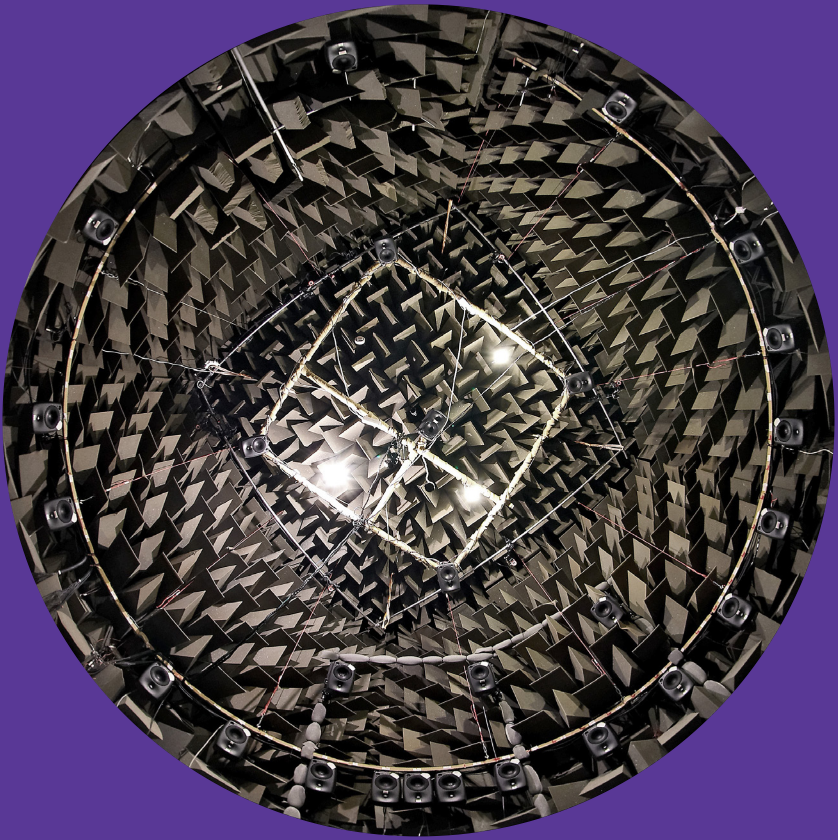
Department of Signal Processing and Acoustics

Functional Modeling of Hearing for Assessment of Spatial Sound Reproduction

Marko Takanen



Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 109/2014

Functional Modeling of Hearing for Assessment of Spatial Sound Reproduction

Marko Takanen

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Electrical Engineering, at a public examination held at the lecture hall S1 of the school on September 5, 2014, at noon.

Aalto University
School of Electrical Engineering
Department of Signal Processing and Acoustics

Supervising professor
Professor Ville Pulkki

Thesis advisor
Professor Ville Pulkki

Preliminary examiners
Dr. Mathias Dietz, Carl von Ossietzky Universität, Germany
Dr. Russell Mason, University of Surrey, U.K

Opponent
Emeritus Professor Ray Meddis, University of Essex, U.K.

Aalto University publication series
DOCTORAL DISSERTATIONS 109/2014

ISBN 978-952-60-5782-8
ISBN 978-952-60-5783-5 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5783-5

Unigrafia Oy
Helsinki 2014

Finland

NORDIC ECOLABEL
441 697
Printed matter

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Marko Takanen

Name of the doctoral dissertation
Functional Modeling of Hearing for Assessment of Spatial Sound Reproduction

Publisher School of Electrical Engineering

Unit Department of Signal Processing and Acoustics

Series Aalto University publication series DOCTORAL DISSERTATIONS 109/2014

Field of research Acoustics and audio signal processing

Manuscript submitted 13 March 2014 **Date of the defence** 5 September 2014

Permission to publish granted (date) 13 May 2014 **Language** English

□ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

Auditory modeling refers to the design of computational models of the human auditory system using digital signal processing algorithms. Such models can potentially be utilized in various applications including development of hearing aids and cochlear implants as well as to explain psychoacoustical phenomena. Another practical application area is the evaluation of sound reproduction, in which the models provide an interesting alternative to the direct use of human subjects in formal listening tests.

This thesis addresses the instrumental evaluation of spatial sound reproduction with a model that emulates the functionality of the auditory pathway based on neurophysiological and psychoacoustical data from the literature. However, the this thesis work also aimed to ensure a more general applicability of the model. The research involved in this work may be divided into two main categories.

The first category consists of developing auditory models and of employing them in the evaluation of sound reproduction. The thesis presents two auditory models with different goals. Namely, one of them is a applicationspecific model designed to evaluate stereophonic sound reproduction capability of small mobile devices. The other demonstrates how several psychoacoustical binaural hearing phenomena may be explained with a more detailed emulation of processing in the auditory pathway. The latter model was also applied to evaluate sound reproduction achieved with both traditional and parametric spatial sound techniques.

The second category focuses on the acquisition of psychoacoustical knowledge. This category provides more insight into how the auditory system analyzes complex auditory scenarios. In addition, this category presents a listening test assessing different binaural synthesis methods in terms of coloration aspects.

Keywords binaural hearing, auditory modeling, sound quality, spatial sound

ISBN (printed) 978-952-60-5782-8 **ISBN (pdf)** 978-952-60-5783-5

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2014

Pages 189 **urn** http://urn.fi/URN:ISBN:978-952-60-5783-5

Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tekijä
Marko Takanen

Väitöskirjan nimi
Kuulon toiminnallisuuden mallintaminen tilaäänentoiston arvioimiseksi

Julkaisija Sähkötekniikan korkeakoulu

Yksikkö Signaalinkäsittelyn ja akustiikan laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 109/2014

Tutkimusala Akustiikka ja äänenkäsittelytekniikka

Käsikirjoituksen pvm 13.03.2014 **Väitöspäivä** 05.09.2014

Julkaisuluvan myöntämispäivä 13.05.2014 **Kieli** Englanti

□ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Digitaaliseen signaalinkäsittelyyn pohjautuvia kuulojärjestelmän laskennallisia malleja kutsutaan auditorisiksi malleiksi, joilla on monia tärkeitä sovelluskohteita mm. kuulolaitteiden ja sisäkorvaistutteiden kehittelyssä sekä psykoakustisten ilmiöiden taustalla olevien prosessien kuvauksessa. Eräs insinööritieteen näkökulmasta tärkeä sovelluskohde koskee äänentoiston laadun arviointia, johon auditoriset mallit voisivat tarjota kiintoisan vaihtoehdon kuuntelukokeiden järjestämiselle.

Tämä väitöskirja käsittelee äänentoiston arviointia auditorisella mallilla, joka simuloi kuulojärjestelmän toiminnallisuutta neurofysiologiaan ja psykoakustiseen tietoon pohjautuen. Lisäksi tavoitteena on taata mallin sovellettavuus myös muihin tarkoituksiin. Tähän taustaan pohjautuen väitöskirjassa esitettävä tutkimus on jaettavissa kahteen osaan.

Ensimmäinen osa koskee auditoristen mallien kehittämistä ja niiden soveltamista äänentoiston laadun arviointiin. Työssä esitellään kaksi auditorista mallia, joista toinen osoittaa kuinka mobiililaitteiden äänentoistoa voidaan arvioida hieman yksinkertaisemmallakin mallilla. Vastaavasti toinen osoittaa kuinka kuulojärjestelmän tarkemmalla kuvauksella kyetään selittämään useita psykoakustisia ilmiöitä. Viimeksi mainittua mallia sovellettin myös sekä perinteisillä tekniikoilla että parametrisillä tilaäänen prosessointi menetelmillä saavutettavan äänentoiston laadun arviointiin.

Väitöskirjan toinen osa keskittyy mallinnuksessa tarvittavan psykoakustisen tiedon hankkimiseen. Tarkemmin ottaen tässä osassa esitellään psykoakustiseen tutkimukseen pohjautuvaa uutta tietoa kuulojärjestelmän tavasta jäsentää monimutkaista äänimaisemaa. Saatua tietoa voidaan soveltaa myös auditoristen mallin kehittämisessä. Lisäksi tämä osa esittelee tutkimuksen, jossa kuuntelukokeen avulla tutkittiin binauraalisten synteesimenetelmien laadullisia ominaisuuksia.

Avainsanat Kuulo, mallintaminen, äänenlaatu ja tilaäänentoisto

ISBN (painettu) 978-952-60-5782-8 **ISBN (pdf)** 978-952-60-5783-5

ISSN-L 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2014

Sivumäärä 189 **urn** http://urn.fi/URN:ISBN:978-952-60-5783-5

Preface

This thesis work was carried out at the Department of Signal Processing and Acoustics at Aalto University School of Electrical Engineering in Espoo, Finland. The work was funded by the Academy of Finland (projects 121252 and 13251770) and Nokia corporation. The thesis work also received support from the Walter Ahlström Foundation and the Finnish Foundation of Technology Promotion.

I wish to extend my utmost gratitude to the two professors who have supervised and instructed me during this thesis work. I started my thesis work under the supervision of Prof. Matti Karjalainen, who passed away in May 2010, and I am grateful to him for inspiring an inexperienced PhD student to try out and learn new things. From 2010, I pursued my research under the supervision of Prof. Ville Pulkki, who's intuitive visions and our countless discussions have continued to raise interesting research questions and ideas for future work. Ville's enthusiasm, in-depth knowledge, and support made this thesis possible. Furthermore, I am grateful to the pre-examiners of this thesis, Dr. Mathias Dietz and Dr. Russell Mason, for their helpful suggestions and comments on the manuscript.

I am also indebted to the co-authors of the papers included in this thesis. Foremost, I wish to thank Olli Santala for the close collaboration during these years, the present work would not have been possible with out it. My special thanks goes to Dr. Gaëtan Lorho who encouraged me to pursue a doctoral degree in the first place as well as instructed me in the beginning of this journey. I am thankful to Dr. Marko Hiipakka for the brainstorming sessions and for demonstrating how efficiently things can be done when there is will to do so. I am also grateful to Hagen Wierstorf and Prof. Alexander Raake for enabling my researcher exchange at the TU Berlin and for the chance to do research with them. I also thank Tuomo Raitio and Prof. Paavo Alku for their contributions and for sharing their expertise on speech processing. In this context, I also wish to thank PhD Nelli Salminen for the indispensable information about neurophysiology and for her enthusiastic attitude towards interdisciplinary research.

I feel privileged that I have been able to work at the acoustics lab and to be part of its great community. For this, I wish to thank my colleagues: Dr. Hannu Pulakka, Dr. Mikko-Ville Laitinen, Dr. Ville Sivonen, and Tapani Pihlajamäki for various sports-related events. Sami Oksanen and Seppo Fagerlund for the hikes. Dr. Okko Räsänen for the discussions over coffee and lunch. Javier Goméz Bolaños and Ilkka Huhtakallio for facilitating the experiments. Prof. Unto Laine for the inspiring discussions. Lea Söderman, Heidi Koponen, Mirja Lemetyinen, Jouni Hynninen, Markku Hietala, and Prof. Jorma Skyttä for handling the administrative and practical issues. Dr. Jukka Ahonen, Juha Vilkamo, Archontis Politis, Symeon Delikaris-Manias, Teemu Koski, Olli Rummukainen, Julia Turku, PhD Catarina Hiipakka, Alessandro Altoè, and all the other colleagues for creating that great atmosphere. I also want to thank Christine Kluge, Janto Skowronek, Michal Soloducha, Katrin Schoenenberg, Dennis Guse, Stefan Hillmann, Matthias Schulz, and other members of the AIPA team as well as the Berlin Broilers for making my researcher exchange at Berlin an unforgettable experience.

I also want to thank my friends for the countless sport and social events that provided me (mostly) healthy means to get my mind away from work. Finally, I express my overwhelming gratitude to my parents Annikki and Olavi, my brothers Mika and Keijo, and my relatives for their love and support during this seemingly endless journey.

Munich, Germany, July 9, 2014,

Marko Takanen

Contents

List of publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Marko Takanen, Olli Santala, and Ville Pulkki. Visualization of functional count-comparison-based binaural auditory model output. *Hearing Research*, Volume 309, pp. 147–163, March 2014.

II Marko Takanen, Olli Santala, and Ville Pulkki. Binaural Assessment of Parametrically Coded Spatial Audio Signals. *The Technology of Binaural Listening*, J. Blauert (Ed.), Springer-Verlag Berlin Heidelberg, Germany, pp. 333–358, 2013.

III Marko Takanen, Hagen Wierstorf, Ville Pulkki, and Alexander Raake. Evaluation of sound field synthesis techniques with a binaural auditory model. In *AES 55th Intl. Conf.*, pp. 1–8, Helsinki, Finland, August 2014.

IV Marko Takanen and Gaëtan Lorho. A Binaural Auditory Model for the Evaluation of Reproduced Stereophonic Sound. In *AES 45th Intl. Conf.*, pp. 1–10, Paper No. 6-6, Helsinki, Finland, March 2012.

V Marko Takanen, Tuomo Raitio, Olli Santala, Paavo Alku, and Ville Pulkki. Fusion of spatially separated vowel formant cues. *J. Acoust. Soc. Am.*, Volume 134, Issue 6, pp. 4508–4517, December 2013.

VI Marko Takanen, Marko Hiipakka, and Ville Pulkki. Audibility of coloration artifacts in HRTF filter designs. In *AES 45th Intl. Conf.*, pp. 1–9, Paper No. 3-3, Helsinki, Finland, March 2012.

Author's contribution

Publication I: "Visualization of functional count-comparison-based binaural auditory model output"

The present author implemented the model that was jointly designed by all authors of the paper, following the original idea of the third author. The later stages of the model contain several examples, such as the onset contrast enhancement, that show the contribution of the present author to the design process. The development of the design concept into implementation was done in close collaboration between the first two authors, who also jointly wrote the manuscript, receiving feedback from the third author.

