Spectral equalization in binaural signals represented by order-truncated spherical harmonics," *J. Acoust. Soc. Am.*, vol. 141, no. 6, pp. 4087–4096, 2017. (Cited on page 15.)

[128] C. Hold, H. Gamper, V. Pulkki, N. Raghuvanshi, and I. J. Tashev, "Improving Binaural Ambisonics Decoding by Spherical Harmonics Domain Tapering and Coloration Compensation," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP*, pp. 261–265, 2019. (Cited on page 15.)

[129] T. Lübeck, H. Helmholz, J. M. Arend, C. Pörschmann, and J. Ahrens, "Perceptual Evaluation of Mitigation Approaches of Impairments Due to Spatial Undersampling in Binaural Rendering of Spherical Microphone Array Data," *J. Audio Eng. Soc.*, vol. 68, no. 6, pp. 428–440, 2020. (Cited on page 15.)

[130] M. .Zaunschirm, M. Frank, and F. Zotter, "Binaural Rendering with Measured Room Responses: First-Order Ambisonic Microphone vs. Dummy Head," *Applied Sciences*, vol. 10, p. 1631, 2020. (Cited on pages 15 and 16.)

[131] F. Zotter and M. Frank, *Ambisonics. A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality.* Springer Nature, 2019, vol. 19. (Cited on pages 15 and 16.)

[132] M. V. Laitinen and V. Pulkki, "Binaural Reproduction For Directional Audio Coding," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 18-21, New Paltz, NY*, pp. 337–340, 2009. (Cited on page 16.)

[133] A. Politis, J. Vilkamo, and V. Pulkki, "Sector-Based Parametric Sound Field Reproduction in the Spherical Harmonic Domain," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 852–866, 2015. (Cited on page 16.)

[134] A. Politis, L. McCormack, and V. Pulkki, "Enhancement of Ambisonic Binaural Reproduction using Directional Audio Coding with Optimal Adaptive

Mixing," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 15-18, New Paltz, NY*, 2017. (Cited on page 16.)

[135] B. D. Van Veen and K. M. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1998. (Cited on page 16.)

[136] J. Bitzer and K. U. Simmer, "Superdirective Microphone Arrays," in *Microphone arrays: Signal processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Berlin Heidelberg: Springer, 2001, pp. 19–38. (Cited on pages 16, 17, 29, and 36.)

[137] H. Cox, R. M. Zeskind, and T. kooij, "Practical Supergain," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, no. 3, pp. 393–398, 1986. (Cited on page 16.)

[138] S. Doclo and M. Moonen, "Design of far-field and near-field broadband beamformers using eigenfilters," *Signal Processing*, vol. 83, pp. 2641–2673, 2003. (Cited on pages 16 and 116.)

[139] ——, "Design of Broadband Beamformers Robust Against Gain and Phase Errors in the Microphone Array Characteristics," *IEEE Transactions on Signal Processing*, vol. 51, no. 10, pp. 2511–2526, 2003. (Cited on pages 17, 27, 28, and 36.)

[140] E. Rasumow, M. Blau, S. Doclo, M. Hansen, S. van de Par, D. Püschel, and V. Mellert, "Least squares versus non-linear cost functions for a virtual artificial head," *Proc. Mtgs. Acoust*, vol. 19, 055082, 2013. (Cited on page 17.)

[141] S. Doclo and M. Moonen, "Superdirective Beamforming Robust Against Microphone Mismatch," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 617–631, 2007. (Cited on pages 17, 28, 29, and 36.)

[142] E. Rasumow, M. Blau, M. Hansen, S. Doclo, S. van de Par, V. Mellert, and D. Püschel, "The Impact of the White Noise Gain (WNG) of a Virtual Artificial Head on the Appraisal of Binaural Sound Reproduction," *Proc. EAA Joint Symposium on Auralization and Ambisonics, April 3-5, Berlin*, 2014. (Cited on pages 17, 25, 29, 30, 34, 36, 38, 41, 60, 79, and 85.)

