Modeling Sound Localization in Sagittal Planes for Human Listeners

PhD Thesis

of

Robert Baumgartner

Date:	June 3, 2015
Matriculation number:	0773071
PhD program:	Sound and Music Computing (V 094 750)
Advisor:	Prof. Dr. Robert Höldrich
Co-advisor:	Prof. Dr. Gerhard Eckel
Co-advisor:	Dr. Piotr Majdak
External reviewer:	Prof. Dr. Steven van de Par

Abstract

Sound localization in sagittal planes includes estimating the source position in up-down and front-back direction, and is based on processing monaural spectral cues. These cues are caused by the directional, acoustic filtering of the sound by the listener's morphology, as described by head-related transfer functions (HRTFs). In order to better understand the listener-specific localization process and to reduce the need for costly psychoacoustic experiments, this PhD project aims at developing a functional model for sound localization in sagittal planes. In the model, directional spectral cues of incoming sounds are compared with internal templates of cues in order to obtain a probabilistic prediction of the listener's directional response and localization performance. As directional cues, the model extracts rising spectral edges from the incoming sound spectrum. The comparison of cues is further influenced by the listener-specific sensitivity, representing the ability to process spectral cues. After extensively evaluating the model's predictive power with respect to effects of HRTF or source-signal modifications on localization performance, particular model components were undertaken hypothesis-driven analyses. Model predictions were used to explain whether the large across-listener variability in localization performance is more attributed to listener-specific HRTFs, a purely acoustic factor, or to listener-specific sensitivities, a purely non-acoustic factor. The results of a systematic permutation of the factors suggest that the non-acoustic factor influences listener-specific localization performance much more than the acoustic factor. In order to investigate the effect of extracting spectral edges as directional cues, model predictions with and without edge extraction were compared. Predictions with edge extraction showed a better correspondence with results from localization experiments and explain the robustness of spectral cues in face of macroscopic variations of the source spectrum. The potential of the model to assess the spatial sound quality of technical devices was addressed. In particular, the microphone position in hearing-assistive devices and the positioning of loudspeakers in surround sound systems applying vector-based amplitude panning was investigated. For the sake of usability and reproducibility, the implementation of the model and the simulated experiments were made publicly available at the Auditory Modeling Toolbox.

Kurzfassung

Schallquellenlokalisation in Sagittalebenen inkludiert sowohl die Schätzung der vertikalen Position einer Quelle als auch die Unterscheidung zwischen Vorne und Hinten, und basiert auf der Auswertung von monauralen spektralen Merkmalen. Diese Merkmale werden verursacht von der richtungsabhängigen, akustischen Filterwirkung der Hörermorphologie (eng.: head-related transfer functions, HRTFs). Um den stark hörerspezifischen Lokalisationsprozess besser verstehen zu lernen und die Notwendigkeit aufwändiger psychoakustischer Lokalisationsexperimente zu reduzieren, zielt dieses Dissertationsprojekt darauf ab, ein funktionales Modell für Lokalisation in Sagittalebenen zu entwickeln. Im Modell werden die spektralen Richtungsmerkmale des einfallenden Schallsignals mit intern gespeicherten Merkmalmuster verglichen, um die Richtungsantwort eines Hörers und dessen Lokalisationsleistung probabilistisch vorher zu sagen. Das Modell extrahiert steigende spektrale Flanken aus dem einfallenden Signalspektrum und verwendet diese als Richtungsmerkmale. Im Merkmalsvergleich wird zudem die hörerspezifische Sensitivität, d.h. die Fähigkeit spektrale Merkmale zu verarbeiten, berücksichtigt. Nach einer umfangreichen Evaluierung des Modells, Effekte von Modifikationen der HRTFs oder des Quellsignals auf die Lokalisationsleistung vorhersagen zu können, folgten hypothesengetriebene Analysen einzelner Modellkomponenten. Modellanalysen wurden durchgeführt um zu klären, ob die starke Variabilität der Lokalisationsleistung zwischen Hörern eher auf die hörerspezifischen HRTFs, einem rein akustischen Faktor, oder auf die hörerspezifischen Sensitivitäten, einem rein nicht-akustischen Faktor, zurück zu führen sind. Die Ergebnisse einer systematischen Faktorenpermutation weisen darauf hin, dass der nicht-akustische Faktor wesentlich mehr Einfluss auf die hörerspezifische Lokalisationsleistung hat als der akustische Faktor. Um die Auswirkung der spektralen Flankenextraktion zur Gewinnung von Richtungsmerkmalen zu untersuchen, wurden Modellvorhersagen mit und ohne Extraktionsstufe verglichen. Vorhersagen mit Extraktionsstufe zeigten dabei deutlich bessere Übereinstimmung mit experimentellen Ergebnissen als ohne. Diese Modellierungsergebnisse weisen darauf hin, dass die Flankenextraktion die Robustheit spektraler Richtungsmerkmale gegenüber makroskopischen Veränderungen des Signalspektrums erhöht. Weiters wurde das Potential des Lokalisationsmodells hinsichtlich technischer Anwendungen aufgezeigt. Konkret wurde die Mikrofonposition bei Hörhilfen sowie die Lautsprecherpositionierung in Raumklangsystemen basierend auf vektororientierter Amplitudengewichtung (eng.: vector-based amplitude panning, VBAP) untersucht. Um die Anwendung des Modells zu vereinfachen und die Reproduzierbarkeit der Modellierungsergebnisse zu gewährleisten, wurden alle Implementierungen in der Auditory Modeling Toolbox veröffentlicht.

kunst
uni
graz

(Name in Blockbuchstaben) (Matrikelnummer)

Erklärung

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

Graz, den ……………………………………….

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

Acknowledgments

I worked on this PhD thesis at the Acoustics Research Institute (ARI) of the Austrian Academy of Sciences in the context of the FWF-funded project LocaPhoto lead by Piotr Majdak. So first and foremost, I offer my sincerest gratitude to Piotr, my co-advisor from the ARI, who has supported me throughout my PhD project with his knowledge and fatherly empathy, whilst pushing my creativity and allowing me the room to work in my own way. He always provided the backing I needed, no matter whether I had been struggling with a scientific problem or a steep rock face. I also want to thank Bernhard Laback, the leader of our working group Experimental Audiology and Psychoacoustics (EAP), for his strong mental and professional support. I greatly value our extensive discussions. Also, I am deeply indebted to Damián Marelli, an Australian visiting scientist at ARI, who helped me to understand the subband technique and to make it run on my computer. He always greeted me with a friendly smile, when I came with my list of questions to his office door. My sincere thanks also go to all my ARI and especially EAP colleagues for lots of fruitful discussions and the friendly working atmosphere. In particular, I want to thank Michael Mihocic for his assistance with conducting the localization experiments.

I am also very grateful to Robert Höldrich and Gerhard Eckel from the Institute of Electronic Music and Acoustics at the University of Music and Performing Arts Graz (KUG) for advising my PhD project from the university side and providing valuable feedback. Moreover, I want to thank the doctoral school at KUG for administrative support and granting my work. My sincere thank also goes to the German Acoustics Association, the European Acoustics Association, and the Association for Research in Otolaryngology for various grants enabling me to present my work at international conferences and exchanging my findings and ideas with experts from the field.

My parents, Anita and Richard, receive my deepest gratitude and love for their dedication and the many years of support. I thank you for making all this possible. I also want to thank Ralf for being a great brother and supporting me whenever I need help. A heartfelt thanks goes out to my marvelous wife, Vera, for all her love, support and patience when I was only thinking about localization cues or, even worse, bugs in my code. You always helped me to restrict my perfectionism and to keep the line. Last but not least, a deep gratitude also goes to Elvis for being a trusty and omnipresent companion. You shared so many working hours with me, always conjured a smile into my face when looking with your true eyes, and often released my mental blockades while taking me for a walk outside.

Contents

Chapter 1

Introduction

The ability to localize sound sources is important and omnipresent in daily life. Accurate sound localization can rescue one's life in traffic, improve speech intelligibility in multi-talker scenarios, or be simply fascinating while listening to spatial music. Compared to visual localization, the functionality of auditory localization is rather complex, but offers certain benefits, namely, it works all around the listener as well as across visual barriers and it operates nonstop, even during sleep. In vision or touch, spatial location is topographically represented by points on the retina or the skin, whereas in audition, spatial location has to be retrieved from the signals sensed by the two ears. This thesis aims to shed some light on how this information is retrieved from the acoustic signals.

Directional features are induced by the acoustic filtering of an incoming sound by the human morphology and they are commonly described by means of head-related transfer functions (HRTFs; Møller et al., 1995). Results from psychoacoustic (Lord Rayleigh or Strutt, 1907; Macpherson and Middlebrooks, 2002) and physiological (May et al., 2008) investigations suggest that the auditory system of normal-hearing listeners processes the directional features quite independently in order to estimate the lateral angle (left/right) and the polar angle (up/down and front/back) of the sound source. Lateral-angle perception is cued by interaural disparities in the time of arrival (ITDs) and sound level (ILDs) and, consequently, is processed by neural networks with pronounced binaural interaction. Polar-angle perception, on the other hand, is cued by monaurally processed spectral features of the HRTFs. The inevitable interference between the source spectrum and the superimposed HRTF (Macpherson and Middlebrooks, 2003), however, renders the monaural cues usually less reliable than the interaural cues. To a certain degree, the auditory system can weight localization cues according to their reliability. For example, on the one hand, unilaterally deaf or plugged listeners manage to use monaural spectral cues to estimate also the lateral angle of the source (Van Wanrooij and Van Opstal, 2007; Kumpik et al., 2010; Agterberg et al., 2014); on the other hand, listeners fail to optimize the weighting of localization cues under reverberant conditions (Ihlefeld and Shinn-Cunningham, 2011).

This PhD project was entirely focused on modeling polar-angle perception in sagittal planes (i.e., orthogonal to the interaural axis) with the goal to better understand how the auditory system decodes directional spectral information. Model stages and parameters were designed to represent important physiologic and psychoacoustic aspects while keeping the model as abstract and simple as possible. Moreover, the model was aimed to reproduce the localization performance of human listeners and not to maximize the computationally achievable performance. Hence, the model was targeted to serve as a human-like evaluation tool for binaural sound reproduction and virtual acoustics.

The following chapters contain articles that describe the details of the model, its evaluation, and several model applications. Chapter 2 starts with a description of salient cues for sound localization in sagittal planes and how those cues are represented by means of HRTFs. Then, this chapter reviews approaches to model sagittal-plane localization and derives the general structure of the proposed model. The model from Langendijk and Bronkhorst (2002) was extended and generalized by considering the bandwidth and spectral shape of the incoming sound, incorporating a more adequate model of cochlear processing, introducing a non-acoustic, listener-specific parameter, there called uncertainty and later called sensitivity, considering a laterally dependent binaural weighting of monaural spectral cues, and proposing a method to derive psychoacoustic performance measures from the model output. Predictions of the extended model were evaluated against listener-specific performance in various listening conditions. Finally, chapter 2 addresses the potential of the model to assess spatial audio applications. Particular application examples include the quality of spatial cues in hearing-assistive devices, the effect of vector-based amplitude panning in sagittal planes, and the effect of non-individualized binaural recordings.

Even for individualized binaural recordings, listeners' localization performance is usually considerably different. For this reason, chapter 3 focuses on listener-specific factors influencing localization performance. Since a recalibration of the auditory system to others' HRTFs can be easily modeled but requires weeks of daily training in experiments with human listeners (Hofman et al., 1998; Carlile, 2014), model simulations were the most elegant way to address this research question. The simulations allowed to directly compare the impact of the HRTFs, representing the acoustic factor, to the impact of the listener-specific uncertainty parameter, representing the non-acoustic factor.

The preceding investigations focused on spectrally flat source spectra. For predictions of various spectral shapes of the sound source, model extensions were required. Chapter 4 provides a detailed mathematical description of an extended version of the model. Extensions include a physiologically inspired feature extraction stage and a representation of the listener's response error induced by a certain psychoacoustic task in addition to the purely perceptual localization error. The model was extensively evaluated to predict the effects of HRTF modifications or source spectrum variations on localization performance. Moreover, this study particularly investigated the role of positive spectral gradient extraction and contralateral spectral cues on sound localization.

In chapter 5, this extended model was applied to design and evaluate a method to efficiently approximate HRTFs for the purpose of virtual binaural acoustics. To this end, we investigated the subband technique, that is, a generalization of fast convolution by allowing for redundancy in the time-frequency domain. The model was used to test various approximation algorithms and to preselect reasonable approximation tolerances subsequently tested in psychoacoustic localization experiments.

Another application of the model is described in chapter 6, which more deeply elaborates on the topic introduced at the end of chapter 2, namely, the effect of vector-based amplitude panning in sagittal planes. In this study, we systematically investigated the effects of the panning ratio and angular span between two loudspeakers in the median plane and evaluated the localization accuracy provided by various loudspeaker arrangements recommended for surround sound systems.

Chapter 7 finally summarizes the results and findings obtained in the context of this PhD project. Limitations of the present model are discussed in order to stimulate future research.

Chapter 2

Assessment of sagittal-plane sound localization performance in spatial-audio applications

This work was published as

Baumgartner, R., Majdak, P., Laback, B. (2013): Assessment of sagittal-plane sound localization performance in spatial-audio applications, in: Blauert, Jens (ed.), The Technology of Binaural Listening. Berlin-Heidelberg-New York, NY (Springer). doi:10.1007/978-3-642-37762-4_4

The initial idea to develop this model came from the second author and was further developed together with the third author. The work was designed by the first two authors. I, as the first author, designed and implemented the model, performed the model simulations, analyzed the data, and generated the figures, while receiving feedback from the other two authors at each of those steps. With the help of the second author, I wrote the manuscript, which then was revised by the third author.

Assessment of Sagittal-Plane Sound Localization Performance in Spatial-Audio Applications

R. Baumgartner, P. Majdak and B. Laback

1 Sound Localization in Sagittal Planes

1.1 Salient Cues

Human normal-hearing, NH, listeners are able to localize sounds in space in terms of assigning direction and distance to the perceived auditory image [26]. Multiple mechanisms are used to estimate sound-source direction in the three-dimensional space. While interaural differences in time and intensity are important for sound localization in the lateral dimension, left/right, [53], monaural spectral cues are assumed to be the most salient cues for sound localization in the sagittal planes, SPs, [27, 54]. Sagittal planes are vertical planes parallel to the median plane and include points of similar interaural time differences for a given distance. The monaural spectral cues are essential for the perception of the source elevation within a hemifield [2, 22, 24] and for front-back discrimination of the perceived auditory event [46, 56]. Note that also the binaural pinna disparities [43], namely, interaural spectral differences, might contribute to SP localization [27].

The mechanisms underlying the perception of lateral displacement are the main topic of other chapters. This chapter focuses on the remaining directional dimension, namely, the one along SPs. Because interaural cues and monaural spectral cues are thought to be processed largely independently of each other [27], the interaural-polar coordinate system is often used to describe their respective contributions in the two dimensions. In the interaural-polar coordinate system the direction of a sound source is described with the lateral angle, $\phi \in [-90°, 90°]$, and the polar angle, $\theta \in [-90°, 270°)$—see Fig. 1, left panel. Sagittal-plane localization refers to the listener's assignment of the polar angle for a given lateral angle and distance of the sound source.

R. Baumgartner · P. Majdak ( ) · B. Laback
Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria
e-mail: piotr@majdak.com

J. Blauert (ed.), *The Technology of Binaural Listening*, Modern Acoustics and Signal Processing, DOI: 10.1007/978-3-642-37762-4_4,

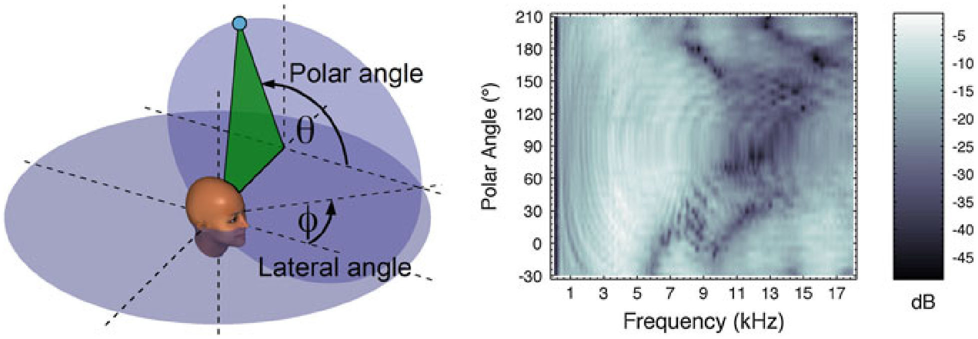


Fig. 1 *Left* Interaural-polar coordinate system. *Right* HRTF magnitude spectra of a listener as a function of the polar angle in the median SP—left ear of NH58

Although spectral cues are processed monaurally, the information from both ears affects the perceived location in most cases [39]. The ipsilateral ear, namely, the one closer to the source, dominates and its relative contribution increases monotonically with increasing lateral angle [12]. If the lateral angle exceeds about 60°, the contribution of the contralateral ear becomes negligible. Thus, even for localization in the SPs, the lateral source position, mostly depending on the broadband binaural cues [27], must be known in order to determine the binaural weighting of the monaural spectral cues.

The nature of the spectral features important for sound localization is still subject of investigations. Due to the physical dimensions, the pinna plays a larger role for higher frequencies [36] and the torso for lower frequencies [1]. Some psychoacoustic studies postulated that macroscopic patterns of the spectral features are important rather than fine spectral details [2, 10, 16, 22–24, 28, 44]. On the other hand, other studies postulated that SP sound localization is possibly mediated by means of only a few local spectral features [17, 37, 52, 56]. Despite a common agreement, according to which the amount of the spectral features can be reduced without substantial reduction of the localization performance, the perceptual relevance of particular features has not been fully clarified yet.

1.2 Head-Related Transfer Functions

The effect of the acoustic filtering of torso, head and pinna can be described in terms of a linear time-invariant system by the so-called head-related transfer functions, HRTFs, [4, 38, 45]. The right panel of Fig. 1 shows the magnitude spectra of the left-ear HRTFs of an exemplary listener, NH58,[1] along the median SP.

HRTFs depend on the individual geometry of the listener and thus listener-specific HRTFs are required to achieve accurate localization performance for binaural

[1] These and all other HRTFs are from http://www.kfs.oeaw.ac.at/hrtf.

synthesis [6, 35]. Usually, HRTFs are measured in an anechoic chamber by determining the acoustic response characteristics between loudspeakers at various directions and microphones inserted into the ear canals. Currently, much effort is put also into the development of non-contact measurement methods for capturing HRTFs like numerical calculation of HRTFs from optically scanned geometry [20, 21] and on customization of HRTFs basing on psychoacoustic tests [16, 34, 46].

Measured HRTFs contain both direction-dependent and direction-independent features and can be thought of as a series of two acoustic filters. The direction-independent filter, represented by the common transfer function, CTF, can be calculated from an HRTF set comprising many directions [34] by averaging the log-amplitude spectra of all available HRTFs of a listener's ear. The phase spectrum of the CTF is the minimum phase corresponding to the amplitude spectrum of the CTF.

In the current study, the topic of interest is the directional aspect. Thus, the directional features are considered, as represented by the directional transfer functions, DTFs. The DTF for a particular direction is calculated by filtering the corresponding HRTF with the inverse CTF. The CTF usually exhibits a low-pass filter characteristic because the higher frequencies are attenuated for many directions due to the head and pinna shadow—see Fig. 2, left panel. Compared to HRTFs, DTFs usually pronounce frequencies and thus spectral features above 4 kHz—see Fig. 2, right panel. DTFs are commonly used to investigate the nature of spectral cues in SP localization experiments with virtual sources [10, 30, 34].

In the following, the proposed model is described in Sect. 2 and the results of its evaluation are presented in Sect. 3, based on recent virtual-acoustics studies that used listener-specific HRTFs. In Sect. 4, the proposed model is applied to predict localization performance for different aspects of spatial-audio applications that involve spectral localization cues. In particular, a focus is put on the evaluation of non-individualized binaural recordings, the assessment of the quality of spatial cues for the design of hearing-assist devices, namely, in-the-ear versus behind-the-ear microphones and the estimation and improvement of the perceived direction of phantom

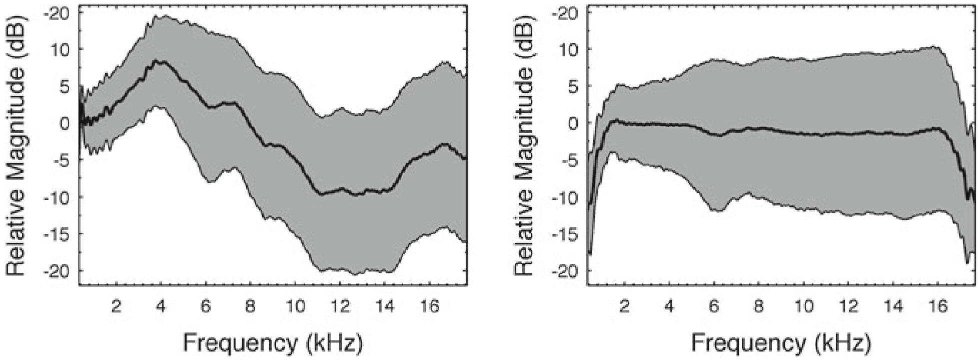


Fig. 2 *Left* Spatial variation of HRTFs around CTF for listener NH58, left ear. *Right* Corresponding DTFs, i.e. HRTFs with CTF removed. *Solid line* Spatial average of transfer function. *Grey area* ± 1 standard deviation

sources in surround-sound systems, namely, 5.1 versus 9.1 versus 10.2 surround. Finally, Sect. 5 concludes with a discussion of the potential of the model for both evaluating audio applications and improving the understanding of human sound-localization mechanisms.

2 Models of Sagittal-Plane Localization

This section considers existing models aiming at predicting listener's polar response angle to the incoming sound. These models can help to explain psychoacoustic phenomena or to assess the spatial quality of audio systems while avoiding the running of costly and time-consuming localization experiments.

In general, machine-learning approaches can be used to predict localization performance. Artificial neural networks, ANNs, have been shown to achieve rather accurate predictions when trained with large datasets of a single listener [19]. However, predictions for a larger subpopulation of human listeners would have required much more effort. Also, the interpretation of the ANN parameters is not straight forward. It is difficult to generalize the findings obtained with an ANN-based model to other signals, persons and conditions and thus to better understand the mechanisms underlying spatial hearing.

Hence, the focus is laid on a *functional* model where model parameters should correspond to physiological and/or psychophysical localization parameters. Until now, a functional model considering both spectral and temporal modulations exists only as a general concept [50]. Note that in order to address a particular research question, models dealing with specific types of modulations have been designed. For example, models for narrow-band sounds [37] were provided in order to explain the well-known effect of directional bands [4]. In order to achieve a sufficiently good prediction as an effect of the modification of the spectral cues, it is assumed that the incoming sound is a *stationary broadband* signal, explicitly disregarding spectral and temporal modulations.

Note that localization models driven by various signal-processing approaches have also been developed [3, 32, 33]. These models are based on general principles of biological auditory systems, they do not, however, attempt to predict human-listener performance—their outcome shows rather the potential of the signal-processing algorithms involved.

In the following, previous developments on modeling SP localization performance are reviewed and a functional model predicting sound localization performance in arbitrary SPs for broadband sounds is proposed. The model is designed to retrieve psychophysical localization performance parameters and can be directly used as a tool to assess localization performance in various applications. An implementation of the model is provided in the AMToolbox, as the `baumgartner2013` model [47].

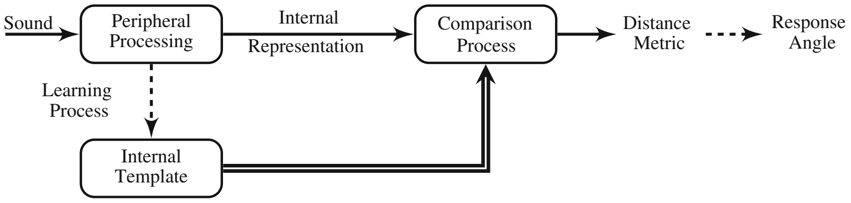


Fig. 3 General structure of a template-based comparison model for predicting localization in SPs

2.1 *Template-Based Comparison*

A common property of existing sound localization models based on spectral cues is that they compare an internal representation of the incoming sound with a template [13, 24, 55]—see Fig. 3. The internal template is assumed to be created by means of learning the correspondence between the spectral features and the direction of an acoustic event [14, 49], based on feedback from other modalities. The localization performance is predicted by assuming that in the sound localization task, the comparison yields a distance metric that corresponds to the polar response angle of the listener. Thus, template-based models include a stage modeling the peripheral processing of the auditory system applied to both the template and incoming sound and a stage modeling the comparison process in the brain.

Peripheral Processing

The peripheral processing stage aims at modeling the effect of human physiology while focusing on directional cues. The effect of the torso, head and outer ear are considered by filtering the incoming sound by an HRTF or a DTF. The effect of ear canal, middle ear and cochlear filtering can be considered by various model approximations. In the early HRTF-based localization models, a parabolic-shaped filter bank was applied [55]. Later, a filter bank averaging magnitude bins of the discrete Fourier transform of the incoming sound was used [24]. Both filter banks, while being computationally efficient, were drastically simplifying the auditory peripheral processing. The Gammatone, GT, filter bank [40] is a more physiology-related linear model of auditory filters and has been used in localization models [13]. A model accounting for the nonlinear effect of the cochlear compression is the dual-resonance nonlinear, DRNL, filter bank [25]. A DRNL filter consists of both a linear and a non-linear processing chain and is implemented by cascading GT filters and Butterworth low-pass filters, respectively. Another non-linear model uses a single main processing chain and accounts for the time-varying effects of the medial-oliviocochlear reflex [57]. All those models account for the contribution of outer hair cells to a different degree and can be used to model the movements of the basilar membrane at a particular frequency. They are implemented in the AMToolbox [47]. In the localization model described in this chapter, the GT filter bank is applied focusing on applications where the absolute sound level plays a minor role.

The filter bank produces a signal for each center frequency and only the relevant frequency bands are considered in the model. Existing models used frequency bands with constant relative bandwidth on a logarithmic frequency scale [24, 55]. In the model described in this chapter, the frequency spacing of the bands corresponds to one equivalent rectangular bandwidth, ERB, [9]. The lowest frequency is 0.7 kHz, corresponding to the minimum frequency thought to be affected by torso reflections [1]. The highest frequency considered in the model depends on the bandwidth of the incoming sound and is maximally 18 kHz, approximating the upper frequency limit of human hearing.

Further in the auditory system, the movements of the basilar membrane at each frequency band are translated into neural spikes by the inner hair cells, IHCs. An accurate IHC model has not been considered yet and does not seem to be vital for SP localization. Thus, different studies used different approximations. In this model, the IHC is modeled as half-wave rectification followed by a second-order Butterworth low-pass with a cut-off frequency of 1 kHz [8]. Since the temporal effects of SP localization are not considered yet, the output of each band is simply temporally averaged in terms of RMS amplitude, resulting in the internal representation of the sound. The same internal representation and therefore peripheral processing is assumed for the template.

Comparison Stage

In the comparison stage, the internal representation of the incoming sound is compared with the internal template. Each entry of the template is selected by a polar angle denoted as template angle. A distance metric is calculated as a function of the template angle and can be interpreted as a potential descriptor for the response of the listener.

An early modeling approach proposed to compare the spectral derivatives of various orders in terms of a band-wise subtraction of the derivatives and then averaging over the bands [55]. The comparison of the first-order derivative corresponds to the assumption that the overall sound intensity does not contribute to the localization process. In the comparison of the second-order derivatives, the differences in spectral tilt between the sound and the template do not contribute. Note that the plausibility of these comparison methods had not been investigated at that time. As another approach, the cross-correlation coefficient has been proposed to evaluate the similarity between the sound and the template [13, 37]. Later, the inter-spectral differences, ISDs, namely, the differences between the internal representations of the incoming sound and the template, calculated for each template angle and frequency band, were used [34] to show a correspondence between the template angle yielding smallest spectral variance and the actual response of human listeners. All these comparison approaches were tested in [24] who, distinguishing zeroth-, first- and second-order derivatives of the internal representations, found that the standard deviation of ISDs best described their results. This configuration corresponds to an average of the first-order derivative from [55], which is robust against changes in the overall level in the comparison process.

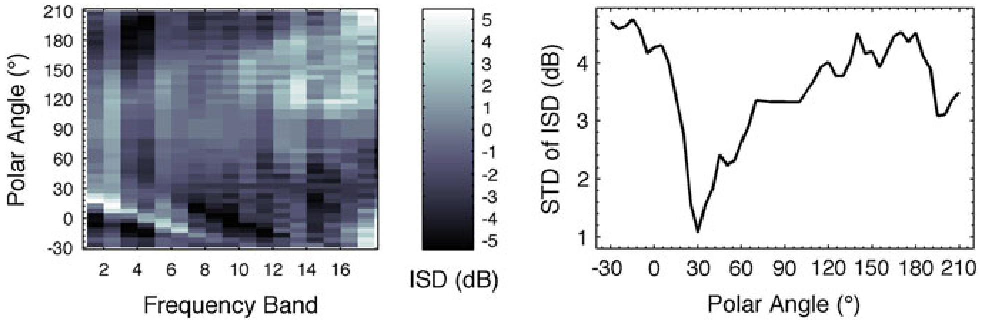


Fig. 4 Example of the comparison process for a target polar angle of 30°. *Left* Inter-spectral differences, ISDs, as a function of the template angle. *Right* Spectral standard deviation, STD, of ISDs as a function of the template angle

The model proposed in this study also relies on ISDs calculated for a template angle and for each frequency band—see Fig. 4, left panel. Then, the spectral standard deviations of ISDs are calculated for all available template angles—see Fig. 4, right panel. For band-limited sounds, the internal representation results in an abrupt change at the cut-off frequency of the sound. This change affects the standard deviation of the ISDs. Thus, in this model, the ISDs are calculated only within the bandwidth of the incoming sound.

The result of the comparison stage is a distance metric corresponding to the prediction of the polar response angle. Early modeling approaches used the minimum distance to determine the predicted response angle [55], which would nicely fit the minimum of the distance metric used in the example reported here—see Fig. 4, right panel. Also, the cross-correlation coefficient has been used as a distance metric and its maximum has been interpreted as the prediction of the response angle [37]. Both approaches represent a deterministic interpretation of the distance metric, resulting in exactly the same predictions for the same sounds. This is rather unrealistic. Listeners, repeatedly listening to the same sound, often do not respond to exactly the same direction [7]. The actual responses are known to be scattered and can be even multimodal. The scatter of one mode can be described by the Kent distribution [7], which is an elliptical probability distribution on the two-dimensional unit sphere.

2.2 Response Probability

In order to model the probabilistic response pattern of listeners, a mapping of the distance metric to polar-response probabilities via similarity indices, SIs, has been proposed [24]. For a particular target angle and ear, they obtained a monaural SI by using the distance metric as the argument of a Gaussian function with a mean of zero and a standard deviation of two—see Fig. 5, $U = 2$. While this choice appears to be somewhat arbitrary, it is an attempt to model the probabilistic relation between the

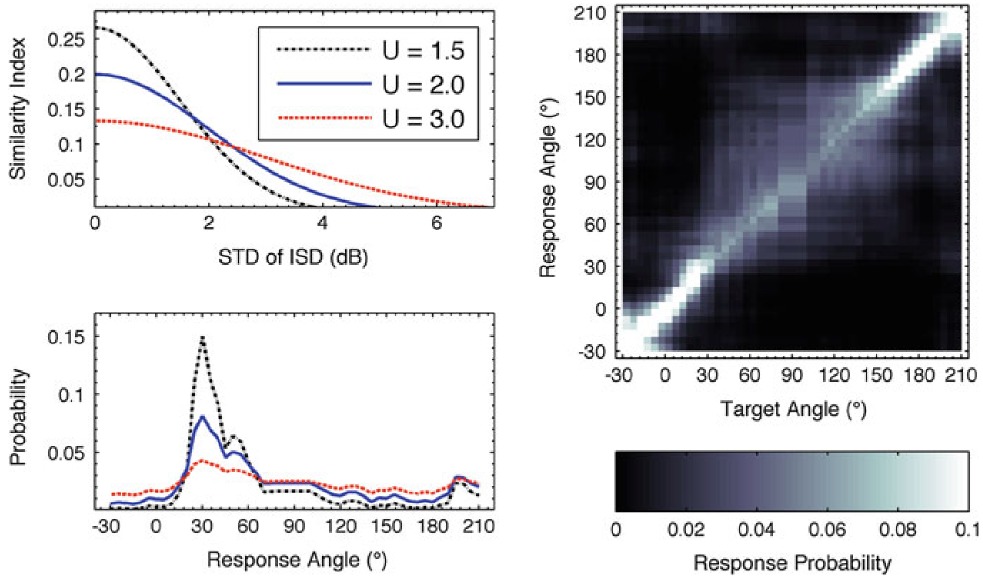


Fig. 5 *Left* Mapping function of similarity index, *top*, for various uncertainties, U, and the resulting PMVs, *bottom*—corresponding to the example shown in Fig. 4. *Right* Predicted response PMV of the localization model as a function of the target angle, i.e. prediction matrix, for the baseline condition in the median SP for listener NH58. Response probabilities are encoded by brightness

distance metric and the probability of responding to a given direction. Note that the resulting SI is bounded by zero and one and valid for the analysis of the incoming sound at one ear only.

The width of the mapping function, U in Fig. 5, actually reflects a property of an individual listener. A listener being more precise in the response to the same sound would need a more narrow mapping than a less precise listener. Thus, in contrast to the previous approach [24], in the model described in this chapter, the width of the mapping function as a listener-specific uncertainty, U, is considered. It accounts for listener-specific localization precision [34, 42, 56] due to factors like training and attention [14, 51]. Note that for simplicity, direction-dependent response precision is neglected. The lower the uncertainty, U, the higher the assumed sensitivity of the listener to distinguish spectral features. In the next section, this parameter will be used to calibrate the model to listener-specific performance.

The model stages described so far are monaural. Thus, they do not consider binaural cues and have been designed for the median SP where the interaural differences are zero and thus binaural cues do not contribute. In order to take into account the contribution of both ears, the monaural model results for both ears are combined. Previous approaches averaged the monaural SIs for both ears [24] and thus were able to consider the contribution of both ears for targets placed in the median SP. In the model described in this chapter, the lateral target range is extended to arbitrary SPs by applying a binaural weighting function [12, 29], which reduces the contribution of the contralateral ear, depending on the lateral direction of the target sound. Thus, the binaural weighting function is applied to each monaural SI, and the sum of the weighted monaural SIs yields the binaural SI.

For an incoming sound, the binaural SIs are calculated for all template entries selected by the template angle. Such a binaural SI as a function of the template angle is related to the listener's response probability as a function of the response angle. It can be interpreted as a discrete version of a probability density function, namely, a probability mass vector, PMV, showing the probability of responding at an angle to a particular target. In order to obtain a PMV, the binaural SI is normalized to have a sum of one. Note that this normalization assumes that the template angles regularly sample an SP. If this is not the case, regularization by spline interpolation is applied before the normalization.

The PMVs, calculated separately for each target under consideration, are represented in a prediction matrix. This matrix describes the probability of responding at a polar angle given a target placed at a specific angle. The right panel of Fig. 5 shows the prediction matrix resulting for the exemplary listener, NH58, in a baseline condition where the listener uses his/her own DTFs, and all available listener-specific DTFs are used as targets. The abscissa shows the target angle, the ordinate shows the response angle and the brightness represents the response probability. This representation is used throughout the following sections. It also allows for a visual comparison between the model predictions and the responses obtained from actual localization experiments.

2.3 Interpretation of the Probabilistic Model Predictions

In order to compare the probabilistic results from the model with the experimental results, likelihood statistics, calculated for actual responses from sound localization experiments and for responses resulting from virtual experiments driven by the model prediction, can be used—see Eq. (1) in [24]. The comparison between the two likelihoods allows one to evaluate the validity of the model, because only for similar likelihoods the model is assumed to yield valid predictions. The likelihood has, however, a weak correspondence with localization performance parameters commonly used in psychophysics.

Localization performance in the polar dimension usually considers local errors and hemifield confusions [35]. Although these errors derived by geometrical aspects cannot sufficiently represent the current understanding of human hearing, they are frequently used and thus enable comparison of results between studies. Quadrant errors, QEs, that is the percentage of polar errors larger or equal to 90°, represent the confusions between hemifields—for instance, front/back or up/down—without considering the local response pattern. Unimodal local responses can be represented as a Kent distribution [7], which, considering the polar dimension only, can be approximated by the polar bias and polar variance. Thus, the local errors are calculated only for local responses within the correct hemifield, namely, without the responses

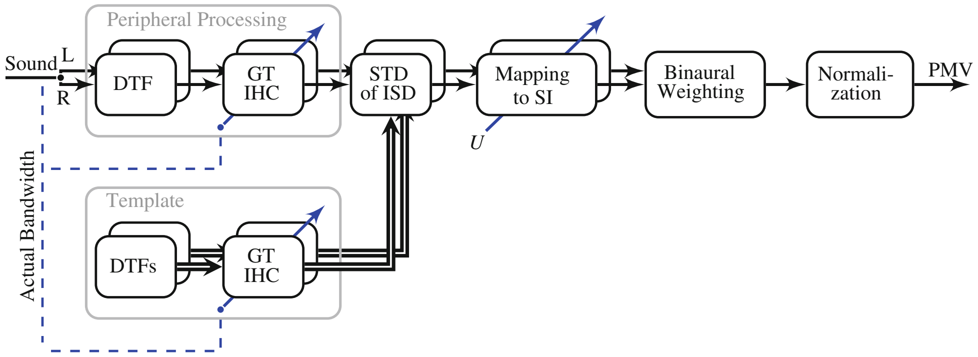


Fig. 6 Structure of the proposed SP localization model—see text for the description of the stages

yielding the QEs. A single representation of the local errors is the local polar RMS error, PE, which combines localization bias and variance in a single metric.

In the proposed model, QEs and PEs are calculated from the PMVs. The QE is the sum of the PMV entries outside the local polar range defined by the response-target difference greater or equal to 90°. The PE is the discrete expectancy value within the local polar range. In the visualization of prediction matrices—see for example right column of Fig. 5—bright areas in the upper left and bottom right corners would indicate large QEs, a strong concentration of the brightness at the diagonal would indicate small PEs. Both errors can be calculated either for a specific target angle or as the arithmetic average across all target angles considered in the prediction matrix.

Figure 6 summarizes the final structure of the model. It requires the incoming signal from a sound source as the input and results in the response probability as a function of response angle, namely PMV, for given template DTFs. Then, from PMVs calculated for the available target angles, QEs and PEs are calculated for a direct comparison with the outcome of a sound-localization experiment.

3 Listener-Specific Calibration and Evaluation

Listeners show an individual localization performance even when localizing broadband sounds in free field [31]. While the listener-specific differences in the HRTFs may play a role, also other factors like experience, attention, or utilization of auditory cues might be responsible for differences in the localization performance. Thus, this section is concerned with the calibration of the model for each particular listener. By creating calibrations for 17 listeners, a pool of listener-specific models is provided. In order to estimate the use of this pool in future applications, the performance of this pool is evaluated in two experiments. In Sect. 4, the pool is applied to various applications.

3.1 Calibration: Pool of Listener-Specific Models

The SP localization model is calibrated to the baseline performance of a listener in terms of finding an optimal uncertainty, U. Recall that the lower the uncertainty, U, the higher the assumed efficiency of the listener in evaluating spectral features. An optimal U minimizes the difference between the predicted and the listener's actual baseline performance in terms of a joint metric of QE and PE, namely, the $\mathcal{L}^2$-norm.

The actual baseline performance was obtained in localization experiments where a listener was localizing sounds using his/her own DTFs presented via headphones. Gaussian white noise bursts with a duration of 500 ms and a fade-in/out of 10 ms were used as stimuli. The acoustic targets were available for elevations from $-30°$ to $80°$ in the lateral range of at least $\pm 30°$ around the median SP. Listeners responded by manually pointing to the perceived direction of a target. For more details on the experimental methods see [10, 30, 51].

The model predictions were calculated considering SPs within the lateral range of $\pm 30°$. The targets were clustered to SPs with a width of 20° each. For the peripheral processing, the lower and upper corner frequency was 0.7 and 18 kHz, respectively, resulting in 18 frequency bands with a spacing of one ERB.

Table 1 shows the values of the uncertainty, U, for the pool of 17 listeners. The impact of the calibration becomes striking by comparing the predictions based on the listener-specific, calibrated pool with the predictions basing on the pool using $U = 2$ for all listeners as in [24]. Figure 7 shows the actual and predicted performance as a comparison with a pool calibrated to $U = 2$ for all listeners and a listener-specific calibrated pool. Note the substantially higher correlation between the prediction with the actual results in the case of the listener-specific calibration. The correlation coefficients in the order of $r = 0.85$ provide evidence for sufficient power in the predictions for the pool.

Table 1 Values of the uncertainty U for the pool of listener-specific models identified by NHn

NHn	12	15	21	22	33	39	41	42	43	46	55	58	62	64	69	71	72
U	1.6	2.0	1.8	2.0	2.3	2.3	3.0	1.8	1.9	1.8	2.0	1.4	2.2	2.1	2.1	2.1	2.2

3.2 Evaluation

In order to evaluate the SP localization model, the experimental data from two studies investigating stationary broadband sounds are modeled and compared to the experimental results. Only two studies were available because both the listener-specific HRTFs and the corresponding responses are necessary for the evaluation. For each of these studies, two predictions are calculated, namely, one for the listeners who actually participated in that experiment and one for the whole pool of listener-specific,

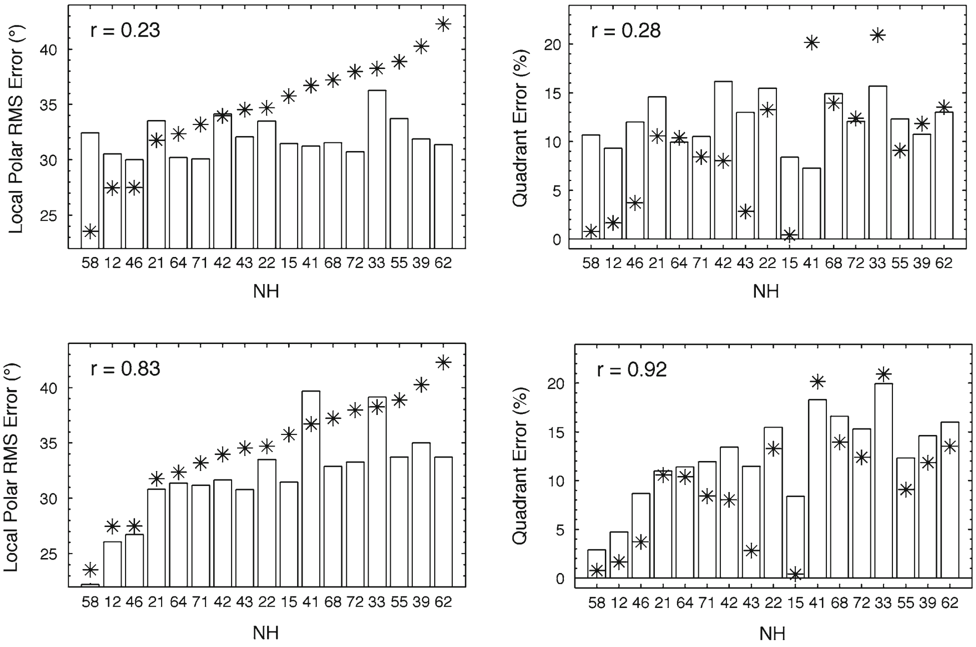


Fig. 7 Localization performance in baseline condition. *Bars* Model predictions. *Asterisks* Actual performance obtained in sound localization experiments. *Top* Model predictions for $U = 2$ as in [24]. *Bottom* Model predictions for listener-specific calibration. r…Pearson's correlation coefficient with respect to actual and predicted performance

calibrated models. For the participants, the predictions are done on the basis of the actual targets, whereas for the pool, all targets are considered by randomly drawing from the available DTFs.