Publication II: "Binaural Assessment of Parametrically Coded Spatial Audio Signals"

The study presents a collaborative work of all the three authors. The present author contributed in the research, for instance, by coming up with the idea of evaluating the techniques in off-sweet-spot-listening conditions as well as by executing the simulations reported in the study. In addition, he designed and performed the informal listening of the different scenarios together with the second author. The first two authors also jointly wrote Sections 4 and 5 of the article.

Publication III: "Evaluation of sound field synthesis techniques with a binaural auditory model"

The study presents a result of a cooperative research. The first two authors simulated the different binaural listening scenarios. The present author performed the computational evaluations as well as compared the model outputs to the perceptual data gathered by the second author. The abstract and Sections 3, 4, 5, 6, and 7 were also written by the present author.

Publication IV: "A Binaural Auditory Model for the Evaluation of Reproduced Stereophonic Sound"

The present author implemented the model while the second author performed the binaural recordings that were utilized in the simulations. The initial draft of the article was mainly written by the present author.

Publication V: "Fusion of spatially separated vowel formant cues"

The present author is mainly responsible for this research. He designed the experiment, generated the stimuli with the help of two co-authors, conducted the experiment, analyzed the results, and wrote the initial draft of the paper. The co-authors provided indispensable feedback at every stage of the study.

Publication VI: "Audibility of coloration artifacts in HRTF filter designs"

The present author was responsible for designing and conducting the listening experiment as well as for the statistical analysis of the results. He also generated the stimuli together with the second author. The abstract and Sections 1, 3, and 4 were written by the present author while Sections 5, and 6 were written jointly by the first two authors.

List of abbreviations

3-D	three-dimensional
ACR	absolute category rating
ASW	apparent source width
BILD	binaural intelligibility level difference
BRIR	binaural room impulse response
CF	characteristic frequency
CN	cochlear nucleus
CV	consensus vocabulary
DirAC	directional audio coding
ERB	equivalent rectangular bandwidth
F0	fundamental frequency
FIR	finite impulse response
GTFB	gammatone filter bank
HARPEX	high angular resolution planewave expansion
HRTF	head-related transfer function
HpTF	headphone transfer function
IACC	interaural cross-correlation
IC	inferior colliculus
ILD	interaural level difference
ITD	interaural time difference
ITU-R	Radiocommunication Sector of the International Telecommunication Union
ITU-T	Telecommunication Standardization Sector of the International Telecommunication Union
IPD	interaural phase difference
IV	individual vocabulary
LEV	listener envelopment
LSO	lateral superior olive

MAA	minimum audible angle
MOS	mean opinion score
MSO	medial superior olive
MUSHRA	multi stimulus test with hidden reference and anchor
PEAQ	perceptual evaluation of audio quality
QESTRAL	quality evaluation of spatial transmission and reproduction using an artificial listener
SC	superior colliculus
VBAP	vector base amplitude panning
WFS	wave field synthesis

1. Introduction

Computational auditory models take digitized signals as input and process them with a series of signal processing algorithms in order to explain different input-output relationships or to emulate the processing in the human auditory pathway. Furthermore, these models may focus on mimicking the functionality of a single organ or nucleus in the pathway or simulating the whole processing chain in order to explain human perception. Successful development of such models can increase the knowledge about the auditory system and the underlying mechanisms facilitating our remarkable hearing ability. This provides an evident motivation for scientific research that has inspired numerous researchers over the past decades.

In addition, several potential application areas exists for auditory models. For instance, these models can be used in the development of hearing aids and cochlear implants that can improve or even restore the hearing ability of a person suffering from conductive or sensorineural hearing impairments. Alternatively, auditory models may be used in conjunction with room acoustical models to estimate in advance how material selection, room dimensions, and other design aspects affect, e.g., the speech intelligibility in a class room. Properties of these models are also applied in audio codecs and parametric audio coding techniques that reduce the transmission data rate while aiming for a perceptually transparent reproduction of the original signal or the original sound scene. Another interesting application area for auditory models consists of evaluation of product sound quality and spatial sound reproduction achieved with a given technique. Moreover, auditory models can provide an interesting instrumental tool for developers of such techniques as such a tool could be used to evaluate in advance how a modification of a specific parameter affects the quality of the sound reproduction.

Current auditory models are able to emulate the processing in great detail. Unfortunately, none of them can simulate a complete set of spatial hearing tasks [1]. This also limits the general applicability of the models. Consequently, this thesis work aims to develop a binaural auditory model that fulfills two requirements: (1) The model should emulate the functionality of the nuclei in the auditory pathway in such detail that allows it to account for human performance in several binaural listening scenarios. (2) The model should be able to assess the performance of spatial sound reproduction techniques.

These two requirements were set to ensure the applicability of the model also in tasks other than in the one addressed in this thesis work. Successful development of such a model comprises multidisciplinary research that combines elements from neurophysiology, psychoacoustics, computational modeling, audio reproduction, and perceptual assessment of sound quality.

- Neurophysiological data provides valuable information about the functionality of the nuclei in the auditory pathway.

- Psychoacoustical studies reveal information about the capabilities and limitations of the human auditory system in auditory scene analysis.

- Signal-driven auditory models are then designed based on the obtained neurophysiological and psychoacoustical data.

- Applicability of the model for the evaluation of spatial sound reproduction needs to be verified by comparing the model outputs to the data acquired from perceptual assessments.

This thesis work collects the primary contributions of the present author to construct a model fulfilling the above-mentioned requirements. In addition to the collection, this thesis contains a literature review covering the basics of the different disciplines in order to make the thesis more readable for professionals of anyone of the disciplines.

2. Hearing

This section describes briefly how the sound emitted by a sound source evokes perception of an auditory image having a specific location and identity. Moreover, the acoustical transfer function up to the listener's ears and the subsequent processing in the auditory pathway are overviewed in separate subsections. In addition, this section reviews the psychacoustical knowledge about the ability of the auditory system to analyze the ear canal signals, e.g. in order to localize individual sound sources or to extract attributes describing the perception. The perception of speech is reviewed in a separate section since speech has a special role in human communication. It should be noted that the presented overviews are limited to the aspects pertaining to the work presented in this thesis.

2.1 Acoustical path from the sound source to the listener's ears

The sound emitted by a sound source in an environment propagates as a sound wave and is received by the ears of the listener. The sound wave may also reflect multiple times on the different surfaces in the environment before reaching the listener. The external ear (consisting of the torso, the head, the pinna, the concha, and the ear canal) of the listener also affects the signal received at the eardrum, and the effect depends on the direction from which the sound arrives at the listener. The directional dependency of the effect results in cues that the auditory system can use to localize the sound source [2, 3].

Moreover, the arrival times of the signals at the two ears are different if the sound wave approaches the listener from the side, and the difference in the arrival times is denoted as the interaural time difference (ITD). In such scenarios, the signal received at the contralateral side is also attenuated due to wave-propagation around the head, and the difference in the levels of the signals at the two ears is denoted as the interaural level difference (ILD). In addition, the direction of arrival has an effect on the manner the sound waves reflect on the torso and the pinna. The pinna flange also attenuates sounds from behind the listener. These direction-dependent characteristics of the acoustical transfer function can be stored in the head-related transfer functions (HRTFs) and binaural room impulse responses (BRIRs) that characterize the propagation path from a point source to the eardrums of the listener in free-field and non-anechoic conditions, respectively [4].

2.2 Auditory pathway

The small variations in the sound pressure at the eardrum result in vibration of the eardrum, and the ossicles (the malleus, incus, and stapes) located in the middle ear transmit the vibration into the fluid inside the cochlea[1]. Thus, the middle ear implements an efficient transmission of sound energy from the low-impedance medium (air) to the much higher-impedance medium (fluid). From a functional point of view, the cochlea acts as a frequency analyzer transforming the mechanical movements into neural impulses. The vibration of the stapes against the oval window generates pressure waves in the fluid inside the cochlea, and as these waves travel inside the cochlea, the basilar membrane and the tectorial membrane start to move vertically and horizontally, respectively [6]. Consequently, the cilia of the inner hair cells bend evoking neural signals that traverse via the auditory nerve to the cochlear nucleus (CN) located in the brainstem [7].

The neural signals from the auditory nerve are then processed in the CN whose various cell types send different kinds of responses to different targets in the auditory pathway [8]. Figure 2.1 illustrates how the ventral cochlear nuclei of the two hemispheres project into the medial superior olives (MSOs) and the lateral superior olives (LSOs) located in the superior olivary complex [9, 10]. On the other hand, the dorsal cochlear nucleus projects directly into the inferior colliculus (IC) [11, 12, 13].

[1] The cochlea is essentially a curved tube that is divided throughout its length into three separate fluid-filled chambers by the Reissner's membrane, the tectorial membrane, and the basilar membrane. On top of the basilar membrane lies the organ of Corti that has two types of receptors (inner and outer hair cells) that are connected to the auditory nerve fibers by their roots and to the tectorial membrane by their fine cilia [5].

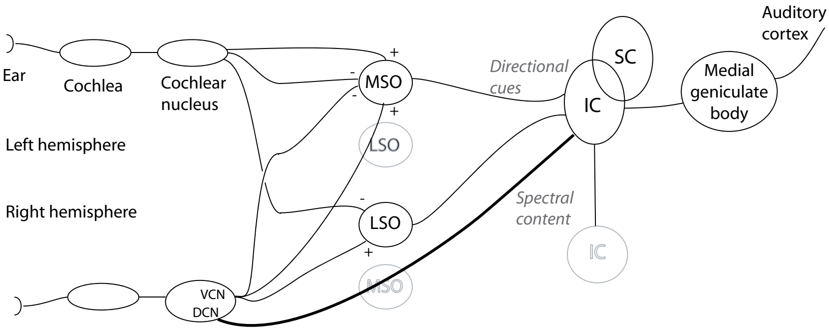


Figure 2.1. Schematic presentation of the mammalian auditory pathway.

The MSO and LSO contribute significantly to the localization and spatial hearing ability of the auditory system as they are sensitive to the binaural cues in the ear canal signals [14]. The main inputs to the MSO consist of the excitation and inhibition arriving from the CNs of the two hemispheres [10], but there is also some evidence for MSO neurons receiving inputs from axons of other MSO neurons [15, 16, 17]. The MSO neurons are sensitive to the ITD [18]. Specifically, the neurons are sensitive to the interaural phase difference (IPD) in such a manner that the neurons sharing the same characteristic frequency (CF) provide their maximum output with an IPD of $\pi/4$ at low frequencies [18].

In each hemisphere, the LSO neurons receive excitation and inhibition from the ipsilateral and contralateral CNs, respectively [9]. The LSO neurons are mostly sensitive to the ILD [19], and they have been found to be capable of acting as fast phase-locked subtractors that can respond to sudden changes in the input signals, having integration times of as low as 2 ms [20]. Such a low integration time may explain why LSO neurons are sensitive also to the fine-structure ITD with low-frequency stimuli [20, 21] and to envelope ITDs in the case of amplitude-modulated sounds [22].

In both hemispheres, the IC then receives the outputs of the CN, the MSO, and the LSO. However, the exact role of the IC is yet somewhat unknown despite the numerous response measurements of the IC neurons [23]. It is known that the IC transmits the spatial information to the auditory cortex and the superior colliculus (SC) and that the information may be modified in the process [23]. The SC, located next to the IC, has multiple layers. For example, layers for visual information as well as layers for sound have been found in the SC [24, 25]. Interestingly, the SC is involved in cross-modal interaction and also in steering the focus of attention towards the stimuli [26, 27]. To facilitate such functionality, the SC includes neurons that react to multimodal stimulation originating from the same spatial location [24, 28], and there are also cells in the SC containing a topographic map of the auditory space that is aligned with the visual map [24, 25].

2.3 Frequency resolution and perception of loudness

The human auditory system can distinguish sound pressure differences in the frequency range covering frequencies approximately from 20 Hz to 20 kHz [29]. Hence, the auditory system covers about ten octaves from the sound spectrum, but its resolution depends on the frequency of the sound. The auditory system dissects the ear canal signal into narrowband components, and the spectral components within each of such critical bands are processed together (see, e.g., [30, 31]). Moreover, the bandwidth of such a critical band increases as the frequency increases [30, 31]. At least two factors of the processing within the cochlea contribute to the frequency resolution: (1) The maximum oscillation of the basilar membrane occurs at different positions depending on the frequency, because the mass, stiffness, and width of the basilar membrane vary along its length [6]. (2) The outer hair cells have been found to implement a dynamic compression and suppression of the sidebands [32] and to pump energy in vibration patterns of the basilar membrane at low stimulus intensities [33].

The so called notched-noise method [34] has proven to be an accurate method to determine the bandwidths of the critical bands. In this method, the subject is presented with a stimulus consisting of a pure tone and wide-band noise having a notch in the spectrum around the frequency of the pure tone. The levels of the pure tone and the masking noise are kept constant and the masked detection threshold of the pure tone is then measured by varying the width of the notch. Thereafter, the width of the notch at the detection threshold can be used to define the width of critical band as an equivalent rectangular bandwidth (ERB)

$$\mathrm{ERB}(f_{\mathrm{c}}) = 24.7(4.37 f_{\mathrm{c}} + 1), \tag{2.1}$$

where f_{c} denotes the center frequency in kHz [35].

An alternative method to estimate the bandwidth of a critical band comprises evaluating the perceived loudness of a stimulus consisting of two narrowband sounds with different spectral content. If the two sounds in the stimulus fall within the same critical band, the stimulus is perceived as softer than when the sounds lie in separate critical bands [36]. Hence, the width of the critical band can be evaluated by varying the spectral content of the sounds. The critical bands estimated with the loudness method have been found to be wider than the ones estimated with the notched-noise method [37].