[143] E. Rasumow, "Synthetic reproduction of head-related transfer functions by using microphone arrays," *Ph.D. thesis, Carl von Ossietzky University, Oldenburg*, 2015. (Cited on pages 17, 27, 28, 29, and 30.)

[144] S. W. Golomb and H. Taylor, "Two-Dimensional Synchronization Patterns for Minimum Ambiguity," *IEEE Transactions on Information Theory*, vol. 28, no. 4, pp. 600–604, 1982. (Cited on pages 18, 43, and 85.)

[145] E. Rasumow, M. Blau, M. Hansen, S. Doclo, S. van de Par, V. Mellert, and D. Püschel, "Robustness of virtual artificial head topologies with respect to microphone positioning," *Proc. Forum Acusticum, Aalborg, Denmark*, pp. 2251–2256, 2011. (Cited on pages 18 and 29.)

[146] S. Sakamoto, J. Kodama, S. Hongo, T. Okamoto, Y. Iwaya, and Y. Suzuki, "A 3D sound-space recording system using spherical microphone array with 252ch microphones," *Proc. 20th International Congress on Acoustics - ICA,*

August 23-27, Sydney, Australia, 2010. (Cited on page 18.)

[147] C. D. Salvador, S. Sakamoto, J. Treviño, and Y. Suzuki, "Design theory for binaural synthesis: Combining microphone array recordings and head-related transfer function datasets," *Acoust. Sci.& Tech.*, vol. 38, no. 2, pp. 51–62, 2017. (Cited on page 18.)

[148] S. Sakamoto, C. Salvador, J. Treviño, and Y. Suzuki, "Binaural synthesis using a spherical microphone array based on the solution to an inverse problem," *Proc. INTER-NOISE, June 16-19, Madrid*, 2019. (Cited on page 18.)

[149] M. Fallahi, M. Hansen, S. Doclo, S. van de Par, D. Püschel, and M. Blau, "Individual binaural reproduction with high spatial resolution using a virtual artificial head with a moderate number of microphones," in preparation. (Cited on pages 21 and 33.)

[150] M. Fallahi, M. Hansen, S. Doclo, S. van de Par, V. Mellert, D. Püschel, and M. Blau, "High spatial resolution binaural sound reproduction using a virtual artificial head," *Proc. Fortschritte der Akustik - DAGA, Kiel*, pp. 1061–1064, 2017. (Cited on pages 21 and 33.)

[151] M. Fallahi, M. Hansen, S. Doclo, S. van de Par, D. Püschel, and M. Blau, "Evaluation of head-tracked binaural auralizations of speech signals generated with a virtual artificial head in anechoic and classroom environments," *Acta Acustica*, vol. 5, no. 30, 2021. (Cited on pages 22 and 57.)

[152] ——, "Binaural Reproduction of Signals captured in a reverberant Room with a Virtual Artificial Head," *Proc. Fortschritte der Akustik - DAGA, Rostock*, pp. 619–622, 2019. (Cited on pages 22 and 57.)

[153] ——, "Individualized dynamic binaural auralization of classroom acoustics using a virtual artificial head," *Proc. 23rd International Congress on Acoustics - ICA, Aachen, Germany*, pp. 731–738, 2019. (Cited on pages 22 and 57.)

[154] ——, "Dynamic Binaural Rendering: The Advantage of Virtual Artificial Heads Over Conventional Ones For Localization With Speech Signals," *Applied Sciences*, vol. 11, p. 6793, 2021. (Cited on pages 23 and 83.)

[155] M. Fallahi, M. Hansen, S. van de Par, S. Doclo, D. Püschel, and M. Blau, "Localization Performance in the Absence of Visual Cues for Binaural Renderings generated with a Virtual Artificial Head," *Proc. Fortschritte der Akustik - DAGA, Hanover*, pp. 106–109, 2020. (Cited on pages 23 and 83.)