Effect of the Number of Spectral Channels

A previous study tested the effect of the number of spectral channels on the localization performance in the median SP [10]. While that study was focused on cochlear-implant processing, the localization experiments were done on listeners with normal hearing using a Gaussian-envelope tone vocoder—for more details see [10]. The frequency range of 0.3–16 kHz was divided into 3, 6, 9, 12, 18, or 24 channels, equally spaced on the logarithmic frequency scale. The top row of Fig. 8 shows three channelized DTFs from an exemplary listener.

The bottom row of Fig. 8 shows the corresponding prediction matrices including the actual responses for this particular listener. Note the correspondence of the localization performance for that particular listener between the actual responses, A, and the model predictions, P. Good correspondence between the actual responses and prediction matrices was found for most of the tested listeners, which is supported by

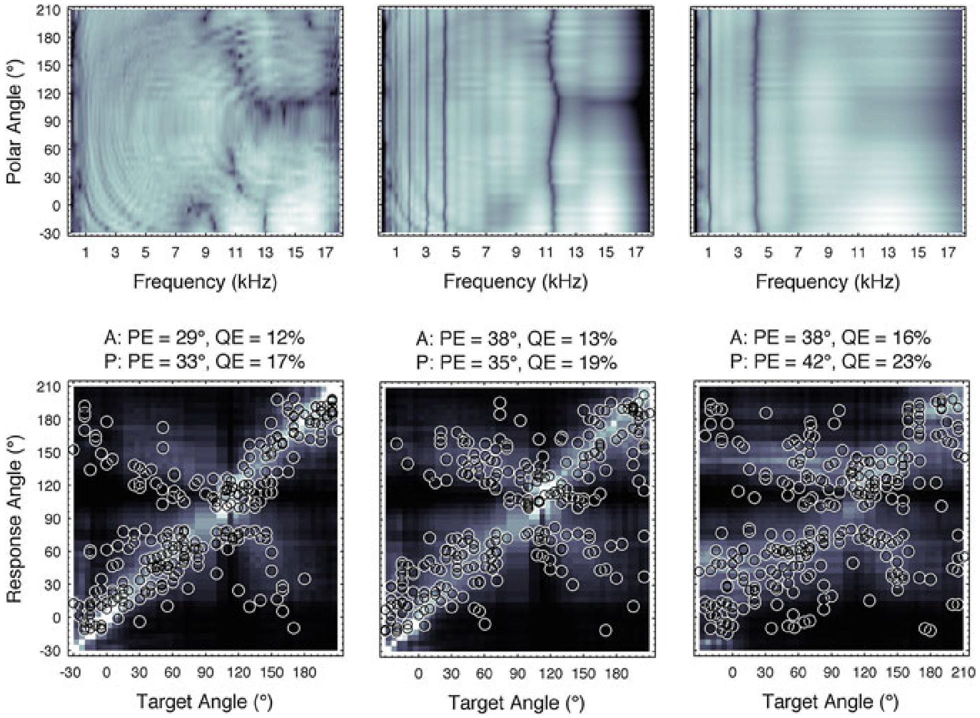


Fig. 8 Effect of the number of spectral channels for listener, NH42. *Top* Channelized left-ear DTFs of median SP with brightness-encoded magnitude as in Fig. 1, *right panel*. *Bottom* Prediction matrices with brightness-encoded probability as in Fig. 5, *right panel*, and actual responses, *open circles*. *Left* Unlimited number of channels. *Center* 24 spectral channels. *Right* 9 spectral channels. A…actual performance from [10], P…predicted performance

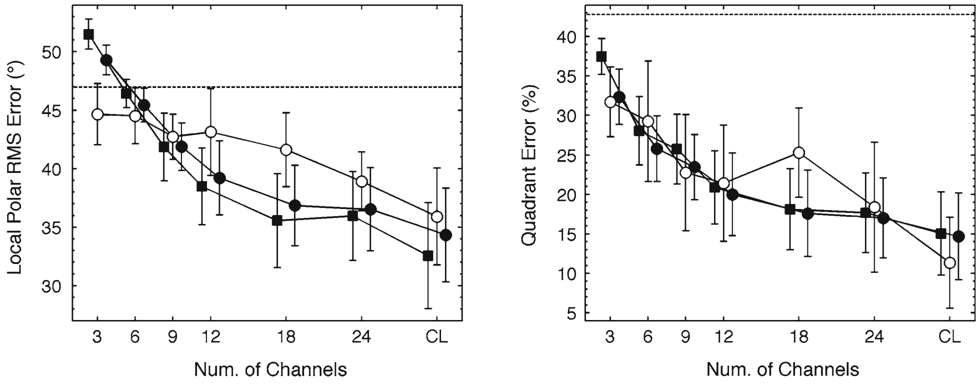


Fig. 9 Localization performance for listener groups as functions of the number of spectral channels. *Open circles* Actual performance of the listeners replotted from [10]. *Filled circles* Performance predicted for the listeners tested in [10] using the targets from [10]. *Filled squares* Performance predicted for the listener pool, using randomly chosen targets. *Error bars* ±1 standard deviations of the average over the listeners. *Dashed line* Chance performance corresponding to guessing the direction of the sound. CL…unlimited number of channels, broadband clicks

the overall response-prediction-correlation coefficients of 0.62 and 0.74 for PE and QE, respectively.

Figure 9 shows the predicted and the actual performance as averages over the listeners. In comparison to the actual performance, the models underestimated the PEs for 12 and 18 channels and overestimated them for 3 channels. The predictions for the pool seem to follow the predictions for the actually tested listeners showing generally similar QEs but slightly smaller PEs. While the analysis of the nature of these errors is outside of the focus of this chapter, both predictions, those for the actual listeners and those for the pool, seem to well represent the actual performance in this localization experiment.

Effect of Band Limitation and Spectral Warping

In another previous study, localization performance was tested in listeners using their original DTFs, band-limited DTFs and spectrally warped DTFs [51]. The band limitation was done at 8.5 kHz. The spectral warping compressed the spectral features in each DTF from the range 2.8–16 kHz to the range 2.8–8.5 kHz. While the focus of that study was to estimate the potential of re-learning sound localization with drastically modified spectral cues in a training paradigm, the experimental *ad-hoc* results from the pre-experiment are used to evaluate the proposed model. Note that, for this purpose, the upper frequency of the peripheral processing stage was configured to 8.5 kHz for the band-limited and warped conditions.

The top row of Fig. 10 shows the DTFs and the bottom row the prediction matrices for the original, band-limited and warped conditions for the exemplary listener, NH12. The actual responses show a good correspondence to the prediction matrices. Figure 11 shows group averages of the experimental results and the corresponding predictions. The group averages show a good correspondence between the actual and predicted performance. The correlation coefficient between the actual responses and predictions was 0.81 and 0.85 for PE and QE, respectively. The predictions of the pool well reflect the group averages of the actual responses.

4 Applications

The evaluation from the previous section shows response-prediction correlation coefficients in the order of 0.75. This indicates that the proposed model is reliable in predicting localization performance when applied with the listener-specific calibrations. Thus, in this section, the calibrated models are applied to predict localization performance in order to address issues potentially interesting in spatial-audio applications.

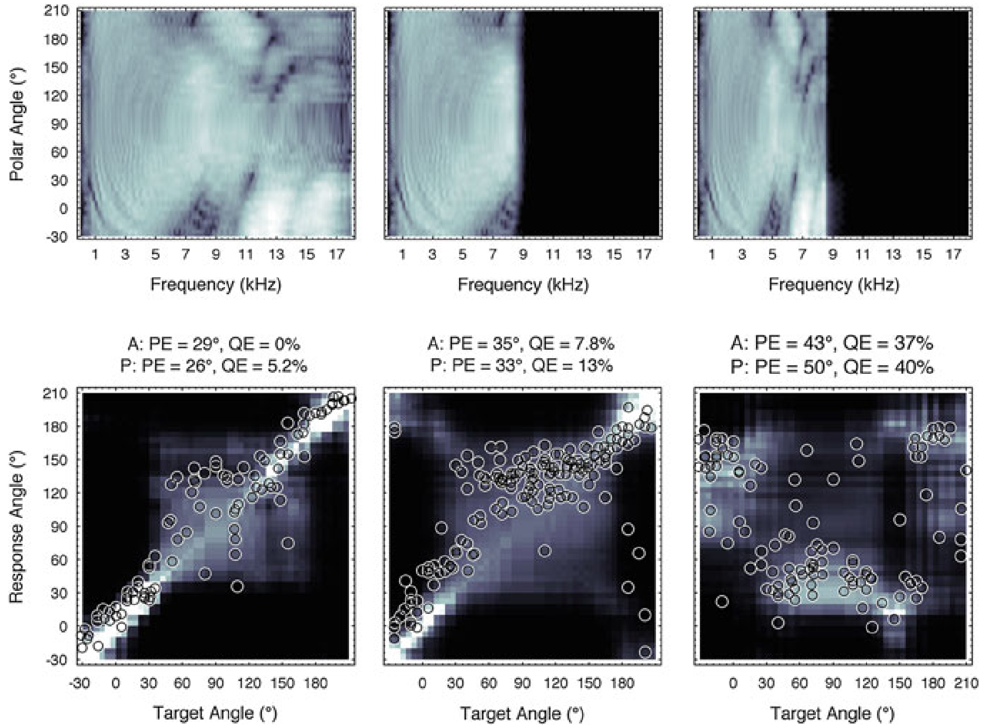


Fig. 10 Listener, NH12, localizing with different DTFs, namely, original, *left column*, band-limited, *center column*, and spectrally warped, *right column*. *Top* Left-ear DTFs in the median SP. *Bottom* Prediction matrices with actual responses from [51], /open circles/. All other conventions are as in Fig. 8

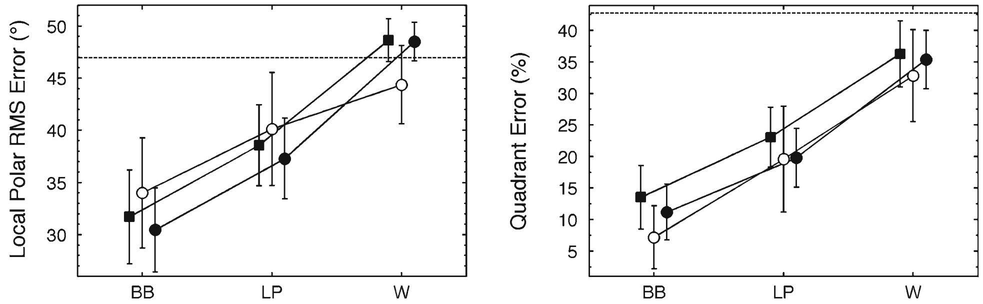


Fig. 11 Localization performance for listener groups in conditions broadband, BB, band-limited, LP, and spectrally warped, W. *Open circles* Actual performance of the tested listeners from [51]. All other conventions are as in Fig. 9

4.1 Non-Individualized Binaural Recordings

Binaural recordings aim at creating a spatial impression when listening via headphones. They are usually created using either an artificial head or mounting microphones into the ear canal of a listener and, thus, implicitly use HRTFs. When listening

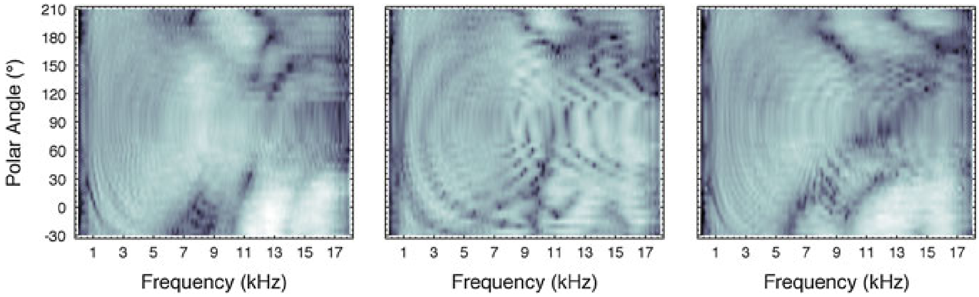


Fig. 12 Left-ear DTFs of different listeners in the median SP. *Left* NH12. *Center* NH58. *Right* NH33. *Brightness* Spectral magnitude—for code see Fig. 1, *right panel*

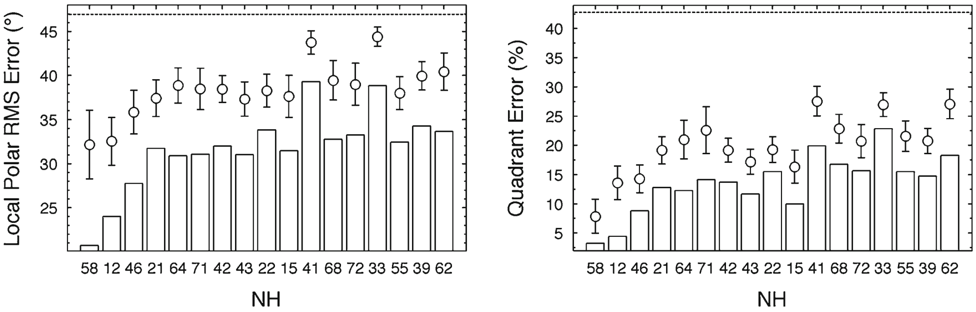


Fig. 13 Listeners' localization performance for non-individualized versus individualized DTFs. *Bars* Individualized DTFs. *Circles* Non-individualized DTFs averaged over 16 DTF sets. *Error bars* ±1 standard deviation of the average. *Dashed line* Chance performance corresponding to guessing the direction of the sound

to binaural recordings, the HRTFs of the listener do not necessarily correspond to those used in the recordings. HRTFs are, however, generally highly listener-specific and the relevant spectral features differ across listeners—see Fig. 12. Usually, SP localization performance degrades when listening to binaural signals created with non-individualized HRTFs [34]. The degree of the performance deterioration can be expected to depend on the similarity of the listener's DTFs with those actually applied. Here, the proposed model is used to estimate the localization performance for non-individualized binaural recordings. Figure 13 compares the performance when listening to individualized recordings with the average performance when listening to non-individualized recordings created from all other 16 listeners. It is evident that, on average, listening with other ears results in an increase of predicted localization errors.

Thus, the question arises of how a pool of listeners would localize a binaural recording from a particular listener, for instance, NH58. Figure 14 shows the listener-specific *increase* in the predicted localization errors when listening to a binaural recording spatially encoded using the DTFs from NH58 with respect to the errors predicted for using individualized DTFs. Some of the listeners like NH22 show only little increase in errors, while others like NH12 show large increase.

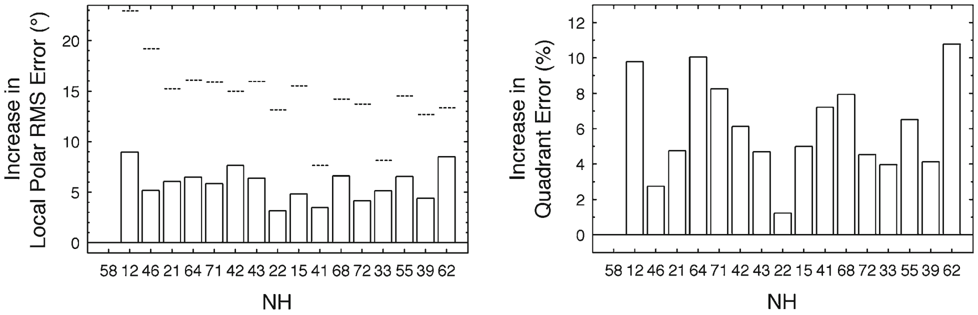


Fig. 14 *Bars* Listener-specific increase in predicted localization errors when listening to the DTFs from NH58 with respect to the errors predicted when listening to individualized DTFs. *Dashed lines* Chance performance, not shown if too large

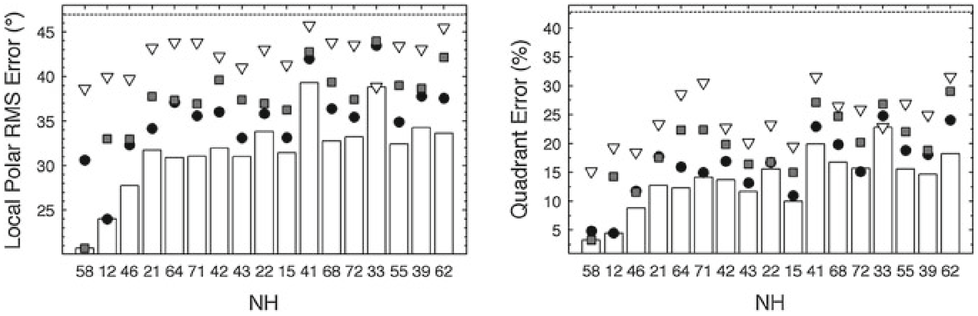


Fig. 15 Localization performance of the pool listening to different DTFs. *Bars* Individualized DTFs. *Circles* DTFs from NH12. *Squares* DTFs from NH58. *Triangles* DTFs from NH33. *Dashed line* Chance performance

Generally, one might assume that the different anatomical shapes of ears produce more or less distinct directional features. Thus, the quality of the HRTFs might vary, having effect on the ability to localize sounds in the SPs. Figure 15 shows the performance of the pool, using the DTFs from NH12, NH58 and NH33. The DTFs from these three listeners provided best, moderate and worst performance, respectively, predicted for the pool listening to binaural signals created with one of those DTF sets.

This analysis demonstrates how to evaluate across-listener compatibility of binaural recordings. Such an analysis can also be applied for other purposes like the evaluation of HRTFs of artificial heads for providing sufficient spatial cues for binaural recordings.

4.2 Assessing the Quality of Spatial Cues in Hearing-Assist Devices

In the development of hearing-assist devices, the casing, its placement on the head, and the placement of the microphone in the casing play an important role for the

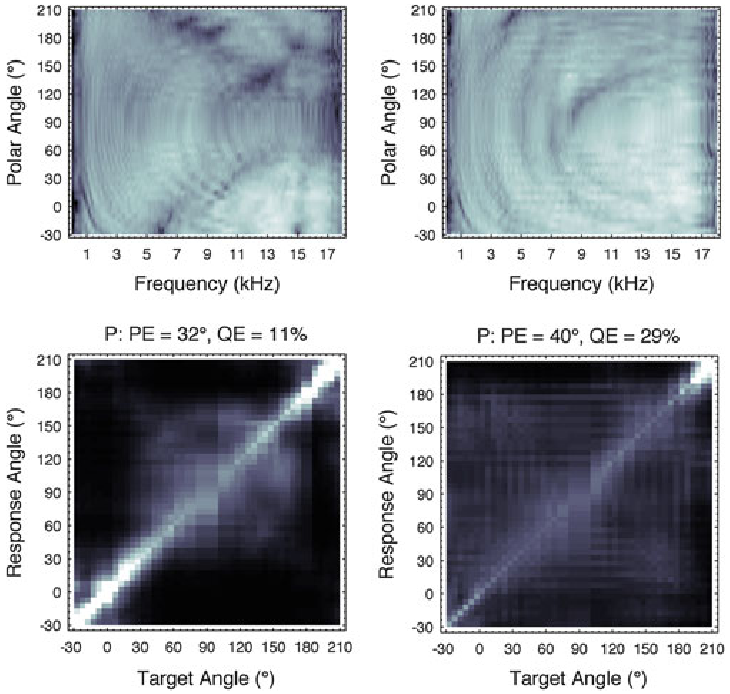


Fig. 16 Impact of the microphone placement. *Top* Left-ear DTFs of median SP from NH10. *Bottom* Prediction matrices. *Left* ITE microphone. *Right* BTE microphone. All other conventions are as in Fig. 8

effective directional cues. The proposed SP localization model can be used to assess the quality of the directional cues picked up by the microphone in a given device. Figure 16 shows DTFs resulting from behind-the-ear, BTE, compared to in-the-ear, ITE, placement of the microphone for the same listener. The BTE microphone was placed above the pinna, pointing to the front, a position commonly used by the BTE processors in cochlear-implant systems. The bottom row of Fig. 16 shows the corresponding prediction matrices and the predicted localization performance, namely, PE and QE. For this particular listener, the model predicts that if NH10 were listening with the BTE DTFs, his/her QE and PE would increase from 12 to 30% and from 32 to 40°, respectively. This can be clearly related to the impact of degraded spatial cues. Note that in this analysis it was assumed that NH10 fully adapted to the particular HRTFs. This was realized by using the same set of DTFs for the targets and the template in the model.

The impact of using BTE DTFs was also modeled for the pool of listeners using the calibrated models. Two cases are considered, namely, *ad-hoc* listening where the listeners are confronted with the DTF set without any experience in using it, and trained listening where the listeners are fully adapted to the respective DTF set. Figure 17 shows the predictions for the pool. The BTE DTFs result in performances close to guessing and the ITE DTFs from the same listener substantially improve the performance. In trained listening, the performance for the ITE DTFs is at the level of

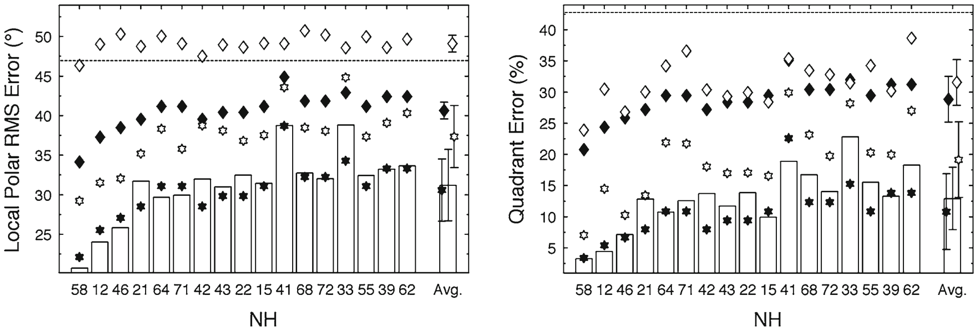


Fig. 17 Localization performance of the pool listening to different DTFs. *Bars* Individualized DTFs. *Open symbols Ad-hoc* listening. *Filled symbols* Trained listening. *Hexagrams* ITE DTFs from NH10. *Diamonds* BTE DTFs from NH10. Avg... average performance over all listeners. *Error bars* ±1 standard deviation. *Dashed line* Chance performance

the individualized DTFs, consistent with the potential of the plasticity of the spectral-to-spatial mapping [13]. The BTE DTFs, however, do not allow performance at the same level as the ITE DTFs, even when full adaptation is considered.

This analysis shows a model-based method to optimize the microphone placement with respect to the salience of directional cues. Such an analysis might be advantageous in the development of future hearing-assist devices.

4.3 Phantom Sources in Surround-Sound Systems

Sound synthesis systems for spatial audio have to deal with a limited number of loudspeakers surrounding the listener. In a system with a small number of loudspeakers, vector-based amplitude panning, VBAP [41], is commonly applied in order to create phantom sources perceived between the loudspeakers. In a surround setup, this method is also commonly used to position the phantom source along SPs, namely, to pan the source from the front to the back [11] or from the eye level to an elevated level [41]. In this section, the proposed model is applied to investigate the use of VBAP within SPs.

Amplitude Panning Along a Sagittal Plane

Now a VBAP setup with two loudspeakers is assumed, which are placed at the same distance, in the horizontal plane at the eye level, and in the same SP. Thus, the loudspeakers are in the front and in the back of the listener, corresponding to polar angles of 0° and 180°, respectively. While driving the loudspeakers with the same signal, the amplitude panning ratio can be varied from 0, front speaker only, to 1, rear speaker only, with the goal of panning the phantom source between the two loudspeakers.

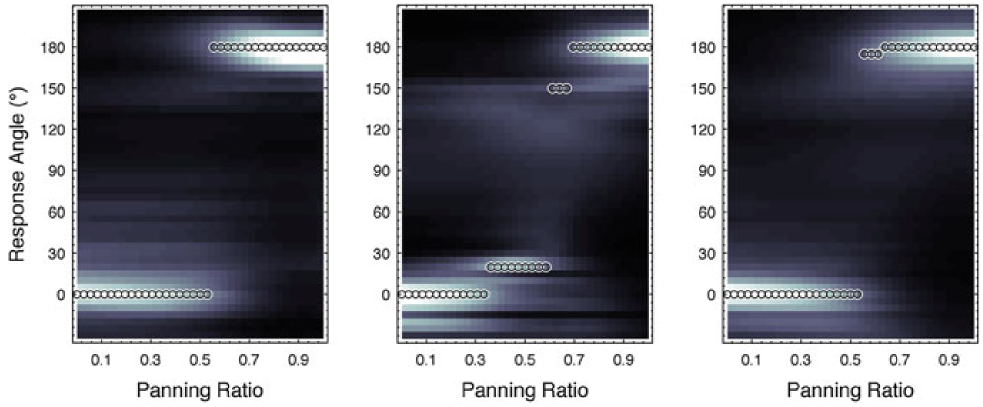


Fig. 18 Predicted response probabilities, PMVs, as a function of the amplitude panning ratio. *Left* Results for NH22. *Center* Results for NH64. *Right* Results for the pool of listeners. *Circle* Maximum of a PMV. Panning ratio of 0: Only front loudspeaker active. Panning ratio of 1: Only rear loudspeaker active. All other conventions are as in Fig. 5, *right panel*

Figure 18 shows the predicted listener-specific response probabilities in terms of the PMV as a function of the panning ratio for two loudspeakers placed at the lateral angle of 30°. The PMVs are shown for two individual listeners and also for the pool of listeners. The directional stability of phantom sources varies across listeners. For NH22, the prediction of perceived location abruptly changes from front to back, being bimodal only around the ratio of 0.6. For NH64, the transition seems to be generally smoother, with a blur in the perceived sound direction. Note that for NH64 and a ratio of 0.5, the predicted direction is elevated even though the loudspeakers were placed in the horizontal plane. The results for the pool predict an abrupt change in the perceived direction from front to back, with a blur indicating a listener-specific unstable representation of the phantom source for ratios between 0.5 and 0.7.

Effect of Loudspeaker Span

The unstable synthesis of phantom sources might be reduced by using a more adequate distance in the SP between the loudspeakers. Thus, it is shown how to investigate the polar span between two loudspeakers required to create a stable phantom source in the synthesis. To this end, a VBAP setup of two loudspeakers placed in the median SP, separated by a polar angle and driven with the panning ratio of 0.5, is used. Note that a span of 0° corresponds to a synthesis with a single loudspeaker and thus to the baseline condition. In the proposed SP localization model, the target angle describes the average of the polar angles of both loudspeakers, which, in VBAP, is thought to correspond to the direction of the phantom source. The model was run for all available target angles resulting in the prediction of the localization performance.

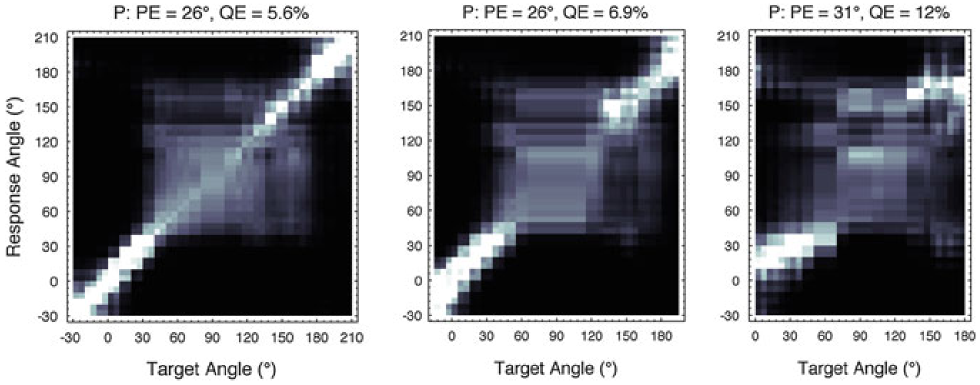


Fig. 19 Predictions for different loudspeaker spans and NH12. *Left* Span of 0°, single-loudspeaker synthesis, baseline condition. *Center* Span of 30°. *Right* Span of 60°. All other conventions are as in Fig. 8

Figure 19 shows prediction matrices and predicted localization performance for NH12 and three different loudspeaker spans. Note the large increase of errors from 30 to 60° of span, consistent with the results from [5]. Figure 20 shows the average increase in localization error predicted for the pool of listeners as a function of the span. The increase is shown relative to the listener-specific localization performance in the baseline condition. Note that not only the localization errors but also the variances across the listeners increase with increasing span.

This analysis shows how the model may help in choosing the adequate loudspeaker span when amplitude panning is applied to create phantom sources. Such an analysis can also be applied when more sophisticated sound-field reproduction approaches like Ambisonics or wave-field synthesis are involved.

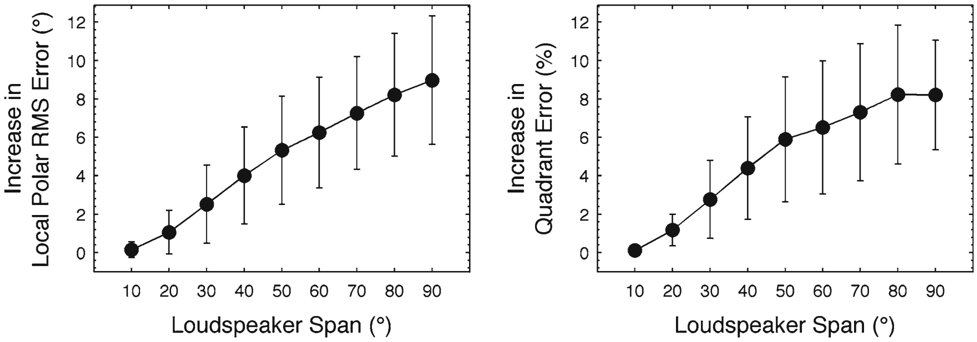


Fig. 20 Increase in localization errors as a function of the loudspeaker span. *Circles* Averages over all listeners from the pool. *Error bars* ±1 standard deviation

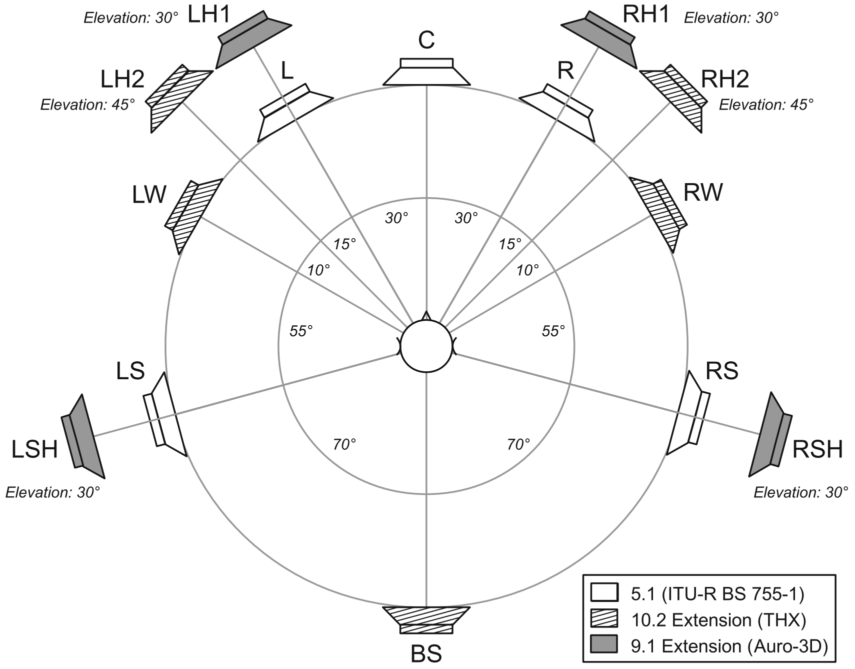


Fig. 21 Loudspeaker positions of three typical surround-sound systems. Drivers for the low-frequency effect, LFE, channels not shown

Results for Typical Surround-Sound Setups

The most common standardized surround-sound setup is known as the 5.1 setup [18]. In this setup, all loudspeakers are placed in the horizontal plane at a constant distance around the listener. Recently, other schemes have been proposed to include elevated speakers in the synthesis systems. The 10.2 setup, known as *Audyssey DSX* [15] and the 9.1 setup, known as *Auro-3D* [48], consider two and four elevated loudspeakers, respectively. Figure 21 shows the positions of the loudspeakers in those three surround-sound setups. The model was applied to evaluate the localization performance when VBAP is used to pan a phantom source at the left hand side from front, L, to back, LS. While in the 5.1 setup only loudspeakers L and LS are available, in 10.2 and 9.1 the loudspeakers LH2 and LH1 & LSH, respectively, may also contribute even to create an elevated phantom source.

VBAP was applied between the closest two loudspeakers by using the law of tangents [41]. For a desired polar angle of the phantom source, the panning ratio was $R = \frac{1}{2} - \frac{\tan(\delta)}{2\tan(0.5\beta)}$ with β denoting the loudspeaker span in polar dimension and δ denoting the difference between the desired polar angle and the polar center angle of the span. The contributing loudspeakers were not always in the same SP, thus, the lateral angle of the phantom source was considered for the choice of the SP in the modeling by applying the law of tangents on the lateral angles of the loudspeakers for the particular panning ratio, R.

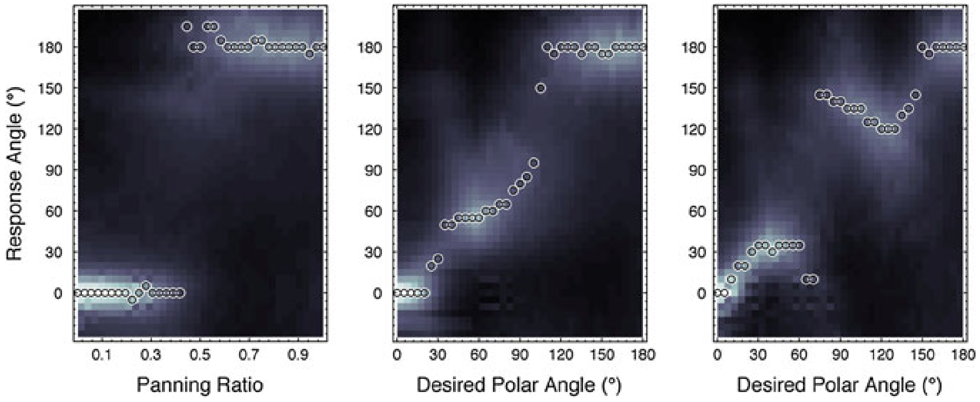


Fig. 22 Predictions for VBAP applied to various surround-sound systems. *Left* 5.1 setup, panning between the loudspeakers L and LS. *Center* 10.2 DSX setup panning from L, polar angle of 0°, via LH2, 55°, to LS, 180°. *Right* 9.1 Auro-3D setup panning from L, 0°, via LH1, 34°, and LSH, 121°, to LS, 180°. *Desired polar angle* Continuous scale representing VBAP across pair-wise contributing loudspeakers. All other conventions are as in Fig. 18

Figure 22 shows the predicted pool averages of the PMVs as a function of the desired polar angle of the phantom source. The improvements due to the additional elevated loudspeakers in the 10.2 and 9.1 setups are evident. Nevertheless, the predicted phantom sources are far from perfectly following the desired angle. Especially for the 9.1 setup, in the rear hemifield, the increase in the desired polar angle, namely, *decrease* in the elevation, resulted in a decrease in the predicted polar angle, namely, *increase* in the elevation.

The proposed model seems to be well-suited for addressing such a problem. It is easy to show how modifications of the loudspeaker setup would affect the perceived angle of the phantom source. As an example, the positions of the elevated loudspeakers in the 9.1 setup were modified in two ways. First, the lateral distance between the loudspeakers, LH1 and LSH, was decreased by modifying the azimuth of LSH from 110 to 140°. Second, both loudspeakers, LSH and LS, were placed to the azimuth of 140°. Figure 23 shows the predictions for the modified setups. Compared to the original setup, the first modification clearly resolves the problem described above. The second modification, while only slightly limiting the lateral range, provides an even better representation of the phantom source along the SP.

5 Conclusions

Sound localization in SPs refers to the ability to estimate the sound-source elevation and to distinguish between front and back. The SP localization performance is usually measured in time-consuming experiments. In order to address this disadvantage, a model predicting SP localization performance of individual listeners has been

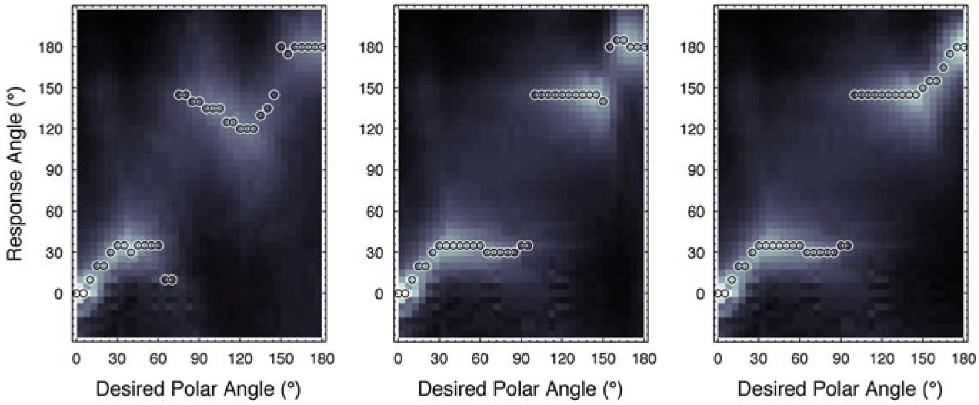


Fig. 23 Predictions for two modifications to the 9.1 Auro 3D setup. *Left* Original setup, loudspeakers LS and LSH at azimuth of 110°. *Center* LSH at azimuth of 140°. *Right* LS and LSH at azimuth of 140°. All other conventions are as in Fig. 22

proposed. Listener-specific calibration was performed for a pool of 17 listeners, and the calibrated models were evaluated using results from psychoacoustic localization experiments. The potential of the calibrated models was demonstrated for three applications, namely,

1. The evaluation of the spatial quality of binaural recordings
2. The assessment of the spatial quality of directional cues provided by the microphone placement in hearing-assist devices
3. The evaluation and improvement of the loudspeaker position in surround-sound systems

These applications are examples of situations where SP localization cues, namely, spectral cues, likely play a role. The model is, however, not limited to those applications and it hopefully will help in assessing spatial quality in other applications as well.

Acknowledgments The authors would like to thank H. S. Colburn and two anonymous reviewers for constructive comments on earlier versions of this manuscript. This research was supported by the Austrian Science Fund, FWF, project P 24124–N13.

References

1. V. R. Algazi, C. Avendano, and R. O. Duda. Elevation localization and head-related transfer function analysis at low frequencies. *J Acoust Soc Am*, 109:1110–1122, 2001.
2. F. Asano, Y. Suzuki, and T. Sone. Role of spectral cues in median plane localization. *J Acoust Soc Am*, 88:159–168, 1990.
3. E. Blanco-Martin, F. J. Casajus-Quiros, J. J. Gomez-Alfageme, and L. I. Ortiz-Berenguer. Estimation of the direction of auditory events in the median plane. *Appl Acoust*, 71:1211–1216, 2010.

4. J. Blauert. *Räumliches Hören (Spatial Hearing)*. S. Hirzel Verlag Stuttgart, 1974.
5. P. Bremen, M. M. van Wanrooij, and A. J. van Opstal. Pinna cues determine orientation response modes to synchronous sounds in elevation. *J Neurosci*, 30:194–204, 2010.
6. A. W. Bronkhorst. Localization of real and virtual sound sources. *J Acoust Soc Am*, 98:2542–2553, 1995.
7. S. Carlile, P. Leong, and S. Hyams. The nature and distribution of errors in sound localization by human listeners. *Hear Res*, 114:179–196, 1997.
8. T. Dau, D. Püschel, and A. Kohlrausch. A quantitative model of the "effective" signal processing in the auditory system. I. Model structure. *J Acoust Soc Am*, 99:3615–3622, 1996.
9. B. R. Glasberg and B. C. J. Moore. Derivation of auditory filter shapes form notched-noise data. *Hear Res*, 47:103–138, 1990.
10. M. J. Goupell, P. Majdak, and B. Laback. Median-plane sound localization as a function of the number of spectral channels using a channel vocoder. *J Acoust Soc Am*, 127:990–1001, 2010.
11. J. Hilson, D. Gray, and M. DiCosimo. *Dolby Surround Mixing Manual*. Dolby Laboratories, Inc., San Francisco, CA, 2005. chapter 5—Mixing techniques.
12. M. Hofman and J. Van Opstal. Binaural weighting of pinna cues in human sound localization. *Exp Brain Res*, 148:458–470, 2003.
13. P. M. Hofman and A. J. V. Opstal. Spectro-temporal factors in two-dimensional human sound localization. *J Acoust Soc Am*, 103:2634–2648, 1998.
14. P. M. Hofman, J. G. A. van Riswick, and A. J. van Opstal. Relearning sound localization with new ears. *Nature Neurosci*, 1:417–421, 1998.
15. T. Holman. *Surround Sound: Up and Running*. Focal Press, 2008.
16. S. Hwang and Y. Park. Interpretations on pricipal components analysis of head-related impulse responses in the median plane. *J Acoust Soc Am*, 123:EL65-EL71, 2008.
17. 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. *Appl Acoust*, 68:835–850, 2007.
18. Int Telecommunication Union, Geneva, Switzerland. *Multichannel stereophonic sound system with and without accompanying picture*, 2012. Recommendation ITU-R BS.775-3.
19. C. Jin, M. Schenkel, and S. Carlile. Neural system identification model of human sound localization. *J Acoust Soc Am*, 108:1215–1235, 2000.
20. B. F. Katz. Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation. *J Acoust Soc Am*, 110:2440–2448, 2001.
21. W. Kreuzer, P. Majdak, and Z. Chen. Fast multipole boundary element method to calculate head-related transfer functions for a wide frequency range. *J Acoust Soc Am*, 126:1280–1290, 2009.
22. A. Kulkarni and H. S. Colburn. Role of spectral detail in sound-source localization. *Nature*, 396:747–749, 1998.
23. A. Kulkarni and H. S. Colburn. Infinite-impulse-response models of the head-related transfer function. *J Acoust Soc Am*, 115:1714–1728, 2004.
24. E. H. A. Langendijk and A. W. Bronkhorst. Contribution of spectral cues to human sound localization. *J Acoust Soc Am*, 112:1583–1596, 2002.
25. E. A. Lopez-Poveda and R. Meddis. A human nonlinear cochlear filterbank. *J Acoust Soc Am*, 110:3107–3118, 2001.
26. F. R. S. Lord Rayleigh. On our perception of sound direction. *Philos Mag*, 13:214–232, 1907.
27. E. A. Macpherson and J. C. Middlebrooks. Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *J Acoust Soc Am*, 111:2219–2236, 2002.
28. E. A. Macpherson and J. C. Middlebrooks. Vertical-plane sound localization probed with ripple-spectrum noise. *J Acoust Soc Am*, 114:430–445, 2003.
29. E. A. Macpherson and A. T. Sabin. Binaural weighting of monaural spectral cues for sound localization. *J Acoust Soc Am*, 121:3677–3688, 2007.