Loudness perception of band-limited noise provides an example of the critical-band-based analysis. Specifically, the perceived overall loudness increases when the bandwidth of the noise is increased to extend over a larger portion of the critical band scale, although the overall level is kept constant [36]. Hence, the overall loudness perception is thought to be formed by integrating the specific loudness values, describing the loudness per critical band [36]. Furthermore, the specific loudness spectrum is very useful for describing spectral aspects of the stimulus (see Sec. 5.2.2). It should be noted that the levels of both ear canal signals affect the binaural loudness perception [38]. Moreover, loudness matching experiments employing band-limited or wide-band noise stimuli have demonstrated that the perceived loudness of such a stimulus depends on the direction from which the stimulus is presented and that such dependencies can be explained with the differences in the HRTFs [39, 40]. The results of these experiments also bolster the idea that the overall loudness follows the 3-dB rule, according to which the perceived loudness can be estimated by summing the powers of the ear canal signals.

2.4 Spatial sound perception

This section discusses the capabilities and limitations of the human spatial hearing. Results of such experiments have thereafter been employed in the design of auditory models of the human auditory pathway.

2.4.1 Acuity of spatial hearing

As mentioned above, the differences in the path of the sound from the sound source to the two ears result in differences between the ear canal signals, and these binaural cues enable the auditory system to localize the sound. In normal environments, all binaural cues (ITD, ILD and envelope ITD) aid the auditory system in the localization task as they provide consistent directional information when a single plane wave arrives at the ears of the listener. However, the individual roles of these binaural cues in localization may be addressed in listening experiments using headphone reproduction. Such experiments have shown that the perceived lateral position of the auditory image can be shifted away from the center by modifying only one of the binaural cues while the values of the other cues is zero [3, 41, 42]. It has also been demonstrated that the ITD dominates the lateralization for broadband and low-pass filtered stimuli when the cues have conflicting non-zero values [43]. On the other hand, the lateralization of high-pass filtered stimuli has been found to be dominated by the ILD [44]. Furthermore, listeners have been able to localize broadband sounds also based on envelope ITDs despite conflicting waveform ITDs [45].

In anechoic conditions, a sound emitted by a point-like sound source evokes the perception of a narrow auditory image [4], while the localization accuracy depends on the direction of the sound source as well as on the type and the duration of the sound [46, 47, 48]. Furthermore, the measured localization accuracy is also influenced by the method employed in the experiment, i.e., measured errors depend on the manner the listeners are required to indicate the perceived direction [49]. Partly due to the last-mentioned dependency, localization accuracy has often been measured as the minimum audible angle (MAA), which describes the minimal angular shift from the original direction that the listener can detect. The MAA resolution is approximately $\pm 1^\circ$ in front and decreases gradually to approximately $\pm 10^\circ$ when the sound is moved to the side on the horizontal plane [50].

In more complex sound scenarios, the task of the auditory system becomes more challenging as the sounds reaching the ears of the listener actually consist of an ensemble of independent signals emitted by multiple sound sources. Such a scenario may also be considered to consist of a target sound and a distracter(s) hindering the listener in the localization of a particular sound from the ensemble. It may also be that there is only one actual sound source in the environment, but the multiple reflections of the sound result in a more challenging localization task. The latter example has been actively studied in experiments of the precedence effect [51], and it has been demonstrated that the listeners are able to localize the sound accurately based on the direction of the direct sound (for a review, see [52]). Multiple factors have an effect on the extent the distracter(s) decrease the localization performance of the listeners. Such factors include, for instance, the number of distracters, the signal types, the frequency content, the signal-to-noise ratio, and the onset and offset times of the target and the distracter(s) [53, 54, 55, 56, 57, 52, 58].

The ability of the listeners to judge the perceived width of the ensemble in such complex environments has also been studied using independent noise bursts. The length of these noise bursts has an effect on whether the ensemble is perceived as point-like or wide [59]. Such an ensemble is perceived as slightly narrower than the actual loudspeaker span, and the center area is perceived less clearly than the ends of the distributed ensemble [60].

2.4.2 Auditory scene analysis and spatial attributes

The human auditory system is thought to form a separate auditory stream for each sound object while analyzing the surrounding auditory scene [61]. In this process, the sounds reaching the ears of the listener are grouped based on several physical cues including spectral relationship [62], common history (i.e., onset and offset times) [63, 64], spatial location [65], and good continuity [66].

The relative impacts of these grouping cues on the auditory scene analysis have also been addressed in previous studies. Considering the scope of this thesis, the most pertaining studies are the ones in which the sound event is split into two or more components that are thereafter presented from different directions around the listener. In such scenarios, the separated components evoke different binaural cues while the other physical cues suggest that these components should be fused together in auditory scene analysis. The ecologically invalid spatial separation does not prevent the auditory system from grouping the components together. The subjects have been able to detect the original stimulus despite the spatial separation between the to-be-fused components in such conditions [67, 68]. Furthermore, the perceived direction of the fused auditory event has been found to fall between the directions from which the separated components were presented [58, 69]. Interestingly, it has been found that the spatial separation may either evoke the perception of an additional sound event [68] or cause the disappearance of a component from the scene if the attention of the subject is directed elsewhere [70]. The latter possible outcome may be interpreted as if the spatial separation prevents some of the otherwise audible sounds from reaching conscious perception [71]. Such an outcome also highlights the impact of focused attention on the perception of the auditory scene [72], which is also influenced by the visual information [73, 74, 75] and head-movements.

The surrounding auditory scene may also be described using spatial attributes such as locations of the individual sound events and the overall spatial impression [76, 77]. The latter attribute may be further divided into the apparent source width (ASW) and the listener envelopment (LEV) attributes [78]. ASW describes the perceived spatial extent of a given sound event, and LEV is related to the spatial impression of the given space itself [79]. The perception of these attributes is related to the reflections of the sound in the space. More precisely, the early lateral reflections contribute to the perception of the ASW, whereas the late reverberation is no longer associated with the direct sound and is consequently associated with the listener envelopment [78]. It should be noted that the spatial impression is also affected by the type(s) of the sound event(s) [80], the properties of the emitted sounds [81], and the acoustical properties of the space and the listening position [82].

2.5 Perception of speech

Humans, like many other animals, communicate with fellow creatures by producing and perceiving sound that carries information. What distinguishes humans from other species is that our communication is based on language, which has evolved to define the meanings of different sound combinations. That is to say, we are able to communicate using spoken language. Humans can also use written language for communication, but due to the fast and interactive nature of speech, the majority of the communication between people takes place using spoken language.

Due to the great importance of speech in human communication, several studies have addressed the perception of speech in different kind of scenarios. The remainder of this section explains the characteristics of speech sounds related to the production of speech as well as gives a brief overview on the studies where the perception of speech has been studied either as a fusion of separate components or in complex sound scenarios containing multiple concurrent sound sources.

2.5.1 Characteristics of speech

Most of the voices in human communication are produced by altering and obstructing the exhaled airflow from the lungs in different parts of the vocal tract. This is accomplished by moving the active articulator (tongue or lower lip) towards the passive articulator (upper lip, teeth, alveolar ridge, hard palate, soft palate, uvula or pharynx) [83]. The movement changes the shape of the vocal tract and consequently alters the resonances of the vocal tract, which are called formants.

The voices in human communication can be divided into two main categories: vowels and consonants. The former are reproduced with an unrestricted airflow in the vocal tract, whereas in the production of consonants, the articulators create a constriction in the vocal tract that obstructs the airflow from the mouth either completely or partially [83]. Consonants can be further categorized based on the location of the constriction in the vocal tract. Furthermore, consonants can also be either voiced or voiceless, whereas all the vowels are voiced. In the case of voiced sound, the oscillation of the vocal folds in the larynx creates a periodic structure in the sound, and consequently the voiced sounds have a harmonic structure with peaks at integer multiplies of the fundamental frequency. In the production of voiceless sounds, the vocal folds do not oscillate, and the glottis remains open letting the air flow directly through the larynx [83].

The manner a given phoneme[2] is articulated in continuous speech depends on the adjacent phonemes in that (or the adjacent) word. The reason for this dependency is that the active articulators cannot jump from one position to another, but they have to move between the positions using continuous trajectories. Therefore, the articulators are in constant movement during the production of speech, as the articulators already start to move to the articulation position of the following phoneme during the production of the current phoneme.

2.5.2 Speech as a fusion of signal components

The auditory system seems to have a tendency to presume a signal with any speech-like character to be speech [84]. For instance, three sinusoidal tones of equal amplitude positioned at the three first formants of a vowel are sufficient for correct identification of the vowel [85], and a three-tone replica of speech can be interpreted as speech [86]. The ability of the auditory system to identify speech does not even require that the different components of speech are presented from the same spatial location.

[2] A phoneme is the smallest linguistically distinctive unit of speech [83].

For instance, the subjects of the experiments in [67, 87] reported hearing only the original speech stimulus despite the fact that the stimulus had been divided into two components, one containing low and the other high frequencies, that were simultaneously presented to different ears of the listener over headphones. However, it was found that the fundamental frequencies (F0s) of the two components must be identical, otherwise the original speech stimulus is not correctly identified [87].

Sometimes a component can contribute to the identification of speech as well as be simultaneously perceived as an additional sound event. For instance, in the experiment by Rand [68], the subjects reported hearing the correct utterance /da/ in one ear and a secondary non-speech sound in the other, when the two components of the speech stimulus were presented simultaneously to the different ears of the listener over headphones. Specifically, one of the components was the formant transition in the beginning of the utterance, and the other was the remaining signal, called the base of the utterance. The contribution of the formant transition in the identification of the utterance is supported by the fact that the base of the utterance was by itself not sufficient for the identification of the utterance [88]. The occurrence of such a duplex perception [89] has been found to depend on the stimulus onset asynchrony between the transition and the base of the utterance [90], the amount of masking noise in the stimulus [90], and the level difference between the transition and the base of the utterance [68, 88].

2.5.3 Perception of speech in complex environments

The astonishing ability of humans to segregate speech in multi-talker situations has been studied actively over the past decades. Both the localization of the different speakers and the identification of the sentences spoken by the different speakers have been addressed in these studies of the so-called cocktail party effect [91]. Traditionally, these studies have employed utterances of meaningless one-syllable words or non-semantic sentences.

In the intelligibility studies, the task of the subject has been to report the utterance he or she heard. The percentage of correctly identified utterances the gives a measure for the intelligibility of speech in the given sound scenario. A pioneering study showing how the intelligibility of speech depends on the locations of the target speech and the masker was conducted by Licklider [92]. He used white noise as the masker and presented the stimuli over headphones to the ears of the listener in a manner that the target and the masker were presented either diotically or dichotically. He reported that a binaural intelligibility level difference (BILD) of approximately 3–3.5 dB is achieved when either the target or the masker is presented with a phase difference of π as compared to the scenario where the both stimuli are in-phase at the two ears [92].

The BILD has been found to depend on the type and the amount of maskers, and on whether the stimuli are presented over headphones or with loudspeakers (for a review, see [4]). For instance, Carhart *et al.* [54] reported that a BILD of 9 dB can be achieved when two competing speech signals are used as maskers and presented either diotically or dichotically when the target speech is presented diotically. Interestingly, identification of speech in multi-talker situations seems to be possible even when the auditory system cannot segregate speech sources based on F0. Moreover, listeners have been found to be able to identify two simultaneously presented whispered vowels with about the same accuracy as simultaneously presented vowels sharing a common F0 [93].

Perhaps due to the communicational significance of speech, the localization of speech in cocktail-party situations has not been addressed as extensively as the intelligibility. Nevertheless, it has been found that listeners are able to localize the target speech with a remarkable accuracy. For instance, the participants of the experiment reported in [94] were able to correctly localize the correctly identified words at least 80% of the time. Similar performances were also reported by Hawley *et al.* [95]. More precisely, they reported an identification accuracy of $\pm 10^\circ$ for the direction of a known target sentence in the presence of one, two or three competing sentence(s) in different direction(s).

3. Spatial sound reproduction

Since the ultimate goal of the auditory modeling work in this thesis is to employ the model in the evaluation of spatial sound reproduction, different spatial sound reproduction methods are briefly described in this section. In addition, an overview of some listening test procedures is given since listening tests are the only reliable method to assess the perceived quality[1] of the spatial sound reproduction.

3.1 Methods for spatial sound reproduction

The common goal in spatial sound reproduction is to provide the listener with the characteristics of a spatial sound scene with or without modifications. This goal may be approached using microphones to capture the spatial characteristics of the sound and processing the obtained signals for loudspeaker or headphone reproduction. Moreover, the traditional approach consists of using one microphone for each loudspeaker, but there are also techniques that aim to extract signals for an arbitrary reproduction method by processing signals recorded at a single position.

3.1.1 Two-channel reproduction

The overview presented in this section describes different methods striving to produce a plausible spatial impression by presenting two-channel signals to the ears of the listener using either loudspeakers or headphones. Different techniques to obtain suitable signals for these reproduction methods are discussed also.

[1] The definition of quality proposed by Lorho [96] is adopted here, and quality is defined as a measure of the distance between the characters of the entity being evaluated and of the target associated with the entity.

Two-channel stereophonic loudspeaker reproduction

The most commonly employed spatial sound reproduction setup consists of two equidistant loudspeakers positioned at directions of $\pm 30^\circ$ in front of the listener, as illustrated in Fig. 3.1(a). In two-channel reproduction, the resulting loudspeaker span of 60° is generally thought to provide the optimal compromise between the contradicting desires to maximize the stereo image width and to facilitate the perception of stable phantom images between the loudspeakers [97]. Moreover, the sound reproduction quality is not greatly sensitive to head movements of the listener.