[156] ——, "Localization Performance for Binaural Signals Generated with a Virtual Artificial Head in the Absence of Visual Cues," *Proc. e-Forum Acusticum, Lyon, France*, pp. 1937–1944, 2020. (Cited on pages 23 and 83.)

[157] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Source of Collinearity.* John Wiley & Sons, 2004. (Cited on page 28.)

[158] M. R. Bai, C. Chung, P. C. Wu, Y. H. Chiang, and C. M. Yang, "Solution Strategies for Linear Inverse Problems in Spatial Audio Signal Processing," *Applied Sciences*, vol. 7, p. 582, 2017. (Cited on pages 29 and 36.)

[159] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust Adaptive Beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-35,

no. 10, pp. 1365–1376, 1987. (Cited on page 29.)

[160] M. R. Bai and C. Lin, "Microphone array signal processing with application in three-dimensional spatial hearing," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2112–2121, 2005. (Cited on page 29.)

[161] M. Geronazzo, S. Spagnol, and F. Avanzini, "Do We Need Individual Head-Related Transfer Functions for Vertical Localization? The Case Study of a Spectral Notch Distance Metric," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1247–1260, 2018. (Cited on page 30.)

[162] B. C. J. Moore, S. R. Oldfield, and G. J. Dooley, "Detection and discrimination of spectral peaks and notches at 1 and 8kHz," *J. Acoust. Soc. Am.*, vol. 85, no. 2, pp. 820–836, 1989. (Cited on page 30.)

[163] E. A. Macpherson and A. T. Sabin, "Binaural weighting of monaural spectral cues for sound localization," *J. Acoust. Soc. Am.*, vol. 121, no. 6, pp. 3677–3688, 2007. (Cited on page 31.)

[164] R. Baumgartner, P. Majdak, and B. Laback, "Modeling sound-source localization in sagittal planes for human listeners," *J. Acoust. Soc. Am.*, vol. 136, no. 2, pp. 791–802, 2014. (Cited on page 31.)

[165] B. Rafaely, "Analysis and Design of Spherical Microphone Arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 135–143, 2005. (Cited on page 36.)

[166] S. Spors, H. Wierstorf, and M. Geier, "Comparison of modal versus delay-and-sum beamforming in the context of data-based binaural synthesis," *Proc. AES 132nd Convention, Budapest, Hungary*, 2012. (Cited on page 36.)

[167] P. Minnaar, J. Plogsties, and F. Christensen, "Directional Resolution of Head-Related Transfer Functions Required in Binaural Synthesis," *J. Audio Eng. Soc.*, vol. 53, no. 10, pp. 919–929, 2005. (Cited on page 39.)

[168] F. Brinkmann, R. Roden, A. Lindau, and S. Weinzierl, "Audibility of head-above-torso orientation in head-related transfer functions," *Proc. Forum Acusticum, September 7-12, Kraków*, 2014. (Cited on page 39.)

[169] F. Brinkmann and S. Weinzierl, "Comparison of head-related transfer functions pre-processing techniques for spherical harmonics decomposition," *Proc. AES Conference on Audio for Virtual and Augmented Reality, August 20-22, Redmond, WA, USA*, 2018. (Cited on page 39.)

[170] R. Bücklein, "The Audibility of Frequency Response Irregularities," *J. Audio Eng. Soc.*, vol. 29, no. 3, pp. 126–131, 1981. (Cited on page 40.)

[171] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. Springer, 2006. (Cited on page 41.)

[172] Eigenmike EM32, mh Acoustics
https://mhacoustics.com (Last viewed June 22, 2021). (Cited on page 43.)

[173] B. Bernschütz, C. Pörschmann, S. Spors, and S. Weinzierl, "SOFiA Sound Field Analysis Toolbox," *Proc. International Conference on Spatial Audio - ICSA, November 10-13, Detmold, Germany*, 2011. (Cited on page 43.)