30. P. Majdak, M. J. Goupell, and B. Laback. 3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training. *Atten Percept Psycho*, 72:454–469, 2010.
31. J. C. Makous and J. C. Middlebrooks. Two-dimensional sound localization by human listeners. *J Acoust Soc Am*, 87:2188–2200, 1990.
32. M. I. Mandel, R. J. Weiss, and D. P. W. Ellis. Model-based expectation-maximization source separation and localization. *IEEE Trans Audio Speech Proc*, 18:382–394, 2010.
33. T. May, S. van de Par, and A. Kohlrausch. A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE Trans Audio Speech Lang Proc*, 19:1–13, 2011.
34. J. C. Middlebrooks. Individual differences in external-ear transfer functions reduced by scaling in frequency. *J Acoust Soc Am*, 106:1480–1492, 1999.
35. J. C. Middlebrooks. Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. *J Acoust Soc Am*, 106:1493–1510, 1999.
36. J. C. Middlebrooks and D. M. Green. Sound localization by human listeners. *Annu Rev Psychol*, 42:135–159, 1991.
37. J. C. Middlebrooks and D. M. Green. Observations on a principal components analysis of head-related transfer functions. *J Acoust Soc Am*, 92:597–599, 1992.
38. H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen. Head-related transfer functions of human subjects. *J Audio Eng Soc*, 43:300–321, 1995.
39. M. Morimoto. The contribution of two ears to the perception of vertical angle in sagittal planes. *J Acoust Soc Am*, 109:1596–1603, 2001.
40. R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice. *An efficient auditory filterbank based on the gammatone function*. APU, Cambridge, UK, 1988.
41. V. Pulkki. Virtual sound source positioning using vector base amplitude panning. *J Audio Eng Soc*, 45:456–466, 1997.
42. B. Rakerd, W. M. Hartmann, and T. L. McCaskey. Identification and localization of sound sources in the median sagittal plane. *J Acoust Soc Am*, 106:2812–2820, 1999.
43. C. L. Searle and I. Aleksandrovsky. Binaural pinna disparity: Another auditory localization cue. *J Acoust Soc Am*, 57:448–455, 1975.
44. M. A. Senova, K. I. McAnally, and R. L. Martin. Localization of virtual sound as a function of head-related impulse response duration. *J Audio Eng Soc*, 50:57–66, 2002.
45. E. A. Shaw. Transformation of sound pressure level from the free field to the eardrum in the horizontal plane. *J Acoust Soc Am*, 56:1848–1861, 1974.
46. R. H. Y. So, B. Ngan, A. Horner, J. Braasch, J. Blauert, and K. L. Leung. Toward orthogonal non-individualised head-related transfer functions for forward and backward directional sound: cluster analysis and an experimental study. *Ergonomics*, 53:767–781, 2010.
47. P. Søndergaard and P. Majdak. The auditory modeling toolbox. In J. Blauert, editor, *The technology of binaural listening*, chapter 2. Springer, Berlin-Heidelberg-New York NY, 2013.
48. G. Theile and H. Wittek. Principles in surround recordings with height. *In Proc. 130th AES Conv. Audio Engr. Soc.*, page Convention Paper 8403, London, UK, 2011.
49. M. M. van Wanrooij and A. J. van Opstal. Relearning sound localization with a new ear. *J Neurosci*, 25:5413–5424, 2005.
50. J. Vliegen and A. J. V. Opstal. The influence of duration and level on human sound localization. *J Acoust Soc Am*, 115:1705–1703, 2004.
51. T. Walder. Schallquellenlokalisation mittels Frequenzbereich-Kompression der Außenohrübertragungsfunktionen (sound-source localization through warped head-related transfer functions). Master's thesis, University of Music and Performing Arts, Graz, Austria, 2010.
52. A. J. Watkins. Psychoacoustical aspects of synthesized vertical locale cues. *J Acoust Soc Am*, 63:1152–1165, 1978.
53. F. L. Wightman and D. J. Kistler. The dominant role of low-frequency interaural time differences in sound localization. *J Acoust Soc Am*, 91:1648–1661, 1992.
54. F. L. Wightman and D. J. Kistler. Monaural sound localization revisited. *J Acoust Soc Am*, 101:1050–1063, 1997.

55. P. Zakarauskas and M. S. Cynader. A computational theory of spectral cue localization. *J Acoust Soc Am*, 94:1323–1331, 1993.
56. P. X. Zhang and W. M. Hartmann. On the ability of human listeners to distinguish between front and back. *Hear Res*, 260:30–46, 2010.
57. M. S. A. Zilany and I. C. Bruce. Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. *J Acoust Soc Am*, 120:1446–1466, 2006.

Chapter 3

Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization

This work was published as

Majdak, P., **Baumgartner, R.**, and Laback, B. (2014): Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization, in: Frontiers in Psychology 5:319. doi:10.3389/fpsyg.2014.00319

The idea of the study was from the first author and me. The work was designed by all authors. I, as the second author, simulated the experiments. The first author and I analyzed the data, created the figures, and drafted the manuscript. The third author revised the manuscript.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 23 April 2014
doi: 10.3389/fpsyg.2014.00319

Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization

Piotr Majdak*, Robert Baumgartner and Bernhard Laback

Psychoacoustics and Experimental Audiology, Acoustics Research Institute, Austrian Academy of Sciences, Wien, Austria

Edited by:
Guillaume Andeol, Institut de Recherche Biomdicale des Armées, France

Reviewed by:
John A. Van Opstal, University of Nijmegen, Netherlands
Simon Carlile, University of Sydney, Australia

****Correspondence:***
Piotr Majdak, Psychoacoustics and Experimental Audiology, Acoustics Research Institute, Austrian Academy of Sciences, Wohllebengasse 12-14, Wien 1040, Austria
e-mail: piotr@majdak.com

The ability of sound-source localization in sagittal planes (along the top-down and front-back dimension) varies considerably across listeners. The directional acoustic spectral features, described by head-related transfer functions (HRTFs), also vary considerably across listeners, a consequence of the listener-specific shape of the ears. It is not clear whether the differences in localization ability result from differences in the encoding of directional information provided by the HRTFs, i.e., an acoustic factor, or from differences in auditory processing of those cues (e.g., spectral-shape sensitivity), i.e., non-acoustic factors. We addressed this issue by analyzing the listener-specific localization ability in terms of localization performance. Directional responses to spatially distributed broadband stimuli from 18 listeners were used. A model of sagittal-plane localization was fit individually for each listener by considering the actual localization performance, the listener-specific HRTFs representing the acoustic factor, and an uncertainty parameter representing the non-acoustic factors. The model was configured to simulate the condition of complete calibration of the listener to the tested HRTFs. Listener-specifically calibrated model predictions yielded correlations of, on average, 0.93 with the actual localization performance. Then, the model parameters representing the acoustic and non-acoustic factors were systematically permuted across the listener group. While the permutation of HRTFs affected the localization performance, the permutation of listener-specific uncertainty had a substantially larger impact. Our findings suggest that across-listener variability in sagittal-plane localization ability is only marginally determined by the acoustic factor, i.e., the quality of directional cues found in typical human HRTFs. Rather, the non-acoustic factors, supposed to represent the listeners' efficiency in processing directional cues, appear to be important.

Keywords: sound localization, localization model, sagittal plane, listener-specific factors, head-related transfer functions

1. INTRODUCTION

Human listeners use monaural spectral cues to localize sound sources in sagittal planes (e.g., Wightman and Kistler, 1997; van Wanrooij and van Opstal, 2005). This includes the ability to assign the vertical position of the source (e.g., Vliegen and van Opstal, 2004) and to distinguish between front and back (e.g., Zhang and Hartmann, 2010). Spectral cues are caused by the acoustic filtering of the torso, head, and pinna, and can be described by means of head-related transfer functions (HRTFs; e.g., Møller et al., 1995). The direction-dependent components of the HRTFs are described by directional transfer functions (DTFs, Middlebrooks, 1999b).

The ability to localize sound sources in sagittal planes, usually tested in psychoacoustic experiments as localization performance, varies largely across listeners (Middlebrooks, 1999a; Rakerd et al., 1999; Zhang and Hartmann, 2010). A factor contributing to the variability across listeners might be the listeners' morphology. The ear shape varies across the human population (Algazi et al., 2001) and these differences cause the DTF features to vary across individuals (Wightman and Kistler, 1997). One might expect that different DTF sets provide different amounts of cues available for the localization of a sound. When listening with DTFs of other listeners, the performance might be different, an effect we refer to in this study as the *acoustic factor* in sound localization.

The strong effect of training on localization performance (Majdak et al., 2010, Figure 7) indicates that in addition to the acoustic factor, also other listener-specific factors are involved. For example, a link between the listener-specific sensitivity to the spectral envelope shape and the listener-specific localization performance has been recently shown (Andéol et al., 2013). However, other factors like the ability to perform the experimental task, the attention paid to the relevant cues, or the accuracy in responding might contribute as well. In the present study, we consolidate all those factors to a single factor which we refer to as the *non-acoustic factor*.

In this study, we are interested in the contribution of the acoustic and non-acoustic factors to sound localization performance. As for the acoustic factor, its effect on localization performance has already been investigated in many studies (e.g., Wightman and Kistler, 1997; Middlebrooks, 1999a; Langendijk and Bronkhorst, 2002). However, most of those studies investigated *ad-hoc* listening with modified DTFs without any re-calibration of the spectral-to-spatial mapping in the auditory system (Hofman et al., 1998). By testing the *ad-hoc* localization performance to modified DTFs, two factors were simultaneously varied: the directional cues in the incoming sound, and their mismatch to the familiarized (calibrated) mapping. The acoustic factor of interest in our study, however, considers changes in the DTFs of the *own* ears, i.e., changes of DTFs without any mismatch between the incoming sound and the calibrated mapping. A localization experiment testing such a condition would need to minimize the mismatch by achieving a re-calibration. Such a re-calibration is indeed achievable in an extensive training with modified DTFs, however, the experimental effort is rather demanding and requires weeks of exposure to the modified cues (Hofman and van Opstal, 1998; Majdak et al., 2013). Note that such a long-term re-calibration is usually attributed to perceptual adaptation, in contrast to the short-term learning found to take place within hours (Zahorik et al., 2006; Parseihian and Katz, 2012).

Using a model of the localization process, the condition of a complete re-calibration can be more easily achieved. Thus, our study is based on predictions from a model of sagittal-plane sound localization (Baumgartner et al., 2013). This model assumes that listeners create an internal template set of their specific DTFs as a result of a learning process (Hofman et al., 1998; van Wanrooij and van Opstal, 2005). The more similar the representation of the incoming sound compared to a template, the larger the assumed probability of responding at the polar angle corresponding to that template (Langendijk and Bronkhorst, 2002). The model from Baumgartner et al. (2013) uses a method to compute localization performance based on probabilistic predictions and considers both acoustic factors in terms of the listener-specific DTFs and non-acoustic factors in terms of an uncertainty parameter U. In Baumgartner et al. (2013), the model has been validated under various conditions for broadband stationary sounds. In that model, the role of the acoustic factor can be investigated by simultaneously modifying DTFs of both the incoming sound and the template sets. This configuration allows to predict sound localization performance when listening with others' ears following a complete re-calibration to the tested DTFs.

In the following, we briefly describe the model and revisit the listener-specific calibration of the model. Then, the effect of the uncertainty representing the non-acoustic factor, and the effect of the DTF set representing the acoustic factor, are investigated. Finally, the relative contributions of the two factors are compared.

2. MATERIALS AND METHODS

2.1. MODEL

In this study, we used the model proposed by Baumgartner et al. (2013). The model relies on a comparison between an internal representation of the incoming sound and an internal template set (Zakarauskas and Cynader, 1993; Hofman and van Opstal, 1998; Langendijk and Bronkhorst, 2002; Baumgartner et al., 2013). The internal template set is assumed to be created by means of learning the correspondence between the spectral features and the direction of an acoustic event based on feedback from other modalities (Hofman et al., 1998; van Wanrooij and van Opstal, 2005). The model is implemented in the Auditory Modeling Toolbox as `baumgartner2013` (Søndergaard and Majdak, 2013).

Figure 1 shows the basic structure of the model from Baumgartner et al. (2013). Each block represents a processing stage of the auditory system in a functional way. The target sound is processed in order to obtain an internal (neural) representation. This target representation is compared to an equivalently processed internal template set consisting of the DTF representations for the given sagittal plane. This comparison process is the basis of a spectral-to-spatial mapping, which yields the prediction probability for responding at a given polar angle.

In general, in this study, we used the model configured as suggested in Baumgartner et al. (2013). In the following, we summarize the model stages and their configuration, focusing on the acoustic and non-acoustic factors in the localization process.

2.1.1. *Peripheral processing*

In the model, the same peripheral processing is considered for the incoming sound and the template. The peripheral processing stage aims at modeling the effect of human physiology while focusing on directional cues. The effect of the torso, head and pinna are considered by filtering the incoming sound by a DTF.

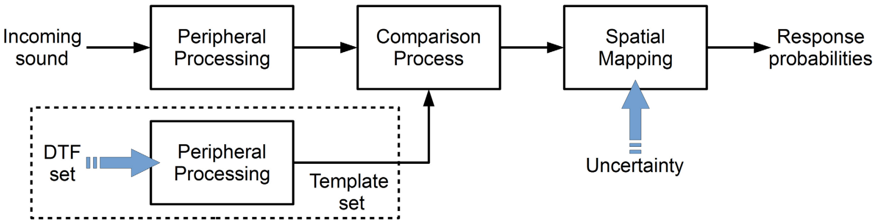


FIGURE 1 | Structure of the sound localization model from Baumgartner et al. (2013). The incoming target sound is peripherally processed and the result is compared to an internal template set. The comparison result is mapped yielding the probability for responding at a given polar angle. The blue arrows indicate the free parameters of the corresponding sections. In the model, the DTF set and the uncertainty represent the acoustic and non-acoustic factors, respectively.

The effect of the cochlear filtering was considered as linear Gammatone filter bank (Patterson et al., 1988). The filter bank produces a signal for each frequency band. 28 frequency bands were considered in the model, determined by the lowest frequency of 0.7 kHz, the highest frequency of 18 kHz, and the frequency spacing of the bands corresponding to one equivalent rectangular bandwidth (Glasberg and Moore, 1990). In the model, the output of each frequency band is half-wave rectified and low-pass filtered (2nd-order Butterworth filter, cut-off frequency of 1 kHz) in order to simulate the effect of the inner hair cells (Dau et al., 1996). The filtered outputs are then temporally averaged in terms of root-mean-square (RMS) amplitude, resulting in the internal representation of the sound.

2.1.2. Comparison stage

In the comparison stage, the internal representation of the incoming sound is compared with the internal template set. Each template is selected by a polar angle denoted as template angle. A distance metric is calculated as a function of the template angle and is interpreted as a descriptor contributing to the prediction of the listener's response.

In the model, the distance metric is represented by the standard deviation (SD) of the inter-spectral differences between the internal representation of the incoming sound and a template calculated across frequency bands. The SD of inter-spectral differences is robust against changes in overall level and has been shown to be superior to other metrics like the inter-spectral cross-correlation coefficient (Langendijk and Bronkhorst, 2002).

2.1.3. Spatial mapping

In the model, a probabilistic approach is used for the mapping of the distance metric to the predicted response probability. For a particular target angle, response angle, and ear, the distance metric is mapped by a Gaussian function to a similarity index (SI), interpreted as a measure reflecting the response probability for a response angle.

The mapping function actually reflects the *non-acoustic factor* of the localization process. In the model, the width of the Gaussian function was considered as a property of an individual listener. Baumgartner et al. (2013) assumed that a listener being more precise in the response to the same sound would need a more narrow mapping than a less precise listener. Thus, the width of the mapping function was interpreted as a listener-specific uncertainty, U. In the model, it accounted for listener-specific localization performance and was a free parameter in the calibration process. In Langendijk and Bronkhorst (2002), the uncertainty parameter has actually also been used (their S), however, it was considered to be constant for all listeners, thus representing a rather general property of the auditory system. The impact of the uncertainty U, representing the non-acoustic factor responsible for the listener variability on the predicted localization performance is described in the following sections.

In the model, the contribution of the two ears was considered by applying a binaural weighting function (Morimoto, 2001; Macpherson and Sabin, 2007), which reduces the contribution of the contralateral ear with increasing lateral angle of the target sound. The binaural weighting function is applied to each monaural SI, and the sum of the weighted monaural SIs yields the binaural SI.

In the model, for a given target angle, the binaural SIs are calculated as a function of the response angle, i.e., for all templates. The SI as a function of response angle is scaled to a sum of one in order to be interpreted as a probability mass vector (PMV), i.e., a discrete version of a probability density function. Such a PMV describes the listener's response probability as a function of the response angle for a given incoming sound.

2.2. EXPERIMENTAL CONDITIONS FOR CALIBRATION

In Baumgartner et al. (2013), the model was calibrated to the actual performance of a pool of listeners for the so-called baseline condition, for which actual data (DTFs and localization responses) were collected in two studies, namely in Goupell et al. (2010) and Majdak et al. (2013). In both studies, localization responses were collected using virtual stimuli presented via headphones. While localization performance seems to be better when using free-field stimuli presented via loudspeakers (Middlebrooks, 1999b), we used virtual stimuli in order to better control for cues like head movements, loudspeaker equalization, or room reflections. In this section, we summarize the methods used to obtain the baseline conditions in those two studies.

2.2.1. Subjects

In total, 18 listeners were considered for the calibration. Eight listeners were from Goupell et al. (2010) and 13 listeners were from Majdak et al. (2013), i.e., three listeners participated in both studies. None of them had indications of hearing disorders. All of them had thresholds of 20-dB hearing level or lower at frequencies from 0.125 to 12.5 kHz.

2.2.2. HRTFs and DTFs

In both Goupell et al. (2010) and Majdak et al. (2013), HRTFs were measured individually for each listener. The DTFs were then calculated from the HRTFs. Both HRTFs and DTFs are part of the ARI HRTF database (Majdak et al., 2010).

Twenty-two loudspeakers (custom-made boxes with VIFA 10 BGS as drivers) were mounted on a vertical circular arc at fixed elevations from −30° to 80°, with a 10° spacing between 70° and 80° and 5° spacing elsewhere. The listener was seated in the center point of the circular arc on a computer-controlled rotating chair. The distance between the center point and each speaker was 1.2 m. Microphones (Sennheiser KE-4-211-2) were inserted into the listener's ear canals and their output signals were directly recorded via amplifiers (FP-MP1, RDL) by the digital audio interface.

A 1729-ms exponential frequency sweep from 0.05 to 20 kHz was used to measure each HRTF. To speed up the measurement, for each azimuth, the multiple exponential sweep method was used (Majdak et al., 2007). At an elevation of 0°, the HRTFs were measured with a horizontal spacing of 2.5° within the range of ±45° and 5° otherwise. With this rule, the measurement positions for other elevations were distributed with a constant spatial angle, i.e., the horizontal angular spacing increased with the elevation. In total, HRTFs for 1550 positions within the full 360° horizontal span were measured for each listener. The measurement procedure lasted for approximately 20 min. The acoustic influence of the equipment was removed by equalizing the HRTFs with the transfer functions of the equipment. The equipment transfer functions were derived from reference measurements in which the microphones were placed at the center point of the circular arc and the measurements were performed for all loudspeakers.

The DTFs (Middlebrooks, 1999b) were calculated. The magnitude of the common transfer function (CTF) was calculated by averaging the log-amplitude spectra of all HRTFs for each individual listener and ear. The phase spectrum of the CTF was set to the minimum phase corresponding to the amplitude spectrum. The DTFs were the result of filtering HRTFs with the inverse complex CTF. Finally, the impulse responses of all DTFs were windowed with an asymmetric Tukey window (fade in of 0.5 ms and fade out of 1 ms) to a 5.33-ms duration.

2.2.3. Stimulus

In Majdak et al. (2013), the experiments were performed for targets in the lateral range of ±60°. In Goupell et al. (2010), the experiments were performed for targets in the lateral range of ±10°. The direction of a target is described by the polar angle ranging from −30° (front, below eye-level) to 210° (rear, below eye-level).

The audio stimuli were Gaussian white noise bursts with a duration of 500 ms, which were filtered with the listener-specific DTFs corresponding to the tested condition. The level of the stimuli was 50 dB above the individually measured absolute detection threshold for that stimulus, estimated in a manual up-down procedure for a frontal eye-leveled position. In the experiments, the stimulus level was randomly roved for each trial within the range of ±5 dB in order to reduce the possibility of using overall level cues for localization.

2.2.4. Apparatus

In both studies, Goupell et al. (2010) and Majdak et al. (2013), the virtual acoustic stimuli were presented via headphones (HD 580, Sennheiser) in a semi-anechoic room. Stimuli were generated using a computer and output via a digital audio interface (ADI-8, RME) with a 48-kHz sampling rate. A virtual visual environment was presented via a head-mounted display (3-Scope, Trivisio). It provided two screens with a field of view of 32° x 24° (horizontal x vertical dimension). The virtual visual environment was presented binocularly with the same picture for both eyes. A tracking sensor (Flock of Birds, Ascension), mounted on the top of the listener's head, captured the position and orientation of the head in real time. A second tracking sensor was mounted on a manual pointer. The tracking data were used for the 3-D graphic rendering and response acquisition. More details about the apparatus are provided in Majdak et al. (2010).

2.2.5. Procedure

For the calibration, the data were collected in two studies using the same procedure. In Goupell et al. (2010), the data were the last 300 trials collected within the acoustic training, see their Sec. II. D. In Majdak et al. (2013), the data were the 300 trials collected within the acoustic test performed at the beginning of the pre-training experiments, see their Sec. II. D. In the following, we summarize the procedure used in the two studies.

In both studies, the listeners were immersed in a spherical virtual visual environment (for more details see Majdak et al., 2010). They were standing on a platform and held a pointer in their right hand. The projection of the pointer direction on the sphere's surface, calculated based on the position and orientation of the tracker sensors, was visualized and recorded as the perceived target position. The pointer was visualized whenever it was in the listeners' field of view.

Prior to the acoustic tests, listeners participated in a visual training procedure with the goal to train them to point accurately to the target. The visual training was a simplified game in the first-person perspective in which listeners had to find a visual target, point at it, and click a button within a limited time period. This training was continued until 95% of the targets were found with an RMS angular error smaller than 2°. This performance was reached within a few hundred trials.

In the acoustic experiments, at the beginning of each trial, the listeners were asked to align themselves with the reference position, keep the head direction constant, and click a button. Then, the stimulus was presented. The listeners were asked to point to the perceived stimulus location and click the button again. Then, a visual target in the form of a red rotating cube was shown at the position of the acoustic target. In cases where the target was outside of the field of view, arrows pointed towards its position. The listeners were asked to find the target, point at it, and click the button. At this point in the procedure, the listeners had both heard the acoustic target and seen the visualization of its position. To stress the link between visual and acoustic location, the listeners were asked to return to the reference position and listen to the same acoustic target once more. The visual feedback was intended to trigger a procedural training in order to improve the localization performance within the first few hundred of trials (Majdak et al., 2010). During this second acoustic presentation, the visual target remained visualized in the visual environment. Then, while the target was still visualized, the listeners had to point at the target and click the button again. An experimental block consisted of 50 targets and lasted for approximately 15 min.

2.3. DATA ANALYSIS

In the psychoacoustic experiments, the errors were calculated by subtracting the target angles from the response angles. We separated our data analysis into confusions between the hemifields and the local performance within the correct hemifield. The rate of confusions was represented by the quadrant error (QE), which is the percentage of responses where the absolute polar error exceeded 90° (Middlebrooks, 1999b). In order to quantify the local performance in the polar dimension, the local polar RMS error (PE) was calculated, i.e., the RMS of the polar errors calculated for the data without QEs.

The listener-specific results from both Goupell et al. (2010) and Majdak et al. (2013) were pooled. Only responses within the lateral range of ±30° were considered because (1) most of the localization responses were given in that range, (2) Baumgartner et al. (2013) evaluated the model using only that range, and (3) recent evaluations indicate that predictions for that range seem to be slightly more accurate than those for more lateral ranges (Baumgartner et al., 2014). For the considered data, the average

QE was 9.3% ± 6.0% and the average PE was 34° ± 5°. This is similar to the results from Middlebrooks (1999b) who tested 14 listeners in virtual condition using DTFs. His average QE was 7.7% ± 8.0% and the average PE was 29° ± 5°.

In the model, targets in the lateral range of ±30° were considered in order to match the lateral range of the actual targets from the localization experiments. For each listener, PMVs were calculated for three lateral segments with a lateral width of 20° each, and these PMVs were evaluated corresponding to the actual lateral target angles. The QE was the sum of the corresponding PMV entries outside the local polar range for which the response-target distance exceeded 90°. The PE was the discrete expectancy value within the local polar range. Both errors were calculated as the arithmetic averages across all polar target angles considered.

3. RESULTS AND DISCUSSION

3.1. MODEL CALIBRATION

In Baumgartner et al. (2013), the model was calibrated individually for each listener by finding the uncertainty U providing the smallest residual in the predictions as compared to the actual performance obtained in the localization experiments.

In our study, this calibration process was revisited. For each listener and all target directions, PMVs were calculated for varying uncertainty U ranging from 0.1 to 4.0 in steps of 0.1. Listener-specific DTFs were used for both the template set and incoming sound. **Figure 2** shows PMVs and the actual localization responses for four exemplary listeners and exemplary uncertainties.

For each listener, the predicted PEs and QEs were calculated from the PMVs, and the actual PEs and QEs were calculated from the experimental results. **Figure 3** shows the predicted QEs and PEs as a function of the uncertainty for the four exemplary listeners. The symbols show the actual QEs and PEs.

In Baumgartner et al. (2013), the uncertainty yielding the smallest squared sum of residues between the actual and predicted performances (PE and QE) was considered as optimal. Using the same procedure, the optimal uncertainties U_k were calculated for each listener k and are shown in **Table 1**. For the

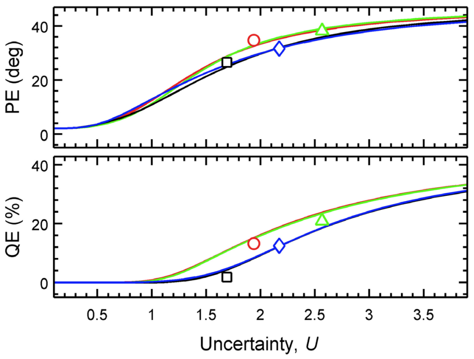


FIGURE 3 | Predicted localization performance depends on the uncertainty. PEs and QEs are shown as functions of U for four exemplary listeners (k = 3: blue squares, k = 9: red triangles, k = 12: green diamonds, k = 15: black circles). Lines show the model predictions. Symbols show the actual performance obtained in the localization experiment (placement on the abscissa corresponds to the optimal listener-specific uncertainty U_k).

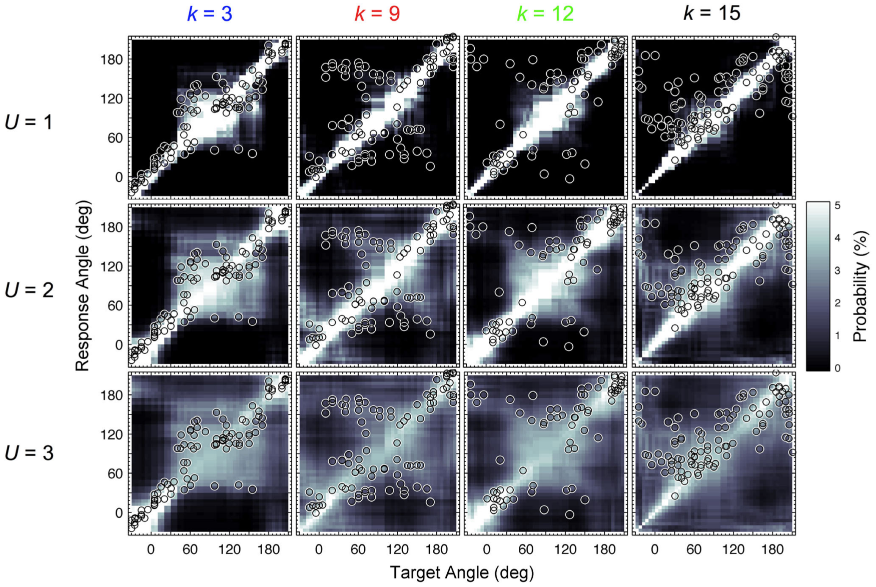


FIGURE 2 | Actual and modeled localization. Actual localization responses (circles) and modeled response probabilities (PMVs, brightness encoded) calculated for three uncertainties U and four exemplary listeners indexed by k.

Table 1 | Uncertainty U_k of individual listener with index *k*.

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
x	58	53	12	42	46	43	15	21	22	71	55	64	72	68	33	39	62	41
U	1.48	1.63	1.68	1.74	1.75	1.83	1.85	1.91	1.94	2.01	2.12	2.22	2.24	2.29	2.33	2.35	2.47	3.05

Listener indexed by k is identified in the ARI HRTF database by NHx$_k$. The listeners are sorted by k corresponding to ascending U_k.

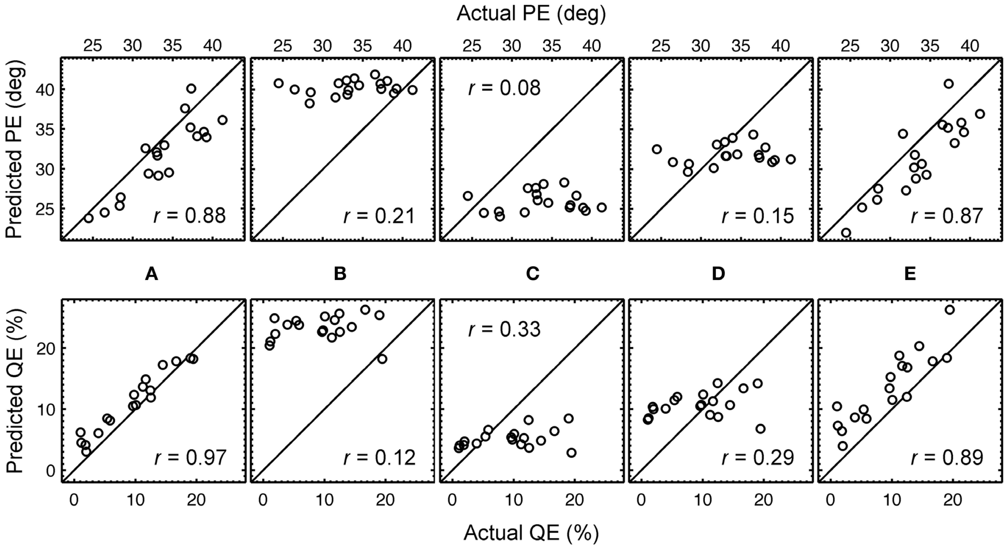


FIGURE 4 | Predicted versus actual localization performance. Predicted PEs and QEs are shown as functions of the actual PEs and QEs, respectively, for each listener. **(A)** Optimal listener-specific uncertainties U_k. **(B)** Listener-constant uncertainty yielding best correlation for PE, $U = 2.89$. **(C)** Listener-constant uncertainty yielding best correlation for QE, $U = 1.87$. **(D)** Listener-constant uncertainty from (Langendijk and Bronkhorst, 2002), $U = 2.0$. **(E)** Listener-specific uncertainties U_k and the same DTF set ($k = 14$) for all listeners (see Section 3.3 for more details). The correlation coefficient is denoted by *r*.

listener group, the average listener-specific uncertainty amounted to 2.05 ($SD = 0.37$).

With the optimal listener-specific uncertainties from **Table 1**, predictions were compared to the actual localization performances. **Figure 4A** shows the correspondence between the actual and predicted QEs and PEs of all listeners when using those listener-specific uncertainties. For the listener group, the correlation coefficient between actual and predicted localization errors was 0.88 for PE and 0.97 for QE. In Baumgartner et al. (2013), the model calibrated with those optimal uncertainties was evaluated in further conditions involving DTF modifications yielding correlation coefficients in the range of 0.75.

3.2. NON-ACOUSTIC FACTOR: LISTENER-SPECIFIC UNCERTAINTY

In Baumgartner et al. (2013), the optimal listener-specific uncertainties were assumed to yield most accurate performance predictions. In Langendijk and Bronkhorst (2002), the effect of spectral cues was modeled by using a parameter corresponding to our uncertainty. Interestingly, that parameter was constant for all listeners and the impact of this listener-specific uncertainty is not clarified yet. Thus, in this section, we investigate the effect of uncertainty being listener-specific as compared to uncertainty being constant for all listeners, when using the model from Baumgartner et al. (2013).

Predictions were calculated with a model calibrated to uncertainty being constant for all listeners. Three uncertainties were used: (1) $U = 2.89$, which yielded largest correlation with the actual PEs of the listeners, (2) $U = 1.87$, which yielded largest correlation with the actual QEs, and (3) $U = 2.0$, which corresponds to that used in Langendijk and Bronkhorst (2002). The DTFs used for the incoming sound and the template set were still listener-specific, representing the condition of listening with own ears. The predictions are shown in **Figures 4B–D**. The corresponding correlation coefficients are shown as insets in the corresponding panels. From this comparison and the comparison to that for listener-specific uncertainties (**Figure 4A**), it is evident that listener-specific calibration is required to account for the listener-specific actual performance.

Our findings are consistent with the results from Langendijk and Bronkhorst (2002) who used a constant calibration for all listeners. The focus of that study was to investigate the change in predictions caused by the variation of spectral cues. Thus, prediction changes for different conditions *within* an individual listener were important, which, in the light of the model from Baumgartner et al. (2013), correspond to the variation of the DTFs used for the incoming sound and not to the variation of the uncertainty. $U = 2.0$ seems to be indeed an adequate choice for predictions for an "average listener". This is supported by the similar average uncertainty of our listener group ($U = 2.05$). It is further supported by the performance predicted with $U = 2.0$, which was similar to the actual group performance. For acurate listener-specific predictions, however, listener-specific uncertainty is required.

The listener-constant uncertainty seems to have largely reduced the predicted performance variability in the listener group. In order to quantify this observation, the group SDs were calculated for predictions with listener-constant U from 1.1 to 2.9 in steps of 0.1 for each listener. For PE, the group SD was 0.96° ± 0.32°. For QE, the group SD was 1.34% ± 0.87%. For comparison, the group SD for predictions with listener-*specific* uncertainties was 4.58° and 5.07% for PE and QE, respectively, i.e., three times larger than those for predictions with the listener-constant uncertainties.

In summary, the listener-specific uncertainty seems to be vital to obtain accurate predictions of the listeners' actual performance. The listener-constant uncertainty drastically reduced the correlation between the predicted and actual performance. Further, the listener-constant uncertainty reduced the group variability in the predictions. Thus, as the only parameter varied in the model, the uncertainty seems to determine to a large degree the baseline performance predicted by the model. It can be interpreted as a parameter calibrating the model in order to represent a good or bad localizer; the smaller the uncertainty, the better the listeners' performance in a localization task. Notably, uncertainty is not associated with any acoustic information considered in the model, and thus, it represents the non-acoustic factor in modeling sound localization.

3.3. ACOUSTIC FACTOR: LISTENER-SPECIFIC DIRECTIONAL CUES

In the previous section, the model predictions were calculated for listeners' own DTFs in both the template set and the incoming sound; a condition corresponding to listening with own ears. With the DTFs of other listeners but own uncertainty, their performance might have been different.

For the investigation of that effect, one possibility would be to vary the quality of the DTF sets along a continuum simultaneously in both the incoming sound and the template set, and analyze the corresponding changes in the predictions. Such an investigation would be, in principle, similar to that from the previous section where the uncertainty was varied and the predicted performance was analyzed. While U represents a measure of the uncertainty, a similar metric would be required in order to quantify the quality differences between two DTF sets. Finding an appropriate metric is challenging. A potentially useful metric is the spectral SD of inter-spectral differences (Middlebrooks, 1999b; Langendijk and Bronkhorst, 2002) as used in the model from (Baumgartner et al., 2013) as the distance metric and thus as basis for the predictions. Being a part of the model, however, this metric is barred from being an independent factor in our investigation.

In order to analyze the DTF set variation as a parameter without any need for quantification of the variation, we systematically replaced the listeners' own DTFs by DTFs from other listeners from this study. The permutation of the DTF sets and uncertainties within the same listener group allowed us to estimate the effect of directional cues relative to the effect of uncertainty on the localization performance of our group.

For each listener, the model predictions were calculated using a combination of DTF sets and uncertainties of all listeners from the group. Indexing each listener by k, predicted PEs and QEs as functions of U_k and D_k were obtained, with U_k and D_k being the uncertainty and the DTF set, respectively, of the k-th listener. **Figure 5** shows the predicted PEs and QEs for all combinations of U_k and D_k. The listener group was sorted such that the uncertainty increases with increasing k and the same sorting order was used for D_k. This sorting order corresponds to that from **Table 1**.

The results reflect some of the effects described in the previous sections. The main diagonal represents the special case of identical k for D_k and U_k, corresponding to listener-specific performance, i.e., predictions for each listener's actual DTFs and optimal listener-specific uncertainty from the calibrated model described

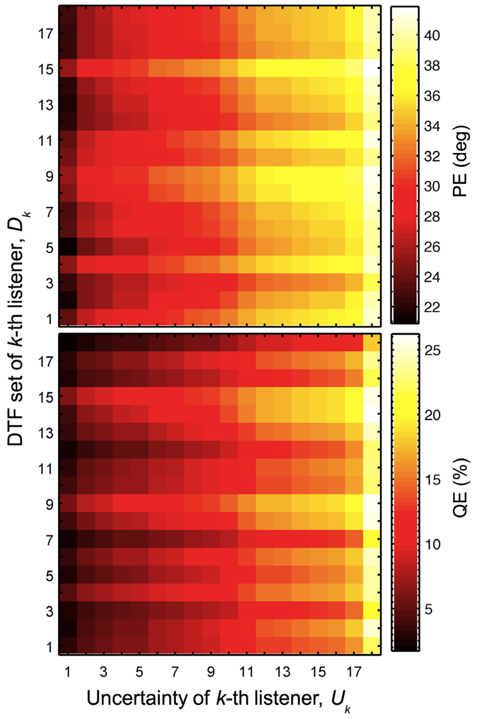


FIGURE 5 | Localization performance depends on the uncertainty and DTF set. Predicted PEs and QEs as functions of the uncertainty of k-th listener (U_k) and DTF set of k-th listener (D_k).

in Section 3.1. Each row, i.e., constant D_k but varying U_k, represents the listener-specific effect of the uncertainty described in Section 3.2, i.e., listening with own ears but having different uncertainties.

In this section, we focus on the results in the columns. Each column describes results for a constant U_k but varying D_k, representing the listener-specific effect of the DTF set. While the predictions show a variation across both columns and rows, i.e., substantial effects of both uncertainty and DTF set, some DTF sets show clear differences to others across all uncertainties. This analysis is, however, confounded by the different baseline performance of each listener and can be improved by considering the performance relative to the listener-specific performance. **Figure 6** shows ΔPEs and ΔQEs, i.e., PEs and QEs relative to the listener-specific PEs and QEs, respectively, averaged over all uncertainties for each DTF set D_k. Positive values represent the performance amount by which our listener group would deteriorate when listening with the DTF set of k-th listener (and being fully re-calibrated). For example, the DTF sets of listeners $k = 9$ and $k = 15$ show such deteriorations. Those DTF sets seem to have provided less accessible directional cues. Further, DTF sets improving the performance for the listeners can be identified, see for example, the DTF sets of listeners $k = 3$ and $k = 12$. These DTF sets seem to have provided more accessible directional cues. The effect of those four DTF sets can be also examined in **Figure 2** by comparing the predictions for constant uncertainties, i.e., across rows.

Thus, variation of the DTF sets had an effect on the predictions suggesting that it also affects the comparison of the predictions with the actual performance. This leads to the question to what extend a constant DTF set across all listeners can explain the actual performances? It might even be the case that listener-specific DTFs are not required for accurate predictions. Thus, similarly to the analysis from Section 3.2 where the impact of listener-specific uncertainty was related to that of a listener-constant uncertainty, here, we compare the impact of listener-specific DTF sets relative to that of a listener-constant DTF set. To this end, predictions were calculated with a model calibrated to the same DTF set for all listeners but with a listener-specific uncertainty. All DTF sets from the pool of available listeners were tested. For each of the DTF sets, correlation coefficients between the actual and predicted performances were calculated. The correlation coefficients averaged over all DTF sets were 0.86 ($SD = 0.007$) for PE and 0.89 ($SD = 0.006$) for QE. Note the extremely small variability across the different DTF sets, indicating only little impact of the DTF set on the predictions. The DTF set from listener $k = 14$ yielded the largest correlation coefficients, which were 0.87 for PE and 0.89 for QE. The corresponding predictions as functions of the actual performance are shown in **Figure 4E**. Note the similarity to the predictions for the listener-specific DTF sets (**Figure 4A**). These findings have a practical implication when modeling the baseline performance of sound localization: for an arbitrary listener, the DTFs of another arbitrary listener, e.g., NH68 ($k = 14$), might still yield listener-specific predictions.

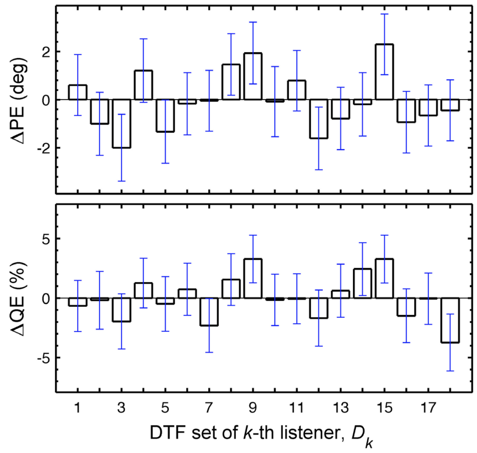


FIGURE 6 | Listener-specific performance depends on the DTF set used in the model. ΔPEs and ΔQEs averaged over all U_ks as a function of D_k. ΔPEs and ΔQEs are the PEs and QEs relative to the listener-specific PEs and QEs, respectively. The whiskers show ±1 SD.

Recall that in our investigation, both the incoming sound and the template set were filtered by the same DTF set, corresponding to a condition where the listener is completely re-calibrated to those DTFs. The highest correlation found for NH68's DTF set does not imply that this DTF set is optimal for *ad-hoc* listening.

In summary, the predicted localization performance varied by a small amount depending on the directional cues provided by the different DTF sets, even when the listener-specific uncertainty was considered. Note that full re-calibration was simulated. This finding indicates that some of the DTF sets provide better access to directional cues than others. Even though the acoustic factor might contribute to the variability in localization performance across listeners, the same DTF set of a single listener (here, NH68) for modeling performance of all listeners yielded still a good prediction accuracy.

3.4. RELATIVE CONTRIBUTIONS OF ACOUSTIC AND NON-ACOUSTIC FACTORS

Both the DTF set and the uncertainty had an effect on the predicted localization performance. However, a listener-constant DTF set provided still acceptable predictions, while a listener-constant uncertainty did not. In this section, we aim at directly comparing the relative contributions of the two factors to localization performance. To this end, we compare the SDs in the predictions as a function of each of the factors. The factor causing more variation in the predictions is assumed to have more impact on sound localization.

We used PEs and QEs predicted for all combinations of uncertainties and DTF sets, as shown in **Figure 5**. For each listener and each performance metric, two SDs were calculated: (1) as a function of the listener-specific DTF set D_k for all available uncertainties, i.e., calculating the SDs across a column separately for each row; and (2) as a function of the listener-specific uncertainty U_k for all available DTF sets, i.e. calculating the SD across

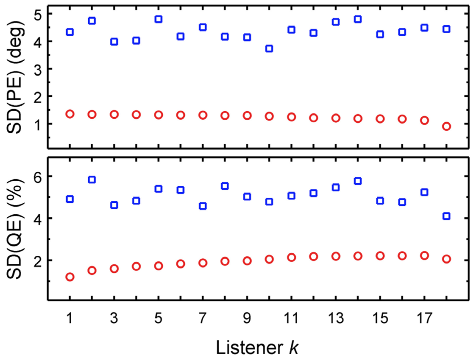


FIGURE 7 | DTF set contributes less than uncertainty to the performance variability of the group. PE SDs and QE SDs as functions of either listener-constant DTF set calculated for listener-specific uncertainties (U_k varied, blue squares) or the listener-constant uncertainty calculated for listener-specific DTF sets (DTF varied, red circles). The abscissa is sorted by the ascending listener-specific uncertainty U_k.

a row separately for each column. **Figure 7** shows these SDs as functions of the *k*-th listener, sorted by ascending listener-specific uncertainty. When U_k was varied, the average SD across listeners was 4.4° ± 0.3° and 5.1% ± 0.4% for PE and QE, respectively. When the DTF set was varied, the average SD was 1.2° ± 0.1° and 1.9% ± 0.3% for PE and QE, respectively. On average, the factor uncertainty caused more than twice as much variability as the factor DTF set.

This analysis shows that while both listener-specific uncertainty and listener-specific DTF set were important for the accuracy in predicted localization performance, the uncertainty affected the performance much more than the DTF set. This indicates that the non-acoustic factor, uncertainty, contributes more than the acoustic factor, DTF set, to the localization performance. This is consistent with the observations of Andéol et al. (2013), where localization performance correlated with the detection thresholds for spectral modulation, but did not correlate with the prominence of the HRTF's spectral shape. The directional information captured by the spectral shape prominence corresponds to the acoustic factor in our study. The sensitivity to the spectral modulations represents the non-acoustic factor in our study. Even though the acoustic factor (DTF set) contributed to the localization performance of an individual listener, the differences *between* the listeners seem to be more determined by a non-acoustic factor (uncertainty).