The loudspeaker signals for a two-channel reproduction may be obtained by mixing individual recordings into a two-channel signal. In this process, the desired spatial positioning of the individual sound event is achieved by means of different panning laws, that is by feeding the corresponding microphone signal with different gains and/or delays to both channels of the resulting loudspeaker input. Alternatively, stereophonic microphone techniques may be employed to record a given auditory environment in such a manner that the microphone signals can be directly used as input to the loudspeaker setup. Several different stereophonic microphone techniques have been presented in the literature including spaced microphone techniques consisting of two identical microphones positioned from about ten centimeters to a few meters apart from each other [98], as illustrated in Fig. 3.2(a), and coincident microphone techniques, such as the Blumlein pair [99], where two directive microphones are positioned at a coincident position but in a manner that their look directions are different, as depicted in Fig. 3.2(b).

Binaural reproduction

Binaural synthesis techniques aim to evoke the desired three-dimensional (3-D) spatial impression by presenting two-channel signals over headphones to the ears of the listener (see Fig. 3.1(b)). However, it is challenging to reproduce an auditory environment in a transparent manner using headphones. As Bauer [101] pointed, signals targeted for two-channel loudspeaker setup cannot be used as such in headphone reproduction since the interaural differences between the ear canal signals would not be the correct ones, which results in the perception of an unnatural stereo image inside the head. An alternative solution is to record the auditory scene with a dummy head [102] having two microphones, one at each ear, as depicted in Fig. 3.2(c). Such a binaural recording opens up the

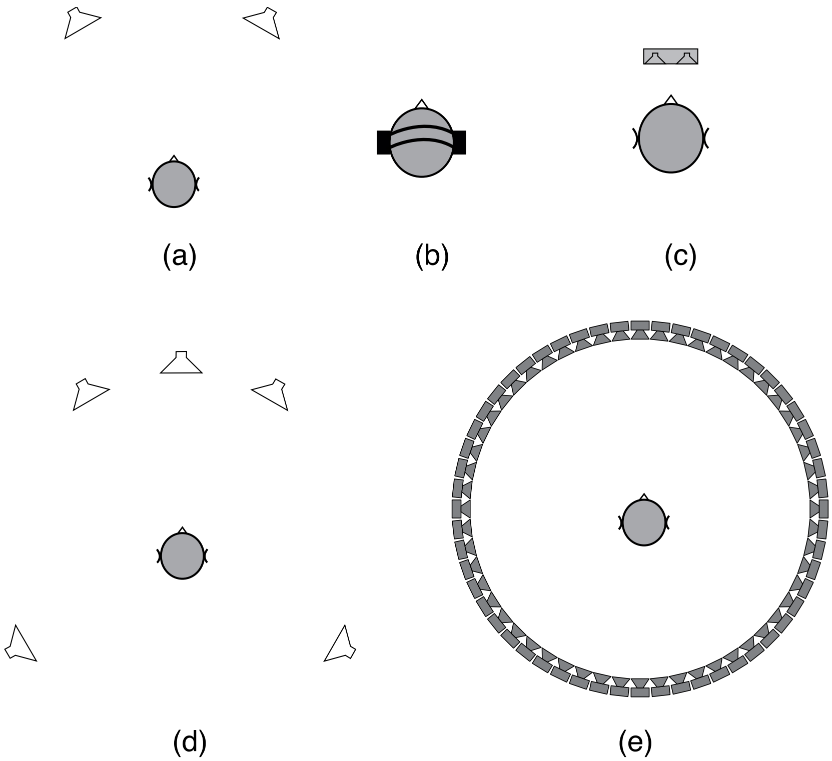


Figure 3.1. Ideal positions of the listener and the loudspeakers when reproducing spatial sound with (a) a two-channel stereophonic loudspeaker setup, (b) headphones, (c) a small portable device, (d) the 5.1 surround system, and (e) the Wave Field Synthesis. Figure 3.1(e) was created with the help of the Sound Field Synthesis toolbox [100].

possibility of presenting the recorded signals as such over headphones once the headphone transfer function (HpTF) has been compensated for [103]. However, binaural recordings have not yet achieved widespread usage, partly due to the need for special recording equipment. In general, dummy heads have also been designed to be used primarily as research tools and not as professional recording instruments. The spatial impression achieved with a reproduction of binaural recordings over headphones is also prone to change due to head movements of the listener.

Therefore, perhaps the best method to produce a perceptually plausible 3-D auditory scene in binaural reproduction is to filter individual monophonic recordings with a set of HRTFs corresponding to the desired spatial positions of the sound events. However, even when individual HRTFs and HpTFs are used in the reproduction, it is still problematic to evoke the illusion of sound sources in the front of the listener (i.e., in the field of view) [104]. The externalization can be improved and the amount of front-

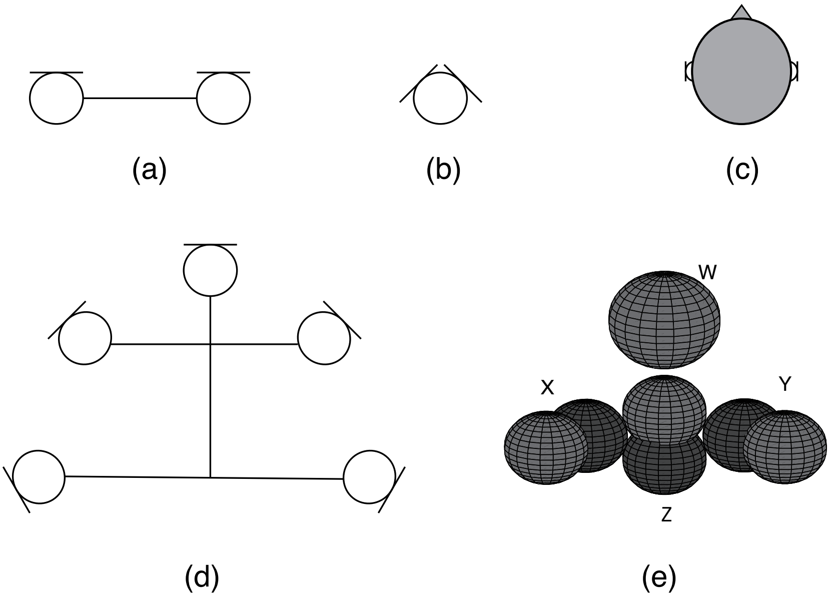


Figure 3.2. Spatial sound recording techniques employing (a) two separated microphones, (b) two directive microphones in a the same place, (c) a dummy head, (d) a five-channel microphone array, and (e) an ideal B-format microphone consisting of an omnidirectional microphone and three orthogonal dipole microphones.

back ambiguities can be reduced when the head-movements of the listener are compensated for using information obtained with head-tracking [105, 106].The use of HRTF filters to provide a more plausible spatial impression may also introduce coloration artifacts in the reproduced sound. Moreover, perceptual studies have demonstrated that listeners may prefer the unprocessed headphone reproduction of stereophonic content over the reproduction employing HRTFs, despite the unnatural stereo image that the former reproduction method creates inside the head [107, 108]. Accurate compensation of HpTFs has proven to be a challenging task [109], which may, at least partially, explain the coloration issues in binaural reproduction employing HRTFs.

Small portable devices

Over the past few years different kinds of mobile phones, tablet computers, and portable gaming devices have become increasingly common among consumers. Although, typically, the sound reproduction with such devices consist of binaural reproduction over headphones, some of these devices also have two small loudspeakers built in them to facilitate stereophonic reproduction. Typically, the small size of such devices does not al-low the loudspeaker layout to reach the optimal spacing of 60° despite the close distance between the loudspeakers and the listener in a typical use scenario (see Fig. 3.1(c)).

So called stereo-enhancement algorithms (for a review, see e.g. [110]) may be used to overcome the barriers of the limited loudspeaker span by creating the illusion of a wider and deeper sound field. Many of these algorithms are based on the cross-talk cancellation algorithm [111], in which the amount of signal transmitted from the left loudspeaker to the right ear is reduced by an interfering signal transmitted from the right loudspeaker and vice versa. However, the interfering signal from the right loudspeaker thereafter needs to be prevented from reaching the left ear by transmitting another interfering signal from the left loudspeaker. As a consequence, a crosstalk-cancelation-based algorithm requires several iterative emissions of interfering signals from the two loudspeakers. Nevertheless, these techniques enable perception of virtual sources also behind the listener, although only within a limited listening area [112].

3.1.2 Multichannel reproduction

In most cases, two-channel loudspeaker reproduction can achieve only a modest spatial impression since sounds are presented only from the front quadrant [97]. The common understanding is that the impression may be enhanced when more loudspeakers are used and they are positioned around the listener. Among the proposed approaches, the most successful by far has been the 5.1 surround system [113]. This section presents an overview of the 5.1 surround system as well as of the wave field synthesis (WFS) technique that aims for authentic replication of the sound field.

Five-channel surround

The 5.1 surround system is based on the recommendation presented by the Radiocommunication Sector of the International Telecommunication Union (ITU-R) [113]. As depicted in Fig. 3.1(d), the recommendation suggests an additional loudspeaker be placed directly in front of the listener in order to improve the stability of phantom images in the front quadrant and two surround loudspeakers be placed at the side of the listener to enhance the reproduction of the ambient characteristics of the given auditory scene. Optionally, a subwoofer may be included to reproduce low-frequency effects [113].

Figure 3.2(d) illustrates a generic layout of a microphone array that can be used to record a given auditory scene for the reproduction with a 5.1 loudspeaker system. The array consists of five directive microphones having specific look-directions. However, the directivity patterns of the microphones and the distances between them in the layout are known to influence the reproduced sound [97], and different kinds of arrays have been presented to cover the aesthetic desires of recording different kinds of acoustical scenes [98]. Alternatively, separate recordings of individual sound events may also be employed in 5.1 reproduction by positioning these sound events around the listener using, e.g., amplitude panning [99]. There are also several upmixing techniques that extract the ambient and direct components from a two-channel stereophonic signal and synthesize the signal for a 5.1 reproduction using the extracted components [114, 115, 116].

The capability of a 5.1 surround system to produce a plausible auditory scene is, however, limited since stable phantom images cannot be created between the front and the surround loudspeakers [97]. The limitation results from the requirement of the recommendation that the 5.1 surround system should be compatible with the two-channel stereophonic content. In addition, the cinema formats of the time when the recommendation was made supported a maximum of six channels [117], and this limitation was recognized already when it was released and so it specified optional loudspeakers be placed between the front and surround speakers. Another limitation of the 5.1 surround system is that the height of the sound events is not considered. This has motivated the development of loudspeaker systems that are backward compatible with the 5.1 surround system as well as capable of reproducing surround sound with height [118, 117]. In such systems, separate recordings of individual sound events may be used in a manner that the sound events are positioned around the listener using, for instance, vector base amplitude panning (VBAP) that generalizes the amplitude panning algorithm for 3-D loudspeaker layouts [119].

Wave field synthesis

WFS [120] is a technique that aims for a perfect reconstruction of the original sound field present in the recording environment. The technique is based on Huygen's principle that states that the original wave at a certain position can be reconstructed by the interference of waves emitted by secondary sources. In spatial sound reproduction, this translates to recording of the original sound field with a dense microphone array and reproducing the captured signals with a matching loudspeaker setup [121]. For instance, linear or spherical microphone and loudspeaker arrays may be used as long as the arrays are equal in shape and size.

If successful, the technique yields a transparent reproduction of the auditory scene within a much larger listening area than what can be achieved with the above-mentioned spatial sound reproduction techniques [122]. WFS can also be used to create focused sources between the loudspeakers and the listener [120]. Currently, the general applicability of WFS is limited mainly due to the requirement of a vast number of loudspeakers and microphones. The spacing between adjacent loudspeakers in the array (see Fig. 3.1(e)) must be smaller than about 3.4 cm in order to reproduce the sound field accurately up to 5 kHz [123]. The same requirement holds for the microphone array, and even denser arrays are needed at higher frequencies. The commonly used approach to circumvent some of these restrictions is to record each sound event with a directive microphone and to take the spatial positioning of the microphones into account when the microphone signals are processed for reproduction with a loudspeaker array [121]. The ambient characteristics are then captured separately with another microphone array [121]. However, the requirements for the loudspeaker array remain the same.

Binaural synthesis provides the means to circumvent the technically challenging and expensive construction of the loudspeaker array required in WFS reproduction. Moreover, HRTFs may be used to emulate the propagation of the sound waves from the loudspeakers to the ears of the listener, and head-tracking can be used to compensate for the head movements of the listener [124, 125]. Although a HRTF database with a very fine angular resolution is necessary in this process, such a database may be obtained by interpolating between HRTFs that have been measured with a lower angular resolution (see, e.g., [126]).

3.1.3 Towards a generic spatial sound format

The existence of several spatial sound reproduction methods introduces challenges in the recording of spatial sound since the same sound scene should ideally be recorded with several different types of microphone arrays (or a dummy head) so that the scene may be later reproduced with different methods. Such challenges could be overcome if there were a generic spatial surround format that can be used to obtain signals for all reproduction methods. This section describes techniques that strive to provide such a solution by processing signals captured at a single position with a coincident microphone array.

Ambisonics

Being introduced already in the 1970s, Ambisonics [127] provided the first approach towards a generic surround audio format that enables reproduction over an arbitrary loudspeaker layout. Ideally, in Ambisonics reproduction, the loudspeakers should be placed evenly around the listener, and at least $2N_{\mathrm{ord}}+1$ loudspeakers should be used in order to achieve accurate spatial sound reproduction in the horizontal plane [128]. Here, N_{ord} denotes the ambisonic order. In principle, Ambisonics processing extracts virtual microphone signals for each loudspeaker via spherical harmonic decomposition of the sound field captured with a coincident microphone array.

Consequently, Ambisonics has inspired designs of novel coincident microphone arrays to enable decomposition of spherical harmonics. For instance, the sound field microphone presented by Farrar [129] allows the extraction of the B-format signals consisting of an omnidirectional signal and three orthogonal dipole signals (see Fig. 3.2(e)). The B-format signals are required for first-order Ambisonics processing. To knowledge of the author, the state-of-the-art coincident microphone arrays enable decomposition of the fourth-order spherical harmonics that are required in fourth-order Ambisonics processing. Design of such microphone arrays has been essential for Ambisonics, since the effective listening area and the accuracy of the reproduction can be enhanced only by increasing jointly the ambisonic order and the number of loudspeakers [130, 131]. When only the number of loudspeakers is increased, these aspects are not improved, but the coloration artifacts in the reproduction are pronounced due to the coherent nature of the loudspeaker signals [132, 131, 133].