[174] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *J. Acoust. Soc. Am.*, vol. 104, no. 5, pp. 3048–3058, 1998. (Cited on page 43.)

[175] ITU P.800, ITU-T Recommendations https://www.itu.int/ITU-T/recommendations/index.aspx (Last viewed June 22, 2021). (Cited on pages 47 and 66.)

[176] P. Chevret and E. Parizet, "An Efficient Alternative to the Paired Comparison Method for the Subjective Evaluation of a Large Set of Sounds," *Proc. 19th International Congress on Acoustics, September 2-7, Madrid*, 2007. (Cited on page 47.)

[177] L. J. Cronbach, "Coefficient Alpha and the Internal Structure of Tests," *Psychometrika*, vol. 16, no. 3, pp. 297–334, 1951. (Cited on pages 48 and 71.)

[178] S. Siegel and N. J. Castellan, *Non parametric statistics for the behavioural sciences*, 2nd ed. McGraw-Hill, Inc., 1988. (Cited on pages 48, 73, and 95.)

[179] H. Jaeger and U. Simmer, "TVOLAP: Time-Variant Overlap-Add in Partitions," 2017, https://github.com/TGM-Oldenburg/TVOLAP (Last viewed June 22, 2021). (Cited on page 55.)

[180] H. Jaeger, J. Bitzer, U. Simmer, and M. Blau, "Echtzeitfähiges binaurales Rendering mit Bewegungssensoren von 3D-Brillen," *Proc. Fortschritte der Akustik - DAGA, Kiel*, pp. 1130–1133, 2017. (Cited on page 55.)

[181] D. S. Brungart, A. J. Kordik, and B. D. Simpson, "Effects of Headtracker Latency in Virtual Audio Displays," *J. Audio Eng. Soc.*, vol. 54, no. 1-2, pp. 32–44, 2006. (Cited on page 56.)

[182] P. Majdak, Y. Iwaya, T. Carpentier, R. Nicol, M. Parmentier, A. Roginska, Y. Suzuki, K. Watanabe, H. Wierstorf, H. Ziegelwanger, and M. Noisternig, "Spatially Oriented Format for Acoustics: A Data Exchange Format Representing Head-Related Transfer Functions," *Proc. AES 134th Convention, Rome, Italy*, 2013. (Cited on pages 56 and 119.)

[183] A. Novák, L. Simon, F. Kadlec, and P. Lotton, "Nonlinear System Identification Using Exponential Sweep-Sine Signal," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 8, pp. 2220–2229, 2010. (Cited on pages 65 and 117.)

[184] "International Phonetic Association and International Phonetic Association Staff: Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet. Cambridge University Press,," 1999. (Cited on page 67.)

[185] W. Gaik, "Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling," *J. Acoust. Soc. Am.*, vol. 94, no. 1, pp. 98–110, 1993. (Cited on page 77.)

[186] M. Blau, "Correlation of Apparent Source Width with Objective Measures in Synthetic Sound Fields," *Acta Acustica united with Acustica*, vol. 90, pp. 720–730, 2004. (Cited on page 77.)

[187] F. Wendt, R. Höldrich, and M. Marschall, "How binaural room impulse responses influence the externalization of speech," *Proc. Fortschritte der Akustik*

- *DAGA, Rostock*, pp. 627–630, 2019. (Cited on page 79.)

[188] F. Völk, F. Heinemann, and H. Fastl, "Externalization in binaural synthesis: effects of recording environment and measurement procedure," *Proc. Acoustics 08, Paris, France*, pp. 6419–6424, 2008. (Cited on page 79.)

[189] G. D. Romigh, D. S. Brungart, and B. D. Simpson, "Free-Field Localization Performance With a Head-Tracked Virtual Auditory Display," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 943–954, 2015. (Cited on page 84.)