Note that the separation of the sound localization process into acoustic and non-acoustic factors in our model assumes a perfect calibration of a listener to a DTF set. It should be considered, though, that listeners might actually be calibrated at different levels to their own DTFs. In such a case, the potentially different levels of calibration would be implicitly considered in the model by different uncertainties, confounding the interpretation of the relative contribution of the acoustic and non-acoustic factors. While the general capability to *re*-calibrate to a new DTF set has been investigated quite well (Hofman and van Opstal, 1998; Majdak et al., 2013), the level of calibration to the own DTF set has not been clarified yet.

4. CONCLUSIONS

In this study, a sound localization model predicting the localization performance in sagittal planes (Baumgartner et al., 2013) was applied to investigate the relative contributions of acoustic and non-acoustic factors to localization performance in the lateral range of ±30°. The acoustic factor was represented by the directional cues provided by the DTF sets of individual listeners. The non-acoustic factor was represented by the listener-specific uncertainty considered to describe processes related to the efficiency of processing the spectral cues. Listener-specific uncertainties were estimated in order to calibrate the model to the actual performance when localizing broadband noises with own ears. Then, predictions were calculated for the permutation of DTF sets and uncertainties across the listener group. Identical DTF sets were used for the incoming sound and the template set, which allowed to simulate the listeners being completely re-calibrated to the tested DTF sets, a condition nearly unachievable in psychoacoustic localization experiments.

Our results show that both the acoustic and non-acoustic factors affected the modeled localization performance. The non-acoustic factor had a strong effect on the predictions, and accounted very well for the differences between the individual listeners. In comparison, the acoustic factor had much less effect on the predictions. In an extreme case of using the same DTF set for modeling performance for all listeners, an acceptable prediction accuracy was still obtained.

Note that our investigation considered only targets positioned in sagittal planes of ±30° around the median plane. Even though we do not have evidence for contradicting conclusions for more lateral sagittal planes, one should be careful when applying our conclusions to more lateral targets. Further, the model assumes direction-static and stationary stimuli presented in the free field. In realistic listening situations, listeners can move their head, the acoustic signals are temporally fluctuating, and reverberation interacts with the direct sound.

An unexpected conclusion from our study is that, globally, i.e., on average across all considered directions, all the tested DTF sets encoded the directional information similarly well. It seems like listener-specific DTFs are not necessarily required for predicting the global listener-specific localization ability in terms of distinguishing between bad and good localizers. What seems to be required, however, is an accurate estimate of the listener-specific uncertainty. One could speculate that, given a potential relation between the uncertainty and a measure of spectral-shape sensitivity, in the future, the global listener-specific localization ability might be predictable by obtaining a measure of the listener-specific uncertainty in a non-spatial experimental task without any requirement of listener-specific localization responses.

ACKNOWLEDGMENTS

We thank Christian Kasess for fruitful discussions on data analysis. This work was supported by the Austrian Science Fund (FWF, projects P 24124–N13 and M1230).

REFERENCES

Algazi, V. R., Avendano, C., and Duda, R. O. (2001). Elevation localization and head-related transfer function analysis at low frequencies. *J. Acoust. Soc. Am.* 109, 1110–1122. doi: 10.1121/1.1349185

Andéol, G., Macpherson, E. A., and Sabin, A. T. (2013). Sound localization in noise and sensitivity to spectral shape. *Hear. Res.* 304, 20–27. doi: 10.1016/j.heares.2013.06.001

Baumgartner, R., Majdak, P., and Laback, B. (2013). "Assessment of sagittal-plane sound localization performance in spatial-audio applications," in *The Technology of Binaural Listening, Modern Acoustics and Signal Processing*, ed J. Blauert (Berlin; Heidelberg: Springer), 93–119.

Baumgartner, R., Majdak, P., and Laback, B. (2014). *Modeling Sound-Source Localization in Sagittal Planes for Human Listeners*. Available online at: http://www.kfs.oeaw.ac.at/research/Baumgartner_et_al_2014.pdf. (Last modified April 10, 2014).

Dau, T., Püschel, D., and Kohlrausch, A. (1996). A quantitative model of the "effective" signal processing in the auditory system. I. Model structure. *J. Acoust. Soc. Am.* 99, 3615–3622. doi: 10.1121/1.414959

Glasberg, B. R., and Moore, B. C. J. (1990). Derivation of auditory filter shapes form notched-noise data. *Hear. Res.* 47, 103–138. doi: 10.1016/0378-5955(90)90170-T

Goupell, M. J., Majdak, P., and Laback, B. (2010). Median-plane sound localization as a function of the number of spectral channels using a channel vocoder. *J. Acoust. Soc. Am.* 127, 990–1001. doi: 10.1121/1.3283014

Hofman, P. M., and van Opstal, J. (1998). Spectro-temporal factors in two-dimensional human sound localization. *J. Acoust. Soc. Am.* 103, 2634–2648. doi: 10.1121/1.422784

Hofman, P. M., van Riswick, J. G. A., and van Opstal, J. (1998). Relearning sound localization with new ears. *Nat. Neurosci.* 1, 417–421. doi: 10.1038/1633

Langendijk, E. H. A., and Bronkhorst, A. W. (2002). Contribution of spectral cues to human sound localization. *J. Acoust. Soc. Am.* 112, 1583–1596. doi: 10.1121/1.1501901

Macpherson, E. A., and Sabin, A. T. (2007). Binaural weighting of monaural spectral cues for sound localization. *J. Acoust. Soc. Am.* 121, 3677–3688. doi: 10.1121/1.2722048

Majdak, P., Balazs, P., and Laback, B. (2007). Multiple exponential sweep method for fast measurement of head-related transfer functions. *J. Audio. Eng. Soc.* 55, 623–637.

Majdak, P., Goupell, M. J., and Laback, B. (2010). 3-D localization of virtual sound sources: effects of visual environment, pointing method, and training. *Atten. Percept. Psycho.* 72, 454–469. doi: 10.3758/APP.72.2.454

Majdak, P., Walder, T., and Laback, B. (2013). Effect of long-term training on sound localization performance with spectrally warped and band-limited head-related transfer functions. *J. Acoust. Soc. Am.* 134, 2148–2159. doi: 10.1121/1.4816543

Middlebrooks, J. C. (1999a). Individual differences in external-ear transfer functions reduced by scaling in frequency. *J. Acoust. Soc. Am.* 106, 1480–1492. doi: 10.1121/1.427176

Middlebrooks, J. C. (1999b). Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. *J. Acoust. Soc. Am.* 106, 1493–1510. doi: 10.1121/1.427147

Møller, H., Sørensen, M. F., Hammershøi, D., and Jensen, C. B. (1995). Head-related transfer functions of human subjects. *J. Audio. Eng. Soc.* 43, 300–321.

Morimoto, M. (2001). The contribution of two ears to the perception of vertical angle in sagittal planes. *J. Acoust. Soc. Am.* 109, 1596–1603. doi: 10.1121/1.1352084

Parseihian, G., and Katz, B. F. G. (2012). Rapid head-related transfer function adaptation using a virtual auditory environment. *J. Acoust. Soc. Am.* 131, 2948–2957. doi: 10.1121/1.3687448

Patterson, R., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (1988). *An Efficient Auditory Filterbank Based on the Gammatone Function*. Cambridge: APU

Rakerd, B., Hartmann, W. M., and McCaskey, T. L. (1999). Identification and localization of sound sources in the median sagittal plane. *J. Acoust. Soc. Am.* 106, 2812–2820. doi: 10.1121/1.428129

Søndergaard, P., and Majdak, P. (2013). "The auditory modeling toolbox," in *The Technology of Binaural Listening, Modern Acoustics and Signal Processing*, ed J. Blauert (Berlin; Heidelberg: Springer), 33–56.

van Wanrooij, M. M., and van Opstal, J. (2005). Relearning sound localization with a new ear. *J. Neurosci.* 25, 5413–5424. doi: 10.1523/JNEUROSCI.0850-05.2005

Vliegen, J., and van Opstal, J. (2004). The influence of duration and level on human sound localization. *J. Acoust. Soc. Am.* 115, 1705–1703. doi: 10.1121/1.1687423

Wightman, F. L., and Kistler, D. J. (1997). Monaural sound localization revisited. *J. Acoust. Soc. Am.* 101, 1050–1063. doi: 10.1121/1.418029

Zahorik, P., Bangayan, P., Sundareswaran, V., Wang, K., and Tam, C. (2006). Perceptual recalibration in human sound localization: learning to remediate front-back reversals. *J. Acoust. Soc. Am.* 120, 343–359. doi: 10.1121/1.2208429

Zakarauskas, P., and Cynader, M. S. (1993). A computational theory of spectral cue localization. *J. Acoust. Soc. Am.* 94, 1323–1331. doi: 10.1121/1.408160

Zhang, P. X., and Hartmann, W. M. (2010). On the ability of human listeners to distinguish between front and back. *Hear. Res.* 260, 30–46. doi: 10.1016/j.heares.2009.11.001

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 October 2013; accepted: 27 March 2014; published online: 23 April 2014.

Citation: Majdak P, Baumgartner R and Laback B (2014) Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization. Front. Psychol. ***5****:319. doi: 10.3389/fpsyg.2014.00319*

This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Psychology.

Chapter 4

Modeling sound-source localization in sagittal planes for human listeners

This work was published as

Baumgartner, R., Majdak, P., Laback, B. (2014): Modeling sound-source localization in sagittal planes for human listeners, in: Journal of the Acoustical Society of America 136, 791-802. doi:10.1121/1.4887447

The idea behind the significant model improvements with regard to the model version described in chapter 2 came from me, the first author. The work was designed by me and the second author. I developed and implemented the model, and benefited from many discussions with the second author. I also evaluated the model, created the figures and drafted the manuscript. The second and third authors revised the manuscript.

Modeling sound-source localization in sagittal planes for human listeners

Robert Baumgartner,[a] Piotr Majdak, and Bernhard Laback
Acoustics Research Institute, Austrian Academy of Sciences, Wohllebengasse 12-14, A-1040 Vienna, Austria

(Received 8 November 2013; revised 12 June 2014; accepted 17 June 2014)

Monaural spectral features are important for human sound-source localization in sagittal planes, including front-back discrimination and elevation perception. These directional features result from the acoustic filtering of incoming sounds by the listener's morphology and are described by listener-specific head-related transfer functions (HRTFs). This article proposes a probabilistic, functional model of sagittal-plane localization that is based on human listeners' HRTFs. The model approximates spectral auditory processing, accounts for acoustic and non-acoustic listener specificity, allows for predictions beyond the median plane, and directly predicts psychoacoustic measures of localization performance. The predictive power of the listener-specific modeling approach was verified under various experimental conditions: The model predicted effects on localization performance of band limitation, spectral warping, non-individualized HRTFs, spectral resolution, spectral ripples, and high-frequency attenuation in speech. The functionalities of vital model components were evaluated and discussed in detail. Positive spectral gradient extraction, sensorimotor mapping, and binaural weighting of monaural spatial information were addressed in particular. Potential applications of the model include predictions of psychophysical effects, for instance, in the context of virtual acoustics or hearing assistive devices.

© *2014 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4887447]

PACS number(s): 43.66.Qp, 43.66.Ba, 43.66.Pn [JFC] Pages: 791–802

I. INTRODUCTION

Human listeners use monaural spectral features to localize sound sources, particularly when binaural localization cues are absent (Agterberg *et al.*, 2012) or ambiguous (Macpherson and Middlebrooks, 2002). Ambiguity of binaural cues usually arises along the polar dimension on sagittal planes, i.e., when estimating the vertical position of the source (e.g., Vliegen and Opstal, 2004) and when distinguishing between front and back (e.g., Zhang and Hartmann, 2010). Head-related transfer functions (HRTFs) describe the acoustic filtering of the torso, head, and pinna (Møller *et al.*, 1995) and thus the monaural spectral features.

Several psychoacoustic studies have addressed the question of which monaural spectral features are relevant for sound localization. It is well known that the amplitude spectrum of HRTFs is most important for localization in sagittal planes (e.g., Kistler and Wightman, 1992), whereas the phase spectrum of HRTFs affects localization performance only for very specific stimuli with large spectral differences in group delay (Hartmann *et al.*, 2010). Early investigations attempted to identify spectrally local features like specific peaks and/or notches as localization cues (e.g., Blauert, 1969; Hebrank and Wright, 1974). Middlebrooks (1992) could generalize those attempted explanations in terms of a spectral correlation model. However, the actual mechanisms of the auditory system used to extract localization cues remained unclear.

Neurophysiological findings suggest that mammals decode monaural spatial cues by extracting spectral gradients rather than center frequencies of peaks and notches. May (2000) lesioned projections from the dorsal cochlear nucleus (DCN) to the inferior colliculus of cats and demonstrated by behavioral experiments that the DCN is crucial for sagittal-plane sound localization. Reiss and Young (2005) investigated in depth the role of the cat DCN in coding spatial cues and provided strong evidence for sensitivity of the DCN to positive spectral gradients. As far as we are aware of, however, the effect of positive spectral gradients on sound localization has not yet been explicitly tested or modeled for human listeners.

In general, existing models of sagittal-plane localization for human listeners can be subdivided into functional models (e. g., Langendijk and Bronkhorst, 2002) and machine learning approaches (e.g., Jin *et al.*, 2000). The findings obtained with the latter are difficult to generalize to signals, persons, and conditions for which the model has not been extensively trained in advance. Hence, the present study focuses on a *functional* model where model parameters correspond to physiologically and/or psychophysically inspired localization parameters in order to better understand the mechanisms underlying spatial hearing in the polar dimension. By focusing on the effect of temporally static modifications of spectral features, we assume the incoming sound (the target) to originate from a single target source and the listeners to have no prior expectations regarding the direction of this target.

The first explicit functional models of sound localization based on spectral shape cues were proposed by Middlebrooks (1992), Zakarauskas and Cynader (1993), as well as Hofman and Opstal (1998). Based on these approaches, Langendijk and Bronkhorst (2002) proposed a probabilistic extension to model their results from localization experiments. All these models roughly approximate peripheral auditory processing in order to obtain internal spectral representations of the

[a]Author to whom correspondence should be addressed. Electronic mail: robert.baumgartner@oeaw.ac.at

incoming sounds. Furthermore, they follow a template-based approach, assuming that listeners create an internal template set of their specific HRTFs as a result of a monaural learning process (Hofman *et al.*, 1998; van Wanrooij and van Opstal, 2005). The more similar the representation of the incoming sound compared to a specific template, the larger the assumed probability of responding at the polar angle that corresponds to this template. Langendijk and Bronkhorst (2002) demonstrated good correspondence between their model predictions and experimental outcomes for individual listeners by means of likelihood statistics.

Recently, we proposed a method to compute psychoacoustic performance measures of confusion rates, accuracy, and precision from the output of a probabilistic localization model (Baumgartner *et al.*, 2013). In contrast to earlier approaches, this model considered a non-acoustic, listener-specific factor of spectral sensitivity that has been shown to be essential for capturing the large inter-individual differences of localization performance (Majdak *et al.*, 2014). However, the peripheral part of auditory processing has been considered without positive spectral gradient extraction, and the model has been able to predict localization responses only in proximity of the median plane but not for more lateral targets. Thus, in the present study, we propose a model that additionally considers positive spectral gradient extraction and allows for predictions beyond the median plane by approximating the motor response behavior of human listeners.

In Sec. II, the architecture of the proposed model and its parameterization are described in detail. In Sec. III, the model is evaluated under various experimental conditions probing localization with single sound sources at moderate intensities. Finally, in Sec. IV the effects of particular model stages are evaluated and discussed.

II. MODEL DESCRIPTION

Figure 1 shows the structure of the proposed sagittal-plane localization model. Each block represents a processing stage of the auditory system in a functional way. First, the spectral auditory processing of an incoming target sound is approximated in order to obtain the target's internal spectral representation. Then, this target representation is compared to a template set consisting of equivalently processed internal representations of the HRTFs for the given sagittal plane. This comparison process is the basis of the spectro-to-spatial mapping. Finally, the impact of monaural and binaural perceptual factors as well as aspects of sensorimotor mapping are considered in order to yield the polar-angle response prediction.

A. Spectral auditory processing

1. Acoustic filtering

In the model, acoustic transfer characteristics are captured by listener-specific directional transfer functions (DTFs), which are HRTFs with the direction-independent characteristics removed for each ear (Middlebrooks, 1999a). DTFs usually emphasize high-frequency components and are commonly used for sagittal-plane localization experiments with virtual sources (Middlebrooks, 1999b; Langendijk and Bronkhorst, 2002; Goupell *et al.*, 2010; Majdak *et al.*, 2013c).

In order to provide a formal description, the space is divided into N_ϕ mutually exclusive lateral segments orthogonal to the interaural axis. The segments are determined by the lateral centers $\phi_k \in [-90°, 90°)$ from the right- to the left-hand side. In other words, all available DTFs are clustered into N_ϕ sagittal planes. Further, let $\theta_{i,k} \in [-90°, 270°)$, from front below to rear below, for all $k = 1,\ldots,N_\phi$ and $i = 1,\ldots,N_\theta[k]$ denote the polar angles corresponding to the impulse responses, $r_{i,k,\zeta}[n]$, of a listener's set of DTFs. The channel index, $\zeta \in \{L, R\}$, represents the left and right ear, respectively. The linear convolution of a DTF with an arbitrary stimulus, $x[n]$, yields a directional target sound,

$$t_{j,k,\zeta}[n] = (r_{j,k,\zeta} * x)[n]. \qquad (1)$$

2. Spectral analysis

The target sound is then filtered using a gammatone filterbank (Lyon, 1997). For stationary sounds at a moderate intensity, the gammatone filterbank is an established approximation of cochlear filtering (Unoki *et al.*, 2006). In the proposed model, the frequency spacing of the auditory filter bands corresponds to one equivalent rectangular bandwidth. The corner frequencies of the filterbank, $f_{\min} = 0.7$ kHz and $f_{\max} = 18$ kHz, correspond to the minimum frequency thought to be affected by torso reflections (Algazi *et al.*, 2001) and, in approximation, the maximum frequency of the hearing range, respectively. This frequency range is subdivided into $N_b = 28$ bands. $G[n,b]$ denotes the impulse response of the bth auditory filter. The long-term spectral profile, $\mathring{\xi}[b]$ in dB, of the stationary but finite sound, $\xi[n]$, with length N_ξ is given by

$$\mathring{\xi}[b] = 10\log_{10}\frac{1}{N_\xi}\sum_{n=0}^{N_\xi-1}(\xi * G[b])^2[n], \qquad (2)$$

for all $b = 1,\ldots,N_b$. With $\xi[n] = t_{j,k,\zeta}[n]$ and $\xi[n] = r_{i,k,\zeta}[n]$ for the target sound and template, respectively, Eq. (2) yields the corresponding spectral profiles, $\mathring{t}_{j,k,\zeta}[b]$ and $\mathring{r}_{j,k,\zeta}[b]$.

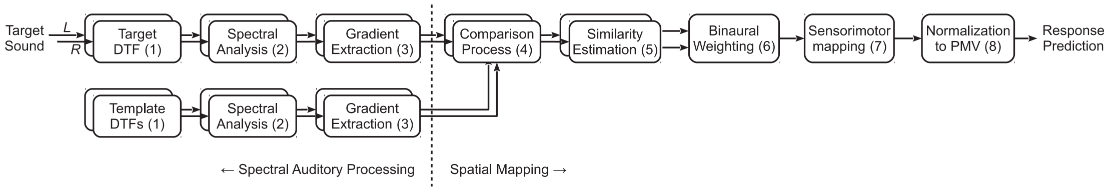


FIG. 1. Structure of the sagittal-plane localization model. Numbers in brackets correspond to equations derived in the text.

3. Positive spectral gradient extraction

In cats, the DCN is thought to extract positive spectral gradients from spectral profiles. Reiss and Young (2005) proposed a DCN model consisting of three units, namely, DCN type IV neurons, DCN type II interneurons, and wideband inhibitors. The DCN type IV neurons project from the auditory nerve to the inferior colliculus, the DCN type II interneurons inhibit those DCN type IV neurons with best frequencies just above the type II best frequencies, and the wideband units inhibit both DCN type II and IV units, albeit the latter to a reduced extent. In Reiss and Young (2005), this model explained most of the measured neural responses to notched-noise sweeps.

Inspired by this DCN functionality, for all $b = 2,\ldots,N_b$, we consider positive spectral gradient extraction in terms of

$$\tilde{\xi}[b] = \max\left(\mathring{\xi}[b] - \mathring{\xi}[b-1], 0\right). \qquad (3)$$

Hence, with $\mathring{\xi}[b] = \mathring{r}_{i,k,\zeta}[b]$ and $\mathring{\xi}[b] = \mathring{t}_{j,k,\zeta}[b]$ we obtain the internal representations $\tilde{r}_{i,k,\zeta}[b]$ and $\tilde{t}_{j,k,\zeta}[b]$, respectively. In relation to the model from Reiss and Young (2005), the role of the DCN type II interneurons is approximated by computing the spectral gradient and the role of the wideband inhibitors by restricting the selection to positive gradients. Interestingly, Zakarauskas and Cynader (1993) already discussed the potential of a spectral gradient metric in decoding spectral cues. However, in 1993, they had no neurophysiological evidence for this decoding strategy and did not consider the restriction to positive gradients.

B. Spatial mapping

1. Comparison process

Listeners are able to map the internal target representation to a direction in the polar dimension. In the proposed model, this mapping is implemented as a comparison process between the target representation and each template. Each template refers to a specific polar angle in the given sagittal plane. In the following, this polar angle is denoted as the polar *response angle*, because the comparison process forms the basis of subsequent predictions of the response behavior.

The comparison process results in a distance metric, $\tilde{d}_{j,k,\zeta}[\theta_{i,k}]$, as a function of the polar response angle and is defined as L_1-norm, i.e.,

$$\tilde{d}_{j,k,\zeta}[\theta_{i,k}] = \sum_{b=2}^{N_b} \left|\tilde{r}_{i,k,\zeta}[b] - \tilde{t}_{j,k,\zeta}[b]\right|. \qquad (4)$$

Since sagittal-plane localization is considered to be a monaural process (van Wanrooij and van Opstal, 2005), the comparisons are processed separately for the left and right ear in the model. In general, the smaller the distance metric, the more similar the target is to the corresponding template. If spectral cues show spatial continuity along a sagittal plane, the resulting distance metric is a smooth function of the polar response angle. This function can also show multiple peaks due to ambiguities of spectral cues, for instance between front and back.

2. Similarity estimation

In the next step, the distance metrics are mapped to similarity indices that are considered to be proportional to the response probability. The mapping between the distance metric and the response probability is not fully understood yet, but there is evidence that in addition to the directional information contained in the HRTFs, the mapping is also affected by non-acoustic factors like the listener's specific ability to discriminate spectral envelope shapes (Andéol *et al.*, 2013).

Langendijk and Bronkhorst (2002) modeled the mapping between distance metrics and similarity indices by somewhat arbitrarily using a Gaussian function with zero mean and a listener-constant standard deviation (their $S = 2$) yielding best predictions for their tested listeners. Baumgartner *et al.* (2013) pursued this approach while considering the standard deviation of the Gaussian function as a listener-specific factor of spectral sensitivity (called uncertainty parameter). Recent investigations have shown that this factor is essential to represent a listener's general localization ability, much more than the listener's HRTFs (Majdak *et al.*, 2014). Varying the standard deviation of a Gaussian function scales both its inflection point and its slope. In order to better distinguish between those two parameters in the presently proposed model, the mapping between the distance metric and the perceptually estimated similarity, $\tilde{s}_{j,k,\zeta}[\theta_{i,k}]$, is modeled as a two-parameter function with a shape similar to the single-sided Gaussian function, namely, a sigmoid psychometric function,

$$\tilde{s}_{j,k,\zeta}[\theta_{i,k}] = 1 - \left(1 + e^{-\Gamma(\tilde{d}_{j,k,\zeta}[\theta_{i,k}] - S_l)}\right)^{-1}, \qquad (5)$$

where Γ denotes the degree of selectivity and S_l denotes the listener-specific sensitivity. Basically, the lower S_l, the higher the sensitivity of the listener to discriminate internal spectral representations. The strength of this effect depends on Γ. A small Γ corresponds to a shallow psychometric function and means that listeners estimate spectral similarity rather gradually. Consequently, a small Γ reduces the effect of S_l. In contrast, a large Γ corresponds to a steep psychometric function and represents a rather dichotomous estimation of similarity, strengthening the effect of S_l.

3. Binaural weighting

Up to this point, spectral information is analyzed separately for each ear. When combining the two monaural outputs, binaural weighting has to be considered. Morimoto (2001) showed that while both ears contribute equally in the median plane, the contribution of the ipsilateral ear increases monotonically with increasing lateralization. The contribution of the contralateral ear becomes negligible at magnitudes of lateral angles beyond 60°. Macpherson and Sabin (2007) further demonstrated that binaural weighting depends on the perceived lateral location, and they quantified the relative contribution of each ear at a ±45° lateral angle.

In order to determine the binaural weighting as a continuous function of the lateral response angle, ϕ_k, we attempted to fit a somehow arbitrarily chosen sigmoid function, $w_L(\phi_k) = (1 + e^{-\phi_k/\Phi})^{-1}$ and $w_R(\phi_k) = 1 - w_L(\phi_k)$ for the left and right ears, respectively, to the anchor points from the

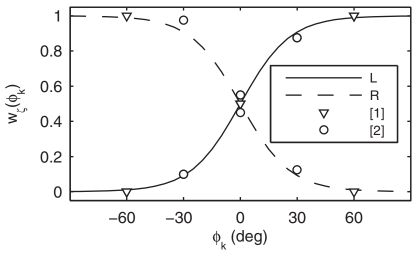


FIG. 2. Binaural weighting function best fitting results from Morimoto (2001) labeled as [1] and Macpherson and Sabin (2007) labeled as [2] in a least squared error sense.

two experiments described above. The lower the choice of the binaural weighting coefficient, Φ, the larger the relative contribution of the ipsilateral ear becomes. Figure 2 shows that choosing $\Phi = 13^\circ$ yields a weighting consistent with the outcomes of the two underlying studies (Morimoto, 2001; Macpherson and Sabin, 2007). The weighted sum of the monaural similarity indices finally yields the overall similarity index,

$$\tilde{\mathrm{s}}_{j,k}[\theta_{i,k}] = \sum_{\zeta:\{L,R\}} \mathrm{w}_\zeta(\phi_k) \cdot \tilde{\mathrm{s}}_{j,k,\zeta}[\theta_{i,k}]. \tag{6}$$

4. Sensorimotor mapping

When asked to respond to a target sound by pointing, listeners map their auditory perception to a motor response. This mapping is considered to result from several subcortical (King, 2004) and cortical (Bizley *et al.*, 2007) processes. In the proposed model, the overall effect of this complex multi-layered process is condensed by means of a scatter that smears the similarity indices along the polar dimension. Since we have no evidence for a specific overall effect of sensorimotor mapping on the distribution of response probabilities, using a Gaussian scatter approximates multiple independent neural and motor processes according to the central limit theorem. The scatter, ε, is defined in the body-centered frame of reference (elevation dimension; Redon and Hay, 2005). The projection of a scatter constant in elevation into the auditory frame of reference (polar dimension) yields a scatter reciprocal to the cosine of the lateral angle. Hence, the more lateral the response, the larger the scatter becomes in polar dimension. In the model, the motor response behavior is obtained by a circular convolution between the vector of similarity indices and a circular normal distribution, $\rho(x;\, \mu,\, \kappa)$, with location $\mu = 0$ and a concentration, κ, depending on ε:

$$\tilde{\mathrm{p}}_{j,k}[\theta_{i,k}] = \tilde{\mathrm{s}}_{j,k}[\theta_{i,k}] \circledast \rho\left(\theta_{i,k}; 0, \frac{\cos^2\phi_k}{\varepsilon^2}\right). \tag{7}$$

The operation in Eq. (7) requires the polar response angle being regularly sampled. Thus, spline interpolation is applied before if regular sampling is not given.

5. Normalization to probabilities

In order to obtain a probabilistic prediction of the response behavior, $\tilde{\mathrm{p}}_{j,k}[\theta_{i,k}]$ is assumed to be proportional to the listener's response probability for a certain polar angle. Thus, we scale the vector of similarity indices such that its sum equals one; this yields a probability mass vector (PMV) representing the prediction of the response probability,

$$\mathrm{p}_{j,k}[\theta_{i,k}] = \frac{\tilde{\mathrm{p}}_{j,k}[\theta_{i,k}]}{\sum\limits_{\iota=1}^{N_\theta[k]} \tilde{\mathrm{p}}_{j,k}[\theta_{\iota,k}]}. \tag{8}$$

Examples of such probabilistic predictions are shown in Fig. 3 for the median plane. For each polar target angle, predicted PMVs are illustrated and encoded by brightness. Actual responses from the three listeners are shown as open circles.

C. Psychoacoustic performance measures

Given a probabilistic response prediction, psychoacoustic performance measures can be calculated by means of expectancy values. In the context of sagittal-plane localization, those measures are often subdivided into measuring either local performance, i.e., accuracy and precision for responses close to the target position, or global performance in terms of localization confusions. In order to define local polar-angle responses, let $\mathcal{A}_k = \{i \in \mathbb{N} : 1 \le i \le \mathrm{N}_\theta[k], |\theta_{i,k} - \vartheta_{j,k}| \bmod 180^\circ < 90^\circ\}$ denote the set of indices corresponding to local responses, $\theta_{i,k}$, to a target positioned at $\vartheta_{j,k}$. Hereafter, we use the quadrant error rate (Middlebrooks, 1999b), denoted as QE, to quantify localization confusions; it measures the rate of non-local responses in terms of

$$\mathrm{QE}_{j,k} = \sum_{i \in \mathcal{A}_k} \mathrm{p}_{j,k}[\theta_{i,k}]. \tag{9}$$

Within the local response range, we quantify localization performance by evaluating the polar root-mean-square error (Middlebrooks, 1999b), denoted as PE, which describes the effects of both accuracy and precision. The expectancy value of this metric in degrees is given by

$$\mathrm{PE}_{j,k} = \sqrt{\frac{\sum\limits_{i \in \mathcal{A}_k} (\theta_{i,k} - \vartheta_{j,k})^2 \mathrm{p}_{j,k}[\theta_{i,k}]}{\sum\limits_{i \in \mathcal{A}_k} \mathrm{p}_{j,k}[\theta_{i,k}]}}. \tag{10}$$

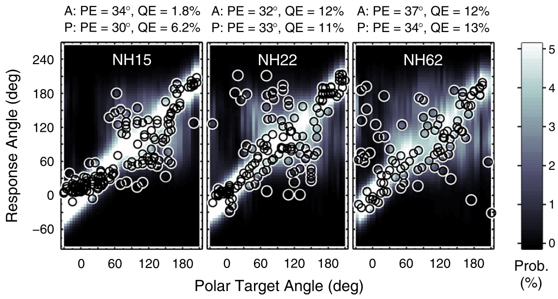


FIG. 3. (Color online) Prediction examples. Actual responses and response predictions for three exemplary listeners when listening to median-plane targets in the baseline condition. Actual response angles are shown as open circles. Probabilistic response predictions are encoded by brightness according to the color bar to the right. A: Actual. P: Predicted.

Alternatively, any type of performance measure can also be retrieved by generating response patterns from the probabilistic predictions. To this end, for each target angle, random responses are drawn according to the corresponding PMV. The resulting patterns are then treated in the same way as if they would have been obtained from real psychoacoustic localization experiments. On average, this procedure yields the same results as computing the expectancy values.

In Fig. 3, actual (A:) and predicted (P:) PE and QE are listed for comparison above each panel. Non-local responses, for instance, were more frequent for NH22 and NH62 than for NH15, thus yielding larger QE. The model parameters used for these predictions were derived as follows.

D. Parameterization

The sagittal-plane localization model contains three free parameters, namely, the degree of selectivity, Γ, the sensitivity, S_l, and the motor response scatter, ε. These parameters were optimized in order to yield a model with the smallest prediction residue, *e*, between modeled and actual localization performance.

The actual performance was obtained in experiments (Goupell *et al.*, 2010; Majdak *et al.*, 2010; Majdak *et al.*, 2013b; Majdak *et al.*, 2013c) with human listeners localizing Gaussian white noise bursts with a duration of 500 ms. The targets were presented across the whole lateral range in a virtual auditory space. The listeners, 23 in total and called the *pool*, had normal hearing (NH) and were between 19 and 46 yrs old at the time of the experiments.

For model predictions, the data were pooled within 20°-wide lateral segments centered at $\phi_k = 0°$, ±20°, ±40°, etc. The aim was to account for changes of spectral cues, binaural weighting, and compression of the polar angle dimension with increasing magnitude of the lateral angle. If performance predictions were combined from different segments, the average was weighted relative to the occurrence rates of targets in each segment.

Both modeled and actual performance was quantified by means of QE and PE. Prediction residues were calculated for these performance measures. The residues were pooled across listeners and lateral target angles in terms of the root mean square weighted again by the occurrence rates of targets. The resulting residues are called the partial prediction residues, e_{QE} and e_{PE}.

The optimization problem can be finally described as

$$\{\Gamma_{\mathrm{opt}}, S_{l\,\mathrm{opt}}, \varepsilon_{\mathrm{opt}}\} = \underset{(\Gamma, S_l, \varepsilon)}{\arg\min}\, e(\Gamma, S_l, \varepsilon), \tag{11}$$

with the joint prediction residue,

$$e(\Gamma, S_l, \varepsilon) = e_{\mathrm{QE}}(\Gamma, S_l, \varepsilon)/\mathrm{QE}^{(\mathrm{c})} + e_{\mathrm{PE}}(\Gamma, S_l, \varepsilon)/\mathrm{PE}^{(\mathrm{c})}.$$

The chance rates, $\mathrm{QE}^{(\mathrm{c})}$ and $\mathrm{PE}^{(\mathrm{c})}$, result from $p_{j,k}[\theta_{i,k}] = 1/N_\theta[k]$ and represent the performance of listeners randomly guessing the position of the target. Recall that the sensitivity, S_l, is considered a listener-specific parameter, whereas the remaining two parameters, Γ and ε, are considered identical for all listeners. S_l was optimized on the basis of targets in the proximity of the median plane (±30°) only, because most of the responses were within this range, because listeners' responses are usually most consistent around the median plane, and in order to limit the computational effort.

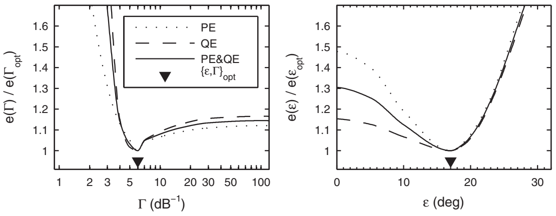


FIG. 4. Model parameterization. Partial (e_{PE}, e_{QE}) and joint (*e*) prediction residues as functions of the degree of selectivity (Γ) and the motor response scatter (ε). Residue functions are normalized to the minimum residue obtained for the optimal parameter value. See text for details.

Figure 4 shows the joint and partial prediction residues scaled by the minimum residues as functions of either Γ or ε. When Γ was varied systematically (left panel), ε and S_l were optimal in terms of yielding minimum *e* for each tested Γ. The joint prediction residue was minimum at $\Gamma_{\mathrm{opt}} = 6\,\mathrm{dB}^{-1}$. For $\Gamma < \Gamma_{\mathrm{opt}}$, the functions for both e_{QE} and e_{PE} steeply decrease, but for $\Gamma < \Gamma_{\mathrm{opt}}$, the increase is less steep, especially for e_{QE}. For ε (right panel), Γ was fixed to the optimum, $\Gamma_{\mathrm{opt}} = 6\,\mathrm{dB}^{-1}$, and only S_l was optimized for each tested ε. In this case, the functions are more different for the two error types. On the one hand, e_{PE} as a function of ε showed a distinct minimum. On the other hand, e_{QE} as a function of ε showed less effect for ε < 17°, because for small ε, non-local responses were relatively rare in the baseline condition. The joint metric was minimum at

TABLE I. Listener-specific sensitivity, S_l, calibrated on the basis of N baseline targets in proximity of the median plane (±30°) with $\Gamma = 6\,\mathrm{dB}^{-1}$ and ε = 17°. Listeners are listed by ID. Actual and predicted QE and PE are shown pairwise (Actual | Predicted).

ID	N	S_l	QE (%)	PE (deg)
NH12	1506	0.26	2.19 | 3.28	27.7 | 26.9
NH14	140	0.58	5.00 | 4.42	25.5 | 26.1
NH15	996	0.55	2.51 | 5.76	33.3 | 30.0
NH16	960	0.63	5.83 | 8.00	31.9 | 28.9
NH17	364	0.76	7.69 | 8.99	33.8 | 32.1
NH18	310	1.05	20.0 | 20.0	36.4 | 36.4
NH21	291	0.71	9.62 | 10.0	34.0 | 33.3
NH22	266	0.70	10.2 | 10.3	33.6 | 33.4
NH33	275	0.88	17.1 | 17.8	35.7 | 34.4
NH39	484	0.86	10.7 | 12.0	37.4 | 35.5
NH41	264	1.02	18.9 | 17.7	37.1 | 39.7
NH42	300	0.44	3.67 | 6.20	30.0 | 27.1
NH43	127	0.44	1.57 | 6.46	34.0 | 28.0
NH46	127	0.46	3.94 | 4.78	28.5 | 27.5
NH53	164	0.52	1.83 | 3.42	26.5 | 24.9
NH55	123	0.88	9.76 | 12.6	38.1 | 33.4
NH57	119	0.97	19.3 | 16.8	28.0 | 33.4
NH58	153	0.21	1.96 | 2.75	24.5 | 23.8
NH62	282	0.98	11.3 | 13.2	38.6 | 35.5
NH64	306	0.84	9.48 | 9.68	33.5 | 33.1
NH68	269	0.76	11.9 | 11.7	32.4 | 32.9
NH71	104	0.76	9.62 | 9.32	33.1 | 33.5
NH72	304	0.79	10.9 | 12.6	38.0 | 35.3

$\varepsilon_{\text{opt}} = 17°$. This scatter is similar to the unimodal response scatter of 17° observed by Langendijk and Bronkhorst (2002).

In conclusion, the experimental data was best described by the model with the parameters set to $\Gamma = 6\,\text{dB}^{-1}$, $\varepsilon = 17°$, and S_l according to Table I. Table I also lists the actual and predicted performance for all listeners within the lateral range of ±30°. On average across listeners, this setting leads to prediction residues of $e_{\text{QE}} = 1.7\%$ and $e_{\text{PE}} = 2.4°$, and the correlations between actual and predicted listener-specific performance are $r_{\text{QE}} = 0.97$ and $r_{\text{PE}} = 0.84$. The same parameter setting was used for all evaluations described in Sec. III.

III. MODEL EVALUATION

Implicitly, the model evaluation has already begun in Sec. II, where the model was parameterized for the listener-specific baseline performance, showing encouraging results. In the present section, we further evaluate the model on predicting the effects of various HRTF modifications. First, predictions are presented for effects of band limitation, spectral warping (Majdak *et al.*, 2013c), and spectral resolution (Goupell *et al.*, 2010) on localization performance. For these studies, listener-specific DTFs, actual target positions, and the corresponding responses were available for all participants. Further, the pool (see Sec. II D) was used to model results of localization studies for which the listener-specific data were not available. In particular, the model was evaluated for the effects of non-individualized HRTFs (Middlebrooks, 1999b), spectral ripples (Macpherson and Middlebrooks, 2003), and high-frequency attenuation in speech localization (Best *et al.*, 2005). Finally, we discuss the capability of the model for target-specific predictions.

A. Effect of band limitation and spectral warping

Majdak *et al.* (2013c) tested band limitation and spectral warping of HRTFs, motivated by the fact that stimulation in cochlear implants (CIs) is usually limited to frequencies up to about 8 kHz. This limitation discards a considerable amount of spectral information for accurate localization in sagittal planes. From an acoustical point of view, all spectral features can be preserved by spectrally warping the broadband DTFs into this limited frequency range. It is less clear, however, whether this transformation also preserves the perceptual salience of the features. Majdak *et al.* (2013c) tested the performance of localizing virtual sources within a lateral range of ±30° for the DTF conditions broad-band (BB), low-pass filtered (LP), and spectrally warped (W). In the LP condition the cutoff frequency was 8.5 kHz and in the W condition the frequency range from 2.8 to 16 kHz was warped to 2.8 to 8.5 kHz. On average, the 13 NH listeners performed best in the BB and worst in the W condition–see their Fig. 6 (pre-training). In the W condition, the overall performance even approached chance rate.

Model predictions were first evaluated on the basis of the participants' individual DTFs and target positions. For the median plane, Fig. 5 shows the actual response patterns and the corresponding probabilistic response predictions for an exemplary listener in the three experimental conditions. Resulting performance measures are shown above each panel. The similarity between actual and predicted performance is reflected by the visual correspondence between actual responses and bright areas. For predictions considering also lateral targets, Fig. 6 summarizes the performance statistics of all participants. The large correlation coefficients and small prediction residues observed across all participants are in line with the results observed for the exemplary listener. However, there is a noteworthy discrepancy between actual and predicted local performance in the W condition. It seems that in this condition the actual listeners managed to access spatial cues, which were not considered in the model and allowed them to respond a little more accurately than predicted. For instance, interaural spectral differences might be potential cues in adverse listening conditions (Jin *et al.*, 2004).

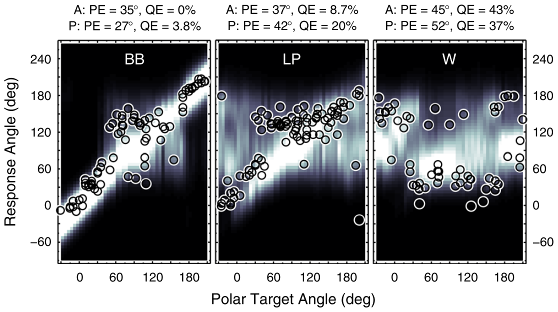


FIG. 5. (Color online) Effect of band limitation and spectral warping. Actual responses and response predictions for listener NH12 in the BB, LP, and W condition from Majdak *et al.* (2013c). Data were pooled within ±15° of lateral angle. All other conventions are as in Fig. 3.

Figure 6 also shows predictions for the listener pool. The predicted performance was only slightly different from that based on the participants' individual DTFs and target positions. Thus, modeling on the basis of our listener pool seems to result in reasonable predictions even if the participants' listener-specific data are unknown. However, the comparison is influenced by the fact that the actual participants are a rather large subset of the pool (13 of 23).

B. Effect of spectral resolution

The localization model was also evaluated on the effect of spectral resolution as investigated by Goupell *et al.* (2010).

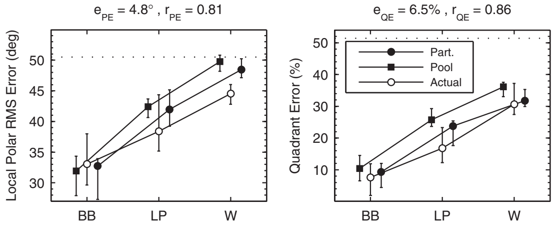


FIG. 6. Effect of band limitation and spectral warping. Listeners were tested in conditions BB, LP, and W. Actual: Experimental results from Majdak *et al.* (2013c). Part.: Model predictions for the actual eight participants based on the actually tested target positions. Pool: Model predictions for our listener pool based on all possible target positions. Symbols and whiskers show median values and inter-quartile ranges, respectively. Symbols were horizontally shifted to avoid overlaps. Dotted horizontal lines represent chance rate. Correlation coefficients, *r*, and prediction residues, *e*, specify the correspondence between actual and predicted listener-specific performance.

This investigation was also motivated by CI listening, because typical CIs suffer from a poor spectral resolution within the available bandwidth due to the spread of electric stimulation. Goupell *et al.* (2010) simulated CI processing at NH listeners, because CI listeners are hard to test as they usually show large inter-individual differences in pathology. For the CI sound simulation, the investigators used a Gaussian-enveloped tone vocoder. In their experiment I, they examined localization performance in the median lateral range (±10°) with the number of vocoder channels as independent variable. To this end, they divided the frequency range from 0.3 to 16 kHz into 3, 6, 9, 12, 18, or 24 channels equally spaced on a logarithmic frequency scale. As a result, the listeners performed worse, the less channels were used—see their Fig. 3.