Nonlinear time-frequency domain techniques

The recently proposed nonlinear time-frequency domain techniques [134, 135] exploit the knowledge about the capabilities and limitations of human spatial hearing and aim to reproduce the sound scenario such that the listener *perceives* being present in the original scene. The emphasis is placed on the perception, since the techniques do not share the goal of the WFS and the Ambisonics techniques to reproduce the actual sound field. Typically, these techniques take B-format signals as input and map them into time-frequency domain, where the spatial analysis is conducted. The spatial characteristics resulting from the analysis are thereafter stored as metadata and transmitted along one or more audio channels for reproduction, where the metadata is used to extract the loudspeaker signals from the transmitted audio channels.

Technique called directional audio coding (DirAC) is based on the assumption that the human auditory system can decode only one cue for direction and another for interaural coherence at each time instant and for each frequency band [134]. Following this assumption, the technique estimates the direction of arrival and diffuseness parameters separately for each time-frequency bin, and uses these parameters to separate the time-frequency bins into non-diffuse and diffuse streams in the reproduction. VBAP is then used to reproduce the non-diffuse stream in order to ensure point-like perception of sound events that have a specific direction, whereas the diffuse stream is reproduced from all loudspeakers after phase decorrelation. Alternatively, the analysis in the technique called high angular resolution planewave expansion (HARPEX) is based on the assumption that the sound field consists of two plane waves arriving from different directions with different amplitudes [135]. Hence, the directions of such plane waves are estimated within each time-frequency bin and used as metadata in the reproduction. HARPEX was designed to extract suitable signals for headphone reproduction from the B-format signals [135, 136]. In contrast, DirAC strives to provide a generic spatial audio format that can be used in various spatial sound applications, such as high-quality reproduction either with an arbitrary loudspeaker setup [134, 133], over headphones [137], as well as in teleconferencing applications requiring low bit-rates [138].

3.2 Assessment of reproduced sound

Listening tests provide an indirect method to assess the perceived quality of sound reproduction in a systematic manner. However, careful design is required when conducting listening tests in order to obtain robust data that can be reproduced by repeating the test and compared to data from other experiments. Such a design involves controlling for the experimental variables, such as the stimuli employed, the type and number of test subjects, and the acoustical characteristics of the test environment [139]. Otherwise, such variables may corrupt the data to the extent that meaningful results are not obtained. For instance, the perception evoked by a loudspeaker reproduction is known to depend on the spatial positioning of the loudspeaker(s) and the nature of the room in which the test is conducted [140, 141, 142, 143]. As a consequence, ITU-R and the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T) have presented several recommendations describing how to assess the sound reproduction quality in a robust manner. As noted also in the recommendations, the test design also dictates the nature of the data obtained, and the remaining parts of this section describe briefly a few of the test procedures.

3.2.1 Absolute category rating

The absolute category rating (ACR) test procedure in the ITU-T recommendation [144], similarly to the other test procedures described in this recommendation, was originally designed for the assessment of speech transmission. However, the ACR procedure may also be employed in the analysis of sound reproduction since it does not require a specific reference to which the evaluated entities should thereafter be compared. Following this procedure to assess spatial sound reproduction, the evaluated reproduction methods are used, one at a time, to present the stimulus to the assessor, who is always asked to rate the quality of the reproduction on a five-point scale ranging from 1 (bad) to 5 (excellent). The mean opinion score (MOS) value for each reproduction method is thereafter obtained by computing the average across the ratings provided by the different assessors. Additionally, different kinds of stimuli may be employed to study the effect of the stimulus type on the MOS values, and repetitions of the different test conditions may be used to measure the panel performance [139].

In the scope of assessment of spatial sound reproduction, the ACR procedure is perhaps most applicable when it is used to assess the performance of methods that aim to present the listener with an artificial auditory scene or to extend an actual sound scene with additional sound events. In such cases, it is difficult to define the reference entity to which the evaluated methods should be compared. However, even when no reference is presented to the assessor, the rating given by the assessor is influenced by his or her expectations and previous experiences [145]. Consequently, the results obtained with the ACR procedure are prone to variation between ratings given by different assessors, which can be controlled by increasing the number of test subjects and/or by selecting a more homogenous panel representing, for instance, a group of potential consumers [139].

3.2.2 Discriminative sensory analysis

Since the results obtained with the ACR procedure are liable to suffer variation between individual ratings, the ACR procedure may fail to detect significant differences between the evaluated methods, especially if it can be assumed that the differences are small. One solution to detect such differences is to conduct paired comparison tests, where the listener is presented with two stimuli and is asked to identify which of the two he or she prefers. The option of a neutral response may also be included. The data resulting from a paired comparison test may be encoded into a preference matrix where a given element describes the number of times the method corresponding to the row index was preferred over the method corresponding to the column index while the elements in the diagonal are zeroes. After obtaining such a matrix, Thurstonian modeling [146] may be employed to map the results on a continuous rating scale describing the ranks of the evaluated methods in a manner that is comparable to the MOS scale.

The general aim in spatial sound reproduction is to achieve a transparent reproduction of an auditory scene by processing signals that have been captured with microphones. Hence, it is sensible to include an unprocessed version of the scene as a reference condition in the listening test in order to evaluate the assumed degradation introduced by the processing. For instance, when evaluating the performance of a method in teleconferencing, the reference condition may be simulated by presenting the utterances of the spatially separated speakers from different loudspeakers around the listener. In such a case, simulation of the processing chain then comprises recording the same scenario with a microphone array, processing the microphone signals with the method, and of presenting the obtained signals with the desired loudspeaker setup. As a consequence, the experimental variables are controlled, and reliable data on the assumed degradation introduced by the processing chain is obtained when the assessor is asked to compare the processed stimulus to the unprocessed one. ITU-R has specified two recommendations for test procedures to conduct such comparisons in order to assess spatial sound reproduction techniques.

The first recommendation is known as the double-blind multi stimulus test with hidden reference and anchor (MUSHRA) procedure, which is applicable when the evaluated methods can be assumed to introduce significant amount of impairments as compared to the unprocessed reference [147]. In each trial of a MUSHRA test, several stimuli are compared at the same time, and the assessor is able to switch on the fly between the indicated reference stimulus, the test stimuli, one hidden reference stimulus, and one hidden anchor stimulus. The assessor is asked to rate the quality of the other stimuli relative to the indicated reference using a continuous scale from 0 to 100 and is instructed to give the rating of 100 for one of the stimuli (i.e., the hidden reference). The anchor stimulus with a known degradation is included in order to provide the low-quality limit that the tested methods are expected to exceed. However, the anchor stimulus needs to be selected carefully so that it is not much worse than all the others, otherwise the assessors may detect the difference between the reference and the anchor stimuli but fail to detect more subtle differences between the tested methods. If applied correctly, the MUSHRA test procedure enables a fast quality evaluation of several (maximum of 15) processed stimuli [147].

The MUSHRA procedure should not be used when the methods introduce only a small amount of impairments. Instead one should use the "double-blind triple stimulus with hidden reference" test paradigm [148] to assess such impairments. In this test paradigm, the assessor is asked to evaluate the impairments of two stimuli in comparison to the indicated reference stimulus, while one of the to-be-evaluated stimuli is the hidden reference. Again, the assessor is able to freely switch between the three stimuli. The impairments are graded using a continuous impairment scale from 5 for imperceptible to 1 for very annoying, with intermediate anchor points as described in [149]. The "double-blind triple stimulus with hidden reference" test paradigm has proven to be an efficient method to detect small differences between the evaluated methods [148], although the different methods are not actually compared to each other. It should be noted that a reference stimulus may be included also as one of the to-be-evaluated stimuli in a paired comparison test without explicitly informing the assessors that one of the entities presents the target characteristics for the other entities.

3.2.3 Descriptive sensory analysis

Neither the above-mentioned discriminative sensory analysis methods nor the ACR test procedure can provide detailed information about the aspects that cause the assessors to prefer one of the spatial sound reproduction methods over another. However, profound knowledge of these aspects would be useful when trying to improve a given spatial sound reproduction technique. For instance, it may be that the technique already provides a very good spatial impression but is perceived as impaired because the technique introduces coloration. In such a case, the perceived quality rating of the technique would probably be improved most efficiently by focusing on the reduction of the coloration artifacts.

Comprehensive knowledge about the characteristics affecting the perceived quality of the the technique can be obtained with a procedure known as preference mapping that looks for relations between independent measurements made on the same object. Moreover, the procedure requires that the assessors are asked to rank the entities in a preference test, and to rate the entities in terms of descriptive attributes, such as loudness, sharpness, and spaciousness. Thereafter, multivariate analysis methods (such as factorial analysis or principal component analysis) may be used to identify the underlying differences between the entities, and to interpret these differences in terms of the attributes [96]. Furthermore, so called "spider web" plots may be used to visualize the multidimensional attribute data in an elegant manner.

The pioneering studies by Nakayama *et al.* [150] and Gabrielsson *et al.* [151] were the first ones where preference mapping procedures were employed to assess sound reproduction. In those studies, the participants were provided with a list of attributes, prepared by the experimenter, and were asked to rate the entities using those attributes. Such pre-defined lists of attributes provide the fastest way of conducting preference mapping experiments, but they impose the risk that the assessors may not interpret the attributes in the desired manner even when the attributes are accompanied with written descriptions. Additionally, the list of attributes may not cover all the important perceptual aspects, or it may contain redundant attributes [96]. The above-mentioned risks can be effectively eliminated when the assessors are able to define the attributes by themselves in an experiment following either the individual vocabulary (IV) or the consensus vocabulary (CV) method. In the IV method, each assessor develops his or her own list of attributes, whereas in the CV method, a panel of selected assessors forms a common list of attributes under the supervision of a panel leader (e.g., the experimenter). In general, the IV method can be seen as the faster method that introduces minimal bias in the individual assessors, while the CV method provides data that can be interpreted and analyzed in a more straight-forward manner [96]. Both the IV and CV methods have been found to offer powerful means to assess spatial sound reproduction in a more detailed manner [152, 153, 154, 155].

4. Computational modeling of binaural hearing

Functional binaural auditory models aim to mimic the remarkable spatial hearing ability of the human auditory system with computational algorithms. The processing in these models comprises emulating the monaural processing in peripheral hearing models of the left and right ears and the subsequent simulation of the binaural processing of the outputs of the two peripheral hearing models. The output of a binaural auditory model is often visualized as a binaural activity map of the surrounding auditory scene. This section presents an overview of the commonly applied approaches in binaural auditory models. The overview is divided into separate sections, each of which describe the different approaches to emulate the processing in the given phase of the processing chain.

4.1 Monaural processing

Modeling of monaural processing begins with a simulation of the propagation of the sound from the sound source to the eardrum of the listener. Thereafter, the impedance matching of the middle-ear may be emulated before the processing in the inner ear is simulated by modeling the frequency analysis in the cochlea and the neural transduction occurring in the inner hair cells and the auditory nerve.

4.1.1 From a sound source to the inner ear

As mentioned earlier in Sec. 2.1, the direction-dependent characteristics of the acoustical transfer function from a point source to the eardrums of the listener are described in the HRTFs or the BRIRs, depending on whether the source and the listener are located in free-field conditions or in reverberant environments, respectively. Therefore, a given acoustical scenario can be simulated in auditory modeling by processing monophonic

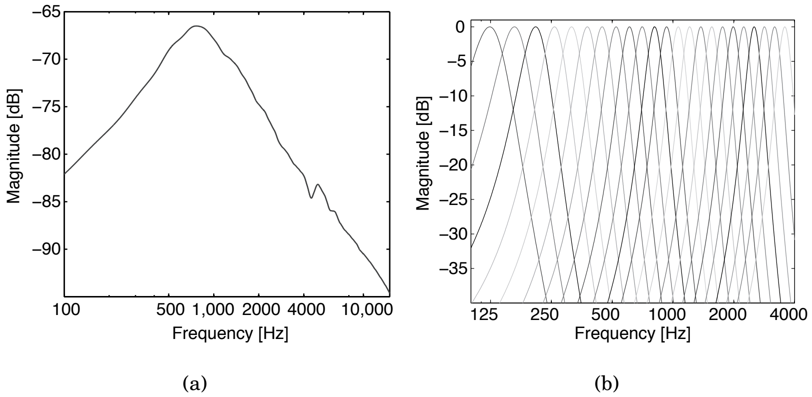


Figure 4.1. Magnitude response of (a) the middle-ear transfer function [161] implemented in [162], and (b) a 16th-order complex-valued gammatone filter bank (GTFB) [163]. The center frequencies in the GTFB were spaced at 1-ERB intervals in the frequency range from 125 to 4000 Hz.

source signals with such transfer functions. The reported differences between HRTFs of individual subjects [156, 157] motivate the use of more generic HRTF and BRIR databases in auditory modeling. Such databases can be obtained by measuring the transfer functions with a dummy head whose physical characteristics represent an average from a large number of people [158]. An alternative approach to the simulation of the auditory scene is to do a binaural recording of the scene, for instance, with a dummy head. On the other hand, headphone listening to monophonic or stereophonic content may be emulated by simply providing the content as input to the auditory model.

The acoustical transfer function of the middle-ear can be emulated by processing the input signal with a finite impulse response (FIR) filter. Moreover, anatomical measurements of temporal bone specimens of human cadavers have revealed that the peak of the displacement of the stapes in the middle-ear depends on the frequency of the pure tone that is used as the excitation [159]. Such a dependency can be emulated with an FIR filter having the desired magnitude response (see Fig. 4.1(a)) [160, 161]. Alternatively, the unprocessed input signal can be used also as input to the cochlea model if the spectral characteristics of the auditory event(s) are not of interest or if the frequency-dependent accuracy of the auditory system is taken into account at later stages.