[190] J. Sandvad, "Dynamic Aspects of Auditory Virtual Environments," *Proc. AES 100th Convention, Copenhagen*, 1996. (Cited on page 84.)

[191] O. Kirkeby and P. A. Nelson, "Digital Filter Design for Inversion Problems in Sound Reproduction," *J. Audio Eng. Soc.*, vol. 47, no. 7-8, pp. 583–595, 1999. (Cited on pages 89, 118, and 119.)

[192] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. II: Psychophysical validation," *J. Acoust. Soc. Am.*, vol. 85, no. 2, pp. 868–878, 1989. (Cited on page 98.)

[193] E. A. Macpherson, "Cue weighting and vestibular mediation of temporal dynamics in sound localization via head rotation," *Proc. Mtgs. Acoust.*, vol. 19, 050131, 2013. (Cited on page 98.)

[194] F. L. Wightman and D. J. Kistler, "Resolution of front-back ambiguity in spatial hearing by listener and source movement," *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2841–2853, 1999. (Cited on pages 104 and 105.)

[195] N. Iyer, E. R. Thompson, and B. D. Simpson, "Response Techniques and Auditory Localization Accuracy," *Proc. 22nd International Conference on Auditory Display - ICAD, July 2-8, Canberra, Australia*, 2016. (Cited on page 107.)

[196] D. J. Folds, "The Elevation Illusion in Virtual Audio," *Proc. HUMAN FACTORS AND ERGONOMICS SOCIETY 50th ANNUAL MEETING*, pp. 1576–1580, 2006. (Cited on page 107.)

[197] J. C. Middlebrooks, "Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1493–1510, 1999. (Cited on page 107.)

[198] H. Helmholz, J. Ahrens, D. L. Alon, S. V. A. Gari, and R. Mehra, "Evaluation of Sensor Self-Noise in Binaural Rendering of Spherical Microphone Array Signals," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP*, pp. 161–165, 2020. (Cited on page 115.)

[199] S. Carlile and J. Leung, "The Perception of Auditory Motion," *Trends in Hearing*, vol. 20, 2016. (Cited on page 116.)

[200] W. O. Brimijoin and M. A. Akeroyd, "The moving minimum audible angle is smaller during self motion than during source motion," *Frontiers in neuroscience*, vol. 8, no. 273, 2014. (Cited on page 116.)

[201] F. Féron, I. Frissen, J. Boissinot, and C. Guastavino, "Upper limits of auditory rotational motion perception," *J. Acoust. Soc. Am.*, vol. 128, no. 6, pp. 3703–3714, 2010. (Cited on page 116.)

[202] S. Spagnol, "On distance dependence of pinna spectral patterns in head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 137, no. 1, pp. EL58–EL64, 2015. (Cited on page 116.)

[203] G. Yu, R. Wu, Y. Liu, and B. Xie, "Near-field head-related transfer-function measurement and database of human subjects," *J. Acoust. Soc. Am.*, vol. 143, no. 3, pp. EL194–EL198, 2018. (Cited on page 116.)

[204] J. M. Arend and C. Pörschmann, "Synthesis of Near-Field HRTFs by Directional Equalization of Far-Field Datasets," *Proc. Fortschritte der Akustik - DAGA, Rostock*, pp. 1454–1457, 2019. (Cited on page 116.)

[205] S. T. Prepeliţă, J. G. Bolaños, V. Pulkki, L. Savioja, and R. Mehra, "Numerical simulations of near-field head-related transfer functions: Magnitude verification and validation with laser spark sources," *J. Acoust. Soc. Am.*, vol. 148, no. 1, pp. 153–166, 2020. (Cited on page 116.)

[206] D. S. Brungart, "Auditory Parallax Effects in the HRTF for Nearby Sources," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 17-20, New Paltz, New York*, pp. 171–174, 1999. (Cited on page 116.)