Our Fig. 7 shows corresponding model predictions for an exemplary listener in three of the seven conditions. Figure 8 shows predictions pooled across listeners for all experimental conditions. Both the listener-specific and pooled predictions showed a systematic degradation in localization performance with a decreasing number of spectral channels, similar to the actual results. However, at less than nine channels, the predicted local performance approached chance rate whereas the actual performance remained better. Thus, it seems that listeners were able to use additional cues that were not considered in the model.

The actual participants from the present experiment, tested on spectral resolution, were a smaller subset of our pool (8 of 23) than in the experiment from Sec. III A (13 of 23), tested on band limitation and spectral warping. Nevertheless, predictions on the basis of the pool and unspecific target positions were again similar to the predictions based on the participants' data. This strengthens the conclusion from Sec. III A that predictions are reasonable even when the participants' listener-specific data are not available.

C. Effect of non-individualized HRTFs

In this section, the model is applied to predict the effect of listening with non-individualized HRTFs, i.e., localizing sounds spatially filtered by the DTFs of another subject. Middlebrooks (1999b) tested 11 NH listeners localizing Gaussian noise bursts with a duration of 250 ms. The listeners were tested with targets spatially encoded by their *own* set of DTFs and also by up to 4 sets of DTFs from *other* subjects (21 cases in total). Targets were presented across the whole lateral range, but for the performance analysis in the polar dimension, only targets within the lateral range of ±30° were considered. Performance was measured by means of QE, PE, and the magnitude of elevation bias in local responses. For the bias analysis, Middlebrooks (1999b) excluded responses from upper-rear quadrants with the argumentation that there the lack of precision overshadowed the overall bias. In general, the study found degraded localization performance for non-individualized HRTFs.

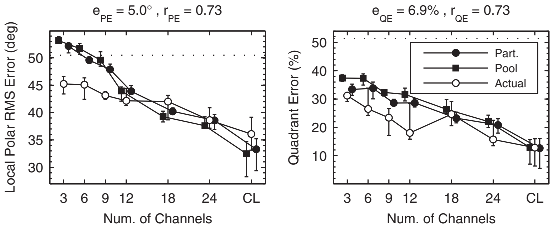


FIG. 8. Effect of spectral resolution in terms of varying the number of spectral channels of a channel vocoder. Actual experimental results are from Goupell *et al.* (2010). CL: Stimulation with broad-band click trains represents an unlimited number of channels. All other conventions are as in Fig. 6.

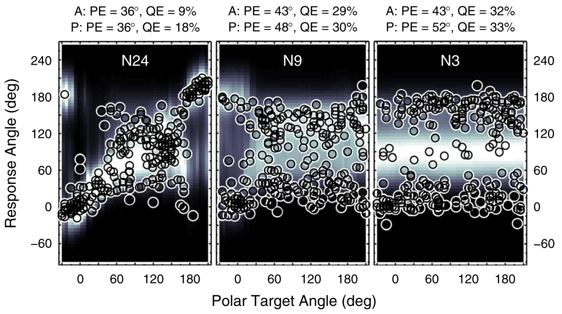


FIG. 7. (Color online) Effect of spectral resolution in terms of varying the number of spectral channels of a channel vocoder. Actual responses and response predictions for exemplary listener NH12. Results for 24, 9, and 3 channels are shown. All other conventions are as in Fig. 3.

Using the model, the performance of each listener of our pool was predicted for the listener's own set of DTFs and for the sets from all other pool members. As the participants were not trained to localize with the DTFs from others, the target sounds were always compared to the listener's own internal template set of DTFs. The predicted magnitude of elevation bias was computed as the expectancy value of the local bias averaged across all possible target positions outside the upper-rear quadrants. Figure 9 shows the experimental results replotted from Middlebrooks (1999b) and our model predictions. The statistics of predicted performance represented by means, medians, and percentiles appear to be quite similar to the statistics of the actual performance.

D. Effect of spectral ripples

Studies considered in Secs. III A–III C probed localization by using spectrally flat source signals. In contrast,

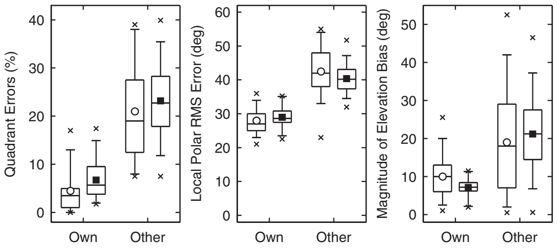


FIG. 9. Effect of non-individualized HRTFs in terms of untrained localization with others' instead of their own ears. Statistics summaries with open symbols represent actual experimental results replotted from Fig. 13 of Middlebrooks (1999b), statistics with filled symbols represent predicted results. Horizontal lines represent 25th, 50th, and 75th percentiles, the whiskers represent 5th and 95th percentiles, and crosses represent minima and maxima. Circles and squares denote mean values.

Macpherson and Middlebrooks (2003) probed localization by using spectrally rippled noises in order to investigate how ripples in the source spectra interfere with directional cues. Ripples were generated within the spectral range of 1 to 16 kHz with a sinusoidal spectral shape in the log-magnitude domain. Ripple depth was defined as the peak-to-trough difference and ripple density as the period of the sinusoid along the logarithmic frequency scale. They tested six trained NH listeners in a dark, anechoic chamber. Targets were 250 ms long and presented via loudspeakers. Target positions ranged across the whole lateral dimension and within a range of $\pm 60°$ elevation (front and back). Localization performance around the median plane was quantified by means of polar error rates. The definition of polar errors relied on an *ad hoc* selective, iterative regression procedure: First, regression lines for responses to baseline targets were fitted separately for front and back; then, responses farther than 45° away from the regression lines were counted as polar errors.

When the ripple depth was kept constant at 40 dB and the ripple density was varied between 0.25 and 8 ripples/octave, participants performed worst at densities around 1 ripple/octave—see their Fig. 6. When the ripple density was kept constant at 1 ripple/octave and the ripple depth was varied between 10 and 40 dB, consistent deterioration with increasing depth was observed—see their Fig. 9. Our Fig. 10 summarizes their results pooled across ripple phases (0 and π), because Macpherson and Middlebrooks (2003) did not observe systematic effects of the ripple phase across listeners. Polar error rates were evaluated relative to listener-specific baseline performance. The statistics of these listener-specific baseline performance are shown in the bottom right panel of Fig. 10.

The effect of spectral ripples was modeled on the basis of our listener pool. The iteratively derived polar error rates were retrieved from the probabilistic response predictions by generating virtual response patterns. The model predicted moderate performance for the smallest densities tested, worst performance for ripple densities between 0.5 and 2 ripples/octave, and best performance for the largest densities tested. Further, the model predicted decreasing performance with increasing ripple depth. Thus, the model seems to qualitatively predict the actual results.

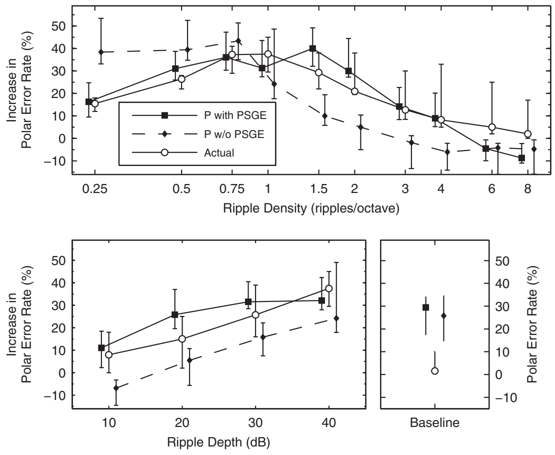


FIG. 10. Effect of spectral ripples. Actual experimental results (circles) are from Macpherson and Middlebrooks (2003). Predicted results (filled circles) were modeled for our listener pool (squares). Either the ripple depth of 40 dB (top) or the ripple density of 1 ripple/octave (bottom) was kept constant. Ordinates show the listener-specific difference in error rate between a test and the baseline condition. Baseline performance is shown in the bottom right panel. Symbols and whiskers show median values and inter-quartile ranges, respectively. Symbols were horizontally shifted to avoid overlaps. Diamonds show predictions of the model without positive spectral gradient extraction (P w/o PSGE), see Sec. IV A.

The baseline performance of the listeners tested by Macpherson and Middlebrooks (2003) were much better than those predicted for our pool members. It seems that in the free field scenario the trained listeners could use spatial cues that might have been absent in the virtual auditory space considered for the model parameterization (see Sec. II D). Moreover, the localization performance might be degraded by potential mismatches between free-field stimuli filtered acoustically and virtual auditory space stimuli created on the basis of measured HRTFs.

E. Effect of high-frequency attenuation in speech

Best *et al.* (2005) tested localization performance for monosyllabic words from a broad-band (0.3 to 16 kHz) speech corpus. The duration of those 260 words ranged from 418 to 1005 ms with an average duration of 710 ms. The speech samples were attenuated by either 0, −20, −40, or −60 dB in the stop band. Broad-band noise bursts with a duration of 150 ms served as baseline targets. They tested five NH listeners in a virtual auditory space. All those listeners had prior experience in sound-localization experiments and were selected in terms of achieving a minimum performance threshold. Localization performance was quantified by means of absolute polar angle errors and QE, albeit QE only for a reduced set of conditions. The results showed gradually degrading localization performance with increasing attenuation of high-frequency content above 8 kHz—see their Fig. 10.

Corresponding model predictions were performed for the median plane by using the same speech stimuli. Absolute polar angle errors were computed by expectancy values. Figure 11 compares the model predictions with the actual results. The predictions represent quite well the relative effect of degrading localization performance with increasing attenuation.

The overall offset between their actual and our predicted performance probably results from the discrepancy in

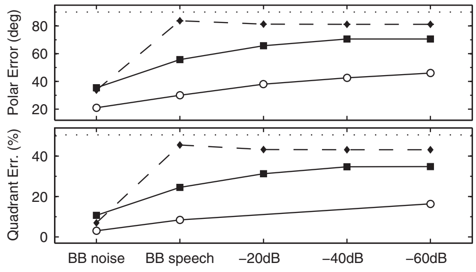


FIG. 11. Effect of high-frequency attenuation in speech localization. Actual experimental results are from Best *et al.* (2005). Absolute polar angle errors (top) and QE (bottom) were averaged across listeners. Circles and squares show actual and predicted results, respectively. Diamonds with dashed lines show predictions of the model without positive spectral gradient extraction—see Sec. IV A. Dotted horizontal lines represent chance rate.

baseline performance; the listeners from Best *et al.* (2005) showed a mean QE of about 3% whereas the listeners from our pool showed about 9%. Another potential reason for prediction residues might be the fact that the stimuli were dynamic and the model integrates the spectral information of the target sound over its full duration. This potentially smears the spectral representation of the target sound and, thus, degrades its spatial uniqueness. In contrast, listeners seem to evaluate sounds in segments of a few milliseconds (Hofman and Opstal, 1998) allowing them to base their response on the most salient snapshot.

F. Target-specific predictions

The introduced performance measures assess the model predictions by integrating over a specific range of directions. However, it might also be interesting to evaluate the model predictions in a very local way, namely, for each individual trial obtained in an experiment. Thus, in this section we evaluate the target-specific predictions, i.e., the correspondence between the actual responses and the predicted response probabilities underlying those responses on a trial-by-trial basis. In order to quantify the correspondence, we used the likelihood analysis (Langendijk and Bronkhorst, 2002).

The likelihood represents the probability that a certain response pattern occurs given a model prediction [see Langendijk and Bronkhorst, 2002, their Eq. (1)]. In our evaluation, the likelihood analysis was used to investigate how well an actual response pattern fits to the model predictions compared to fits of response patterns generated by the model itself. In the comparison, two likelihoods are calculated for each listener. First, the actual likelihood was the log-likelihood of actual responses. Second, the expected likelihood was the average of 100 log-likelihoods of random responses drawn according to the predicted PMVs. The average accounted for the randomness in the model-based generation of response patterns from independently generated patterns. Finally, both actual and expected likelihoods were normalized by the chance likelihood in order to obtain likelihoods independent of the number of tested targets. The chance likelihood was calculated for actual responses given the same probability for all response angles, i.e., given a model without any directional information. Hence, expected likelihoods can range from 0 (model of unique response) to 1 (non-directional model). The actual likelihoods should be similar to the expected likelihoods, and the more consistent the actual responses are across trials, the smaller actual likelihoods can result.

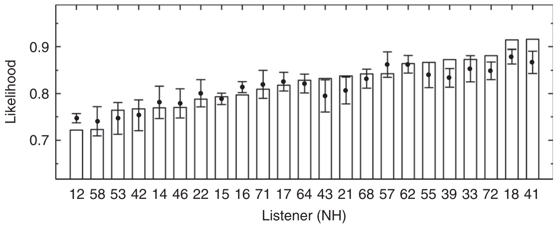


FIG. 12. Listener-specific likelihood statistics used to evaluate target-specific predictions for the baseline condition. Bars show actual likelihoods, dots show mean expected likelihoods, and whiskers show tolerance intervals with 99% confidence level of expected likelihoods.

Figure 12 shows the likelihood statistics for all listeners of the pool tested in the baseline condition. Expected likelihoods are shown by means of tolerance intervals with a confidence level of 99% (Langendijk and Bronkhorst, 2002). For 15 listeners, the actual likelihoods were within the tolerance intervals of the expected likelihoods, indicating valid target-specific predictions; examples were shown in Fig. 3. For eight listeners, the actual likelihoods were outside the tolerance intervals indicating a potential issue in the target-specific predictions. From the PMVs of those latter eight listeners, three types of issues can be identified. Figure 13 shows examples for each of the three types in terms of PMVs and actual responses from exemplary listeners. NH12 and NH16, listeners of the first type (left panel), responded too seldom at directions for which the model predicted a high probability. A smaller motor response scatter, ε, might be able to better represent such listeners, suggesting listener-specific ε. On the other hand, NH39, NH21, and NH72, listeners of the second type (center panel), responded too often at directions for which the model predicted a low probability. These listeners actually responded quite inconsistently, indicating some procedural uncertainty or inattention during the localization task, an effect not captured by the model. NH18, NH41, and NH43, listeners of the third type, responded quite consistently for most of the target angles, but for certain regions, especially in the upper-front quadrant, actual responses clearly deviate from high probability regions.

The reasons for such deviations are unclear, but it seems that the spectro-to-spatial mapping is incomplete or misaligned somehow.

In summary, for about two-thirds of the listeners, the target-specific predictions were similar to actual localization responses in terms of likelihood statistics. For one-third of the listeners, the model would gain from more individualization, e.g., a listener-specific motor response scatter and the consideration of further factors, e.g., a spatial weighting accounting for a listener's preference of specific directions.

In general, it is noteworthy that the experimental data we used here might not be appropriate for the current analysis. One big issue might be, for instance, that the participants

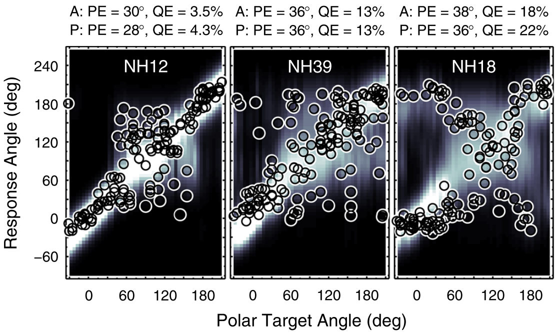


FIG. 13. (Color online) Exemplary baseline predictions. Same as Fig. 3 but for listeners where actual likelihoods were outside the tolerance intervals. See text for details.

were not instructed where to point when they were uncertain about the position. Some listeners might have used the strategy to point simply to the front and others might point to a position randomly chosen from case to case. The latter strategy is most consistent with the model assumption. Further, the likelihood analysis is very strict. A few actual responses being potentially outliers can have a strong impact on the actual likelihood, even in cases where most of the responses are in the correctly predicted range. Our performance measures considering a more global range of directions seem to be more robust to such outliers.

IV. EVALUATION AND DISCUSSION OF PARTICULAR MODEL COMPONENTS

The impact of particular model components on the predictive power of the model is discussed in the following. Note that modifications of the model structure require a new calibration of the model's internal parameters. Hence, the listener-specific sensitivity, S_l, was recalibrated according to the parameterization criterion in Eq. (11) in order to assure optimal parameterization also for modified configurations.

A. Positive spectral gradient extraction

Auditory nerve fibers from the cochlea form synapses in the cochlear nucleus for the first time. In the cat DCN, tonotopically adjacent fibers are interlinked and form neural circuits sensitive to positive spectral gradients (Reiss and Young, 2005). This sensitivity potentially makes the coding of spectral spatial cues more robust to natural macroscopic variations of the spectral shape. Hence, positive spectral gradient extraction should be most important when localizing spectrally non-flat stimuli like, for instance, rippled noise bursts (Macpherson and Middlebrooks, 2003) or speech samples (Best *et al.*, 2005).

To illustrate the role of the extraction stage in the proposed model, both experiments were modeled with and without the extraction stage. In the condition without extraction stage, the L_1-norm was replaced by the standard deviation (Middlebrooks, 1999b; Langendijk and Bronkhorst, 2002; Baumgartner *et al.*, 2013) because overall level differences between a target and the templates should not influence the distance metric. With the extraction stage, overall level differences are ignored by computing the spectral gradient. Figure 10 shows the effect of ripple density for model predictions either with or without extraction stage. The model without extraction stage (dashed lines) predicted the worst performance for densities smaller or equal than 0.75 ripples/octave and a monotonic performance improvement with increasing density. This deviates from the actual results from Macpherson and Middlebrooks (2003) and the predictions obtained by the model with extraction stage, both showing improved performance for macroscopic ripples below 1 ripple/octave. This deviation is supported by the correlation coefficient calculated for the 14 actual and predicted median polar error rates. The coefficient decreased from 0.89 to 0.73 when removing the extraction stage. Figure 11 shows a similar comparison for speech samples. The model with extraction stage predicted a gradual degradation with increasing attenuation of high frequency content, whereas the model without extraction stage failed to predict this gradual degradation; predicted performance was close to chance performance even for the broad-band speech. Thus, the extraction of positive spectral gradients seems to be an important model component in order to obtain plausible predictions for various spectral modifications.

TABLE II. The effects of model configurations on the prediction residues. PSGE: Model with or without positive spectral gradient extraction. MBA: Model with or without manual bandwidth adjustment to the stimulus bandwidth. Prediction residues (e_{PE}, e_{QE}) between actual and predicted PE and QE are listed for acute performance with the BB, LP, and W conditions of the experiments from Majdak *et al.* (2013c)

		BB		LP		W	
PSGE	MBA	e_{PE}	e_{QE}	e_{PE}	e_{QE}	e_{PE}	e_{QE}
Yes	no	3.4°	2.9%	4.5°	7.6%	6.2°	7.7%
Yes	yes	3.4°	2.9%	5.6°	7.8%	4.8°	7.4%
No	no	2.1°	2.8%	10.1°	23.9%	5.3°	12.6%
No	yes	2.1°	2.8%	3.9°	7.7%	5.3°	8.1%

Band limitation in terms of low-pass filtering is also a macroscopic modification of the spectral shape. Hence, the extraction stage should have a substantial impact on modeling localization performance for low-pass filtered sounds. In Baumgartner *et al.* (2013), a model was used to predict the experimental results for band-limited sounds from Majdak *et al.* (2013c) For that purpose, the internal bandwidth considered in the comparison process was manually adjusted according to the actual bandwidth of the stimulus. This means that prior knowledge of stimulus characteristics was necessary to parameterize the model. In the proposed model, the extraction stage is supposed to automatically account for the stimulus bandwidth. To test this assumption, model predictions for the experiment from Majdak *et al.* (2013c) were performed by using four different model configurations: With or without extraction stage and with or without manual bandwidth adjustment. Table II lists the resulting partial prediction residues. Friedman's analysis of variance was used to compare the joint prediction residues, e, between the four model configurations. The configurations were separately analyzed for each experimental condition (BB, LP, and W). The differences were significant for the LP ($\chi^2 = 19.43$, $p < 0.001$) condition, but not significant for the BB ($\chi^2 = 5.77$, $p = 0.12$) and W ($\chi^2 = 2.82$, $p = 0.42$) conditions. The Tukey's honestly significant difference *post hoc* test showed that in the LP condition the model without extraction stage and without bandwidth adjustment performed significantly worse than all other model configurations ($p < 0.05$). In contrast, the model with the extraction stage and without manual adjustment yielded results similar to the models with adjustment ($p \gg 0.05$). This shows the relevance of extracting positive spectral gradients in a sagittal-plane localization model in terms of automatic bandwidth compensation.

B. Sensorimotor mapping

The sensorimotor mapping stage addresses the listeners' sensorimotor uncertainty in a pointing task. For instance, two spatially close DTFs might exhibit quite large spectral

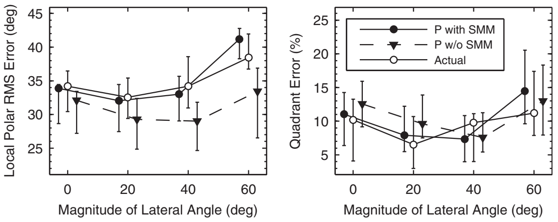


FIG. 14. Baseline performance as a function of the magnitude of the lateral response angle. Symbols and whiskers show median values and interquartile ranges, respectively. Open symbols represent actual and closed symbols predicted results. Symbols were horizontally shifted to avoid overlaps. Triangles show predictions of the model without the sensorimotor mapping stage (P w/o SMM).

differences. However, even though the listener might perceive the sound at two different locations, he/she would unlikely be able to consistently point to two different directions because of a motor error in pointing the direction. In general, this motor error also increases in the polar dimension with increasing magnitude of the lateral angle.

The importance of the sensorimotor mapping stage is demonstrated by comparing baseline predictions with and without response scatter, i.e., $\varepsilon = 17°$ and $\varepsilon = 0°$, respectively. Baseline performance was measured as a function of the lateral angle and results were pooled between the left- and right-hand side. Figure 14 shows that the actual performance was worse at the most lateral angles. The performance predicted with the model including the mapping stage followed this trend. The exclusion of the mapping stage, however, degraded the prediction accuracy; prediction residues rose from $e_{\mathrm{PE}} = 3.4°$ to $e_{\mathrm{PE}} = 5.5°$ and $e_{\mathrm{QE}} = 3.4\%$ to $e_{\mathrm{QE}} = 3.9\%$, and correlation coefficients dropped from $r_{\mathrm{PE}} = 0.72$ to $r_{\mathrm{PE}} = 0.64$ or stayed the same in the case of QE ($r_{\mathrm{QE}} = 0.81$). Figure 14 further shows that the exclusion particularly degraded the local performance predictions for most lateral directions. Hence, the model benefits from the sensorimotor mapping stage especially for predictions beyond the median plane.

C. Binaural weighting

Several studies have shown that the monaural spectral information is weighted binaurally according to the perceived lateral angle (Morimoto, 2001; Hofman and Van Opstal, 2003; Macpherson and Sabin, 2007). It remained unclear, however, whether the larger weighting of the ipsilateral information is beneficial in terms of providing more spectral cues or whether it is simply the larger gain that makes the ipsilateral information more reliable. We investigated this question by performing baseline predictions for various binaural weighting coefficients, Φ. Three different values of Φ were compared, namely, $\Phi = 13°$ according to the optimal binaural weighting coefficient found in Sec. II D, and $\Phi \to \pm 0°$ meaning that only the ipsilateral resp. contralateral information is considered.

Table III shows the prediction residues, correlation coefficients, and predicted average performance for the three configurations. The differences between configurations are surprisingly small for all parameters. The negligible differences of residues and correlations show that realistic binaural weighting is rather irrelevant for accurate model predictions, because the ipsi- and contralateral ear seem to contain similar spatial information. The small differences in average performance also indicate that, if at all, the contralateral path provides only little less spectral cues than the ipsilateral path. Consequently, a larger ipsilateral weighting seems to be beneficial mostly in terms of larger gain.

TABLE III. Performance predictions for binaural, ipsilateral, and contralateral listening conditions. The binaural weighting coefficient, Φ, was varied in order to represent the three conditions: Binaural: $\Phi = 13°$; ipsilateral: $\Phi \to +0°$; contralateral: $\Phi \to -0°$. Prediction residues (e_{PE}, e_{QE}) and correlation coefficients (r_{PE}, r_{QE}) between actual and predicted results are shown together with predicted average performance ($\overline{\mathrm{PE}}$, $\overline{\mathrm{QE}}$).

	e_{PE}	e_{QE}	r_{PE}	r_{QE}	$\overline{\mathrm{PE}}$	$\overline{\mathrm{QE}}$
Binaural	3.4°	3.4%	0.72	0.81	32.6°	9.4%
Ipsilateral	3.4°	3.4%	0.72	0.80	32.5°	9.2%
Contralateral	3.3°	4.7%	0.71	0.77	32.6°	10.6%

V. CONCLUSIONS

A sagittal-plane localization model was proposed and evaluated under various experimental conditions, testing localization performance for single, motionless sound sources with high-frequency content at moderate intensities. In total, predicted performance correlated well with actual performance, but the model tended to underestimate local performance in very challenging experimental conditions. Detailed evaluations of particular model components showed that (1) positive spectral gradient extraction is important for localization robustness to spectrally macroscopic variations of the source signal, (2) listeners' sensorimotor mapping is relevant for predictions especially beyond the median plane, and (3) contralateral spectral features are only marginally less pronounced than ipsilateral features. The prediction results demonstrated the potential of the model for practical applications, for instance, to assess the quality of spatial cues for the design of hearing assistive devices or surround-sound systems (Baumgartner *et al.*, 2013).

However, there are several limitations of the current model. To name a few, the model cannot explain phase effects on elevation perception (Hartmann *et al.*, 2010). Also the effects of dynamic cues like those resulting from moving sources or head rotations were not considered (Vliegen *et al.*, 2004; Macpherson, 2013). Furthermore, the gammatone filterbank used to approximate cochlear filtering is linear and thus, the present model cannot account for known effects of sound intensity (Vliegen and Opstal, 2004). Future work will also need to be done in the context of modeling dynamic aspects of plasticity due to training (Hofman *et al.*, 1998; Majdak *et al.*, 2010, 2013c) or the influence of cross-modal information (Lewald and Getzmann, 2006).

The present model concept can serve as a starting point to incorporate those features. The first steps, for instance, toward modeling effects of sound intensity, have already been taken (Majdak *et al.*, 2013a). Reproducibility is inevitable in order to reach the goal of a widely applicable model. Thus, the implementation of the model (`baumgartner2014`) and the modeled experiments (`exp_baumgartner2014`) are provided in the Auditory Modeling Toolbox (Søndergaard and Majdak, 2013).

ACKNOWLEDGMENTS

We thank Virginia Best and Craig Jin for kindly providing their speech data. We also thank Katherine Tiede for proofreading and Christian Kasess for fruitful discussions on statistics. This research was supported by the Austrian Science Fund (FWF, projects P 24124 and M 1230).

Agterberg, M. J., Snik, A. F., Hol, M. K., Wanrooij, M. M. V., and Opstal, A. J. V. (**2012**). “Contribution of monaural and binaural cues to sound localization in listeners with acquired unilateral conductive hearing loss: Improved directional hearing with a bone-conduction device,” Hear. Res. **286**, 9–18.

Algazi, V. R., Avendano, C., and Duda, R. O. (**2001**). “Elevation localization and head-related transfer function analysis at low frequencies,” J. Acoust. Soc. Am. **109**, 1110–1122.

Andéol, G., Macpherson, E. A., and Sabin, A. T. (**2013**). “Sound localization in noise and sensitivity to spectral shape,” Hear. Res. **304**, 20–27.

Baumgartner, R., Majdak, P., and Laback, B. (**2013**). “Assessment of sagittal-plane sound localization performance in spatial-audio applications,” in *The Technology of Binaural Listening*, edited by J. Blauert (Springer, Berlin-Heidelberg), Chap. 4.

Best, V., Carlile, S., Jin, C., and van Schaik, A. (**2005**). “The role of high frequencies in speech localization,” J. Acoust. Soc. Am. **118**, 353–363.

Bizley, J. K., Nodal, F. R., Parsons, C. H., and King, A. J. (**2007**). “Role of auditory cortex in sound localization in the mid-sagittal plane,” J. Neurophysiol. **98**, 1763–1774.

Blauert, J. (**1969**). “Sound localization in the median plane,” Acta Acust. Acust. **22**, 205–213.

Goupell, M. J., Majdak, P., and Laback, B. (**2010**). “Median-plane sound localization as a function of the number of spectral channels using a channel vocoder,” J. Acoust. Soc. Am. **127**, 990–1001.

Hartmann, W. M., Best, V., Leung, J., and Carlile, S. (**2010**). “Phase effects on the perceived elevation of complex tones,” J. Acoust. Soc. Am. **127**, 3060–3072.

Hebrank, J., and Wright, D. (**1974**). “Spectral cues used in the localization of sound sources on the median plane,” J. Acoust. Soc. Am. **56**, 1829–1834.

Hofman, M., and Van Opstal, J. (**2003**). “Binaural weighting of pinna cues in human sound localization,” Exp. Brain Res. **148**, 458–470.

Hofman, P. M., and Opstal, A. J. V. (**1998**). “Spectro-temporal factors in two-dimensional human sound localization,” J. Acoust. Soc. Am. **103**, 2634–2648.

Hofman, P. M., van Riswick, J. G. A., and van Opstal, A. J. (**1998**). “Relearning sound localization with new ears,” Nature Neurosci. **1**, 417–421.

Jin, C., Corderoy, A., Carlile, S., and van Schaik, A. (**2004**). “Contrasting monaural and interaural spectral cues for human sound localization,” J. Acoust. Soc. Am. **115**, 3124–3141.

Jin, C., Schenkel, M., and Carlile, S. (**2000**). “Neural system identification model of human sound localization,” J. Acoust. Soc. Am. **108**, 1215–1235.

King, A. J. (**2004**). “The superior colliculus,” Curr. Biol. **14**, R335–R338.

Kistler, D. J., and Wightman, F. L. (**1992**). “A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction,” J. Acoust. Soc. Am. **91**, 1637–1647.

Langendijk, E. H. A., and Bronkhorst, A. W. (**2002**). “Contribution of spectral cues to human sound localization,” J. Acoust. Soc. Am. **112**, 1583–1596.

Lewald, J., and Getzmann, S. (**2006**). “Horizontal and vertical effects of eye-position on sound localization,” Hear. Res. **213**, 99–106.

Lyon, R. F. (**1997**). “All-pole models of auditory filtering,” in *Diversity in Auditory Mechanics*, edited by E. R. Lewis, G. R. Long, R. F. Lyon, P. M. Narins, C. R. Steele, and E. Hecht-Poinar (World Scientific Publishing, Singapore), pp. 205–211.

Macpherson, E. A. (**2013**). “Cue weighting and vestibular mediation of temporal dynamics in sound localization via head rotation,” POMA **19**, 050131.

Macpherson, E. A., and Middlebrooks, J. C. (**2002**). “Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited,” J. Acoust. Soc. Am. **111**, 2219–2236.

Macpherson, E. A., and Middlebrooks, J. C. (**2003**). “Vertical-plane sound localization probed with ripple-spectrum noise,” J. Acoust. Soc. Am. **114**, 430–445.

Macpherson, E. A., and Sabin, A. T. (**2007**). “Binaural weighting of monaural spectral cues for sound localization,” J. Acoust. Soc. Am. **121**, 3677–3688.

Majdak, P., Baumgartner, R., and Laback, B. (**2014**). “Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization,” Front Psychol. **5**:319, doi:10.3389/fpsyg.2014.00319.

Majdak, P., Baumgartner, R., Necciari, T., and Laback, B. (**2013a**). “Sound localization in sagittal planes: Modeling the level dependence,” poster presented at the 36th Mid-Winter Meeting of the ARO, Baltimore, MD, http://www.kfs.oeaw.ac.at/doc/public/ARO2013.pdf (Last viewed June 13, 2014).

Majdak, P., Goupell, M. J., and Laback, B. (**2010**). “3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training,” Atten. Percept. Psycho. **72**, 454–469.

Majdak, P., Masiero, B., and Fels, J. (**2013b**). “Sound localization in individualized and non-individualized crosstalk cancellation systems,” J. Acoust. Soc. Am. **133**, 2055–2068.

Majdak, P., Walder, T., and Laback, B. (**2013c**). “Effect of long-term training on sound localization performance with spectrally warped and band-limited head-related transfer functions,” J. Acoust. Soc. Am. **134**, 2148–2159.

May, B. J. (**2000**). “Role of the dorsal cochlear nucleus in the sound localization behavior of cats,” Hear. Res. **148**, 74–87.

Middlebrooks, J. C. (**1992**). “Narrowband sound localization related to external ear acoustics,” J. Acoust. Soc. Am. **92**, 2607–2624.

Middlebrooks, J. C. (**1999a**). “Individual differences in external-ear transfer functions reduced by scaling in frequency,” J. Acoust. Soc. Am. **106**, 1480–1492.

Middlebrooks, J. C. (**1999b**). “Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency,” J. Acoust. Soc. Am. **106**, 1493–1510.

Møller, H., Sørensen, M. F., Hammershøi, D., and Jensen, C. B. (**1995**). “Head-related transfer functions of human subjects,” J. Audio Eng. Soc. **43**, 300–321.

Morimoto, M. (**2001**). “The contribution of two ears to the perception of vertical angle in sagittal planes,” J. Acoust. Soc. Am. **109**, 1596–1603.

Redon, C., and Hay, L. (**2005**). “Role of visual context and oculomotor conditions in pointing accuracy,” Neuro Report **16**, 2065–2067.

Reiss, L. A. J., and Young, E. D. (**2005**). “Spectral edge sensitivity in neural circuits of the dorsal cochlear nucleus,” J. Neurosci. **25**, 3680–3691.

Søndergaard, P., and Majdak, P. (**2013**). “The auditory modeling toolbox,” in *The Technology of Binaural Listening*, edited by J. Blauert (Springer, Berlin-Heidelberg), Chap. 2.

Unoki, M., Irino, T., Glasberg, B., Moore, B. C. J., and Patterson, R. D. (**2006**). “Comparison of the roex and gammachirp filters as representations of the auditory filter,” J. Acoust. Soc. Am. **120**, 1474–1492.

van Wanrooij, M. M., and van Opstal, A. J. (**2005**). “Relearning sound localization with a new ear,” J. Neurosci. **25**, 5413–5424.

Vliegen, J., and Opstal, A. J. V. (**2004**). “The influence of duration and level on human sound localization,” J. Acoust. Soc. Am. **115**, 1705–1713.

Vliegen, J., Van Grootel, T. J., and Van Opstal, A. J. (**2004**). “Dynamic sound localization during rapid eye-head gaze shifts,” J. Neurosci. **24**, 9291–9302.

Zakarauskas, P., and Cynader, M. S. (**1993**). “A computational theory of spectral cue localization,” J. Acoust. Soc. Am. **94**, 1323–1331.

Zhang, P. X., and Hartmann, W. M. (**2010**). “On the ability of human listeners to distinguish between front and back,” Hear. Res. **260**, 30–46.

Chapter 5

Efficient approximation of head-related transfer functions in subbands for accurate sound localization

This work was published as

Marelli, D., **Baumgartner, R.**, Majdak, P. (2015): Efficient approximation of head-related transfer functions in subbands for accurate sound localization, in: IEEE/ACM Transactions on Audio, Speech and Language Processing 23, 1130-1143. doi:10.1109/TASLP.2015.2425219

The idea behind this work came from the first and third authors. In general, the work can be subdivided into a mathematical part and a psychoacoustic part. The mathematical part includes the development and implementation of the approximation algorithms as well as the numerical experiments, and was done by the first author. The psychoacoustic part includes evaluating the algorithms by means of model simulations and psychoacoustic experiments, and was done by me, as the second author, well-advised by the third author. In order to evaluate the algorithms, I also had to implement the sound synthesis via subband processing. The methodology of the localization experiments was already established at the time of the experiments. The manuscript was written by the first author and me while sections were divided according to the topical separation described above. All authors revised the whole manuscript.

Efficient Approximation of Head-Related Transfer Functions in Subbands for Accurate Sound Localization

Damián Marelli, Robert Baumgartner, and Piotr Majdak

***Abstract*—Head-related transfer functions (HRTFs) describe the acoustic filtering of incoming sounds by the human morphology and are essential for listeners to localize sound sources in virtual auditory displays. Since rendering complex virtual scenes is computationally demanding, we propose four algorithms for efficiently representing HRTFs in subbands, i.e., as an analysis filterbank (FB) followed by a transfer matrix and a synthesis FB. All four algorithms use sparse approximation procedures to minimize the computational complexity while maintaining perceptually relevant HRTF properties. The first two algorithms separately optimize the complexity of the transfer matrix associated to each HRTF for fixed FBs. The other two algorithms jointly optimize the FBs and transfer matrices for complete HRTF sets by two variants. The first variant aims at minimizing the complexity of the transfer matrices, while the second one does it for the FBs. Numerical experiments investigate the latency-complexity trade-off and show that the proposed methods offer significant computational savings when compared with other available approaches. Psychoacoustic localization experiments were modeled and conducted to find a reasonable approximation tolerance so that no significant localization performance degradation was introduced by the subband representation.**

***Index Terms*—Head-related transfer functions (HRTFs), subband signal processing, sparse approximation, sound localization, virtual acoustics.**

I. Introduction

HEAD-RELATED transfer functions (HRTFs) describe the acoustic filtering of incoming sounds by the torso, head, and pinna, using linear-time-invariant systems [1]. Listeners can be immersed into a virtual auditory environment, by filtering sounds with listener-specific HRTFs [2]. Complex environments involve multiple virtual sources and room reflections. Strictly speaking, a correct representation of such environments requires the filtering of virtual sources and their reflections with the corresponding HRTFs, which is a computationally demanding procedure. Although perceptually motivated methods for more efficient reverberation simulation have been proposed (e.g., static filter for late reflections and HRTFs for only up to second-order reflections [3]), the computational requirements on HRTF filtering remain demanding and calls for the need of an efficient approximation of HRTFs.

Efficient HRTF filtering is classically achieved by using the overlap-add (OA) or overlap-save (OS) method [4,§ 5.3.2]. For static sound sources, an even more efficient implementation can be achieved by using pole-zero (PZ) models of the HRTFs [5]. However, when processing moving sound sources, the commutation of pole filter coefficients is problematic, because their update may produce an inconsistent internal filter state, which yields audible artifacts, or even unstable filters [6]. This problem can be tackled by filtering the audio signal in parallel pipelines and cross-fading between them. This, however, severely degrades the computational efficiency. While the OA or OS methods are always stable [4,§ 5.3.2] and simpler to handle in the case of moving sources, they introduce a certain latency. This latency can be reduced using zero- or low-delay fast convolution (ZDFC, LDFC) [7]. The OS, OA, ZDFC and LDFC methods, which we collectively call segmented fast Fourier transform (SFFT) methods, permit accommodating a trade-off between computational complexity and latency [8]. It was recently shown that a better trade-off can be achieved using a subband (SB) approximation of HRTFs [9], [10], if certain approximation tolerance can be allowed.

In the SB approach, an HRTF is represented as the concatenation of an analysis filterbank (FB), followed by a transfer matrix, called subband model (SBM), and a synthesis FB. This scheme already leads to major computational savings, for a given latency. However, these savings can be further improved if the analysis and synthesis FBs are chosen to be equal for all HRTFs within a set. In such case, we can make use of the following two properties: 1) When a number of reflections of a single audio input channel, i.e., virtual source signal, is to be simulated, the output of the analysis FB stage associated to each reflection is the same. Hence, an analysis FB needs to be evaluated only once per virtual source signal, regardless of its number of reflections. Thus, the complexity of the analysis FB stage is inversely proportional to the number of reflections per source signal. 2) The

Manuscript received September 30, 2014; revised February 02, 2015; accepted April 07, 2015. Date of publication April 21, 2015; date of current version May 08, 2015. This work was supported in part by the Austrian Science Fund (FWF, projects M1230-N13 and P24124). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Woon-Seng Gan.

D. Marelli is with the School of Electrical Engineering and Computer Science, University of Newcastle, Callaghan, NSW 2308, Australia, and also with the Acoustics Research Institute, Austrian Academy of Sciences, Austria (e-mail: damian.marelli@newcastle.edu.au).

R. Baumgartner and P. Majdak are with the Acoustics Research Institute, Austrian Academy of Sciences, 1040 Vienna, Austria (e-mail: robertbaumgartner@icloud.com; piotr@majdak.com).

This paper has supplemental downloadable multimedia material available at http://ieeexplore.ieee.org provided by the authors. The Supplementary Materials contain supplementary figures in .png format. This material is 1 MB in size.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2015.2425219

final step in the spatialization task consists in adding together all spatialized sources and reflections. In view of the linearity of FB operations, the output of all SBMs can be added together before applying the synthesis FB stage. In this way, the synthesis FB needs to be computed only once per audio output channel, i.e., ear signal. Hence, its complexity is minor.

In this paper, we propose algorithms for efficient approximation of HRTFs in subbands, considering features being perceptually relevant in hearing. In particular, we focus on the sound localization in human listeners. In general, listeners localize sound sources on the basis of monaural and binaural cues. Binaural cues, like the interaural time and level differences, are basically used to determine the lateral direction (defined from left to right) of a sound [11]. Monaural spectral cues are used to determine the polar direction of a sound in sagittal planes (ranging from down via front and top to rear directions) [12]. Thus, we aim at approximating HRTFs while preserving both interaural and monaural cues, in order to maintain a listener's localization performance in the three-dimensional space. To this end, we approximate HRTFs using a criterion based on logarithmic amplitude responses, which is an approximate measure for loudness [13], [14]. We also apply a frequency weighting corresponding to the bandwidth of auditory filters [15], [16]. Our approximation criterion considers both, amplitude and phase, of HRTFs. In psychoacoustic experiments, we evaluate the resulting HRTF approximations, for various approximation tolerances, on the listeners' performance in localizing virtual sound sources.

We propose four algorithms, which we call greedy, relaxation, SBM-shrink and FB-shrink. The greedy and relaxation algorithms rely on an *a priori* fixed FB design, and minimize the complexity (i.e., the number of non-zero entries) of the SBM, for a particular HRTF. The greedy algorithm is the only one which does not require an initialization (i.e., an initial "guess" of the SBM). Hence, it is used to provide an initial approximation for the SMB. For an improved result, the SBM yielded by the greedy algorithm can be used to initialize the relaxation algorithm, which further minimizes the complexity within the fixed FB assumption. In contrast to these two algorithms, the SBM-shrink and FB-shrink algorithms optimize the choice of FBs. Both are initialized using the SBMs produced by the relaxation algorithm. The SBM-shrink algorithm jointly minimizes the support of all SBMs of an HRTF set, while keeping unchanged the supports of the FBs. It does so by jointly optimizing the FBs together with the SBMs. The rationale behind this algorithm is that the increase in optimization flexibility obtained by optimizing the FBs permits achieving further complexity reductions. The FB-shrink algorithm, being complementary to the SBM-shrink, reduces the support of the FBs, for a given set of SBMs, while keeping the support of the SBMs unchanged. All algorithms offer computational efficiency while 1) keeping the accuracy of the HRTF approximation within a certain prescribed approximation tolerance; and 2) keeping the latency of the filtering process within a certain prescribed threshold. This paper is based on the preliminary work reported in [17].

Notation 1. Given a time sequence $x(t)$, $t \in \mathbb{Z}$, we use $x(\omega)$, $\omega \in (-\pi, \pi]$ to denote its discrete-time Fourier transform. Also, when it is clear from the context, we use x to denote either $x(t)$ or $x(\omega)$. The i-th entry of vector $\mathbf{a}$ is denoted by $[\mathbf{a}]_i$ and the (i, j)-th entry of matrix $\mathbf{A}$ by $[\mathbf{A}]_{i,j}$.