4.1.2 Cochlear functionality

The traditional approach to emulate the frequency selectivity of the basilar membrane is to process the input signal with a filter bank consisting of a set of bandpass filters spaced at equal intervals on the ERB scale. Notched-noise measurements (see Sec. 2.2) have revealed that the human auditory filters have the assymetric shape of a roex filter function [164, 35]. However, the phase response of a roex filter function is not defined [165]. Hence, a gammatone filter is typically used in auditory modeling, since a gammatone filter provides an excellent match to the impulse response of the primary neurons in a cat, and the shape of the response is very similar to that of the roex filter function [165]. Figure 4.1(b) illustrates the magnitude responses of a gammatone filter bank (GTFB).

However, a linear filter bank, such as the GTFB, cannot account for the level-dependencies in the functionality of the cochlea. An increase in the stimulus level has been shown to result in increased asymmetry as well as in reduced gain of the auditory filter [166, 35]. These suppressive and compressive functionalities of the cochlear amplifier have been emulated in more advanced filter bank models in which each bandpass filter actually consists of two filter blocks in parallel [167, 168, 161] or in cascade [169, 170]. One of these filter blocks is linear, emulating the passive cochlear amplifier at high stimulus levels, whereas the other accounts for the increased non-linearity at lower stimulus levels due to the active role of the cochlear amplifier [168, 170].

On one hand, the functionality of the cochlear amplifier shows also time-dependent non-linear characteristics [171]. Accurate simulation of such dynamic non-linearities requires physical-based modeling of the cochlea with a transmission-line model [172]. Such a model represents the basilar membrane as a cascade of coupled mass-spring-damper systems where the active role of the cochlear amplifier may be simulated by including negative damping elements in the model [173]. The parameters of the mass-spring-damper systems at different positions along the basilar membrane can be derived by measuring otoacoustic emissions that reflect the active mechanisms in the cochlea [174]. Since the biophysical properties of the cochlea are taken into account in transmission line models, they can simulate both forward and reverse traveling waves inside the cochlea, and they may therefore be used to simulate the otoacoustic emissions of listeners with normal and impaired hearing. In auditory modeling, transmission-line models may be employed to derive the velocity and the displacement of the basilar membrane at certain positions specified by the probe frequencies. The first transmission-line models of the cochlea were developed in the 1980s (see, e.g., [175, 176, 177]), and recent models include, for instance, the models by Verhulst *et al.* [178] and Hudde & Becker [179].

4.1.3 Hair-cells, auditory nerve, and cochlear nucleus

The inner hair-cells and the auditory nerve fibers convert the displacement of the basilar membrane into neural signals that are further processed in the CN and at higher stages of the auditory pathway (Sec. 2.2). The rate of impulses in the neural signal originating from a single inner hair-cell can be described as a stochastic Poisson process where the expected number of pulses within a certain period of time is affected by the displacement of the basilar membrane at the position of the hair-cell and the amount of transmitter material in the hair-cell [180]. The firing rate of an auditory nerve fibre depends non-linearly on the rate and intensity of the input [181], and it has been demonstrated that also such a non-linear behavior can be modeled as a stochastic process [182]. Detailed anatomical knowledge has enabled the design of stochastic processes so that the inner hair-cell and auditory nerve models can accurately replicate results of neurophysiological measurements [183, 184, 161, 182].

In the CN, the neural signals from the auditory nerve fibers evoke activity in the dorsal and ventral CN cells [8]. Also, inhibitory connections between ventral and dorsal CN cells have been found [185], which may facilitate monaural echo suppression in the CN and, consequently, contribute to the echo suppression in the precedence effect [51]. In 2007, Bürk & van Hemmen [186] presented a mathematical model of the CN where the firing rate of a ventral CN was emulated as a stochastic Poisson process influenced by the rate of excitatory and inhibitory pulses arriving from the auditory nerve fiber and the dorsal CN, respectively. The firing rate of the dorsal CN was also emulated similarly, and it was demonstrated that the model can account for monaural echo suppression [186].

Since the functionality of a single inner hair-cell or auditory nerve fiber is stochastic, the commonly used approach in functional auditory models is to emulate the neural transduction with a series of signal processing operations. Moreover, the input from the cochlea model is typically half-wave rectified first and then filtered with a lowpass filter [187, 188, 189].

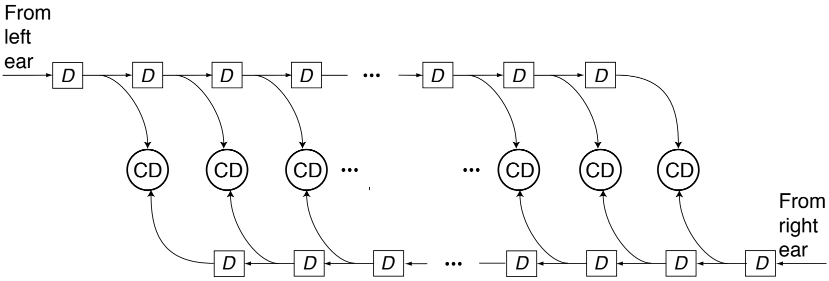


Figure 4.2. Schematic presentation of the coincidence detector model proposed by Jeffress [193], where the delay lines represent axons that connect the ear canal inputs to the coincidence detector neuron (CD). Here, D denotes the unit delay.

The output of such processing may then be thought to represent an average firing rate of the auditory nerve fibers sharing the common CF. Hence, the computational complexity is reduced at the expense of a less detailed description of the functionality of the nuclei. Optionally, the operations may be extended with automatic gain-control loops with different time-constants to emulate the non-linear adaptation of the auditory nerve fibers [190, 191]. Such loops also enable the emulation of temporal integration in loudness perception [192].

4.2 Models of binaural interaction

The majority of the binaural processing algorithms are based on the coincidence detection model proposed by Jeffress [193]. This model suggests that the receptive fields in the brain are narrowly tuned to specific locations and that the perceived location of a sound event is determined in the brain by analyzing the relative arrival time of the sound at the two ears. As illustrated in Fig. 4.2, the model consists of an array of coincidence-detector neurons receiving excitatory signals from both ears, and delay lines are used to represent axons connecting the left and right cochlear nuclei to the neuron. The highest activity is then received from the coincidence detector neuron whose input connections effectively cancel out the ITD between the ear canal signals. Such processing can be elegantly emulated by computing the normalized interaural cross-correlation (IACC) [194]

$$\gamma(t,\tau) = \frac{\int_{T=t}^{t+\Delta t} x_{\mathrm{l}}(T-\tau/2)x_{\mathrm{r}}(T+\tau/2)\mathrm{d}T}{\sqrt{\int_{T=t}^{t+\Delta t} x_{\mathrm{l}}^2(T)\mathrm{d}T + \int_{T=t}^{t+\Delta t} x_{\mathrm{r}}^2(T)\mathrm{d}T}}, \tag{4.1}$$

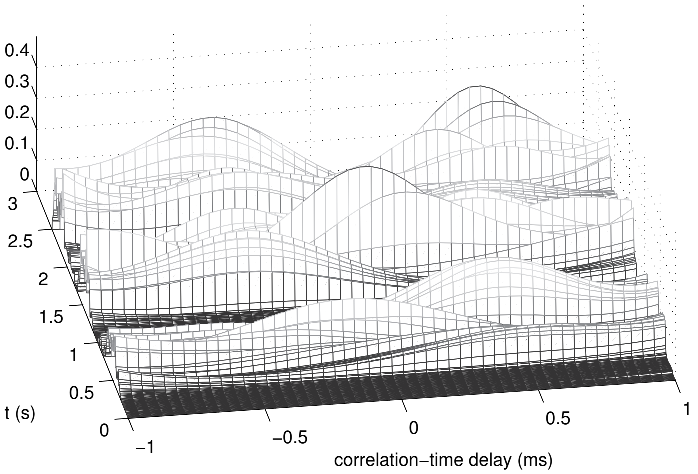


Figure 4.3. Cross-correlogram type binaural activity map for a scenario of two simultaneous talkers at ±30° azimuth directions. The activity map was obtained with the implementation [162] of the Lindemann model [201].

where t denotes the time instant, τ is the interaural delay, Δt denotes the length of the integration window, and x_l and x_r are the signals from the left and right ears, respectively. An estimate of the ITD can then be obtained as the interaural delay of the maximum of the IACC function. The output of the IACC computation may also be used to visualize the auditory scene as a cross-correlogram-type binaural activity map [195] that shows the outputs of the coincidence detector neurons at different time instants (see Fig. 4.3).

The original coincidence detection model has been extended with additional operations so that the model can more accurately account for psychoacoustical phenomena. The idea of computing the IACC separately in different auditory frequency bands was introduced in the work by Stern & Colburn [196] and Blauert & Cobben [197]. The resulting improvement in the frequency resolution has brought across the need to resolve problems caused by the spatial aliasing phenomenon. Problems arise because the wavelength becomes shorter than the head size at frequencies above 700 Hz. As a consequence, in the case of a pure tone signal, the value of the IACC function is equal at more than one interaural delay. It is believed that the auditory system resolves the ambiguity caused by spatial aliasing by simply selecting the alternative closest to the median plane [198, 45]. In coincidence detection models, this can be emulated by limiting the distribution of best delays as a function of frequency [199, 200].

One important extension of the coincidence detection model was presented by Lindemann [201]. In his model, the IACC computation was extended with two monaural detectors that were able to shift the peak of the IACC function towards the stronger ear canal input. Additionally, he introduced the idea of contralateral inhibition that effectively suppressed the activities of other coincidence detectors whenever one particular coincidence detector detected the signal. Contralateral inhibition enabled the model to account also for echo suppression in the precedence effect as the suppressive effect of the inhibition decreased gradually to zero within a 10-ms-long time window [202]. Similar functionality may also be achieved by including a multiplication with a forgetting factor into the IACC computation [203].

Another significant extension of the coincidence detection model was presented by Breebaart *et al.* [204]. In their model, the delay lines were connected to a chain of attenuators, and each coincide detector of the original model (see Fig. 4.2) was replaced with two excitation-inhibition cells, one receiving the excitation from the left ear and inhibition from the right ear and the other with opposite connections. Effectively, they extended the coincidence detection model to also account for ILD sensitivity. For a binaural input signal, the model outputs an activity map having local minima around the positions corresponding to the ITD and ILD values, and the depths of the troughs depend on the interaural coherence between the ear canal signals. It was also shown that the functionality of the model is in good accordance with human perception in several binaural signal detection scenarios [204, 205, 206]. Later, Braasch & Blauert [207] found that the precedence effect phenomenon is most accurately explained when the ITD cues are estimated with the Lindemann model [201], the ILD cues are estimated with the Breebaart model [204], and such models are extended with temporal inhibition processes. Recently, Braasch [208] presented a new model, specifically aiming to explain the precedence effect phenomenon and showed that it can explain the precedence effect in even greater detail.

Other types of binaural processing algorithms have been presented as well. The equalization-cancellation model [209] was designed to account for binaural signal detection in the presence of masking noise, and no attempts were made to emulate processing in the auditory pathway. In this model, the left and right inputs are first filtered with a set of bandpass filters so that the narrowband target can be more easily separated from the masker. Thereafter, the masker signal components are equalized in the two ears by adjusting the ITD and ILD values, and the ear canal signals are subtracted from each other, which ideally eliminates the masker from the signal.

Another interesting approach was recently presented by Dietz *et al.* [210]. In their model, the left and right ear outputs of the hair-cell processing are both provided as inputs to two separate analysis mechanisms that extract fine-structure IPD and envelope IPD information from the inputs. Both of these analysis mechanisms employ complex-valued gammatone filters (see Fig. 4.1(b)). In the fine-structure IPD extraction, the center frequency and the bandwidth of the filter depend on the CF of the frequency band, whereas envelope IPD extraction employs the same modulation-frequency-dependent filter in all frequency bands [210]. In both analysis mechanisms, processing with a complex-valued filter results in a complex-valued signal that is characterized by the amplitude a and the phase ϕ. Consequently, an estimate of the instantaneous IPD may be obtained as

$$\hat{\Phi}(t) = \arg\left(\frac{a_{\mathrm{l}}(t)}{a_{\mathrm{r}}(t)} e^{i(\phi_{\mathrm{l}}(t)-\phi_{\mathrm{r}}(t))}\right), \tag{4.2}$$

where $a \exp{(i\phi)}$ expresses the complex-valued input in polar form. Dietz *et al.* have also extended the model with additional operations that enable visualization of the auditory scene as a binaural activity map [211]. Moreover, the fine-structure IPD and the envelope IPD estimates are first mapped separately onto topographically-organized maps that are thought to consist of a set of neurons that each respond maximally to a specific IPD in the range from -2π to 2π. The two maps are then combined such that the impacts of the fine-structure and envelope information on the resulting binaural activity map are controlled with adjustable weights associated with them [211]. The localization performance of the model has been shown to improve when only reliable IPD estimates are used [212], following the binaural cue selection idea proposed by Faller & Merimaa [203]. Overall, the Dietz model has been shown to localize sound events accurately also when the auditory scene consists of an ensemble of individual sound events [211, 212].