[207] J. Poppitz, M. Blau, and M. Hansen, "Entwicklung und Evaluation eines Systems zur Messung individueller HRTFs in privater Wohn-Umgebung," *Proc. Fortschritte der Akustik - DAGA, Aachen*, pp. 812–815, 2016. (Cited on page 118.)

[208] A. Budnik, "Untersuchungen zur dynamischen binauralen Auralisierung von Klassenraumakustik unter Berücksichtigung des Quellenabstrahlverhaltens," *Master Thesis - Carl von Ossietzky University Oldenburg*, 2019. (Cited on page 118.)

[209] M. Fallahi, M. Blau, M. Hansen, S. Doclo, S. van de Par, and D. Püschel, "Constrained optimization for binaural sound reproduction using a virtual artificial head," *Proc. Fortschritte der Akustik - DAGA, Munich*, pp. 691–694, 2018. (Cited on page 121.)

[210] M. Fallahi, M. Hansen, S. Doclo, S. van de Par, D. Püschel, and M. Blau, "Individual binaural reproduction of music recordings using a virtual artificial head," *Proc. AES Conference on Spatial Reproduction, August 6-9, Tokyo, Japan*, 2018. (Cited on page 121.)

[211] C. Avendano, R. O. Duda, and V. R. Algazi, "Modeling the Contralateral HRTF," *Proc. AES 16th International Conference on Spatial Sound Reproduction*, 1999. (Cited on page 121.)

[212] K. Watanabe, R. Kodama, S. Sato, S. Takane, and K. Abe, "Influence of flattening contralateral head-related transfer functions upon sound localization performance," *Acoust. Sci.& Tech.*, vol. 32, no. 3, pp. 121–124, 2011. (Cited on page 121.)

[213] H. Ziegelwanger, W. Kreuzer, and P. Majdak, "A-priori mesh grading for the numerical calculation of the head-related transfer functions," *Applied Acoustics*, vol. 114, pp. 99–110, 2016. (Cited on pages 121 and 132.)

[214] F. Denk, S. M. A. Ernst, S. D. Ewert, and B. Kollmeier, "Adapting Hearing Devices to the Individual Ear Acoustics: Database and Target Response Cor-

rection Functions for Various Device Styles," *Trends in Hearing*, vol. 22, 2018. (Cited on page 125.)

LIST OF PUBLICATIONS

The following publications are related to the work in this thesis.

Journal Papers

[J3] **M. Fallahi**, M. Hansen, S. Doclo, S. van de Par, D. Püschel, and Matthias Blau, "Individual binaural reproduction with high spatial resolution using a virtual artificial head with a moderate number of microphones," in preparation.

[J2] **M. Fallahi**, M. Hansen, S. Doclo, S. van de Par, D. Püschel, and Matthias Blau, "Evaluation of head-tracked binaural auralizations of speech signals generated with a virtual artificial head in anechoic and classroom environments," *Acta Acustica*, vol. 5, no. 30, 2021.

[J1] **M. Fallahi**, M. Hansen, S. Doclo, S. van de Par, D. Püschel, and Matthias Blau, "Dynamic Binaural Rendering: The Advantage of Virtual Artificial Heads Over Conventional Ones For Localization With Speech Signals," *Applied Sciences*, vol. 11, pp. 6793, 2021.

Conference Papers

[C7] **M. Fallahi**, M. Hansen, S. van de Par, S. Doclo, D. Püschel, and M. Blau, "Localization Performance for Binaural Signals Generated with a Virtual Artificial Head in the Absence of Visual Cues," in *Proc. e-Forum Acusticum*, Lyon, France, pp. 1937–1944, 2020.

[C6] **M. Fallahi**, M. Hansen, S. van de Par, S. Doclo, D. Püschel, and M. Blau, "Localization Performance in the Absence of Visual Cues for Binaural Renderings generated with a Virtual Artificial Head," in *Proc. Fortschritte der Akustik - DAGA*, Hanover, Germany, pp. 106–109, 2020.