II. Approximation of Linear Systems Using Subbands

A. Problem Description

The input/output relation of a linear system with frequency response $g(\omega)$ is given by

$$y(\omega) = g(\omega)x(\omega). \tag{1}$$

The same system can be approximately implemented in the subband domain as follows [9]:

$$\boldsymbol{\xi}(\omega) = \downarrow_D \{\mathbf{h}(\omega)x(\omega)\} \tag{2}$$

$$\hat{\boldsymbol{\psi}}(\omega) = \mathbf{S}(\omega)\boldsymbol{\xi}(\omega), \tag{3}$$

$$\hat{y}(\omega) = \mathbf{f}^*(\omega)\uparrow_D \left\{\hat{\boldsymbol{\psi}}(\omega)\right\}, \tag{4}$$

where $\mathbf{h}(\omega) = [h_1(\omega), \cdots, h_m(\omega), \cdots, h_M(\omega)]^T$ and $\mathbf{f}(\omega) = [f_1(\omega), \cdots, f_m(\omega), \cdots, f_M(\omega)]^T$ denote the analysis and synthesis filters, respectively, M denotes the number of subbands, $\downarrow_D\{\cdot\}$ denotes the downsampling operation with factor $D \leq M$ (i.e., keeping one out of D samples), $\uparrow_D\{\cdot\}$ denotes the upsampling operation of factor D (i.e., inserting $D - 1$ zero-valued samples between every two samples), $\mathbf{S}(\omega)$ denotes the SBM, $\boldsymbol{\xi}(\omega)$ and $\hat{\boldsymbol{\psi}}(\omega)$ denote the subband representation of the input $x(\omega)$ and the approximated output $\hat{y}(\omega)$, respectively, and * denotes transpose conjugation. We choose $h_m(\omega) = h(\omega - 2\pi\frac{m-1}{M})$ and $f_m(\omega) = f(\omega - 2\pi\frac{m-1}{M})$, using prototype finite impulse response filters h and f of tap size l_h and l_f, respectively. We call (2) the analysis stage and (4) the synthesis stage.

B. Polyphase Representation

Using the polyphase representation [18], we can write (1)–(4) as

$$\mathbf{y}(\omega) = \mathbf{G}(\omega)\mathbf{x}(\omega) \tag{5}$$

$$\boldsymbol{\xi}(\omega) = \mathbf{H}(\omega)\mathbf{x}(\omega), \tag{6}$$

$$\hat{\boldsymbol{\psi}}(\omega) = \mathbf{S}(\omega)\boldsymbol{\xi}(\omega), \tag{7}$$

$$\hat{\mathbf{y}}(\omega) = \mathbf{F}^*(\omega)\hat{\boldsymbol{\psi}}(\omega), \tag{8}$$

where the D-dimensional vectors $\mathbf{x}(\omega)$, $\mathbf{y}(\omega)$ and $\hat{\mathbf{y}}(\omega)$ denote the polyphase representations of $x(\omega)$, $y(\omega)$ and $\hat{y}(\omega)$, respectively. They are defined by

$$[\mathbf{x}(t)]_d = x(tD - d + 1),$$

for all $d = 1, \cdots, D$, and similarly for $\mathbf{y}(\omega)$ and $\hat{\mathbf{y}}(\omega)$. Also, the $D \times D$ matrix $\mathbf{G}(\omega)$ is the polyphase representation of the target $g(\omega)$ and is defined by

$$[\mathbf{G}(t)]_{d,e} = g(tD + e - d), \tag{9}$$

for all $d, e = 1, \cdots, D$. The $M \times D$ matrices $\mathbf{H}(\omega)$ and $\mathbf{F}(\omega)$ are the polyphase representations of the analysis and synthesis stages, respectively. The matrix $\mathbf{H}(\omega)$ is defined by

$$[\mathbf{H}(t)]_{m,d} = h_m(tD + d),$$

for all $m = 1, \cdots, M$ and $d = 1, \cdots, D$, and $\mathbf{F}(t)$ is defined similarly.

Let

$$\hat{\mathbf{G}}(\omega) = \mathbf{F}^{*}(\omega)\mathbf{S}(\omega)\mathbf{H}(\omega). \tag{10}$$

Then, from (6)–(8) we have

$$\hat{\mathbf{y}}(\omega) = \hat{\mathbf{G}}(\omega)\mathbf{x}(\omega). \tag{11}$$

In view of (5) and (11), we can think of $\hat{\mathbf{G}}(\omega)$ as the approximation of the polyphase representation $\mathbf{G}(\omega)$ of the target $g(\omega)$. Consider the set $\hat{\mathbf{g}}(t) = [\hat{g}_1(t), \cdots, \hat{g}_D(t)]^T$ of D impulse responses defined by

$$\hat{g}_d(tD + e - d) = \left[\hat{\mathbf{G}}(t)\right]_{d,e}, \tag{12}$$

for all $d, e = 1, \cdots, D$. It is straightforward to verify that

$$\hat{y}(t) = \sum_{\tau \in \mathbb{Z}} \hat{g}_{t \bmod D}(\tau) x(t - \tau),$$

i.e., $\hat{y}(t)$ can be obtained by cyclically sampling the outputs of the filters $\hat{\mathbf{g}}(\omega)$. Hence, the subband approximation (2)–(4) behaves as a cyclostationary set of D filters.

Notation 2. In view of (12), we define the polyphase map by $\Phi : \hat{\mathbf{g}} \mapsto \hat{\mathbf{G}}$, where $\hat{\mathbf{g}}$ and $\hat{\mathbf{G}}$ stand for their frequency representations $\hat{\mathbf{g}}(\omega)$ and $\hat{\mathbf{G}}(\omega)$, respectively.

C. Diagonal Solution

It follows from [19, Theorem 1] that, if the support of the prototypes $h(\omega)$ and $f(\omega)$ are contained in $[-\pi/D, \pi/D]$, then the implementation (2)–(4) can be carried out with zero error using a diagonal $\mathbf{S}(\omega)$. A particular choice is to choose $f(\omega) = h(\omega)$ being root raised cosine windows with inflection angular frequency $\omega_0 = \pi/M$ and roll-off factor $\beta = M/D - 1$, i.e.,

$$f(\omega) = h(\omega) = \begin{cases} \sqrt{D}, & |\omega| \leq \frac{1-\beta}{M}\pi, \\ \sqrt{\frac{D}{2}\left(1+\cos\left(\frac{\pi}{\beta\omega_0}\left(\omega_0 - \omega\right)\right)\right)}, & \frac{1-\beta}{M}\pi < |\omega| \leq \frac{1+\beta}{M}\pi, \\ 0, & \frac{1-\beta}{M}\pi < |\omega|. \end{cases}$$

This choice has the property that, if $g(\omega) = 1$, then $\mathbf{S}(\omega) = \mathbf{I}_M$ (i.e., the identity matrix of dimension M).

D. Latency

The implementation (2)–(4) introduces a latency. This latency has three components. The first one is that introduced by the analysis FB, when the prototype impulse response $h(t)$ is non-causal. More precisely, if $\delta_h = \max\{t : h(-t) \neq 0\}$ is the non-causality of $h(t)$, then the analysis stage introduces a latency of δ_h samples. The second component is the one introduced by the non-causality of the SBM $\mathbf{S}(t)$. Since the SBM is applied on the downsampled signals $\boldsymbol{\xi}(t)$, its induced latency is $\delta_S = D\max\{t : \mathbf{S}(-t) \neq 0\}$. The last component is introduced by the synthesis FB. Again, since this FB is applied on the upsampled signal $\boldsymbol{\psi}(t)$, its induced latency is

$$\delta_f = D\left\lfloor \frac{\max\{t : f(t) \neq 0\}}{D} \right\rfloor,$$

where $\lfloor x \rfloor$ denotes the largest integer smaller than or equal to x.

The expressions above indicate that, while δ_S and δ_f can be adjusted in steps of D samples, the adjustments of δ_h can be done in steps of single samples. In view of this, we fix the non-causality of $\mathbf{S}(t)$, so that δ_S is fixed. We also choose $f^{*}(t)$ to be causal (i.e., $f(t)$ to be anti-causal), so that $\delta_S = 0$, and choose $h(t)$ to be anti-causal, leading to $\delta_h = l_h - 1$. With these choices, the non-causality of the whole scheme (in samples) is

$$\Delta = D\delta_S + l_h - 1, \tag{13}$$

which can be adjusted in steps of single samples by tuning the value of l_h. Notice that this latency can be arbitrarily reduced by choosing a negative value of δ_S, provided that this choice is compatible with the desired approximation.

E. Computational Complexity

Since $\mathbf{h}(\omega)$ is of Gabor type, i.e., consists of modulated versions $h_m(\omega) = h(\omega - 2\pi\frac{m-1}{M})$, $m = 1, \cdots, M$, of a prototype filter $h(\omega)$, and $h(t)$ has tap size l_h, then its polyphase representation is given by [20]

$$\mathbf{H}(\omega) = \mathcal{W}_M \mathbf{L}_2 \Lambda_h \mathbf{L}_1(\omega), \tag{14}$$

where $\mathcal{W}_M \in \mathbb{C}^{M \times M}$ is the DFT matrix, i.e., $[\mathcal{W}_M]_{k,l} = e^{-j\frac{2\pi}{M}kl}$, for all $k, l = 0, \cdots, M-1$. Also,

$$\mathbf{L}_1^T(\omega) = [e^{j(a-1)\omega}\mathbf{I}_D, \cdots, \mathbf{I}_D]_{:,\mathrm{end}-l_h+1:\mathrm{end}},$$
$$\mathbf{L}_2 = [\underbrace{\mathbf{I}_M, \mathbf{I}_M \cdots, \mathbf{I}_M}_{b\mathrm{times}}]_{:,\mathrm{end}-l_h+1:\mathrm{end}},$$
$$\Lambda_h = \mathrm{diag}\{h(1 - l_h), \cdots, h(0)\},$$

with $a = \lceil \frac{l_h}{D} \rceil$, $b = \lceil \frac{l_h}{M} \rceil$, $[\mathbf{A}]_{:,\mathrm{end}-l_h+1:\mathrm{end}}$ denoting the matrix formed with the last l_h columns of $\mathbf{A}$, and $\mathrm{diag}\{x_1, \cdots, x_l\}$ denoting the diagonal matrix with elements x_i, $i = 1, \cdots, I$, in its main diagonal. All the same applies to $\mathbf{f}(\omega)$, with $f(\omega)$ denoting its prototype and l_f the tap size of $f(t)$.

Using (14), and assuming that M is a power of two, so that an M-point FFT requires $M\log_2 M$ (real) multiplications [7], the implementation of the analysis FB requires

$$\Psi_{\mathrm{FB}}(h) = \frac{l_h + M\log_2 M}{D},$$

real multiplications per (fullband) sample. The same applies to the synthesis FB, with l_f replacing l_h. Also, assuming that the input signal $x(t)$ is real valued, only half of the SBM $\mathbf{S}(t)$ entries need to be computed. Then,

$$\Psi_{\mathrm{SBM}}(\mathbf{S}) = \frac{\#\{\mathbf{S}\}}{D}, \tag{15}$$

where $\#\{\mathbf{S}\} = \|\Re\{\mathbf{S}\}\|_0 + \|\Im\{\mathbf{S}\}\|_0$ denotes the number of non-zero entries of $\mathbf{S}(t)$, considering the real and imaginary parts of complex entries as two different coefficients.

III. Problem Description

Let $g^{(l)}(\omega)$, $l = 1, \cdots, L$, be a set of HRTFs. For each l, we want to approximate $g^{(l)}$ using (2)–(4). Suppose that h and f are given, and that for each $l = 1, \cdots, L$, we have an SBM $\mathbf{S}^{(l)}$.

Let $\mathbf{S} = \{\mathbf{S}^{(1)}, \cdots, \mathbf{S}^{(L)}\}$ and $\Xi = [\mathbf{S}, h, f]$. The error $\Upsilon^{(l)}(\Xi)$ in approximating $g^{(l)}$ is given by

$$\Upsilon^{(l)}(\Xi) = \frac{1}{2\pi D} \sum_{d=1}^{D} \int_{-\pi}^{\pi} w(\omega) \left| 20 \log_{10} g^{(l)}(\omega) - 20 \log_{10} \hat{g}_d^{(l)}(\omega) \right|^2 d\omega, \tag{16}$$

where $w(\omega)$ is a frequency weighting function motivated by the frequency selectivity of the auditory system. In combination with the magnitude deviation in decibels, this frequency weighting leads to a rough approximation of the bandwidth of auditory filters [15], [16]. As a consequence of using complex logarithms in 16, it is straightforward to obtain that an amplitude deviation of 1 dB between $g^{(l)}(\omega)$ and $\hat{g}_d^{(l)}(\omega)$, is weighted equally to a phase deviation of 0.12 rad. The consideration of the phase is important because the auditory system is sensitive to phase deviations [21], especially in terms of binaural incoherence [22]. We would then like to solve

$$\begin{aligned} &\text{find} && \Xi_\star = \arg\min_{\Xi} \Psi(\Xi) \\ &\text{subject to} && \Upsilon^{(l)}(\Xi) \leq \epsilon^2, \text{forall } l \in \{1, \cdots, L\}, \\ & && \Delta^{(l)}(\Xi) \leq \tau, \text{for all } l \in \{1, \cdots, L\}, \end{aligned} \tag{17}$$

where ϵ denotes the approximation tolerance, τ the latency constraint, and, for some given $a, b, c \geq 0$,

$$\Psi(\Xi) = a \sum_{l=1}^{L} \Psi_{\text{SBM}}(\mathbf{S}) + \beta \Psi_{\text{FB}}(h) + c \Psi_{\text{FB}}(f), \tag{18}$$

is a measure of the complexity of the whole scheme and $\Delta^{(l)}(\Xi)$ denotes the latency of the subband implementation of $g^{(l)}$, computed using (13).

About the choice of $w(\omega)$: Suppose that we want to do the optimization in the Bark scale [16], i.e.,

$$\Upsilon^{(l)}(\Xi) = \frac{1}{D} \sum_{d=1}^{D} \frac{1}{b_h - b_l} \int_{b_l}^{b_h} \left| 20 \log_{10} g(\omega(b)) - 20 \log_{10} \hat{g}_d^{(l)}(\omega(b)) \right|^2 db,$$

where $\omega(b)$ denotes the conversion from the Bark scale to angular frequency, and b_l and b_h denote the integration limits in the Bark scale. Then, we need to choose

$$w(\omega) = \frac{f_s}{2(b_h - b_l)} \chi_{[\omega(b_l), \omega(b_h)]}(|\omega|)\, b'\left(\frac{f_s}{2\pi}\omega\right),$$

where $b(\cdot)$ denotes conversion from Hertz to Bark, $b'(\cdot)$ denotes its first derivative, f_s denotes the sampling frequency in Hertz and $\chi_{[a,b]}(|\omega|)$ denotes the indicator function on $[a, b]$ (i.e., $\chi_{[a,b]}(|\omega|) = 1$ if $|\omega| \in [a, b]$ and $\chi_{[a,b]}(|\omega|) = 0$ otherwise).

The functions $\Psi(\Xi)$ and $\Upsilon^{(l)}(\Xi)$, $l = 1, \cdots, L$, are neither convex, nor quasi-convex. Hence, the problem (17) cannot be solved using standard optimization algorithms, and convergence to the global optimal solution cannot be guaranteed. Thus, we propose algorithms for approximately solving (17). In the greedy and relaxation algorithms, presented in Section V, the SBMs $\mathbf{S}^{(l)}$, $l = 1, \cdots, L$, are designed assuming that h and f are given. The resulting SBMs can then be used to initialize the SBM-shrink and FB-shrink algorithms presented in Section VI, which aim at designing the complete set of parameters Ξ. These algorithms require computing the derivatives of $\Upsilon^{(l)}(\Xi)$ with respect to the entries of $\mathbf{S}$, h and f. These derivatives are given in Section IV.

IV. Derivatives of the Approximation Error

From (16), we have

$$\Upsilon^{(l)}(\Xi) = \frac{K}{2\pi D} \sum_{d=1}^{D} \int_{-\pi}^{\pi} w(\omega) \left| \tilde{c}_d^{(l)}(\omega) \right|^2 d\omega \tag{19}$$

with $K = 400/\log^2 10$ being a scaling constant required to convert $\log(\cdot)$ into $20\log_{10}(\cdot)$ and

$$\begin{aligned} \tilde{c}_d^{(l)}(\omega) &= c^{(l)}(\omega) - \hat{c}_d^{(l)}(\omega), \\ c^{(l)}(\omega) &= \log g^{(l)}(\omega), \\ \hat{c}_d^{(l)}(\omega) &= \log \hat{g}_d^{(l)}(\omega), \\ \hat{g}_d^{(l)}(\omega) &= \left[\Phi^{-1}(\mathbf{F}^*\mathbf{SH})(\omega)\right]_d. \end{aligned} \tag{20}$$

For each entry $[\mathbf{S}^{(l)}]_{m,n}(k)$ of $\mathbf{S}^{(l)}(t)$, we consider its real and imaginary components separately. Hence, we define a *subband index* as a quartet (m, n, k, ρ), where $\rho \in \{\mathfrak{R}, \mathfrak{I}\}$ indicates whether the index corresponds to the real or the imaginary component of $[\mathbf{S}^{(l)}]_{m,n}(k)$. For each $1 \leq m \leq M$, let $\overline{m} = \operatorname{mod}(M + 1 - m, M) + 1$. Then, for each (m, n, k, ρ), we define its conjugate index by $\overline{(m, n, k, \rho)} = (\overline{m}, \overline{n}, k, \rho)$. We say that an index i is self-conjugate if $i = \overline{i}$. To each subband index $i = (m, n, k, \rho)$, we associate a real coefficient $\sigma_i^{(l)} = \rho\{[\mathbf{S}^{(l)}(k)]_{m,n}\}$. Since the impulse response $g(t)$ is real valued, the coefficient $\sigma_{\overline{i}}^{(l)}$ associated to the conjugate of index i is given by

$$\sigma_{\overline{i}}^{(l)} = \begin{cases} \sigma_i^{(l)}, & \rho = \mathfrak{R}, \\ -\sigma_i^{(l)}, & \rho = \mathfrak{I}. \end{cases} \tag{21}$$

Hence, we only consider indexes $i = (m, n, k, \rho)$ with $1 \leq m, n \leq M/2 + 1$ or $2 \leq m \leq M/2$, and such that $\rho = \mathfrak{R}$ whenever i is self-conjugate. We call such indexes, *essential subband indexes*. We use $\mathcal{E}$ to denote the set of essential subband indexes, $\mathcal{S} = \{i \in \mathcal{E} : i = \overline{i}\}$ to denote the set of self-conjugate indexes in $\mathcal{E}$, and $\mathcal{S}^c$ to denote its complement in $\mathcal{E}$. We also use $\mathcal{R} = \{(m, n, k, \rho) \in \mathcal{E} : \rho = \mathfrak{R}\}$ and $\mathcal{I} = \{(m, n, k, \rho) \in \mathcal{E} : \rho = \mathfrak{I}\}$ to denote the set of real and imaginary indexes in $\mathcal{E}$, respectively. Notice that, in view of (21), $\mathcal{S} \subseteq \mathcal{R}$.

In view of (21), we associate to each index $i \in \mathcal{E}$ a SBM $\mathbf{U}_i(t)$, defined by

$$\mathbf{U}_i(t) = \begin{cases} \mathbf{E}_{m,n,k}(t), & i \in \mathcal{S}, \\ \mathbf{E}_{m,n,k}(t) + \mathbf{E}_{\overline{m},\overline{n},k}(t), & i \in \mathcal{S}^c \cap \mathcal{R}, \\ j\left(\mathbf{E}_{m,n,k}(t) - \mathbf{E}_{\overline{m},\overline{n},k}(t)\right), & i \in \mathcal{S}^c \cap \mathcal{I}, \end{cases} \tag{22}$$

where the impulse response $\mathbf{E}_{m,n,k}(t)$ is given by

$$[\mathbf{E}_{m,n,k}(\kappa)]_{\mu,\nu} = \begin{cases} 1, & (m, n, k) = (\mu, \nu, \kappa) \\ 0, & \text{otherwise} \end{cases}.$$

We then have the following lemma.

Lemma 3: For $l = 1, \cdots, L$, let $\Upsilon^{(l)}(\Xi)$ be given by (16) and $\theta \in \mathbb{R}$ be a scalar parameter upon which either $\mathbf{S}$, h or f depend. Then,

$$\frac{d}{d\theta}\Upsilon^{(l)}(\Xi) = \frac{K}{2\pi D}\sum_{d=1}^{D}\int_{-\pi}^{\pi} w(\omega)\Re\left\{\frac{\overline{\tilde{\mathfrak{c}}_{\mathfrak{d}}^{(\mathfrak{l})}(\omega)}}{\hat{\mathfrak{g}}_{\mathfrak{d}}^{(\mathfrak{l})}(\omega)}\frac{\mathfrak{d}}{\mathfrak{d}\theta}\hat{\mathfrak{g}}_{\mathfrak{d}}^{(\mathfrak{l})}(\omega)\right\}\mathfrak{d}\omega. \quad (23)$$

Furthermore, the derivatives $\frac{d}{d\theta}\hat{g}_d^{(l)}$ are given by:

1) *If θ is a coefficient of $\mathbf{S}^{(l)}$ and $i = (m, n, k, \rho)$ is its associated essential subband index, then,*

$$\frac{d}{d\theta}\hat{g}_d^{(l)} = \begin{cases} \left[\Phi^{-1}\left(\mathbf{J}_{m,n,k}\right)\right]_d, & i \in \mathcal{S}, \\ 2\left[\Phi^{-1}\left(\mathbf{J}_{m,n,k}^{\Re}\right)\right]_d, & i \in \mathcal{S}^c \cap \mathcal{R}, \\ -2\left[\Phi^{-1}\left(\mathbf{J}_{m,n,k}^{\Im}\right)\right]_d, & i \in \mathcal{S}^c \cap \mathcal{I}, \end{cases} \quad (24)$$

where $\mathbf{J}_{m,n,k}^{\Re}(\omega)$ and $\mathbf{J}_{m,n,k}^{\Im}(\omega)$ are, respectively, the Fourier transforms of the real and imaginary parts of

$$\mathbf{J}_{m,n,k}(\omega) = \mathbf{F}^*(\omega)\mathbf{E}_{m,n,k}(\omega)\mathbf{H}(\omega).$$

2) *If $\theta = h(t)$, for some $t \in \{1 - l_h, \cdots, 0\}$* (i.e., *the coefficient θ corresponds to the t-th entry of the impulse response of h), then*

$$\frac{d}{d\theta}\hat{g}_d^{(l)} = \left[\Phi^{-1}\left(\mathbf{F}^*\mathbf{S}^{(l)}\mathcal{W}_M\mathbf{L}_2\mathbf{D}_{t+l_h}\mathbf{L}_1\right)\right]_d, \quad (25)$$

where $\mathbf{D}_n$ is the diagonal matrix with a one in the n–th entry of its main diagonal, and zero everywhere else.

3) *If $\theta = f(t)$, for some $t \in \{1 - l_f, \cdots, 0\}$, then*

$$\frac{d}{d\theta}\hat{g}_d^{(l)} = \left[\Phi^{-1}\left(\mathbf{L}_1^*\mathbf{D}_{t+l_f}\mathbf{L}_2^*\mathcal{W}_M^*\mathbf{S}^{(l)}\mathbf{H}\right)\right]_d. \quad (26)$$

I) PROOF: See Appendix A. ■

V. ALGORITHMS WITH FIXED FILTERBANKS

In view of (15), for each $l = 1, \cdots, L$, we need to minimize the number $\#\{\mathbf{S}^{(l)}\}$ of non-zero entries of $\mathbf{S}^{(l)}(t)$. To this end, following the discussion in Section II-C, we choose the FB prototypes h and f so that the entries of each $\mathbf{S}^{(l)}(t)$ are concentrated on the main diagonal as much as possible. To this end, we design $h = f$ as root raised cosine windows with $\omega_0 = \pi/M$ and $\beta = M/D - 1$, which are symmetrically truncated so that their relative energy outside the band $[-\pi/D, \pi/D]$ is below a certain threshold ϑ, i.e.,

$$1 - \frac{\int_{-\pi/D}^{\pi/D}|h(\omega)|^2 d\omega}{\int_{-\pi}^{\pi}|h(\omega)|^2 d\omega} \leq \vartheta.$$

With this choice of prototypes, the last two terms in (18) are fixed. Hence, the design of each SBM can be addressed separately. Thus, for each $l = 1, \cdots, L$, we solve

$$\begin{array}{ll} \text{find} & \hat{\mathbf{S}}^{(l)} = \underset{\mathbf{S}^{(l)}}{\arg\min}\,\#\left\{\mathbf{S}^{(l)}\right\} \\ \text{subject to} & \Upsilon^{(l)}(\Xi) \leq \epsilon^2, \\ & \Delta^{(l)}(\Xi) \leq \tau. \end{array} \quad (27)$$

We propose below two algorithms for solving (27). The first one is called *greedy*, and is described in Section V-A. It consists of an iterative procedure, which at each iteration increases by one the number of non-zero entries of $\mathbf{S}^{(l)}$, until the constraint $\Upsilon^{(l)}(\Xi) \leq \epsilon^2$ is met. This is done while respecting the constraint $\Delta^{(l)}(\Xi) \leq \tau$ at each iteration (obviously, the iterations will never end if both constraints are such that the problem is unfeasible). This algorithm chooses the support of $\mathbf{S}^{(l)}$ in a greedy fashion, i.e., choosing at each iteration the 'best' next entry. However, there is no guarantee that the set of chosen entries at the end of the iterations is the best one. Hence, the goal of the second algorithm is to remedy this drawback. We call this algorithm *relaxation*, and describe it in Section V-B. This algorithm is initialized by the SBMs $\mathbf{S}^{(l)}$ resulting from the greedy algorithm, and then solves a sequence of constrained optimization problems, aiming at reducing the support of $\mathbf{S}^{(l)}$. Therefore, these two algorithm can be considered as two stages of a single design method.

Since the design of each SBM can be addressed separately, to simplify the notation, in the remainder of this section we assume that there is only one HRTF to be approximated, i.e., $L = 1$ and $\mathbf{S} = \mathbf{S}^{(1)}$.

Remark 4. The greedy algorithm is inspired by the algorithm proposed in [9, S5]. Both algorithms consist in an iterative procedure, thus, they may at first sight appear to be similar. There is, however, an essential difference between them: The algorithm in [9, S5] was originally designed to minimize an approximation error defined using the *linear amplitude* scale. Then, in order to achieve a minimization in the logarithmic amplitude scale, that algorithm has to be iteratively applied, each time minimizing the linear amplitude error with a different frequency weighting $w(\omega)$. In contrast, the greedy algorithm proposed here is a different approach aiming at the direct minimization of the *logarithmic amplitude* error. Thus, only one run of the greedy algorithm is necessary in order to achieve the desired solution; and, as shown in Section VII-B2, it produces SBMs of significantly lower complexity.

A. Greedy Algorithm

The greedy algorithm proceeds in iterations. Let $\mathbf{S}_q$ denote the subband model at the q-th iteration, and $\hat{g}_{d,q}$ and $\hat{\mathbf{G}}_q$ be defined as in (10)–(12) by using $\mathbf{S}_q$ in place of $\mathbf{S}$. We also use $\mathcal{H}_q = \operatorname{supp}(\mathbf{S}_q)$ to denote the *support* of $\mathbf{S}_q$, i.e., the set of essential subband indexes $(m, n, k, \rho) \in \mathcal{E}$ such that $\rho\{[\mathbf{S}(k)]_{m,n}\} \neq 0$.

Notice that, in view of (13), the delay constraint in (27) requires that $\delta_s \leq \frac{\tau - l_h + 1}{D}$. We can then devise the following iterative algorithm. Each iteration carries out two main steps, which we call *support update* and *optimization*. The detailed description of these two steps are given in Sections V-A1 and V-A2, respectively.

Greedy Algorithm.

The inputs of the algorithm are M, D, ϵ, τ and ϑ. Design $h = f$ using a root raised cosine window with $\omega_0 = \pi/M$ and $\beta = M/D - 1$, truncated so that the energy outside the band $[-\pi/D, \pi/D]$ is below ϑ. Put $\mathbf{S}_0 = 0$. Then, at the q-th iteration, the algorithm carries out the following steps:

1) **Support update:**Pick a new subband index $(m_q, n_q, k_q, \rho_q) \in \mathcal{S}$, with $k_q \geq -\frac{\tau - l_h + 1}{D}$, and add it to the current support, i.e., $\mathcal{H}_q = \mathcal{H}_{q-1} \cup \{(m_q, n_q, k_q, \rho_q)\}$.
2) **Optimization:** Use an unconstrained optimization method, initialized by $\mathbf{S}_{q-1}$ and $\rho_q\{[\mathbf{S}(k_q)]_{m_q,n_q}\} = 0$, to solve
$$\mathbf{S}_q = \underset{\mathrm{supp}(\mathbf{S})=\mathcal{H}_q}{\arg\min}\ \Upsilon(\mathbf{S}, h, f). \tag{28}$$
3) Stop if $\Upsilon(\mathbf{S}_q, h, f) \leq \epsilon^2$ or a maximum number of iterations is reached.

The output of the algorithm is $\mathbf{S}_q$ or *unfeasible if* the maximum number of iterations was reached.

We provide the details of each step below.

1) Support update: Let $\tilde{c}_{d,q}(\omega)$, $d = 1, \cdots, D$ be the values of (20) at the q-th iteration. Let $\tilde{\mathbf{c}}_q^{(l)}(\omega) = [\tilde{c}_{1,q}^{(l)}(\omega), \cdots, \tilde{c}_{D,q}^{(l)}(\omega)]^T$ and $\tilde{\mathbf{C}}_q = \Phi(\tilde{\mathbf{c}}_q \mathbf{1}_D)$ ($\mathbf{1}_D$ is a D-dimensional column vector of ones) be its polyphase representation. It is straightforward to see that

$$\Upsilon(\Xi) = \frac{1}{D}\left\|\tilde{\mathbf{C}}_q\right\|_{\mathbf{W}}^2, \tag{29}$$

where $\mathbf{W} = \Phi(w\mathbf{1}_D)$ and

$$\|\mathbf{X}\|_{\mathbf{W}}^2 = \langle \mathbf{X}, \mathbf{X}\rangle_{\mathbf{W}}$$
$$\langle \mathbf{X}, \mathbf{Y}\rangle_{\mathbf{W}} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{Tr}\{\mathbf{X}(\omega)\mathbf{W}(\omega)\mathbf{Y}^\star(\omega)\}\, d\omega.$$

Now, at iteration q, we have

$$\tilde{\mathbf{C}}_q(\omega) = \left(\mathbf{G}(\omega) - \hat{\mathbf{G}}_q(\omega)\right)\mathbf{Z}_q(\omega),$$
$$\mathbf{Z}_q(\omega) = \left(\mathbf{G}(\omega) - \hat{\mathbf{G}}_q(\omega)\right)^{-1}\tilde{\mathbf{C}}_q(\omega). \tag{30}$$

Approximating $\mathbf{Z}_q$ with $\mathbf{Z}_{q-1}$, we can write (29) in a linear least-squares form as follows

$$\Upsilon_{\mathrm{lls}}(\mathbf{S}_q) = \frac{1}{D}\|\mathbf{G}\mathbf{Z}_{q-1} - \mathbf{F}^*\mathbf{S}_q\mathbf{H}\mathbf{Z}_{q-1}\|_{\mathbf{W}}^2$$
$$= \frac{1}{D}\|\mathbf{G} - \mathbf{F}^*\mathbf{S}_q\mathbf{H}\|_{\mathbf{R}_{q-1}}^2. \tag{31}$$

with $\mathbf{R}_q(\omega) = \mathbf{Z}_q(\omega)\mathbf{W}(\omega)\mathbf{Z}_q^*(\omega)$.

To choose the next subband index, for each $i \in \mathcal{E}$, we define $\mathbf{V}_i(\omega)$ to be the polyphase representation of the D-cyclostationary system induced by $\mathbf{U}_i(\omega)$ (22), i.e.,

$$\mathbf{V}_i(\omega) = \mathbf{F}^*(\omega)\mathbf{U}_i(\omega)\mathbf{H}(\omega).$$

Then, in view of (31), we choose the index $i_q \in \mathcal{E}$ for which the correlation (weighted by $\mathbf{R}_{q-1}$) between $\mathbf{V}_i(\omega)$ and the current residual $\mathbf{G}(\omega) - \hat{\mathbf{G}}_{q-1}(\omega)$ is maximized, i.e.,

$$i_q = \underset{i\in\mathcal{E}}{\arg\max}\ \frac{\left|\left\langle \mathbf{G} - \hat{\mathbf{G}}_{q-1}, \mathbf{V}_i\right\rangle_{\mathbf{R}_{q-1}}\right|}{\|\mathbf{V}_i\|_{\mathbf{R}_{q-1}}}. \tag{32}$$

To compute the inner products in (32) in an efficient manner, we use

$$\langle \mathbf{G} - \hat{\mathbf{G}}_{q-1}, \mathbf{V}_{m,n,k,\rho}\rangle_{\mathbf{R}_{q-1}}$$
$$= \langle \mathbf{X}_{q-1}, \mathbf{U}_{m,n,k,\rho}\rangle_{\mathbf{I}}$$
$$= \begin{cases} \Re\left([\mathbf{X}_{\mathfrak{q}-1}(\mathfrak{k})]_{\mathfrak{m},\mathfrak{n}}\right), & (m,n,k,\rho) \in \mathcal{S}, \\ 2\Re\left([\mathbf{X}_{\mathfrak{q}-1}(\mathfrak{k})]_{\mathfrak{m},\mathfrak{n}}\right), & (m,n,k,\rho) \in \mathcal{S}^c \cap \mathcal{R}, \\ -2\Im\left([\mathbf{X}_{\mathfrak{q}-1}(\mathfrak{k})]_{\mathfrak{m},\mathfrak{n}}\right), & (m,n,k,\rho) \in \mathcal{S}^c \cap \mathcal{I}, \end{cases}$$

(recall that $\mathbf{I}$ denotes the identity matrix), and

$$\mathbf{X}_q(\omega)$$
$$= \mathbf{F}(\omega)\left(\mathbf{G}(\omega) - \hat{\mathbf{G}}_q(\omega)\right)\mathbf{Z}_q(\omega)\mathbf{W}(\omega)\mathbf{Z}_q^*(\omega)\mathbf{H}^*(\omega).$$

2) Optimization: The unconstrained optimization problem (28) can be solved using any gradient search method. We use the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method described in [23]. This requires the computation of the derivatives of $\Upsilon(\Xi)$, with respect to the entries of $\mathbf{S}$, which are given in Lemma 3.

B. Relaxation Algorithm

From the greedy algorithm we obtain a feasible SBM $\mathbf{S}$ (i.e., one which satisfies the constraints in (27)), together with its support set $\mathcal{H} = \mathrm{supp}(\mathbf{S})$. The relaxation algorithm described in this section aims to further reduce the size $\#\{\mathbf{S}\}$ of $\mathcal{H}$, while staying within the feasible region (i.e., respecting the same constraints).

We have

$$\#\{\mathbf{S}\} = \sum_{(m,n,k,\rho)\in\mathcal{H}} \chi\left(\rho\left\{[\mathbf{S}(k)]_{m,n}\right\}\right),$$

where

$$\chi(z) = \begin{cases} 0, & z = 0, \\ 1, & z \neq 0. \end{cases}$$

Clearly, it is very difficult to minimize $\#\{\mathbf{S}\}$ using numerical optimization methods, because χ is constant almost everywhere. To go around this, following [24], we choose an $\alpha > 0$, and replace χ by

$$r_\alpha(z) = 1 - e^{-\frac{z^2}{2\alpha^2}},$$

which is a smooth function for each $\alpha > 0$ and converges in a point-wise manner to χ. This leads us to the following algorithm, which solves a sequence of constrained optimization problems with decreasing values of α.

Step 1 of relaxation algorithm requires solving a constrained optimization problem. To this end, in this work we use the barrier method [25, S11.3], which requires the derivatives of Υ with respect to the entries of $\mathbf{S}$. These are given in Lemma 3.

VI. Algorithms Designing the Filterbanks

The relaxation algorithm described in Section V outputs a set of SBMs $\mathbf{S} = [\mathbf{S}^{(1)}, \cdots, \mathbf{S}^{(L)}]$, together with their supports $\mathcal{H} = \{\mathcal{H}^{(1)}, \cdots, \mathcal{H}^{(L)}\}$, satisfying the constraints in (17), for given choices of h and f. We can then use these SBMs for

Relaxation algorithm

The inputs of the algorithm are M, D, ϵ, τ, ϑ and a $\varsigma > 0$. Run the greedy algorithm to obtain $\mathbf{S}$, h and f. Put $\mathbf{S}_0 = \mathbf{S}$, $\mathcal{H}$ $=$ supp$(\mathbf{S})$ and

$$\alpha_0 = \|\mathbf{S}\|_\infty \triangleq \max_{m,n,k,\rho} \left| \rho \left\{ [\mathbf{S}]_{m,n}(k) \right\} \right|.$$

Then, for each $q \in \mathbb{N}$, iterate over the following steps:

1) Use a constrained optimization method, initialized with $\mathbf{S}_{q-1}$, to solve

$$\begin{aligned} \mathbf{S}_q = & \quad \arg\min_{\mathbf{S}} R_{\alpha_{q-1}} (\mathbf{S}), \\ \text{subject to} & \quad \operatorname{supp}(\mathbf{S}) = \mathcal{H}, \\ & \quad \Upsilon (\mathbf{S}, h, f) \leq \epsilon^2, \end{aligned} \tag{33}$$

with

$$R_\alpha (\mathbf{S}) = \sum_{(m,n,k,\rho)\in\mathcal{H}} r_\alpha \left(\rho \left\{ [\mathbf{S}(k)]_{m,n} \right\} \right).$$

2) Put $\alpha_q = 0.5\alpha_{q-1}$ and stop if $\alpha_q < \varsigma\|\mathbf{S}_q\|_\infty$. Upon termination, make zero all entries $\rho\{[\mathbf{S}_q]_{m,n}(k)\}$ for which $|\rho\{[\mathbf{S}_q]_{m,n}(k)\}| \leq \varsigma\|\mathbf{S}_q\|_\infty$. The output of the algorithm is $\mathbf{S}_q$, or *unfeasible if* the greedy algorithm returned with *unfeasible*.

SBM-shrink Algorithm

The inputs of the algorithm are M, D, ϵ, τ, ϑ and $\varsigma > 0$. Run the relaxation algorithm, for each $l = 1, \cdots, L$, to obtain $\mathbf{S}$, h and f. Put $\mathbf{S}_0 = \mathbf{S}$, $h_0 = h$, $f_0 = f$, $\mathcal{H} =$ supp$(\mathbf{S})$, $\mathcal{I}_h =$ supp(h), $\mathcal{I}_f =$ supp(f) and

$$\alpha_0 = \max_{l\in\{1,\cdots,L\}} \left\| \mathbf{S}^{(l)} \right\|_\infty.$$

Then, for each $q \in \mathbb{N}$, iterate over the following steps:

1) Use a constrained optimization method, initialized with $\mathbf{S}_{q-1}$, h_{q-1} and f_{q-1}, to solve

$$\begin{aligned} [\mathbf{S}_q, h_q, f_q] = & \quad \arg\min_{\mathbf{S},h,f} \textstyle\sum_{l=1}^{L} R^{(l)}_{\alpha_{q-1}} \left(\mathbf{S}^{(l)}\right), \\ \text{subject to} & \quad \operatorname{supp}(\mathbf{S}, h, f) = \{\mathcal{H}, \mathcal{I}_h, \mathcal{I}_f\}, \\ & \quad \Upsilon^{(l)} (\mathbf{S}, h, f) \leq \epsilon^2, \text{forall } l, \\ & \quad \|h\|_2 \leq 1, \\ & \quad \|f\|_2 \leq 1, \end{aligned}$$

with

$$R^{(l)}_\alpha (\mathbf{S}) = \sum_{(m,n,k,\rho)\in\mathcal{H}^{(l)}} r_\alpha \left(\rho \left\{ \left[\mathbf{S}^{(l)}(k) \right]_{m,n} \right\} \right).$$

2) Put $\alpha_q = 0.5\alpha_{q-1}$.

3) Stop if $\alpha_q < \varsigma \min_{l\in\{1,\cdots,L\}} \|\mathbf{S}^{(l)}_q\|_\infty$.

Upon termination, for each $l = 1, \cdots, L$, make zero all entries $\rho\{[\mathbf{S}^{(l)}_q]_{m,n}(k)\}$ for which $|\rho\{[\mathbf{S}^{(l)}_q]_{m,n}(k)\}| \leq \varsigma\|\mathbf{S}^{(l)}_q\|_\infty$. The outputs of the algorithm are $\mathbf{S}_q$, h_q and f_q, or *unfeasible if* the relaxation algorithm returned with *unfeasible*, for some l $= 1, \cdots, L$.

FB-shrink algorithm:

The inputs of the algorithm are M, D, ϵ, τ and ϑ. Run the relaxation algorithm, for each $l = 1, \cdots, L$, to obtain $\mathbf{S}$, h and f. Put $\mathbf{S}_0 = \mathbf{S}$, $h_0 = h$, $f_0 = f$, $\mathcal{H} =$ supp$(\mathbf{S})$, $\mathcal{I}_{h,0} =$ supp(h) and $\mathcal{I}_{f,0} =$ supp(f). Then, for each $q \in \mathbb{N}$, iterate over the following steps:

1) Shrink the supports $\mathcal{I}_{h,q-1}$ and $\mathcal{I}_{f,q-1}$ by removing their first entry (recall that h and f are anti-causal).

2) If $\Upsilon^{(l)}(\mathbf{S}_{q-1}, h_{q-1}, f_{q-1}) > \epsilon^2$, for some l (i.e., the solution becomes unfeasible), use a constrained optimization method, initialized with $\mathbf{S}_{q-1}$, h_{q-1}, f_{q-1} and

$$\zeta > \max_{l=1,\cdots,L} \Upsilon^{(l)} (\mathbf{S}_{q-1}, h_{q-1}, f_{q-1}) - \epsilon^2,$$

to solve

$$\begin{aligned} [\mathbf{S}_q, h_q, f_q, \zeta_q] = & \quad \arg\min_{\mathbf{S},h,f,\zeta} \zeta + \left(\|h\|^2 - 1\right)^2 + \\ & \quad +\left(\|f\|^2 - 1\right)^2, \\ \text{subject to} & \quad \operatorname{supp}(\mathbf{S}, h, f) = \{\mathcal{H}, \mathcal{I}_{h,q}, \mathcal{I}_{f,q}\}, \\ & \quad \Upsilon^{(l)} (\mathbf{S}, h, f) \leq \epsilon^2 + \zeta, \text{forall } l. \end{aligned}$$

Else, put $\{\mathbf{S}_q, h_q, f_q\} = \{\mathbf{S}_{q-1}, h_{q-1}, f_{q-1}\}$.

3) If $\zeta_q > 0$, (i.e., if the solution is still unfeasible), stop. The outputs of the algorithm are $\mathbf{S}_q$, h_q and f_q or *unfeasible if* the relaxation algorithm returned with *unfeasible*, for some $l = 1, \cdots, L$.

initializing an algorithm which optimizes the complete parameter set $\Xi = [\mathbf{S}, h, f]$. In Sections VI-A and VI-B, we introduce two algorithms for doing so. The SBM-shrink algorithm in Section VI-A aims to reduce the supports in $\mathcal{H}$, while the FB-shrink algorithm in Section VI-B does it with the supports $\mathcal{I}_h = \{-l_h + 1, \cdots, 0\}$ and $\mathcal{I}_f = \{-l_f + 1, \cdots, 0\}$ (recall that the filters h and f are anti-causal) of h and f, respectively.

A. Algorithm to Reduce the Support of the Subband Models

The SBM-shrink algorithm proposed in this section aims at reducing the supports in $\mathcal{H}$ resulting from the relaxation algorithm, while keeping the supports $\mathcal{I}_h$ and $\mathcal{I}_f$ unchanged. The SBM-shrink algorithm is similar to the relaxation algorithm, with the difference in that, instead of a single SBM $\mathbf{S}^{(l)}$, it jointly tunes $\mathbf{S}$, h and f.