Current neurophysiological knowledge questions whether the aforementioned binaural processing algorithms emulate accurately how the binaural cues are decoded in the auditory pathway and how the surrounding auditory space is represented in the brain. Nevertheless, this does not diminish the validity of these approaches, as many of them have successfully explained the binaural hearing phenomena in great detail. More-

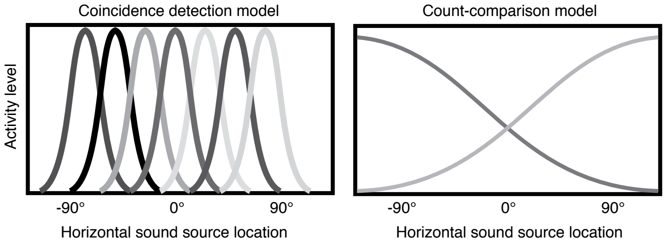


Figure 4.4. Effect of horizontal sound source location on the activity levels of different receptive fields in the brain according to the coincidence detection and count-comparison models.

over, they are often computationally less demanding than neurophysiology–based models, which also makes them appealing for use as research tools in several application areas.

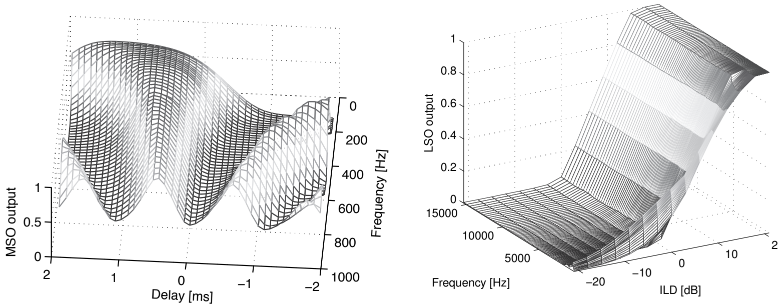
4.2.1 Neurophysiology-based models

As mentioned in Sec. 2.2, the binaural cues in the ear canal signals are decoded by the MSOs and LSOs located in the superior olivary complex. Moreover, the MSO neurons have been shown to be sensitive to the ITD [18], whereas the LSO neurons are mostly sensitive to the ILD at all frequencies, but also to the ITD at low frequencies [19, 20, 21]. The manner in which these nuclei decode the binaural cues in the human auditory system is still under debate, since their functionality cannot be measured non-invasively. However, alternative theories have been proposed based either on direct neurophysiological measurements in other species, or on the analysis of human data from IC and cortical activity measurements. One of the prevailing theories is the coincidence detection model [193] described above, and another is the count-comparison model [41, 6, 213]. According to the latter theory, the nuclei in the two cerebral hemispheres encode the spatial direction of sound simply in the rate of the output. Such processing results in two wide, receptive fields spanning an entire hemifield (see Fig. 4.4), and the spatial location is then indicated by the relative activation rates of populations in the two hemispheres [214].

The LSO seems to follow the count-comparison model since it provides a higher output when the excitation from the ipsilateral CN has a higher level than the inhibition arriving from the contralateral CN [215]. In the case of the MSO, there is evidence supporting both theories. Some results suggest that the ITD encoding in the MSO follows the coincidence detection model [216], whereas, more recently, others claim that the encoding follows the count-comparison principle [217, 218]. Furthermore, neural coding of the ITD in the human cortex seems to follow the count-comparison model [219]. The remainder of this section provides a brief overview of different computational approaches to model the functionality of the MSO and LSO nuclei. The overview is mainly limited to describing functional models of the nuclei that aim to simulate the pooled response of neurons sharing the same CF. A more detailed review also including descriptions of physiological models of the nuclei can be found, for instance, in [220].

The first MSO models may be considered as pure coincidence detection models since their inputs consist only of excitation signals coming from both hemispheres, and they contain a set of topographically-organized neurons that each have a unique best ITD [221, 222]. In each neuron, the excitations from the ipsilateral and contralateral sides are integrated over a short period of time and the neuron fires if the cumulative activation exceeds a threshold value [222]. Following the neurophysiological findings of inhibitory inputs to the MSO [10], Brughera *et al.* [223] presented a Jeffress-type MSO model that receives a phase-locked excitation and a slightly-delayed inhibitory input from both sides. However, they concluded that the functionality of the MSO neuron of a dog can be emulated with only excitatory signals since the inhibition affected only the level of the output, not its shape [223]. The opposite conclusion was reached by Brand *et al.* [224], who modeled the MSO so that the excitation from one hemisphere was immediately preceded by phase-locked inhibition from the other hemisphere. Such a modeling approach was found to result in an accurate match with neurophysiological recordings from MSO neurons of a gerbil. Since the best ITD of a coincidence detection neuron was modulated by the amount of inhibition using very short time constants [224], their model contradicts the fixed internal delay hypothesis of the Jeffress model [193].

Another interesting variation of the Jeffress-type MSO model was presented by Hancock & Delgutte [225]. They emulated coincidence detection by computing the IACC separately at each CF but varied the delays employed in the IACC computation depending on the CF. Consequently, the neurons with the lowest CF had the broadest ITD tuning curves with the maximum at the longest ITD, while the neurons with the highest CF had the narrowest tuning curves with peaks closer to the zero ITD. Thus, their model may be considered as a count-comparison-based model,



(a)

(b)

Figure 4.5. Output (a) of the MSO model [226] for a Gaussian white noise signal as a function of ITD and CF, and (b) of the LSO model [226] for a Gaussian white noise signal as a function of ILD and CF.

although the coincidence counting in the MSO was emulated with the IACC computation. An alternative functional model of the MSO following the count-comparison principle has been presented by Pulkki & Hirvonen [226]. In their model, the contralateral input is first delayed in a frequency-dependent manner, and the coincidence counting is thereafter emulated with a simple multiplication operation. In the end, the output of the coincidence counting is self-normalized with the help of the contralateral input so that the output of the model as such already indicates the direction of the sound [226]. The two models [225, 226] were shown to provide a good match with the neurophysiological measurements of ITD tuning curves in anesthetized cats [225] and guinea-pigs [227], respectively. As an example, Fig. 4.5(a) illustrates the output of the MSO model presented in [226] for a broadband sound as a function of ITD and CF.

The LSO is known to receive its excitation from the ipsilateral CN and inhibition from the contralateral CN [9]. Consequently, perhaps the simplest approach to emulate the ILD sensitivity of an LSO neuron is to subtract the level of the contralateral input from that of the ipsilateral input. Such an approach was effectively employed in the model by Reed & Blum [228], where the LSO was modeled to consist of topographically-organized neurons that are excited depending on the level of their input. Moreover, the high-threshold ipsilateral neurons are paired with the low-threshold contralateral neurons and vice versa. Such a structure is analogous to the coincidence detection model [193]. In each neuron pair, the activity of the contralateral neuron is subtracted from the activity of the ipsilateral neuron, which results in decoding of the ILD in the model, since more neuron pairs fire above their spontaneous rate when the level difference between the ipsilateral and contralateral inputs increases [228]. An alternative approach is to compute the instantaneous level difference in dB between the ipsilateral and contralateral inputs. Yue & Johnson [229] exploited the latter approach and presented an LSO model where the activity of an LSO neuron was modeled as a stochastic process. There, the expected firing rate of a neuron depended on the level difference between the ipsilateral and contralateral inputs but saturated when the level difference exceeded a threshold value [229]. The instantaneous level difference approach was also exploited in the functional model presented in [230, 226], where the output of the LSO model was also limited to between 0 and 1, as illustrated in Fig. 4.5(b).

5. Instrumental evaluation of reproduced sound

This section gives an overview of the existing computational algorithms that may be applied to evaluate (spatial) sound reproduction. The focus is placed on those algorithms that emulate (at least to some extent) the processing in the human auditory pathway in order to either predict the overall quality rating or to provide instrumental metrics related to sensory attributes affecting the overall quality impression.

5.1 Overall quality evaluation

In general, computational evaluation of the overall quality comprises comparing a given signal to the reference signal associated with it. Moreover, methods belonging to this category take both the signals as input, process them separately with an auditory model, and compute a set of metrics describing the differences between the processed signals. Thereafter, calibrated regression models or trained neural networks are used to integrate the metrics into a single value describing the perceived impairment of quality on a continuous scale from 5, meaning imperceptible, to 1, meaning very annoying. Ideally, the value provides an accurate estimate about the perceived quality score such as would have been obtained by conducting a formal listening test.

5.1.1 PEAQ

The perceptual evaluation of audio quality (PEAQ) algorithm was originally developed for evaluating impairments introduced by audio codecs on monophonic or stereophonic audio files [231]. The algorithm was designed so that it could reliably evaluate codecs that are assumed to introduce only a small amount of impairments [232]. Consequently, the algorithm could provide an alternative to the test procedure in [148](see Sec. 3.2.2) that should be used to assess such impairments in perceptual studies.

The quality evaluation in the PEAQ algorithm is based on five instrumental metrics that are thought to be related to nonlinear distortion, linear distortion, difference in harmonic structure, differences in masked thresholds, and changes in modulations. Such metrics are derived from the psychoacoustical model that first computes the excitation patterns separately for the signal being evaluated and the corresponding reference signal and thereafter extracts the metrics based on the time-aligned excitation patterns of the two signals. Specifically, the excitation pattern is obtained by emulating the frequency analysis of the basilar membrane with a linear filter bank and by simulating the neural transduction by extracting the low-pass filtered envelopes of the filter bank outputs. The psychoacoustical model also includes steps to emulate the transfer function of the middle ear (see Fig. 4.1(a)) and the level-dependent characteristics of the cochlea amplifier (see Sec. 4.1.2). Hence, the excitation patterns may be thought to present neural signals traversing via the auditory nerve to the cochlear nucleus.

The metrics provided by the psychoacoustical model are then provided as input to an artificial neural network trained to derive an estimate of the perceived quality rating based on the instrumental metrics. Moreover, the "backward propagation of errors" method was used to iteratively optimize the parameters of the network to predict the desired output from a set of inputs. Data sets from several previously conducted listening tests were used in the training so that the quality ratings were used as the desired output values, while each set of inputs consisted of the metrics provided by the pshychoacoustical model for the stimulus associated with a given quality rating [231].

However, the PEAQ algorithm does not account for spatial artifacts when evaluating the perceived quality. In contrast, the algorithm computes separate quality ratings for both channels of a two-channel audio signal based on the above-mentioned metrics. Consequently, the algorithm cannot be employed as such to evaluate spatial sound reproduction, nor can it be employed to evaluate impairments introduced by audio codecs[1] that compress multichannel audio files in the encoding phase and render the compressed audio files to the original format in the decoding

[1] The aim of these codecs (see, e.g., [233, 234]) is to reduce the data rate in the transmission of multichannel audio content, like, e.g., in the 5.1 surround audio signals, without introducing perceivable artifacts.

phase. In order to overcome the aforementioned limitations, several methods have been proposed to extend the original algorithm [235, 236, 237]. All of the proposed methods process binaural signals that are obtained by simulating the multichannel reproduction of the signals using HRTFs corresponding to the directions of the loudspeakers in the reproduction system. The binaural input signals are fed to the two psychoacoustical models, one on each side, that are implemented identically as in the original PEAQ algorithm. The motivation behind the identical implementation lies in the desire to obtain the same monaural metrics. In addition to the monaural processing, the proposed methods also employ a Jeffress-type model (see Sec. 4.2) to extract the IACC, ITD, and ILD values from the excitation patterns of the left and right ear signals. Additional metrics related to the spatial aspects are then acquired when the IACC, ITD, and ILD values obtained for the signal under evaluation are compared to the ones obtained for the reference signal. These additional metrics and the monaural metrics are then fed either to a regression model [237] or to an artificial neural network [236] that outputs an overall quality rating for the evaluated signal. The overall performance of the proposed methods is similar, while they differ in the nature of the stimuli that they cannot evaluate reliably [237]. This also explains why, to the knowledge of the author, the standardization of the PEAQ algorithm is still ongoing, and none of the proposed methods have been included in a revised version of the standard.

5.1.2 QESTRAL

In contrast to the PEAQ algorithm, the QESTRAL (quality evaluation of spatial transmission and reproduction using an artificial listener) method was specifically designed to evaluate spatial sound reproduction based on metrics obtained from a binaural auditory model [238]. Similarly as in the proposed multichannel extensions of the PEAQ algorithm, the binaural input signals to the model are obtained by simulating a given listening scenario using HRTFs corresponding to the directions of the loudspeakers, as seen from the listener's point of view. Again, the listening scenario is simulated for both the reference condition and the condition under evaluation, where the latter condition contains the impairments introduced by, e.g., processing with a multichannel audio codec, or a deviation from the ideal configuration of the loudspeakers and the listener for the given reproduction setup (see Fig. 3.1).

After obtaining the binaural input signals, they are processed with a peripheral hearing model that emulates the frequency selectivity of the cochlea with a linear GTFB (see Fig. 4.1(b)), while the neural transduction occurring in the inner hair-cells and the auditory nerve fibers is modeled with a half-wave rectification and subsequent lowpass filtering of the outputs of the GTFB. Thereafter, the resulting signals for the left and right ears are fed to a Jeffress-type cross-correlation unit that derives the IACC, ITD, and ILD values separately for each frequency band. Afterwards, the ITD and ILD values are mapped to azimuth angles using a lookup table containing reference ITD and ILD values for each horizontal direction. Subsequently, the energy-weighted averages are computed from the IACC values and the direction estimates obtained for the ITD and ILD values and the resulting average values are used to derive several metrics that the binaural model provides as the output [239]. Moreover, the metrics are thought to describe the spatial characteristics in the binaural input signal such that the metrics are expected to be related to localization angles, the apparent source width, and listener envelopment [238, 239].

Upon receiving such metrics for both the reference condition and the condition under inspection, differences in the metrics between the two conditions are computed, and a regression model is applied to integrate the obtained difference metrics into a global measure of spatial quality on a MOS scale [238]. The original QESTRAL method [238] employes separate regression models and different metrics for sweet-spot[2] and off-sweet-spot listening scenarios, where the models were calibrated using data from listening tests conducted in the corresponding scenarios [241]. The limitations of such an approach are addressed in the revised version of the method [242] that uses only one regression model for the integration of the metrics, while being able to derive relatively accurate estimates of the spatial quality across the listening area.