[C5] **M. Fallahi**, M. Hansen, S. Doclo, S. van de Par, D. Püschel, and M. Blau, "Individualized dynamic binaural auralization of classroom acoustics using a virtual artificial head," in *Proc. 23rd International Congress on Acoustics* -

ICA, Aachen, Germany, pp. 731–738, 2019.

[C4] **M. Fallahi**, M. Hansen, S. Doclo, S. van de Par, D. Püschel, and M. Blau, "Binaural Reproduction of Signals captured in a reverberant Room with a Virtual Artificial Head," in *Proc. Fortschritte der Akustik - DAGA*, Rostock, Germany, pp. 619–622, 2019.

[C3] **M. Fallahi**, M. Hansen, S. Doclo, S. van de Par, D. Püschel, and M. Blau, "Individual binaural reproduction of music recordings using a virtual artificial head," in *Proc. AES Conference on Spatial Reproduction*, Tokyo, Japan, 2018.

[C2] **M. Fallahi**, M. Blau, M. Hansen, S. Doclo, S. van de Par, and D. Püschel, "Constrained optimization for binaural sound reproduction using a virtual artificial head," in *Proc. Fortschritte der Akustik - DAGA*, Munich, Germany, pp. 691–694, 2018.

[C1] **M. Fallahi**, M. Hansen, S. Doclo, S. van de Par, V. Mellert, D. Püschel, and M. Blau, "High spatial resolution binaural sound reproduction using a virtual artificial head," in *Proc. Fortschritte der Akustik - DAGA*, Kiel, Germany, pp. 1061–1064, 2017.

ACKNOWLEDGMENTS

Foremost, I would like to express my gratitude to Prof. Dr. ir. Simon Doclo and Prof. Dr.-Ing. Matthias Blau for giving me the opportunity to work with them as Ph.D. candidate. It was an honor for me.

I am very grateful to Matthias Blau for the years of supervision, support, and constructive guidance throughout this work. I would like to thank him for putting his trust in me, for his expert advice, and for all the valuable things which he taught me with kindness and patience.

I would like to express my gratitude to Simon Doclo for his supervision, continuous support, immense knowledge, and many valuable suggestions during the whole work. I am thankful to him for having patience with me and teaching me a lot.

I would like to thank Doz. Dr. Piotr Majdak for taking time to read and evaluate my thesis and for participating in my thesis committee and Prof. Dr. Steven van de Par as a further member of the examination committee. Furthermore, I also like to express my sincere thanks to Steven van de Par for the worthy discussions and the thoughtful suggestions.

I am greatly thankful to Prof. Dr. Martin Hansen for the numerous fruitful discussions, for his support and for his kind encouragement.

I am also very thankful to Dr. Dirk Püschel and Prof. Dr. Volker Mellert for the valuable discussions we had. I would like to thank Dirk Püschel and Akustik Technology Göttingen for the partial financial support of the project.

The financial support by the Bundesministerium für Bildung und Forschung (grant no. 03FH021IX5) is greatly acknowledged.

Special thanks to Ali Aroudi, Bastian Bechtold, Prof. Dr. Jörg Bitzer, Armin Budnik, Holger Groenewold, Hagen Jaeger, Theresa Nüsse, Eugen Rasumow, Reinhild Roden, Tobias Sankowsky-Rothe, Dr.-Ing. Uwe Simmer, Steffen Vogl, Florian Wiese and Marco Wilmes, for their support during my Ph.D. Many special thanks to all members of the Institut für Hörtechnik und Audiologie at Jade Hochschule Oldenburg. I would also like to thank all subjects who participated in my experiments.

In particular, I like to thank all my family members, especially my parents and my sisters Delaram and Bita, and my friend Dietmar, for their mental and emotional support, encouragement, and for being always there for me.

Oldenburg, September 2021
Mina Fallahi