As with the relaxation algorithm, we use the barrier method [25, S11.3] to solve Step 1 of the SBM-shrink algorithm. The required derivatives of $\Upsilon^{(l)}$, $l = 1, \cdots, L$, with respect to the entries of $\mathbf{S}$, h and f are given in Lemma 3.

B. Algorithm to Reduce the Support of the Filterbanks

The FB-shrink algorithm proposed in this section reduces the supports $\mathcal{I}_h$ and $\mathcal{I}_f$, while keeping $\mathcal{H}$ unmodified. The basic idea is to sequentially shrink $\mathcal{I}_h$ and $\mathcal{I}_f$ until the problem becomes unfeasible. Notice that the objective of this algorithm is complementary to that of the SBM-shrink algorithm.

As with the SBM-shrink algorithm, we use the barrier method [25, S11.3] to solve Step 2 of the FB-shrink algorithm.

VII. Experiments

In this section we aim at finding a convenient parameterization (i.e., the values of M, D, ϑ and ϵ) for the proposed subband approximation algorithms, yielding good numerical efficiency and still accurate sound localization performance. In order to evaluate the latter, we make use of both, simulated localization experiments using a model for sagittal-plane sound localization

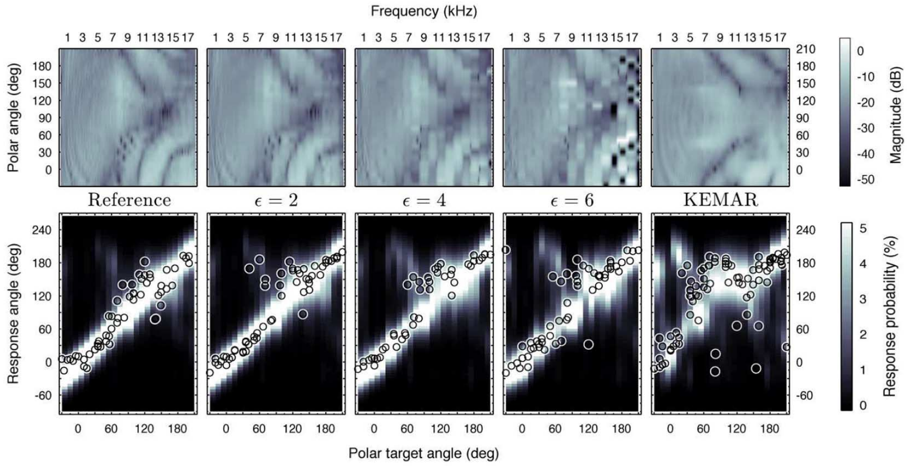


Fig. 1. Experimental data from the most sensitive listener NH14. Top row: Magnitude characteristics of left-ear DTFs in the median plane. Bottom row: Actual responses and predicted response probabilities of the model. Magnitudes and response probabilities are encoded by brightness. Experimental conditions are shown in columns for the reference DTFs, various subband representations ($\epsilon \in \{2, 4, 6\}$), and non-individualized DTFs (KEMAR).

performance [26], as well as psychoacoustic sound localization experiments with human subjects. In Section VII-A we describe the HRTFs used in our experiments, the model used to predict localization performance, and the methodology of the psychoacoustic experiments. In Section VII-B we first evaluate the effect of various approximation tolerances on localization performance. Then, we compare the performance of the different proposed approximation algorithms. Finally, we compare the latency-complexity trade-offs offered by the subband technique to those offered by SFFT techniques.

A. Methods

HRTFs: We based our experiments on free-field HRTFs of eight listeners (NH12, NH14, NH15, NH39, NH43, NH64, NH68, and NH72) taken from the ARI database at http://sofaconventions.org. These HRTFs were measured for elevation angles ranging from $-30°$ to $80°$, azimuth angles ranging all around the listener, and frequencies up to 18 kHz. For more details on the measurement setup and procedure see [27], [28]. For our study, the head-related impulse responses were resampled to a sampling rate of 36 kHz, and consist of 192 samples. In order to reduce the impact of direction-independent components in HRTFs, directional transfer functions (DTFs) were calculated [28], [29]. To this end, the log-magnitude spectra of all HRTFs of a listener were averaged across directions, and then the HRTFs were filtered with the inverse minimum-phase representation of this average. In the ARI database, these DTFs are available as `dtfB` files. The magnitude spectra of DTFs of an exemplary listener (NH14) as a function of the polar angle, ranging from $-30°$ (front, below eye-level) via 90° (top) to 210° (rear, below eye-level) are shown in the top row of Fig. 1.

Furthermore, in order to consider the condition of listening with non-individualized HRTFs, i.e., through different ears, we calculated DTFs from HRTFs measured on a mannequin (KEMAR) [7].

Sagittal-plane Localization Model: The input of the sagittal-plane localization model consists of the so-called template DTFs, target DTFs, and listeners' sensitivities [26]. The template DTFs are the listener-specific DTFs, i.e., we simulate the listener's auditory system as being tuned to the acoustically measured DTFs. The target DTFs represent the DTFs under test. The listeners' sensitivities are used to consider the listeners' individual localization performance. In the sound localization process, they are attributed to non-acoustic listener-specific factors (i.e., others than those attributed to DTFs) distinguishing between poor and good localizers [30]. The listeners whose DTFs are considered in this experiment already participated in previous sound localization studies, thus their corresponding sensitivities are known from [26]. The output of the model for a certain target is a probability mass vector (PMV), describing the predicted probabilities of response angles in the polar dimension (down, up, front, back). Examples of such PMVs as functions of the polar target angle are shown in Fig. 1 (bottom row, with the response probability encoded by brightness). See [26] for more details on the process for obtaining PMVs. From the PMVs, common psychoacoustic measures of localization performance can be derived [12]. Quadrant error rates are the percentage of hemisphere confusions, i.e., deviations between polar response and target angles exceeding 90°. When a hemisphere confusion does not occur, then the resulting response is called local response. The local polar error is defined as the root mean square of polar-angle deviations between local responses and the corresponding targets. Hence, this error comprises both the accuracy (response bias) and precision (response variability) of local responses. As suggested by [12], we evaluate sagittal-plane localization performance only within $\pm 30°$ lateral angle, because the polar-angle dimension is increasingly compressed for more lateral positions.

Psychoacoustic Localization Experiment: Listeners NH14, NH15, NH39, NH62 and NH68 participated in the sound localization experiments. None of the tested listeners had indication of hearing disorders. All of them had thresholds of 20 dB hearing level or lower, at frequencies from 0.125 kHz to 12.5 kHz.

Virtual acoustic stimuli were generated by filtering Gaussian white noise bursts with a duration of 500 ms with the DTFs corresponding to the tested direction. The presentation level was 50 dB above the individually measured absolute detection threshold for that stimulus, estimated in a manual up-down procedure for a frontal eye-leveled position. In the experiments, the stimulus level was randomly roved for each trial within the range of ± 5 dB, in order to reduce the possibility of using overall level cues for localization. Experimental target directions were randomly selected in the range from $-30°$ to $80°$ elevation, and covered the full azimuthal range.

In the experiments, the listeners were immersed in a virtual visual environment, presented via a head-mounted display (Oculus Rift). They were asked to respond to stimuli by using a manual pointer. We tracked both the head of the listener and the pointer, to render the environment in real time, and to collect the response directions. For more details on the apparatus see [28].

The procedure of the sound localization task was identical to that from [31]. Prior to the acoustic tests, listeners performed a visual and an acoustic training. The goal of the visual training was to train subjects to perform accurately within the virtual environment. The training was completed when the listeners were able to point within 4 seconds to a visual target with a directional error smaller than 2. In the subsequent acoustic training, listeners localized virtual acoustic stimuli with visual feedback. The goal of the acoustic training was to settle a stable localization performance of the subjects. The acoustic training consisted of 6 blocks, with 50 acoustic targets each, and lasted 2 hours.

In the actual acoustic tests, in each trial, the listeners had to align their head to the front, press a button to start the stimulus, then point to the perceived direction, and click a button. During the presentation, the listeners were instructed not to move. Each DTF condition was tested in three blocks of 100 trials each, with a fixed DTF condition in a block. Each block lasted approximately 15 minutes and after each block, subjects had a pause. The presentation order of blocks was randomized across listeners. More details on this task can be found in [28], [31].

The localization performance in the polar dimension was measured by means of quadrant error rates and local polar errors. The localization performance in the lateral dimension was measured by means of the lateral error, i.e., the root mean square of the target-response deviations.

B. Results

Effect of the Approximation Tolerance ϵ : To investigate the effect of the approximation tolerance ϵ on the localization performance, we simulate localization experiments using the sagittal-plane localization model. We focus on localization in sagittal planes, because this is the dimension where spectral modifications of HRTFs are known to be most critical for sound localization [11].

We base the configuration of the different subband approximation algorithms in that of [9, § VI-B], where $M = 32$, $D = 20$, $\vartheta = -32$ dB and $\epsilon \simeq 3$ was used. However, since we want to produce approximations with values of ϵ as small as 1, we reduce the value of ϑ to $\vartheta = -36$ dB. Also, we used the greedy algorithm due to the practical reason that, across its iterations, it produces a sequence of intermediate approximations covering a range of tolerances ϵ. To assure that all the other proposed algorithms would produce very similar localization performances, provided that they have the same configuration (i.e., the values of M, D, ϵ, etc.), we compare the algorithm performances in Section VII-B2. To prevent the greedy algorithm from choosing an unnecessarily large number of coefficients, we constrain the set of subband indexes to be on the main diagonal, i.e., in the support update step of the greedy algorithm, we consider subband indexes $(m, n, k, \rho) \in \mathcal{S}$ having $m = n$. Finally, we choose $\varsigma = 10^{-6}$, and we use a latency constraint of $\tau = 180$ samples (i.e., 5 ms).

We evaluate the predicted sagittal-plane localization performance for the eight listeners, considering 150 target directions, randomly selected within a lateral range of $\pm 30°$. The target DTFs were 1) subband-approximated DTFs using the greedy algorithm with ϵ ranging from 1 to 10, 2) the original DTFs representing the reference condition, and 3) the non-individualized DTFs (KEMAR). As examples for the target DTFs, Fig. 1 (top row) shows the magnitude spectra of the selected target DTFs of an exemplary listener (NH14). Notice that, for $\epsilon \geq 6$, the subband-approximated spectra show gaps (i.e., unsupported subbands), which might be perceptually relevant.

The target DTFs were applied to the sagittal-plane sound localization model and the corresponding PMVs were calculated. Fig. 1 (bottom row) shows the predicted response probabilities for the most sensitive listener NH14. In the reference condition, regions of large probabilities are highly concentrated toward response-target deviations of 0. With increasing ϵ, this concentration gradually diminishes. Least concentration and large-probability regions far away from the polar target angle (c.f., quadrant errors) are obtained with the non-individualized DTFs. Fig. 2 shows the predicted localization performance for all listeners. As just shown for the exemplary listener NH14, the performance was better for the reference DTFs and worst with the non-individualized DTFs (KEMAR). Across all listeners (represented as the median), the performance obtained for the approximated DTFs degraded consistently with increasing ϵ. The approximation tolerance of $\epsilon = 3$ appears to yield a small degradation only. For $\epsilon < 2$, performance seems to approach that for the reference condition. For $\epsilon > 6$, the performance degradation seems to stagnate, at least for the local polar error. Interestingly, even for the largest approximation tolerance tested ($\epsilon = 10$), the predicted performance was still better than that obtained for the non-individualized DTFs (KEMAR). Thus, ϵ from 2 to 6 seems to provide a reasonable range for further tests.

We now evaluate the localization performance of subband-approximated DTFs, in psychoacoustic sound localization experiments with human subjects. We do so for the values of ϵ within the selected interval $2 \leq \epsilon \leq 6$. The goal is to confirm the performance predictions from our preceding experiment. For each listener, five DTF sets were tested. Two of these sets were the original DTFs (reference) and a non-individualized DTF set (KEMAR), respectively. The other three DTF sets were obtained from subband approximations with tolerances $\epsilon = 2, 4,$ and 6. We choose these values because the subband-approximated DTFs sometimes show spectral gaps for $\epsilon \geq 6$ (as can be seen for example in Fig. 1), and the predicted localization per-

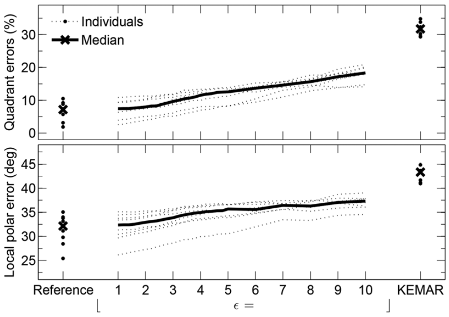


Fig. 2. Predicted sagittal-plane localization performance as a function of approximation tolerance ϵ. Notice that performance for $\epsilon = 1$ is close to the one for individually measured DTFs (Reference), whereas performance for $\epsilon = 10$ remains better than for non-individualized DTFs (KEMAR).

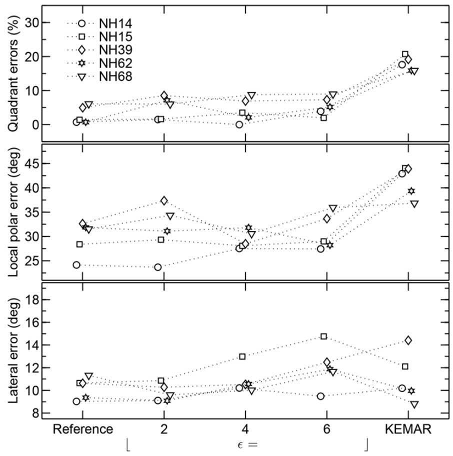


Fig. 3. Listener-specific sagittal-plane localization performance obtained in psychoacoustic experiments. Note the robustness of each listener's performance across various ϵ.

formance shown in Fig. 2 suggests that ϵ from 2 to 6 might yield a good localization.

Fig. 1 (bottom row) shows the responses (open circles) of NH14, together with the target-specific response predictions described in Section VII-A2.[1] Consistently with the model predictions, responses were most accurate (i.e., concentrated to the diagonal) for the reference condition, slightly less accurate for the subband conditions ($\epsilon \in \{2, 4, 6\}$), and less accurate for the non-individualized condition. Actual responses mostly coincided with high probability regions indicating reasonable correspondence between predicted and actual results.

Fig. 3 shows the actual localization performance of all tested listeners. Subband approximations yielded generally better performance than the non-individualized DTFs and the performance seems not to differ significantly between the reference and subband conditions. In comparison with the predictions shown in Fig. 2, the actual listener-specific psychoacoustic outcomes were less consistent across conditions, and the listeners tended to perform better than predicted. In particular, the actual outcomes suggest that the subband approximations preserved the perceptually relevant spectral features quite well even for $\epsilon = 6$.

Fig. 3 also shows the lateral error of the listeners in the experimental conditions. The performance seems to be similar for $\epsilon \leq 4$, and degraded for $\epsilon = 6$.

In conclusion, the actual and predicted evaluation results suggest that subband approximations maintained accurate localization performance for a large range of ϵ. While in sagittal planes, $\epsilon = 6$ seems to be sufficient, in horizontal planes, $\epsilon \leq 4$ seems to be required. In view of this, $\epsilon = 3$ seems to be a conservative choice, from the point of view of sound localization, for future applications.

Comparison of Subband Approximation Algorithms: In this section we compare the four proposed algorithms in terms of localization performance and complexity. We also include in the comparison the algorithm proposed in [9, V]. For this and the subsequent comparisons, we use approximations of 24 median-plane DTFs of the left ear of the subject NH68. We do so because the listener-specific sensitivity of NH68 is, on the average, indicating that NH68 is representing a typical listener. Furthermore, we use the algorithm configuration described in Section VII-B1, with $\epsilon = 3$. Table I shows the predicted localization performance for the four algorithms. The predicted performance obtained without subband approximation, i.e., reference performance, is 13.8% quadrant errors and 33.9 local polar error. Also, the predicted localization performance for the non-individualized DTFs (KEMAR) is 31.8% quadrant errors and 41.4 local polar error (notice that the reference as well as the non-individualized performances obtained here differ from the one in Fig. 2, because now we are working with a more restricted DTF set). The predicted performance differs only marginally across the tested algorithms, being also close to that obtained for the reference performance. This suggests that conclusions on localization performance drawn by using one particular algorithm can be generalized to any other algorithm.

TABLE I
Comparison of Different Subband Approximation Algorithms, For $\vartheta = -36$ dB

Algorithm	Ψ_{SBM}	Ψ_{FB}	l_h	QE	LPE
Greedy	2.271	14.95	139	16.5%	34.3°
Relaxation	2.198	14.95	139	16.4%	34.3°
SBM-shrink	2.033	14.95	139	16.9%	35.1°
FB-shrink	2.198	11.3	66	17.0%	34.9°
Alg. in [9]	3.712	14.95	139	15.9%	36.1°

Table II also shows the amount of multiplications per sample per SBM (on average) Ψ_{SBM} and per FB Ψ_{FB}, as well as the tap-size l_h of the analysis FB prototype, resulting from the tested algorithms. All proposed algorithms clearly outperform the algorithm proposed in [99, §V], in terms of complexity per SBM. Also, the FB-shrink algorithm yields a significant reduction in terms of FB complexity and prototype tap size. On the other hand, we see that differences in terms of SMB complexity between the four proposed algorithms are only marginal. However, these differences become more significant in Table II, where FB prototype energy leakage thresholds

[1] DTFs, psychoacoustic response patterns, and predicted response probabilities, for all other listeners, are provided as supplementary material.

TABLE II
COMPARISON OF DIFFERENT SUBBAND APPROXIMATION ALGORITHMS, FOR $\vartheta = -20$ dB

Algorithm	Ψ_{SBM}	Ψ_{FB}	l_h	QE	LPE
Greedy	2.954	10.25	45	15.5%	34.4°
Relaxation	2.694	10.25	45	15.4%	34.4°
SBM-shrink	2.264	10.25	45	15.4%	34.8°
FB-shrink	2.694	9.85	37	14.8%	34.7°
Alg. in [9]	4.646	10.25	45	15.8%	34.6°

is increased to $\vartheta = -20$ dB. In particular, we see that the relaxation algorithm yields a sensible advantage over the greedy one, and SBM-shrink yields even further advantages. The drawback of the SBM-shrink and FB-shrink algorithms is that they require the joint optimization of all FB prototypes and SBM coefficients. Hence, they do not scale conveniently for approximating large HRTF sets. Thus, their computational effort for the approximation of large sets might not always justify the computational complexity advantage that they offer. Hence, we conclude that, while the SBM-shrink and FB-shrink algorithms may be preferred choices for approximating small HRFT sets, the relaxation algorithm seems to be the most convenient choice for large sets.

Latency-Complexity Trade-Offs of SB and SFFT Methods: In this section we study the trade-offs between latency and complexity offered by both, SB and SFFT methods (see Appendix B for a detailed description of SFFT methods). As in Section VII-B2, we use the median-plane DTFs of the left ear of subject NH68. In view of our above conclusions, we focus our study in the relaxation algorithm with $\epsilon = 3$. Also, since we want to see the dependence of the latency on other design parameters, we do not use a latency constraint, i.e., we set $\tau = -\infty$, and we replace, in the relaxation and greedy algorithms, the SB index constraint $k_q \geq -\frac{\tau - l_h + 1}{D}$ by $k_q \geq 0$. We then do the approximation for several values of the number of subbands M, downsampling factor D and FB prototype energy leakage threshold ϑ.

It follows from our discussion in Section II-E that choosing the number of subbands M to be a power of two leads to a reduced complexity in the computation of an FFT/IFFT. Hence, we constrain our search to $M \in \{4, 8, 16, 32\}$. Also, the FB prototype design described in Section II-C is only valid for values of D in the range $M/2 \leq D \leq M$. Hence, to avoid complicating the design of this prototype, we should in principle constrain our search for D to these values. However, practical evidence indicates that the choice $D = M$ leads to a very poor design. Hence, we constrain our search to $M/2 \leq D < M$. Finally, we constrain our search for ϑ to $\vartheta \in \{-20\text{ dB}, -30\text{ dB}, -40\text{ dB}\}$.

Fig. 4 shows the dependencies on D of the average SBM complexity per HRTF, the complexity of a FB stage (either analysis or synthesis), and the overall latency, respectively. The dependencies are shown for different values of ϑ. Notice that the value of M is not shown in the plots, because it can be inferred from D. All curves show peaks at the values of D when it becomes close to M. Hence, we conclude that these values are undesirable choices. We also see that a decrease in ϑ produces a decrease of the SBM complexity, but an increase of the FB complexity and the overall latency. This means that there exists a trade-off between latency and FB complexity on the one hand, and SBM complexity on the other.

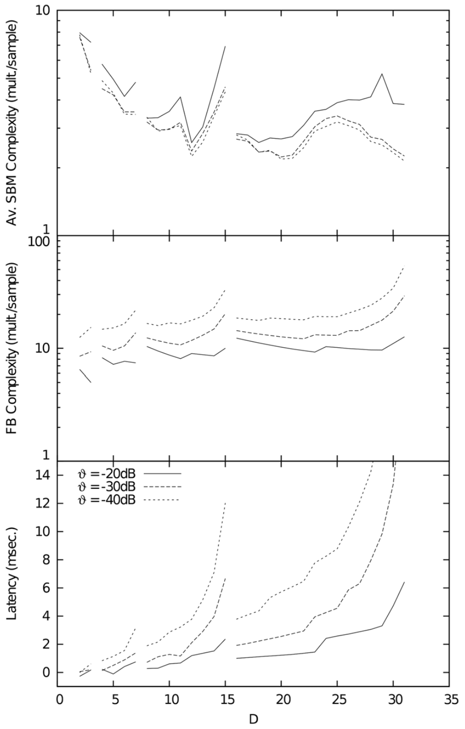


Fig. 4. Average complexity per HRTF of the SBMs (top), complexity of either the analysis or the synthesis FB (center) and implementation latency (bottom) vs. down-sampling factor D, for various FB prototype energy leakage thresholds ϑ.

The optimal choices of M, D and ϑ depend on the proportion of FB complexity that is associated to the computation of each subband model. Notice that regardless of the number of sound sources and reflections, the synthesis FB stage needs to be computed only once per ear. This applies also to SFFT methods, where the synthesis stage is formed by a set of IFFT operations. Hence, we neglect the contribution of the synthesis stage in the complexity of both, SB and SFFT methods. Also, notice that, in the SB method, the analysis FB stage needs to be computed only once per sound source, regardless of the number of reflections to be simulated. This is because time delays can be represented in SBMs provided that we have one SBM for each delayed version of the HRTF, with delays in the range $\{0, D-1\}$. In contrast, this does not apply to SFFT methods, as this would require having SFFT representations for all possible delays, and long delays would require increasing the size of FFT segments. Hence, the latency-complexity trade-off offered by the SB method depends on the number of reflections per sound source considered for 3D sound rendering. Fig. 5 shows this trade-off, for different values of ϑ and two extreme scenarios, namely, for the scenario of free-field listening, i.e., without reflections, and for the scenario of listening in a highly reverberant space, i.e., as the number of reflections approaches infinity. The

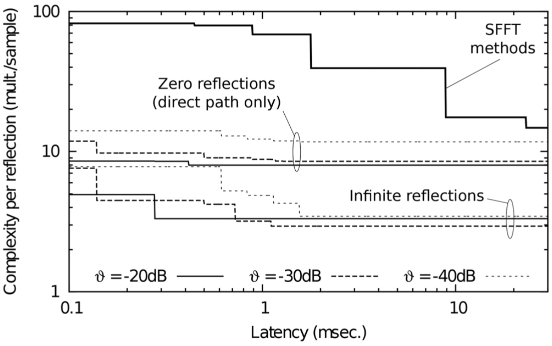


Fig. 5. Latency-complexity trade-offs for different values of ϑ and number of reflections.

latter scenario also applies to the case when virtual source signals can be pre-processed, i.e., filtered by an analysis FB, so that their SB representations do not need to be computed at the time of 3D rendering. In this case, the computational complexity of the analysis FB can be ignored. Notice that this pre-processing is only possible for the SB method, since, as explained before, time delays cannot be represented in SFFT methods after the analysis stage has been processed. In order to avoid showing values resulting from undesirable choices of D (such as those resulting in the aforementioned peaks), we only plot the lower envelopes resulting from each trade-off, i.e., for each latency value, we plot the minimum complexity resulting from all values of D yielding smaller or equal latencies. Recall, that the SB method introduces a certain approximation error ϵ, whereas the SFFT methods do not. Also, notice that a constant value of ϵ does not necessarily guarantee that the localization performance remains unchanged for all combinations of D and ϑ. We did not validate the perceptual performance of all the configurations shown in Fig. 5 using psychoacoustic experiments. Nevertheless, predictions indicate that localization performance varies only within $17\% \pm 3\%$ quadrant error rate and 36 ± 2 local polar error. Also, we do not expect substantial perceptual differences in the presence of reflections, because we have no evidence that reflections need to be more accurately represented than the direct path. Taking this into consideration, Fig. 5 serves the purpose of illustrating the major advantage offered by the SB method over SFFT methods, in terms of latency-complexity trade off.

As mentioned in Section I, PZ models do not introduce latency, and their numerical efficiency is maximized in applications with static sources. We point out that, if a minor latency can be afforded, the SB method also largely outperforms PZ models in terms of numerical efficiency. More precisely, in the comparison of Fig. 5, the use of PZ models requires 75.83 multiplications per sample and reflection (i.e., an average model order of 37.42). This suggests that the SB method may still be the best option in applications with static sources.

VIII. Conclusions

The subband approximation of HRTFs allows scaling the computational effort when rendering spatial sound in virtual binaural acoustics. Our psychoacoustic results indicate that the proposed algorithms preserve the salience of spatial cues, even for relatively high approximation tolerances, yielding computationally very efficient implementations. Especially, but not only, for low implementation latencies, the subband approach is much more efficient than SFFT methods. Moreover, the complexity of the directional part of the filtering process has only little effect on the overall computational effort. Hence, while the subband approach already outperforms SFFT methods for sound sources presented in free-field, its computational efficiency becomes even more advantageous when considering additional room reflections filtered with their corresponding HRTFs. Hence, the method appears to be very well-suited for real-time applications of virtual auditory displays considering multiple room reflections.

This work can be considered as a first step towards the application of subband methods to virtual binaural acoustics. More experiments are still required to evaluate the performance of this approach in a variety of practical situations. For example, in a real-time virtual auditory display, due to listener and sound source movements, the implementation of HRTFs requires the processing of time-variant filters. Also, the implementation of room acoustics requires filtering delayed versions of the direct sound by the HRTFs corresponding to the reflections' directions. Moreover, virtual binaural acoustics involve other aspects of hearing, apart from spatialization, like timbre and externalization. These aspects were not considered in the present work, and remain to be evaluated. Nevertheless, since the subband method is a generalization of SFFT methods, we expect that many of the properties of SFFT methods, for implementing HRTFs in the aforementioned situations, will be also enjoyed by subband methods.

The source code for the approximation of HRTFs is available at http://sf.net/p/sbhrtf.

Appendix A
Proof of Lemma 3

Proof of Lemma 3: From (19), we have

$$\begin{aligned}
&\frac{d}{d\theta}\Upsilon^{(l)}\left(\mathbf{S},h,f,M,D\right)\\
&= \frac{K}{2\pi D}\sum_{d=1}^{D}\int_{-\pi}^{\pi} w(\omega)\frac{d}{d\theta}\left|\tilde{c}_{d}^{(l)}(\omega)\right|^{2}d\omega\\
&= \frac{K}{\pi D}\sum_{d=1}^{D}\int_{-\pi}^{\pi} w(\omega)\Re\left\{\overline{\tilde{\mathfrak{c}}_{\mathfrak{d}}^{(\mathfrak{l})}(\omega)}\frac{\mathfrak{d}}{\mathfrak{d}\theta}\tilde{\mathfrak{c}}_{\mathfrak{d}}^{(\mathfrak{l})}(\omega)\right\}\mathfrak{d}\omega\\
&= \frac{-K}{\pi D}\sum_{d=1}^{D}\int_{-\pi}^{\pi} w(\omega)\Re\left\{\frac{\overline{\tilde{\mathfrak{c}}_{\mathfrak{d}}^{(\mathfrak{l})}(\omega)}}{\hat{\mathfrak{g}}_{\mathfrak{d}}^{(\mathfrak{l})}(\omega)}\frac{\mathfrak{d}}{\mathfrak{d}\theta}\hat{\mathfrak{g}}_{\mathfrak{d}}^{(\mathfrak{l})}(\omega)\right\}\mathfrak{d}\omega.
\end{aligned}$$

and (23) follows. For (24), we have

$$\hat{g}_{d}^{(l)} = \left[\Phi^{-1}\left(\mathbf{F}^{*}\mathbf{S}^{(l)}\mathbf{H}\right)\right]_{d},$$

where $\mathbf{F}^{*}(t) = \overline{\mathbf{F}(-t)^{T}}$. Now,

$$\mathbf{S}^{(l)}(t) = \sum_{i\in\mathcal{E}}\sigma_{i}^{(l)}\mathbf{U}_{i}(t).$$

Then,

$$\frac{d}{d\theta}\hat{g}_d = \left[\Phi^{-1}\left(\mathbf{F}^*\mathbf{U}_i\mathbf{H}\right)\right]_d,$$

and the result follows after noticing that $\mathbf{J}_{\overline{m,n,k}}(t) = \overline{\mathbf{J}_{m,n,k}}(t)$.

For (25), from (14), we have

$$\begin{aligned}\frac{d}{d\theta}\hat{g}_d^{(l)} &= \left[\Phi^{-1}\left(\mathbf{F}^*\mathbf{S}^{(l)}\frac{d}{d\theta}\mathbf{H}\right)\right]_d \\ &= \left[\Phi^{-1}\left(\mathbf{F}^*\mathbf{S}^{(l)}\mathcal{W}_M\mathbf{L}_2\frac{d}{d\theta}\Lambda_h\mathbf{L}_1(\omega)\right)\right]_d \\ &= \left[\Phi^{-1}\left(\mathbf{F}^*\mathbf{S}^{(l)}\mathcal{W}_M\mathbf{L}_2\mathbf{D}_i\mathbf{L}_1(\omega)\right)\right]_d,\end{aligned}$$

and (26) follows similarly. ■

Appendix B
Segmented FFT methods

In this work we use the term segmented FFT (SFFT) to refer to four methods for achieving fast convolution with low latency. Two of these methods are the overlap-add (OA) and overlap-save (OS) [4, §5.3.2]. The slight difference between them is not relevant for the purposes of this work, so we do not differentiate between them, and we jointly refer to both as the OA/OS method. It consists of splitting the input signal into a sequence of overlapping segments of samples. Then, the filtering operation is separately applied to each segment, in the frequency domain, using FFT. More precisely, let n denote the length of the impulse response $g(t)$ of the filter and N be the length of each segment. This technique uses an overlap length of $n-1$ samples. The complexity, in number of real multiplications per sample, of the resulting implementation is

$$\Psi_{\text{OA/OS,FFT}} = \Psi_{\text{OA/OS,IFFT}} = \frac{N}{N-n+1}\log_2 N,$$

for the FFT and IFFT stages, and

$$\Psi_{\text{OA/OS,Filt}} = \frac{2N}{N-n+1},$$

for the filtering stage in the frequency domain. Also, its implementation latency, in samples, is

$$\Delta_{\text{OA/OS}} = N - n.$$

The third method that we consider within the SFFT family is the low-delay fast convolution (LDFC) technique, proposed in [7]. This method splits the n-tap impulse response $g(t)$ into a number of non-overlapping blocks, each of which is processed using the OA/OS method. The impulse response splitting is done such that the first two blocks have the same length (which must equal a power of two), and the length of every other block is twice the one of its predecessor. Then, the OA/OS method that is applied to each block uses a segment whose length is twice that of the block. The resulting complexity is the addition of the complexity of each OA/OS stage, and the implementation latency is

$$\Delta_{\text{LDFC}} = \frac{N_1}{2},$$

where N_1 denotes the length of the first block. Hence, this method permits reducing the latency of the OA/OS method, at the expense of an increase in complexity.

The fourth method within the SFFT family is the zero-delay fast convolution (ZDFC) method, also proposed in [7]. This method is similar to the LDFC one, with the difference in that the first block is implemented in the time domain using convolution. While this method yields a zero latency implementation, we do not consider it in our experiments, due to its high complexity.

References

[1] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-related transfer functions of human subjects," *J. Audio Eng Soc*, vol. 43, pp. 300–321, 1995.

[2] B.-s. Xie*, Head-Related Transfer Function and Virtual Auditory Display*. Plantation, FL, USA: J Ross, 2013.

[3] L. Savioja, J. Houpaniemi, T. Lokki, and R. Väänäen, "Creating interactive virtual acoustic environments," *J. Audio Eng. Soc.*, vol. 47, no. 9, pp. 675–705, 1999.

[4] J. G. Proakis and D. G. Manolakis*, Digital signal processing: Principles, algorithms, and applications*, 3rd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 1996.

[5] M. Blommer and G. Wakefield, "Pole-zero approximations for head-related transfer functions using a logarithmic error criterion," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 278–87, May 1997.

[6] J. Jot, V. Larcher, and O. Warusfel, "Digital signal processing issues in the context of binaural and transaural stereophony," in *Proc. AES Conv.*, 1995.

[7] W. Gardner, "Efficient convolution without input-output delay," *J. Acoust. Soc. Amer.*, vol. 43, no. 3, pp. 127–136, 1995.

[8] M. Vorländer*, Auralization: Fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality (RWTHedition)*, 1st ed. New York, NY, USA: Springer, 11, 2008.

[9] D. Marelli and M. Fu, "A recursive method for the approximation of lti systems using subband processing," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1025–1034, Mar. 2010.

[10] P. Chanda and S. Park, "Immersive rendering of coded audio streams using reduced rank models of subband-domain head-related transfer functions," in *Proc. ICASSP*, 2006, pp. 345–348.

[11] E. A. Macpherson and J. C. Middlebrooks, "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," *J. Acoust. Soc. Amer.*, vol. 111, pp. 2219–2236, 2002.

[12] J. C. Middlebrooks, "Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency," *J. Acoust. Soc. Amer.*, vol. 106, pp. 1493–1510, 1999.

[13] H. Fletcher and W. Munson, "Loudness, its definition, measurement and calculation," *J. Acoust. Soc. Amer.*, vol. 5, no. 2, pp. 82–108, 1933.

[14] B. Moore, B. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness, and partial loudness," *J. Audio Eng. Soc*, vol. 45, no. 4, pp. 224–240, 1997.

[15] B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, pp. 103–138, 1990.

[16] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Amer.*, vol. 68, no. 5, pp. 1523–1525, 1980.

[17] D. Marelli, R. Baumgartner, and P. Majdak, "Efficient representation of head-related transfer functions in subbands," in *Proc. EUSIPCO*, 2014.

[18] P. Vaidyanathan*, Multirate Systems and Filterbanks*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[19] D. Marelli and M. Fu, "Performance analysis for subband identification," *IEEE Trans. Signal Process.*, vol. 51, no. 12, pp. 3128–3142, Dec. 2003.

[20] S. Weiss and R. Stewart, "Fast implementation of oversampled modulated filter banks," *Electron Lett.*, vol. 36, no. 17, pp. 1502–1503, 2000.

[21] B. Moore and B. Glasberg, "Difference limens for phase in normal and hearing-impaired subjects," *J. Acoust. Soc. Amer.*, vol. 86, p. 1351, 1989.

[22] M. J. Goupell and W. M. Hartmann, "Interaural fluctuations and the detection of interaural incoherence: Bandwidth effects," *J. Acoust. Soc. Amer.*, vol. 119, no. 6, pp. 3971–3986, 2006.

[23] R. Fletcher, *Practical Methods of Optimization*, 2nd ed. New York, NY, USA: Wiley, 1987.

[24] H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed l0 norm," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 289–301, Jan. 2009.

[25] S. Boyd and L. Vandenberghe*, Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[26] R. Baumgartner, P. Majdak, and B. Laback, "Modeling sound-source localization in sagittal planes for human listeners," *J. Acoust. Soc. Amer.*, vol. 136, pp. 791–802, 2014.

[27] 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, pp. 623–637, 2007.

[28] P. Majdak, M. J. Goupell, and B. Laback, "3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training," *Atten. Percept. Psycho.*, vol. 72, pp. 454–469, 2010.

[29] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *J. Acoust. Soc. Amer.*, vol. 91, no. 3, pp. 1637–1647, 1992.

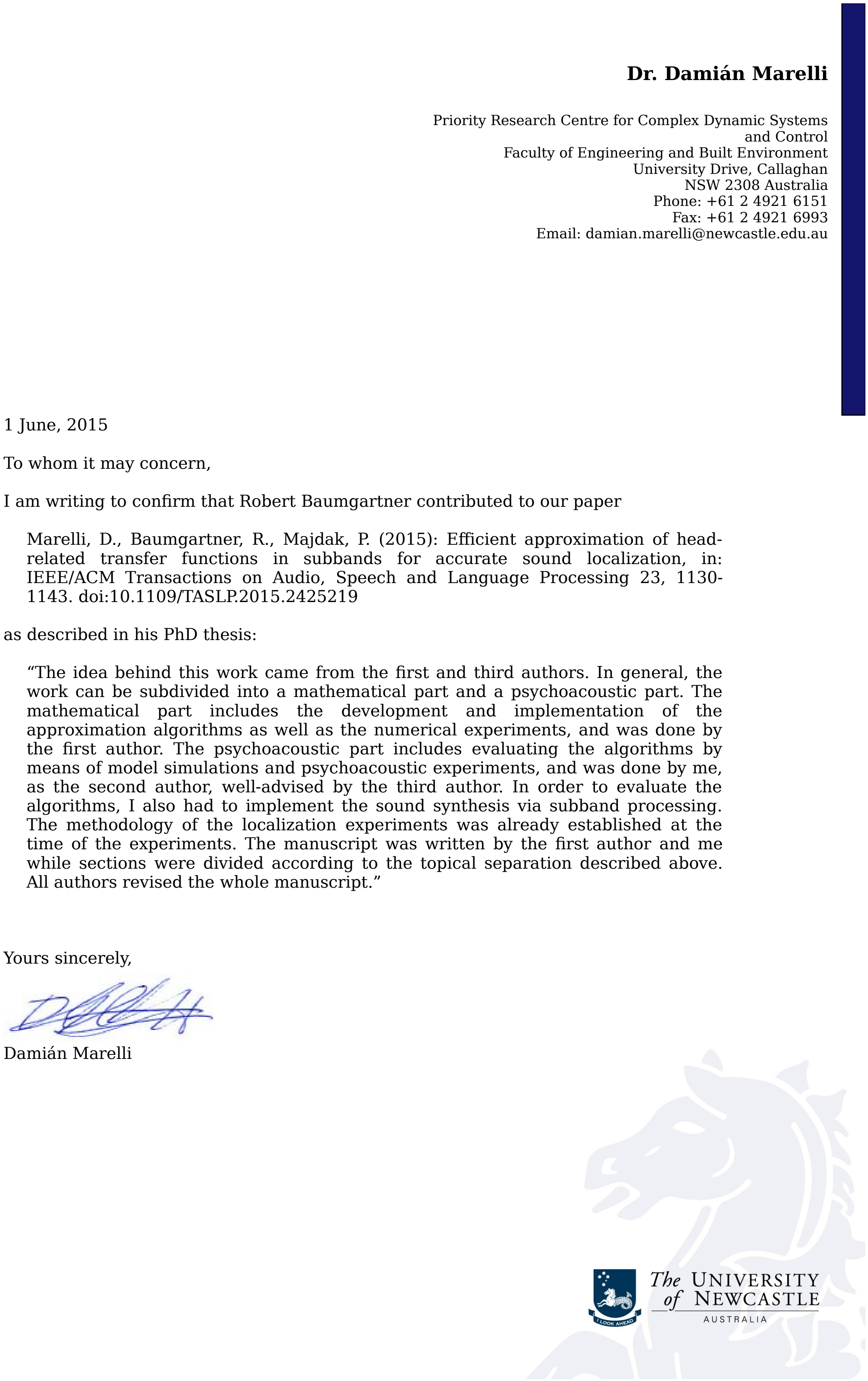
[30] P. Majdak, R. Baumgartner, and B. Laback, "Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization," *Front. Psychol.*, vol. 5, no. 319, pp. 1–10, 2014.

[31] P. Majdak, B. Masiero, and J. Fels, "Sound localization in individualized and non-individualized crosstalk cancellation systems," *J. Acoust. Soc. Amer.*, vol. 133, no. 4, pp. 2055–68, Apr. 2013, http://view.ncbi.nlm.nih.gov/pubmed/23556576.

Damián Marelli received his Bachelors degree in electronics engineering from the Universidad Nacional de Rosario, Argentina, in 1995 and a Ph.D. degree in electrical engineering and a Bachelor (Honous) degree in mathematics from the University of Newcastle, Australia, in 2003. From 2004 to 2005, he held a postdoctoral position at the Laboratoire d'Analyse Topologie et Probabilités, CNRS/Université de Provence, France. Since 2005, he has been a Research Academic at the School of Electrical Engineering and Computer Science at the University of Newcastle, Australia. In 2007, he received a Marie Curie Postdoctoral Fellowship, hosted at the Faculty of Mathematics, University of Vienna, Austria, and in 2011 he received a Lise Meitner Senior Fellowship, hosted at the Acoustics Research Institute of the Austrian Academy of Sciences. His main research interests include time–frequency analysis, system identification, statistical signal processing and sensor networks.

Robert Baumgartner is a Research Assistant at the Acoustics Research Institute of the Austrian Academy of Sciences in Vienna and a Ph.D. candidate at the University of Music and Performing Arts Graz (KUG). In 2010 and 2012, he received his B.Sc. and M.Sc. degrees, respectively, in electrical engineering and audio engineering, a inter-university program at the University of Technology Graz (TUG) and KUG. His master thesis about modeling sound localization in sagittal planes with the application to subband-encoded HRTFs was honored by the German Acoustic Association (DEGA) with a student award. His research interests include spatial hearing and spatial audio.

Piotr Majdak was born in Opole, Poland, in 1974. He studied electrical engineering and audio engineering at the University of Technology and at the University of Music and Performing Arts, both in Graz, Austria, and received an M.Sc. degree in 2002. Since 2002, he has been working at the Acoustics Research Institute (ARI) of the Austrian Academy of Sciences, and is involved in projects including binaural signal processing, localization of sounds, and cochlear implants. In 2008, he received his Ph.D. degree on lateralization of sounds based on interaural time differences in cochlear-implant listeners. He is member of the Acoustical Society of America, Association for Research in Otolaryngology, the Austrian Acoustic Association, and the president of the Austrian section of the Audio Engineering Society (AES).



Dr. Damián Marelli

Priority Research Centre for Complex Dynamic Systems
and Control
Faculty of Engineering and Built Environment
University Drive, Callaghan
NSW 2308 Australia
Phone: +61 2 4921 6151
Fax: +61 2 4921 6993
Email: damian.marelli@newcastle.edu.au

1 June, 2015

To whom it may concern,

I am writing to confirm that Robert Baumgartner contributed to our paper

> Marelli, D., Baumgartner, R., Majdak, P. (2015): Efficient approximation of head-related transfer functions in subbands for accurate sound localization, in: IEEE/ACM Transactions on Audio, Speech and Language Processing 23, 1130-1143. doi:10.1109/TASLP.2015.2425219

as described in his PhD thesis:

> “The idea behind this work came from the first and third authors. In general, the work can be subdivided into a mathematical part and a psychoacoustic part. The mathematical part includes the development and implementation of the approximation algorithms as well as the numerical experiments, and was done by the first author. The psychoacoustic part includes evaluating the algorithms by means of model simulations and psychoacoustic experiments, and was done by me, as the second author, well-advised by the third author. In order to evaluate the algorithms, I also had to implement the sound synthesis via subband processing. The methodology of the localization experiments was already established at the time of the experiments. The manuscript was written by the first author and me while sections were divided according to the topical separation described above. All authors revised the whole manuscript.”

Yours sincerely,

Damián Marelli

The UNIVERSITY of NEWCASTLE
AUSTRALIA

Chapter 6

Effect of vector-based amplitude panning on sound localization in sagittal planes

On 4 May 2015, this work was submitted as

Baumgartner, R., Majdak, P. (2015): Amplitude panning and sound localization in sagittal planes: a modeling study, submitted to: Journal of the Audio Engineering Society (JAES).

The idea behind this work came from me as the first author. The work was designed by me and the co-author. I performed the model simulations, created the figures, and drafted the manuscript. The co-author revised an earlier version of this manuscript.

Effect of vector-based amplitude panning on sound localization in sagittal planes *

ROBERT BAUMGARTNER, *AES Member* **AND PIOTR MAJDAK,** *AES Member*

Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria

Localization of sound sources in sagittal planes, including the front-back and up-down dimensions, relies on listener-specific monaural spectral cues. A functional model, approximating human processing of spectro-spatial information, was applied to assess sagittal-plane localization performance for sound events created by vector-based amplitude panning (VBAP) of coherent loudspeaker signals. First, we assessed VBAP between two loudspeakers in the median plane. The model predicted a strong dependence on listeners' individual head-related transfer functions and a systematic degradation with increasing polar-angle span between the loudspeakers. Nevertheless, on average across listeners and directions, the VBAP principle seems to work reasonably well for spans up to 40° as indicated by a regression analysis. Then, we investigated the directional dependence of the performance in several loudspeaker arrangements designed in layers of constant elevation. The simulations emphasized the critical role of the loudspeaker elevations on localization performance.

0 INTRODUCTION

Spatial audio systems have to deal with a limited number of loudspeakers. In order to create sound events localized between the loudspeaker positions, so-called phantom sources, systems often apply VBAP between coherent loudspeaker signals. The VBAP technique determines loudspeaker gains in terms of regulating the summed sound intensity vector generated by a pair or triplet of loudspeakers, depending on whether the rendered sound field consists of one or two directional dimensions, respectively [18].

We use the interaural-polar coordinate system shown in Fig. 1 to distinguish the effects of VBAP on sound localization. In the lateral-angle dimension (left-right), VBAP introduces interaural differences in level (ILD) and time (ITD) and thus, perceptually relevant localization cues [20]. In the polar-angle dimension, however, monaural spectral features at high frequencies cue sound localization [11], a mechanism generally not captured in the broadband concept of VBAP.

Polar-angle perception is based on a monaural learning process in which spectral features, that are characteristic for the listener's morphology, are related to certain directions [5]. Due to the monaural processing, the use of spectral features can be disrupted by spectral irregularities superimposed by the source spectrum [12]. The use of spectral features is limited to high frequencies (above around 0.7-4 kHz), because the spatial variance of head-related transfer functions (HRTFs) increases with frequency [17]. Sounds lasting only a few milliseconds can evoke already a strong polar-angle perception [4, 10]. If sounds last longer, listeners can also use dynamic localization cues introduced by head rotations of 30° azimuth or wider in order to estimate the polar angle of the source [16]. However, if high frequencies are available and the source spectrum is reasonably smooth, spectral cues dominate polar-angle perception [9]. In the present study, we explicitly focus on monaural spectral localization cues and thus, consider the most strict condition of static broadband sounds and non-moving listeners.

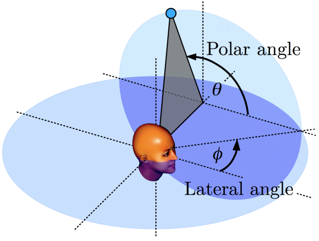


Fig. 1. Interaural-polar coordinate system with lateral angle, $\phi \in [-90°, 90°]$, and polar angle, $\theta \in [-90°, 270°)$.

For a simple arrangement with two loudspeakers in the median plane, Fig. 2 illustrates the spectral mismatch between the HRTF of a targeted phantom-source direction and the corresponding spectrum obtained by VBAP. The

[*] To whom correspondence should be addressed: robert.baumgartner@oeaw.ac.at

loudspeakers were placed at polar angles of 0° and 40°, respectively, and the targeted direction was centered between the real sources, i. e., at 20°. The actual HRTF for 20° is shown for comparison. In this example, the spectrum obtained by VBAP is clearly different than the actual HRTF. Since HRTFs vary substantially across listeners [23], the spectral mismatch also varies from case to case. Psychoacoustic localization experiments with similar loudspeaker arrangements showed that amplitude panning in the median plane works reasonably well for some unspecified listeners, but the derivation of a generic amplitude-panning law like VBAP that is adequate for all listeners seems impossible [19].

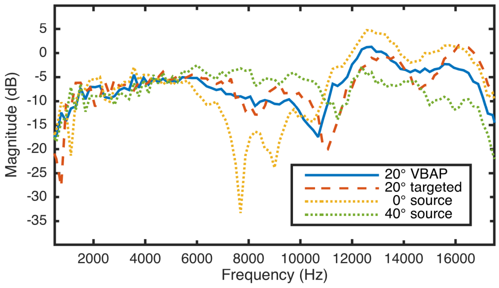


Fig. 2. Example showing the spectral discrepancies obtained by VBAP. The targeted spectrum is the HRTF for 20° polar angle. The spectrum obtained by VBAP is the superposition of two HRTFs from directions 40° polar angle apart of each other with the targeted source direction centered in between.

In this study, we aim at more systematic and objective investigation of the limits and effects of VBAP in sagittal planes. To this end, we applied an extensively evaluated model of sagittal-plane sound localization for human listeners. The model-based investigation is subdivided in two parts. In Sec. 2, we considered an arrangement with two loudspeakers placed in the median plane where binaural disparities are negligible and VBAP basically affects monaural spectral cues. With this reduced setup, localization performance of phantom sources was investigated systematically as a function of panning angle and loudspeaker span. In Sec. 3, we simulated arrangements for surround sound systems consisting of various numbers of loudspeakers in the upper hemisphere and evaluated their spatial quality in terms of localization accuracy.

1 GENERAL METHODS

The sagittal-plane localization model aims at predicting the polar-angle response probability for a given target sound. Figure 3 shows the template-based model structure. In stage 1, the filtering of the incoming sound by the torso, head and pinna is represented by directional transfer functions (DTFs), that is, HRTFs with the direction-independent part removed [8]. Then, the spectral analysis of the cochlea is approximated by a Gammatone filterbank in stage 2. It results in a spectral magnitude profile with center frequencies ranging from 0.7 to 18 kHz. In stage 3, the positive spectral gradients of the profile are considered as monaural spectral cues and are compared with equivalently processed direction-specific templates of cues in stage 4. The outcome of stage 4 is an internal distance metric as a function of the polar angle. In stage 5, these distances are mapped to similarity indices that are proportional to the predicted probability of the listener's polar-angle response. The shape of the mapping curve is determined by a listener-specific sensitivity parameter, which represents the listener-specific localization performance to a large degree [14]. In stage 6, monaural spatial information is combined binaurally whereby a binaural weighting function accounts for a preferred contribution of the ipsilateral ear [13]. After stage 7 emulates the response scatter induced by sensorimotor mapping (SMM), the combined similarity indices are normalizated to a probability mass vector (PMV) in stage 8. The PMV provides all information necessary to calculate commonly used measures of localization performance. The implementation of the model is incorporated in the Auditory Modeling Toolbox as the `baumgartner2014` model [21]. Also, the simulations of the present study are provided in the AMT.

The predictive power of the model was evaluated under several HRTF modifications and variations of the source spectrum [1]. For that evaluation, the SMM stage was important to mimic the listeners' localization responses in a psychoacoustic task using a manual pointing device. Model-based investigations can benefit from the possibility to remove the usually considerably large task-induced scatter (17° in [1]), which might hide some perceptual details. Hence, for studying the effect of amplitude panning, we did not use the SMM stage, because we aimed at modeling the listeners' perceptual accuracy without any task-induced degradation. Predictions were performed for the same 23 normal-hearing listeners (14 female, 9 male, 19-46 years old) whose data were used also for the model evaluation [1]. Targets were stationary sounds with a white frequency spectrum.

Localization accuracy of phantom sources was evaluated by means of the root mean square (RMS) of local (i.e., localized within the correct hemisphere) polar response errors, in the following called *polar error*. Note that the polar error measures both localization blur and bias.

Amplitude panning was applied according to the VBAP principle [18], briefly described as follows. Loudspeakers at neighboring directions, defined by Cartesian-coordinate vectors $\mathbf{l}_i$ of unity length, were combined to triplets, $\mathbf{L} = [\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3]$. In order to create a phantom source in the direction of the unity-length vector $\mathbf{p}$, the amplitudes of the coherent loudspeaker signals were weighted by $\mathbf{g} = \mathbf{p}^T\mathbf{L}^{-1}/\|\mathbf{p}^T\mathbf{L}^{-1}\|$. In case of a two-loudspeaker setup, $\mathbf{g}$ and $\mathbf{L}$ have only two rows. We call the polar-angle component of $\mathbf{p}$ the *panning angle* and the amplitude ratio ΔL between two loudspeakers the *panning ratio*.

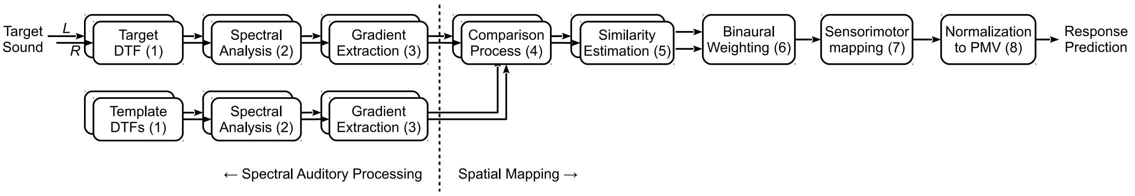


Fig. 3. Structure of the sagittal-plane localization model used for simulations. Reprinted with permission from [1]. © Acoustical Society of America.

2 PANNING BETWEEN TWO LOUDSPEAKERS IN THE MEDIAN PLANE

2.1 Effect of panning angle

Two loudspeakers were simulated in the median plane and in front of the listener, one at a polar angle of $-15°$ and the other one at 30°, essentially reproducing the setup used in experiment I from [19]. For this setup, we simulated localization experiments for all listeners and panning angles in steps of 5°. Figure 4 shows the response predictions for two specific listeners and predictions pooled across all listeners. Each column in a panel shows the color-encoded PMV predicted for a target sound. In general, response probabilities were largest for panning angles close to a loudspeaker direction. Panning angles far from loudspeaker directions were less congruent with large response probabilities. Those angles seemed to evoke sounds localized quite likely at the loudspeaker directions or at the back of the listener. Front-back reversals were often present in the case of listener NH62. Predictions for other listeners, like NH71, were more consistent with the VBAP principle and indicate some correlation between the panning angle and probabilities of localization responses.

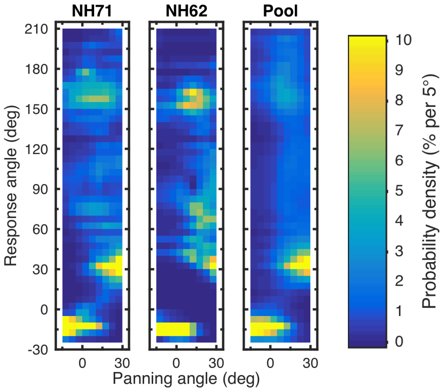


Fig. 4. Response predictions to sounds created by VBAP with two loudspeakers in the median plane positioned at polar angles of $-15°$ and 30°, respectively. Predictions for two exemplary listeners and pooled across all listeners. Each column of a panel shows the predicted PMV of polar-angle responses to a certain sound. Note the inter-individual differences and the generally small probabilities at response angles not occupied by the loudspeakers.

The effect of VBAP on localization performance seems to be indeed very listener-specific in this experiment. We analyzed the across-listener variability in polar error based on the same simulations. For this analysis, each polar error was subtracted from the listener's error predicted for a real source at the direction of the panning angle, because performance is different across listeners and directions in general. Figure 5 shows how each listener's performance degrades as a function of the panning angle. Listener-specific increases in polar error ranged up to 50°, that is, even more than the angular span between the loudspeakers, while the variability across listeners is considerably large with up to 40° of increase in polar error.

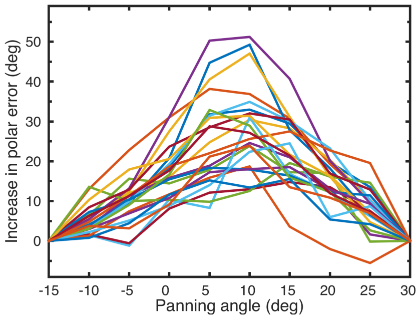


Fig. 5. Listener-specific increases in polar error as a function of the panning angle. Increase in polar error defined as difference between polar error obtained by VBAP source and polar error obtained by real source at corresponding panning angle. Same loudspeaker arrangement as for Fig. 4. Note the large inter-individual differences and the increase in polar error being largest at panning angles centered between the loudspeakers, i.e., at panning ratios around $\Delta L = 0$.

The large variability across listeners is consistent with the listeners' experiences reported in [19]. We thus aimed to reproduce the psychoacoustic results from [19]. In that experiment, Pulkki used the same loudspeaker arrangement as above with two additional loudspeakers at polar angles of 0° and 15° considered as reference sources. The listeners were asked to adjust the panning between the outer two loudspeakers (at $-15°$ and 30°) such that they hear a phantom source that coincides best with a reference source. Since all loudspeakers were visible, this most probably focused the spatial attention of the listeners to the range between $-15°$ and 30° polar angle. Hence, we restricted the range of response PMVs to the same range.

Response PMVs for various panning angles were interpolated to obtain a 1°-sampling of panning angles.

The model simulates localization experiments where listeners respond to a sound by pointing to the perceived direction. It is not clear how these directional responses are related to the adjustments performed in [19]. We considered two possible response strategies of the listeners in the adjustment task. According to the first strategy, called probability maximization (PM), listeners focused their attention to a very narrow angular range and adjusted the panning so that they would most likely respond in this range. For simulating the PM strategy, the panning angle with the largest response probability at the targeted direction was selected as the adjusted angle. The second strategy, called centroid matching (CM), was inspired by a listener's experience described in [19], namely, that "he could adjust the direction of the *center of gravity* of the virtual source"(p. 758). For modeling the CM strategy, we selected the panning angle that yielded a centroid of localization responses closest to the reference source direction.

We predicted the adjusted panning angles according to the two different strategies for our pool of listeners and for both reference sources. We also retrieved the panning angles obtained by [19] from his Fig. 8. Figure 6 shows the descriptive statistics of panning angles from [19] together with our simulation results. Pulkki observed a mean panning angle that was about 5° too high for the reference source at 0°, but significantly larger and quite close to the reference source at 15°. The across-listener variability was at least 20° of panning angle for both reference sources. This considerably large variability was obtained although two of 16 listeners were removed in [19], as they reported to perceive the sounds inside their head. For the PM strategy, medians and interquartile ranges were mostly similar to the experimental results from [19], including the 5°-offset for reference source at 0°. However, the marginals (whiskers) and the upper quartile for the reference source at 15° were spread too widely. Predictions based on the CM strategy yielded a similar median for the reference source at 0°, a median too small for the reference source at 15°, and a smaller inter-individual spread in general. Despite all those differences to the results from [19], both simulations confirmed in a single-sided paired-sample t-test that adjusted panning angles significantly increased with increasing angle of the reference source ($p < .001$).

In order to quantify the goodness of fit (GOF) between the actual and simulated results, we applied the following analysis. First, we estimated a parametric representation of the actual results from [19]. To this end, we calculated the sample mean $\hat{\mu}$ and the square root of the unbiased estimate of the sample variance $\hat{\sigma}$ for the two reference sources. Then, we quantified the GOF between the actual and simulated results by means of p-values from one-sample Kolmogorov-Smirnov tests [15] performed with respect to the estimated normal distributions.

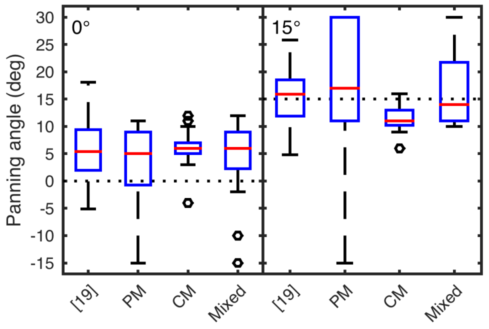


Fig. 6. Panning angles for loudspeaker arrangement of Fig. 4 judged best for reference sources at polar angles of 0° or 15° in the median plane. Comparison between experimental results from [19] and simulated results based on various response strategies: PM, CM, and both mixed – see text for descriptions. Dotted horizontal line: polar angle of the reference source. Red horizontal line: median; blue bar: inter-quartile range (IQR); whisker: within quartile ±1.5∗IQR; black circle: outlier. Note that the simulations predicted an offest similar to the results from [19] for the reference source at 0°.

Table 1 lists the estimated means and standard deviations for the two different source angles together with the GOF for each simulation. The GOF was also evaluated for the results from [19] to show that the estimated normal distributions adequately represent these results. For the reference source at 15°, the $p < 0.01$ indicates that our predictions represent those from [19] significantly different. For the reference source at 0°, the predictions cannot be statistically rejected, but they were still poorly representing the actual data from [19].

			Goodness of fit			
Ref.	$\hat{\mu}$	$\hat{\sigma}$	[19]	PM	CM	Mixed
0°	6.0°	5.1°	.77	.15	.14	.85
15°	15.6°	4.8°	.70	$< .01$	$< .01$	.24

Table 1. Means, standard deviations, and GOF estimated for the two reference sources (Ref.). Goodness of fit for results from [19] and simulations of various response strategies. Note the relatively large GOFs for the simulations based on mixed response strategies indicating a reasonable correspondence between actual and predicted results.

Thus, we attempted to better represent the pool of listeners from [19] by assuming that listeners chose one of the two strategies and used it for both reference sources. To this end, we created a mixed strategy pool by assigning one of the two strategies to each of the listeners individually so that the sum of the two GOFs is maximum. This mixing procedure assigned 17 listeners to the PM strategy and six listeners to the CM strategy. Both GOFs for the mixed case were larger than 0.24 and, thus, indicated a moderate and good correspondence between the simulated and actual results for the reference source at 0° and 15°, respectively. Again consistent with [19], a single-sided paired-sample t-test confirmed that simulated panning angles significantly increase with increasing angle of the reference source ($p < .001$).

In summary, simulating individual response strategies of listeners was required to explain the across-listener variability observed in [19]. Furthermore, the model was able to explain the 5° offset in median panning angle for a reference source at 0°. Hence, this offset seems to be caused by a general acoustic property provided by human HRTFs.

2.2 Effect of loudspeaker span

The loudspeaker arrangement in the previous experiment yielded quite large polar errors, especially for panning angles around the center between the loudspeakers. Reducing the span between the loudspeakers in the sagittal plane, is expected to improve the localization accuracy of phantom sources. For analyzing the effect of loudspeaker span, the two loudspeakers in the median plane were simulated with equal amplitude ratio ($\Delta L = 0$) in order to obtain a worst-case estimate. We systematically varied the loudspeaker span within 90° in steps of 10°. For each listener and span, we averaged the predicted polar errors across target polar angles ranging from $-25°$ to $205°$. Figure 7 shows the predicted increase in average polar errors as a function of the loudspeaker span. With increasing span the predicted errors increased consistently accompanied by a huge increase in inter-listener variability. This effect is consistent with the findings from [2], who showed that the closer two sources are positioned in the median plane the stronger the weighted average of the two positions is related to the perceived location.

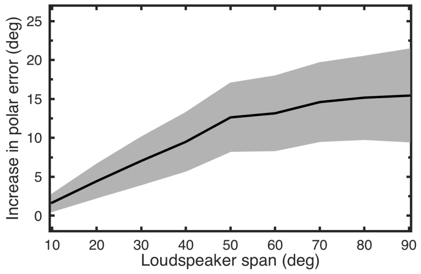


Fig. 7. Increase in polar error as a function of loudspeaker span in the median plane with amplitude ratio $\Delta L = 0$. Increase in polar defined as for Fig. 5; mean ±1 *SD* across listeners. Note the systematic increase in polar error with increasing loudspeaker span.

In order to directly compare our results with those from [2], we evaluated the correlation between panning angles and predicted response angles for the panning ratios tested in [2]: $\Delta L \in \{-13, -8, -3, 2, 7\}$ dB. For each listener and panning ratio, predicted response angles were obtained by first applying the model to predict the response PMV and then generating 100 response angles by a random process following this PMV. Since listeners in [2] were tested only in the frontal part of the median plane, we evaluated the correlations for frontal target directions, while restricting the response range accordingly. For further comparison, we additionally analyzed predictions for rear as well as front and rear targets, restricting the response ranges accordingly. Figure 8 shows the predicted results together with those replotted from [2]. For all angular target ranges, the predicted coefficients decrease monotonically with increasing loudspeaker span. The results for the front show a strong quantitative correspondence with those from [2]. Compared to the front, panning angles at overall and rear directions correlate less well with localization responses up to loudspeaker spans of 70° and 90°, respectively. For the overall target range, our simulations show that for loudspeaker spans up to 40°, the VBAP principle can explain at least 50% of the localization variance in a linear regression model.

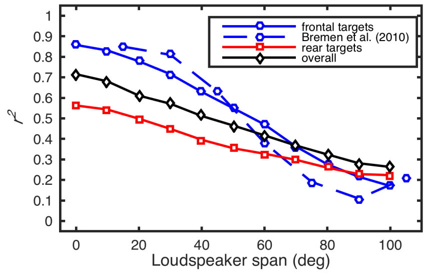


Fig. 8. Effect of loudspeaker span in the median plane on determination of phantom source direction by VBAP principle, analyzed separately for frontal, rear, and overall (frontal and rear) targets. Data pooled across listeners. Note the correspondence with the results obtained by [2].

3 PANNING WITHIN SURROUND SOUND SYSTEMS

For surround sound systems including elevated loudspeakers, plenty recommendations of specific loudspeaker arrangements exist. As shown in Sec. 2.2, the loudspeaker span strongly affects the localization accuracy of elevated phantom sources. Thus, for a given number of loudspeakers, one can optimize the localization accuracy of source positions either for all possible directions or for preferred areas. In this section, we analyzed the spatial distribution of predicted localization accuracy for some exemplary setups.

3.1 Selected systems

We selected an exemplary set of six recommendations for loudspeaker arrangements in the following denoted as systems $A - F$, which are sorted by decreasing number of incorporated loudspeakers (Tables 2 and 3). The loudspeaker directions of all systems are organized in layers with constant elevation. In addition to the horizontal layer at the ear level, system A has two elevated layers at 45° and 90° elevation, systems $B - D$ have one elevated layer at 45°, system E has elevated layers at 30° and 90° elevation, and system F has one elevated layer at 30°.

System A represents the 22.2 Multichannel Sound System developed by the NHK Science & Technical Research Laboratories in Tokio [3], in our present study investigated without the bottom layer consisting of three loudspeakers below ear level. Systems B and C represent the 11.2 and 10.2 Vertical Surround System (VSS), respectively, devel-

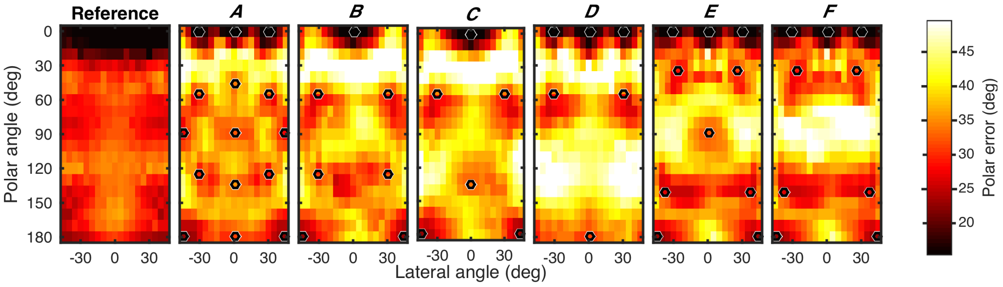


Fig. 9. Predicted polar error as a function of the lateral and polar angle of a phantom source created by VBAP in various surround sound systems. System specifications are listed in Table 2. Open circles indicate loudspeaker directions. Reference shows polar error predicted for a real source placed at the phantom source directions investigated for systems $A - F$.

oped by Samsung [7]. System *D* represents a 10.2 surround sound system developed by the Integrated Media Systems Center at the University of Southern California (USC) [6]. Systems *E* and *F* represent the Auro-3D 10.1, and 9.1 listening format, respectively [22].

Ele.	Azi.	Pol.	Lat.	*A*	*B*	*C*	*D*	*E*	*F*
0°	0°	0°	0°	•	•	•	•	•	•
	30°	0°	30°	••			••	••	••
	60°	0°	60°	••	••	••	••		
	90°	0°	90°	••	••	••	••		
	135°	180°	45°	••	••	••		••	••
	180°	180°	0°	•			•		
30°	30°	34°	26°					••	••
	135°	141°	38°					••	••
45°	0°	45°	0°	•					
	45°	55°	30°	••	••	••	••		
	90°	90°	45°	••					
	135°	125°	30°	••	••				
	180°	135°	0°	•		•			
90°	0°	90°	0°	•				•	

Table 2. Loudspeaker directions of considered surround sound systems. Dots indicate occupied directions. Double dots indicate that corresponding directions to the right hand side (negative azimuth and lateral angle) are occupied as well. Ele.: elevation; Azi.: azimuth; Pol.: polar angle; Lat.: lateral angle.

3.2 Methods and results

Following the standard VBAP approach [18], the number of active loudspeakers depends on the desired direction of the phantom source and may vary between one, two or three speakers according to whether the desired source direction coincides with a loudspeaker direction, is located directly between two loudspeakers, or lies within a triplet of loudspeakers, respectively. Since sagittal-plane localization is most important for sources near the median plane, we investigated only phantom sources within the range of ±45° lateral angle.

We predicted polar errors as a function of the lateral and polar angle of a targeted phantom source direction for the different arrangements and a reference system containing loudspeakers at all considered directions. Figure 9 shows the across-listener averages of the predicted polar errors. The simulation of the reference system shows that, in general, listeners perceive the location of sources in the front most accurately. In the various surround sound systems, polar errors appeared to be smaller at directions close to loudspeakers (open circles), a relationship already observed in Sec. 2.1. Consequently, one would expect that the overall polar error increases with decreasing number of loudspeakers, but this relationship does not completely apply to all cases. System *A* with the largest number of loudspeakers resulted in a quite regular spatial coverage of accurately localized phantom source directions. Systems $B - D$ covered generally less directions. Systems *B* and *C* showed only minor differences in the upper rear hemisphere, where system *D* yielded strong diffuseness. For systems *E* and *F*, which have a lower elevated layer and thus a smaller span to the horizontal layer, the model predicted very accurate localization of phantom sources targeted in the frontal region. Hence, positioning the elevated layer at 30° seems to be a good choice when synthesized auditory scenes are focused to the front, which might be frequent especially in the context of multimedia presentations. Note that 30° elevation at 30° azimuth corresponds to a polar angle of about 34°, whereas 45° elevation at 45° azimuth corresponds to a polar angle of about 55°, that is, a span for which larger errors are expected – see Sec. 2.2.

Table 3 summarizes the predicted degradation in localization accuracy in terms of the increase of errors relative to the reference and averaged across listeners. We distinguish between the mean degradation, Δe_{mean}, as indicator for the general system performance, and the maximum degradation, Δe_{max}, across directions as an estimate of the worst-case performance. The predicted degradations confirm our previous observations, namely, that systems with less loudspeakers and higher elevated layers yield phantom sources that appear to provide less localization accuracy. Due to the lower elevation of the second layer, systems *E* and *F* seem to provide the best trade-offs between number of loudspeakers and localization accuracy.

Our results seem to be consistent with directional quality evaluations from [7]. In that study, the overall spatial quality of system *A* was rated best, no quality differences between system *B* and *C* were reported, and system *D* was rated worse. Systems *E* and *F* were not tested in this study.

System	N	Ele.	Δe_{mean}	Δe_{max}
A	19	45°	6.8°	28.8°
B	11	45°	8.9°	38.6°
C	10	45°	10.0°	38.6°
D	10	45°	11.4°	31.3°
E	10	30°	6.4°	29.3°
F	9	30°	7.3°	29.3°

Table 3. Predicted increase in polar errors as referred to reference and averaged across listeners. Distinction between mean and maximum degradation across directions. N: Number of loudspeakers. Ele.: Elevation of second layer. Notice that the elevation of the second layer seems to have larger effect on Δe_{mean} and Δe_{max} than N.

4 CONCLUSIONS

A localization model was used to simulate the effect of VBAP on the localization accuracy in sagittal planes. In VBAP, monaural spectral localization cues encoding different source directions are superimposed with a frequency-independent weighting. This may lead to conflicting encoded information on the phantom source direction. Simulations of an arrangement with two loudspeakers in the median plane showed pronounced inter-individual variability caused by the differences in the listeners' HRTFs. Predicted localization accuracy was quite good for some listeners, but poor for many others. Hence, there is minor evidence for a generic panning law that is adequate for all listeners. For loudspeaker spans of up to 40° polar angle, however, listener-specificities are, statistically seen, small enough to render the VBAP principle suitable for sound spatialization. The loudspeaker span was also striking in the assessment of various surround sound systems. In our simulations, systems with layers at 30° elevation provided more accurate representations of phantom sources than those with layers at 45° elevation. In the future investigations, the localization model can easily be applied to other loudspeaker arrangements or sound spatialization techniques.

5 ACKNOWLEDGEMNT

This study was supported by the Austrian Science Fund (FWF, P 24124).

6 REFERENCES

[1] R. Baumgartner, P. Majdak, and B. Laback. Modeling sound-source localization in sagittal planes for human listeners. *The Journal of the Acoustical Society of America*, 136(2):791–802, 2014.

[2] P. Bremen, M. M. van Wanrooij, and A. J. van Opstal. Pinna cues determine orientation response modes to synchronous sounds in elevation. *J Neurosci*, 30:194–204, 2010.

[3] K. Hamasaki, K. Hiyama, and R. Okumura. The 22.2 multichannel sound system and its application. In *Audio Engineering Society Convention 118*, May 2005.

[4] P. M. Hofman and A. J. V. Opstal. Spectro-temporal factors in two-dimensional human sound localization. *J Acoust Soc Am*, 103:2634–2648, 1998.

[5] P. M. Hofman, J. G. A. van Riswick, and A. J. van Opstal. Relearning sound localization with new ears. *Nature Neurosci*, 1:417–421, 1998.

[6] T. Holman. *Surround Sound: Up and Running*. Focal Press, 2008.

[7] S. Kim, Y. W. Lee, and V. Pulkki. New 10.2-channel Vertical Surround System (10.2-VSS); Comparison study of perceived audio quality in various multichannel sound systems with height loudspeakers. In *Proceedings of the 129th Convention of the Audio Engineering Society*, page Convention Paper 8296, 2006.

[8] D. J. Kistler and F. L. Wightman. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *J. Acoust. Soc. Am.*, 91 (3):1637–1647, 1992.

[9] E. A. Macpherson. Cue weighting and vestibular mediation of temporal dynamics in sound localization via head rotation. In *Proceedings of meetings of acoustics*, volume 133, pages 3459–3459. Acoust Soc Am, 2013.

[10] E. A. Macpherson and J. C. Middlebrooks. Localization of brief sounds: Effects of level and background noise. *J Acoust Soc Am*, 108:1834–1849, 2000.

[11] E. A. Macpherson and J. C. Middlebrooks. Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *J Acoust Soc Am*, 111:2219–2236, 2002.

[12] E. A. Macpherson and J. C. Middlebrooks. Vertical-plane sound localization probed with ripple-spectrum noise. *J Acoust Soc Am*, 114:430–445, 2003.

[13] E. A. Macpherson and A. T. Sabin. Binaural weighting of monaural spectral cues for sound localization. *J Acoust Soc Am*, 121:3677–3688, 2007.

[14] P. Majdak, R. Baumgartner, and B. Laback. Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization. *Front Psychol*, 5(319):pages not available yet, doi:10.3389/fpsyg.2014.00319, 2014.

[15] F. J. Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[16] K. I. McAnally and R. L. Martin. Sound localization with head movement: implications for 3-d audio displays. *Frontiers in neuroscience*, 8, 2014.

[17] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen. Head-related transfer functions of human subjects. *J Audio Eng Soc*, 43:300–321, 1995.

[18] V. Pulkki. Virtual sound source positioning using vector base amplitude panning. *J Audio Eng Soc*, 45:456–466, 1997.

[19] V. Pulkki and M. Karjalainen. Localization of amplitude-panned virtual sources I: stereophonic panning. *J. Audio Eng. Soc.*, 49(9):739–752, 2001.

[20] V. Pulkki, M. Karjalainen, and J. Huopaniemi. Analyzing virtual sound source attributes using a binaural auditory model. *J Audio Eng Soc*, 47(4):203–217, 1999.

[21] P. Søndergaard and P. Majdak. The auditory-modeling toolbox. In J. Blauert, editor, *The Technology of Binaural Listening*, chapter 2. Springer, Berlin–Heidelberg–New York NY, accepted for publication, 2013.

[22] W. van Baelen. Challenges for spatial audio formats in the near future. In *26. VDT International Audio Convention, Leipzig*, pages 196–205, 2010.

[23] B.-s. Xie. *Head-related transfer function and virtual auditory display*. J Ross, 2013.

THE AUTHORS

Robert Baumgartner

Piotr Majdak

Robert Baumgartner is research assistant at the Acoustics Research Institute (ARI) of the Austrian Academy of Sciences (OeAW) in Vienna and a PhD candidate at the University of Music and Performing Arts Graz (KUG). In 2012, he received his MSc degree in Electrical and Audio Engineering, a inter-university program at the University of Technology Graz (TUG) and the KUG. His master thesis about modeling sound localization in sagittal planes with the application to subband-encoded HRTFs was honored by the German Acoustic Association (DEGA) with a student award. His research interests include spatial hearing and spatial audio. Robert is a member of the DEGA, a member of the Association for Research in Otolaryngology (ARO), and the secretary of the Austrian section of the AES.

●

Piotr Majdak, born in Opole, Poland, in 1974, studied Electrical and Audio Engineering at the TUG and KUG, and received his MSc degree in 2002. Since 2002, he has been working at the ARI, and is involved in projects including binaural signal processing, localization of sounds, and cochlear implants. In 2008, he received his PhD degree on lateralization of sounds based on interaural time differences in cochlear-implant listeners. He is member of the Acoustical Society of America, the ARO, the Austrian Acoustic Association, and the president of the Austrian section of the AES.

Chapter 7

Concluding remarks

In this PhD project, a model for sound localization in sagittal planes was proposed. The model compares monaural spectral cues with an internal template and, to this end, requires the individual HRTFs and sensitivities to predict listeners' individual localization performance. If listeners' HRTFs and sensitivities were not available, a representative subpopulation of 23 listeners was used to predict average data. Predictions of the model were successfully evaluated against listener-specific performance in psychoacoustic experiments testing:

- baseline performance as a function of lateral eccentricity (Pearson's correlation coefficients in the range of $r > .72$),
- the effect of HRTF modifications induced by various spectral resolutions ($r = .73$) or band limitation in combination with spectral warping ($r > .81$), and
- the effect of different source spectra such as spectrally rippled noise bursts or speech syllables (r not applicable because only average data predicted).

These evaluations were focused on predictions of localization performance averaged across many target directions. Response predictions for specific target directions were evaluated only exemplarily, because the fact that listeners were forced to respond to each target, regardless of whether they really perceived the source as being located somewhere in space or not, likely distorts this evaluation. A further evaluation of the target-specific model predictions would require localization experiments where listeners are instructed to indicate targets that are hard or impossible to localize.

Mechanisms of sound localization were investigated on the basis of model predictions. The most important findings are summarized in the following.

1. Sound localization in sagittal planes appears to be cued by positive spectral gradients rather than spectral notches and/or peaks. This gradient extraction was crucial to adequately model listeners' robustness in localization performance against macroscopic variations of the source spectrum.

2. Large differences in listener-specific localization performance seems to be predominantly caused by individual differences in auditory processing efficiency rather than acoustic circumstances. In a linear regression model, a non-acoustic factor, representing the listener's sensitivity to differentiate spectral cues, explained much more variance of listener-specific localization performance than the acoustic factor of listener-specific HRTFs.

3. Spectral cues provided by the contralateral ear appear to provide similar spatial uniqueness as those provided by the ipsilateral ear, since unilateral localization performance was predicted similar for both ears. However, due to the head shadow, diffuse background noise masks contralateral cues more than ipsilateral cues. Hence, the generally larger ipsilateral weighting of spectral cues, independent of the actual presence of any background noise, seems to be a neural processing strategy optimized for noisy environments.

The model served as a useful tool to evaluate spatial audio systems and hearing-assistive devices. It confirmed that behind-the-ear casing in hearing-assistive devices severely degrades localization performance and that binaural recordings provide best localization accuracy if recorded in someone's own ears. Model predictions were also essential in evaluating algorithms for subband approximation of HRTFs and for determining an appropriate approximation tolerance. Subsequent psychoacoustic experiments confirmed the validity of the model predictions. Moreover, the model was applied to show that inter-individual differences in HRTFs cause a strong listener-specific effect of vector-based amplitude panning in sagittal planes. On average, the localization accuracy degrades with increasing polar-angle distance between a targeted phantom source direction and the closest loudspeaker direction. As a consequence, polar-angle distances of about 40° between loudspeaker directions turned out to be a good trade-off between playback-system complexity and obtained localization accuracy.

The applicability of the present model is limited in several respects. Hence, depending on the requirements of future applications, the model complexity should be increased by the following aspects. First, in the present model, cochlear processing is approximated by a linear Gammatone filterbank. Hence, the model cannot represent the well-known interaction between intensity and duration of the sound affecting localization performance (Hartmann and Rakerd, 1993; Hofman and Opstal, 1998; Macpherson and Middlebrooks, 2000; Vliegen and Opstal, 2004). A non-linear model of the auditory periphery, like the dual-resonance non-linear filterbank (Lopez-Poveda and Meddis, 2001) or the model from Zilany et al. (2014), needs to be incorporated in order to address this interaction. The model from Zilany et al. (2014) would additionally provide the possibility to study hearing impairment in terms of dysfunctions of inner and/or outer hair cells and a loss of auditory nerve fibers with specific spontaneous firing rates. Second, the present model does not consider temporal variations. The bandpass signals obtained from the cochlear filterbank are temporally integrated across the whole duration of the stimulus. Thus, the effect of large group delays, for instance, occurring for slowly frequency-modulated tones (Hofman and Opstal, 1998; Hartmann et al., 2010), are disregarded. The duration of temporal integration windows (Hofman and Opstal, 1998) and time constants of lateral spectral inhibition (Hartmann et al., 2010), presently being a part of the spectral gradient extraction, in the range of about 5 ms were discussed as potential reasons for this effect. Third, the model considers both the listener and the source as non-moving objects. A time-dependent model framework as outlined above would be required to represent moving sources. Head rotations cause lateral dynamics and change interaural localization cues. Combined with proprioceptive and vestibular feedback, these cues can provide useful information especially for front-back discrimination (McAnally and Martin, 2014; Macpherson, 2013). Hence, a motion-sensitive model of sagittal-plane localization requires an adequate processing of interaural cues. To this end, a near choice would be to combine the present model with an existing model for horizontal-plane localization (Dietz et al., 2011; Takanen et al., 2014). Finally, there are many more influencing factors not yet addressed by the model. The position of the eyes, for instance, systematically modulates sound localization (Lewald and Getzmann, 2006). In summary, it is still a long way to accurately represent realistic environments with multiple auditory objects in reverberant acoustic environments, but the modular structure of the present model should provide a good starting point for developing a more general localization model addressing those various aspects. Future accessibility and reproducibility of the present model investigations is guaranteed by providing the corresponding implementations in the Auditory Modeling Toolbox (Søndergaard and Majdak, 2013).

References

Agterberg, M. J. H., Hol, M. K. S., Van Wanrooij, M. M., Van Opstal, A. J., and Snik, A. F. M. (2014). Single-sided deafness and directional hearing: contribution of spectral cues and high-frequency hearing loss in the hearing ear. *Auditory Cognitive Neuroscience*, 8:188.

Carlile, S. (2014). The plastic ear and perceptual relearning in auditory spatial perception. *Frontiers in Neuroscience*, 8:237.

Dietz, M., Ewert, S. D., and Hohmann, V. (2011). Auditory model based direction estimation of concurrent speakers from binaural signals. *Speech Communication*, 53:592–605.

Hartmann, W. M., Best, V., Leung, J., and Carlile, S. (2010). Phase effects on the perceived elevation of complex tones. *The Journal of the Acoustical Society of America*, 127:3060–3072.

Hartmann, W. M. and Rakerd, B. (1993). Auditory spectral discrimination and the localization of clicks in the sagittal plane. *The Journal of the Acoustical Society of America*, 94:2083–2092.

Hofman, P. M. and Opstal, A. J. V. (1998). Spectro-temporal factors in two-dimensional human sound localization. *The Journal of the Acoustical Society of America*, 103:2634–2648.

Hofman, P. M., van Riswick, J. G. A., and van Opstal, A. J. (1998). Relearning sound localization with new ears. *Nature Neuroscience*, 1:417–421.

Ihlefeld, A. and Shinn-Cunningham, B. G. (2011). Effect of source spectrum on sound localization in an everyday reverberant room. *The Journal of the Acoustical Society of America*, 130:324–333.

Kumpik, D. P., Kacelnik, O., and King, A. J. (2010). Adaptive reweighting of auditory localization cues in response to chronic unilateral earplugging in humans. *The Journal of Neuroscience*, 30:4883–4894.

Langendijk, E. H. A. and Bronkhorst, A. W. (2002). Contribution of spectral cues to human sound localization. *The Journal of the Acoustical Society of America*, 112:1583–1596.

Lewald, J. and Getzmann, S. (2006). Horizontal and vertical effects of eye-position on sound localization. *Hearing Research*, 213:99–106.

Lopez-Poveda, E. A. and Meddis, R. (2001). A human nonlinear cochlear filterbank. *The Journal of the Acoustical Society of America*, 110:3107–3118.

Lord Rayleigh or Strutt, F. R. (1907). On our perception of sound direction. *Philosophical Magazine Series 6*, 13:214–232.

Macpherson, E. A. (2013). Cue weighting and vestibular mediation of temporal dynamics in sound localization via head rotation. In *Proceedings of Meetings of Acoustics*, volume 133, pages 3459–3459. Acoustical Society of America.

Macpherson, E. A. and Middlebrooks, J. C. (2000). Localization of brief sounds: Effects of level and background noise. *The Journal of the Acoustical Society of America*, 108:1834–1849.

Macpherson, E. A. and Middlebrooks, J. C. (2002). Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *The Journal of the Acoustical Society of America*, 111:2219–2236.

Macpherson, E. A. and Middlebrooks, J. C. (2003). Vertical-plane sound localization probed with ripple-spectrum noise. *The Journal of the Acoustical Society of America*, 114:430–445.

May, B. J., Anderson, M., and Roos, M. (2008). The role of broadband inhibition in the rate representation of spectral cues for sound localization in the inferior colliculus. *Hearing Research*, 238:77–93.

McAnally, K. I. and Martin, R. L. (2014). Sound localization with head movement: implications for 3-d audio displays. *Auditory Cognitive Neuroscience*, 8:210.

Møller, H., Sørensen, M. F., Hammershøi, D., and Jensen, C. B. (1995). Head-related transfer functions of human subjects. *Journal of the Audio Engineering Society*, 43:300–321.

Søndergaard, P. and Majdak, P. (2013). The auditory modeling toolbox. In Blauert, J., editor, *The Technology of Binaural Listening*, chapter 2. Springer, Berlin–Heidelberg–New York NY.

Takanen, M., Santala, O., and Pulkki, V. (2014). Visualization of functional count-comparison-based binaural auditory model output. *Hearing Research*, 309:147–163.

Van Wanrooij, M. M. and Van Opstal, A. J. (2007). Sound localization under perturbed binaural hearing. *Journal of Neurophysiology*, 97:715–26.

Vliegen, J. and Opstal, A. J. V. (2004). The influence of duration and level on human sound localization. *The Journal of the Acoustical Society of America*, 115:1705–1713.

Zilany, M. S. A., Bruce, I. C., and Carney, L. H. (2014). Updated parameters and expanded simulation options for a model of the auditory periphery. *The Journal of the Acoustical Society of America*, 135:283–286.