5.2 Instrumental metrics

The overall spatial sound perception is influenced by several attributes, such as timbre, spatial impression, loudness, and temporal characteris-

[2] The term sweet spot is used to refer to the limited listening area within which the most accurate spatial sound reproduction is achieved with a given reproduction method [240].

tics. Although, the above-mentioned PEAQ and QESTRAL algorithms also employ metrics related to these attributes, the algorithms have been optimized for predicting the overall quality impression and not to provide accurate instrumental metrics for the different attributes. Such instrumental metrics would be useful, for instance, when developing techniques for spatial sound reproduction since the metrics could be used to predict the results of a descriptive sensory analysis experiment. The remaining parts of this section give a brief description about the existing computational algorithms designed to provide such metrics. The aim is not to describe all models, but rather to present examples how metrics related to the different attributes can be derived.

5.2.1 Loudness

Several computational algorithms have been developed to predict the loudness as perceived by an average, normal-hearing test subject [243, 244, 245, 160]. These algorithms share a common basic structure. That is to say, the transmission of the sound through the external and middle ear is first emulated by filtering the signal with linear filters. Thereafter, excitation patterns at different auditory frequency bands are computed from the filtered signal, and the excitation patterns are transformed into a specific loudness spectrum. Finally, the overall loudness prediction is acquired as the sum of the specific loudness values. Typically, the excitation patterns are derived from the physical spectrum of the signal. Such an approach yields an accurate prediction of the loudness for steady-state signals, but not for time-variant signals such as speech. Consequently, revised versions of the models have been presented.

In the revised version of Zwicker's loudness model [244], the excitation patterns are computed in the time domain [246]. The computation consists of filtering the signal first with a linear filter bank, after which the resulting signals are full-wave rectified and lowpass filtered. Eventually, the revised approach results in a continuous signal representing the loudness as a function of time, and it was proposed that the overall loudness may be predicted based on the peak values in such a signal [246]. Glasberg & Moore [247] used an alternative approach when they revised their original model [160]. Their revised model divides the signal into overlapping time frames, and a short-term spectrum is computed for each time frame. Thereafter, a separate excitation pattern is derived from each short-term spectrum similarly as in [160]. As a consequence, the revised model [247] derives short-term loudness values for each time frame, while the overall loudness is obtained by integrating over the short-term loudness values.

It should be noted that the aforementioned models are based on the idea that the loudness values are computed separately for each ear, and the overall binaural loudness perception is predicted by summing the obtained loudness values. According to this idea, a given stimulus should be reproduced with a 6-dB higher level in monaural reproduction in order to achieve the same loudness perception as is acquired when the stimulus is presented to the both ears. As mentioned previously in Sec. 2.3, recent loudness matching experiments indicate that the required increase is only about 3 dB. Such results have motivated the design of new models that can account for the binaural loudness phenomenon. For instance, Moore & Glasberg have presented a model [248] where the specific loudness values are computed separately for the two ears, while the loudness values of the left ear are designed to be able to inhibit the corresponding values of the right ear, and vice versa. The inhibitory effect of a given specific loudness value is thought to spread to adjacent frequency bands as well. The inhibited specific loudness values are then summed at each ear to acquire separate loudness values for the two ears. Eventually, a prediction of the overall binaural loudness is obtained by summing the acquired loudness values. Such an approach provides a good approximation of the above-mentioned 3-dB rule with a diotic/monaural ratio of 1.5.

5.2.2 Distortion aspects

The processing involved in a spatial sound reproduction technique is bound to introduce some amount of linear and nonlinear distortion in the reproduced sound. Linear distortions are often perceived as coloration resulting from differences in the amplitude spectrum. One one hand, a sound reproduction suffering from nonlinear distortions may be described as noisy or rough, since the processing introduces frequency components that are not present in the original signal. As the human auditory system analyzes the characteristics of the sound separately in each auditory frequency band, audibility of such distortions may be evaluated by inspecting differences in the specific loudness values or the excitation patterns in different frequency bands. This idea was harnessed in the studies that predicted the perceived impairments caused by linear [249] and nonlinear distortions [250].

The algorithm presented in the former study used the loudness model by Moore *et al.* [160] to compute excitation patterns for a reference signal and its impaired version. Then, first- and second-order differences between the excitation patterns were computed and averaged across different frequency bands to obtain corresponding difference metrics. Moreover, the averaging contained a multiplication with a weighting function to emulate the relative impact of distortion in a given frequency band. Finally, a weighted sum between the first- and second-order difference metrics was computed to provide a prediction of the perceived impairment caused by linear distortions.

The algorithm presented in the latter study [250] employed a 40-band GTFB and a middle ear compensation filter (see Figs. 4.1(b) and 4.1(a)) to obtain continuous excitation patterns for the two time-aligned signals, i.e. the reference signal and its impaired version. The continuous excitation patterns were then divided into non-overlapping time frames, and a normalized cross-correlation was computed between the frames associated with the two signals. The authors stated that the maximum value of the normalized cross-correlation function can be used as a measure of the amount of distortion, since the maximum value is closer to zero, the greater the influence of the distortion. Consequently, their method computed an energy-weighted average across different frequency bands to obtain a unitary distortion measure for each time frame. Eventually, an average value was computed across the measures obtained for different time frames, and the resulting value was used as a global measure for impairments caused by nonlinear distortions.

The perceptual quality ratings obtained in a previously conducted listening experiment [251] were used to calibrate the parameters of the two algorithms [249, 250]. On completion of such calibration processes, the two algorithms were each able to accurately predict the quality ratings obtained for new stimuli subjected to linear and nonlinear distortions, respectively.

5.2.3 Spatial aspects

As discussed previously in Sec. 2.4.2, attributes relating to the spatial impression include directions of individual sound events, their apparent source widths, and listener envelopment. Binaural auditory models may be used to provide metrics related to these attributes and, consequently, to evaluate spatial sound reproduction techniques. Many of the binaural models (see Sec. 4.2) can be categorized as localization models that aim to mimic the human ability to localize sound events. Hence, a relatively straightforward application for such models is to evaluate whether a given technique can preserve the directional characteristics of the virtual sound sources. This idea was exploited by Pulkki *et al.* [252] who used a binaural auditory model to evaluate whether amplitude panning may be used to generate such a virtual sound source that evokes the same localization cues as the corresponding real sound source does. The evaluation was made separately for each frequency band based on IACC, ITD, and ILD values derived with a Jeffress-type binaural auditory model from the binaural input signals that were generated using HRTFs. Using such an approach, they demonstrated why amplitude panning suffers from difficulties in generating plausible virtual sources at the side of the listener.

The directions of the virtual sound sources in stereophonic two-channel reproduction were addressed also by Braasch [253] who evaluated the performance of different stereo microphone techniques. Moreover, the techniques were simulated to record several sound scenarios to obtain signals for a two-channel reproduction. Subsequently, binaural listening of the original scenarios and the corresponding reproductions were simulated using HRTFs to acquire binaural signals that were processed with a binaural auditory model. The performance of the technique were then evaluated by inspecting the outputs provided by the model for the different binaural input signals. Specifically, the ITD and ILD values were estimated following the Lindemann [201] and Breebaart [204] algorithms, respectively. Using the ITD and ILD values, Braasch was able to illustrate how the direction and the extent of the virtual sound source depends on the selection of the microphone technique.

The directional accuracy of elevated virtual sound sources has also been evaluated in a recent work by Baumgartner *et al.* [254]. Similarly as in the above-mentioned studies, the authors simulated binaural listening of different sound scenarios using HRTFs and used a binaural model to estimate the direction of the virtual sound source from the resulting binaural input signals. The simulated multichannel reproduction systems were designed to employ VBAP in the positioning of the virtual sources at different elevation angles while the lateral angle was limited to between $\pm 45^\circ$. After computing the errors between the desired and the estimated directions of the virtual sources, the authors were able to show how the error depends on the desired direction of the virtual sound source and on the loudspeaker layout used in the reproduction. Furthermore, the differences were found to be in accordance with the results of a perceptual study [255].

Psychoacoustical experiments have revealed that the perception of the ASW and LEV are related to fluctuations of the ITD values [256, 43, 79, 80]. As mentioned above (see also Sec. 4.2), many binaural processing algorithms derive accurate estimates of the ITD values from the binaural input signals. Hence, an analysis of the variance of the estimated ITD values across time provides a direct method to evaluate the ASW in binaural auditory models. Although this idea was originally proposed in [257], the ASW was not actually evaluated there. In contrast, a binaural auditory model was used to evaluate how the perceived direction of a wide-band noise stimulus depends on the bandwidth and the length of the stimulus. To the knowledge of the author, the study by Hess & Blauert [258] is the first where the ASW was evaluated by inspecting the variance of the ITD values. Moreover, they generated a set of frequency-modulated wide-band noise stimuli with a specific ITD. The stimuli were then employed in a perceptual study where the perceived location and the ASW of the evoked auditory image were measured separately for each stimulus. Such metrics were also estimated with a binaural auditory model based on the Lindemann algorithm [201]. The estimated metrics were found to be in good agreement with the results of the listening experiment.

Recently, van Dorp Schuitman *et al.* [259] presented a novel binaural auditory model that can evaluate the spatial impression based on several metrics related to reverberance, clarity, apparent source width, and listener envelopment. Their model is based on the binaural processing algorithm [204] that receives the excitatory and inhibitory inputs from the left and right ear peripheral processors (see Sec. 4.2). In [259], the metrics are derived in a central processing unit that receives the outputs of the two peripheral processors and the ITD values estimated with the binaural processing algorithm. The unit first divides the peripheral processor outputs into direct and reverberant streams, respectively, depending on whether the level of the output at a given time instant exceeds a specific frequency-dependent threshold value or not. Then, standard deviations of the ITD values are computed separately for the values associated with the two streams. Thereafter, metrics related to reverberance and clarity are derived from the proportions of the reverberant and direct streams in the input signal, while the standard deviations of the ITD values associated with those streams are analyzed to derive metrics related to the LEV and ASW, respectively. Their model contains several parameter values that were optimized with a genetic algorithm to provide the most accurate match to the corresponding perceptual ratings obtained with a listening experiment. On completion of such an optimization, their model was able to make accurate predictions of perceptual ratings of three other listening experiments.

It should be noted that the QESTRAL method has also been applied to evaluate listener envelopment [260]. In that study, the above-mentioned IACC, ITD, and ILD-based metrics, evaluated with a binaural auditory model, were extended with several other metrics acquired by analyzing the interchannel differences between the loudspeaker channels and by inspecting the recorded B-format signals. As only some of the metrics were derived with a binaural auditory model, the approach is not comparable to the aforementioned approaches. Nevertheless, the method presented in [260] was able to predict the LEV ratings obtained in a listening experiment after the regression model employed in the method had been optimized using data from another listening experiment.

6. Summary of publications

This section summarizes the contents of the publications included in this thesis.

Publication I: "Visualization of functional count-comparison-based binaural auditory model output"

The human spatial hearing ability is enabled by the binaural cue encoding occurring in the MSO and LSO. According to the count-comparison principle, these nuclei encode the left/right direction of sound in the rate of the output, and the spatial direction is determined at the higher stages of the auditory pathway by comparing the activation rates in the two hemispheres. Moreover, the SC has been found to contain a topographic map of the auditory space that is aligned with the visual map. Such neurophysiological data provided the motivation for PI that presented a computational model where the functionality of the MSO and LSO nuclei were emulated following the count-comparison principle, and the outputs of the nuclei models were combined in order to form a topographically organized binaural activity map of the auditory space.

Specifically, the presented model contains methods that merge the outputs of the MSO and LSO models together to form two *where* cues, one in each hemisphere. These methods also emulated the tendency of the auditory system to emphasize onsets in localization of sound events. Thereafter, the *where* cues were employed to steer the *what* cues originating from the periphery model onto a one-dimensional binaural activity map. The resulting map is thought to consist of a set of left/right organized neurons, each of which are assumed to be sensitive to a specific frequency area, and distinctive colors are used for each frequency area in order to ease the visual inspection of the map.

It was shown in PI that the binaural activity map provided by the model matches with human spatial perception in several binaural listening scenarios. As a consequence, the study demonstrated that common binaural phenomena can be explained when the functionality and the topology of the nuclei in the auditory pathway are taken into account in a signal-driven binaural auditory model.

Publication II: "Binaural assessment of parametrically coded spatial audio signals"

Parametric audio coding techniques exploit the assumption that the accuracy in the reproduction of the sound field may be compromised without introducing audible artifacts in the reproduced sound. Hence, the spatial characteristics are extracted in a time-frequency domain analysis of the microphone signals, stored as metadata in the encoding phase, and utilized in the reproduction of the microphone signals. Typically, the analysis employs several parameters that are known to affect the performance of the technique. Ideally, the values of these parameters should be selected on a signal basis, which is not possible in practice. Consequently, the parameter values are optimized during the development process. However, despite the careful optimization, some artifacts may still be audible with critical input signals.

Publication II presents an overview of the different parametric audio coding techniques and demonstrates how the performance of these techniques may be evaluated with the binaural auditory model described in PI. Moreover, several spatial artifacts that are specific to these techniques are described, including dynamically or statically biased directions, spatially too narrow auditory images, and effects of off-sweet-spot listening. Using simulated B-format microphone recordings of artificially generated sound scenarios, the techniques were employed to obtain signals for reproduction scenarios. Several spatial sound reproduction scenarios introducing the above-mentioned artifacts were then simulated using HRTFs, and the binaural auditory model was used to derive binaural activity maps for the different scenarios from the binaural input signals. The resulting binaural activity maps were then inspected, and it was found that the artifacts as well as various differences between the techniques are visible in the maps. Furthermore, the findings were found to be in line with results obtained from previously conducted listening experiments or, lacking such experiments, observations found in informal listening. As a consequence, the study demonstrates that the binaural auditory model can be used to evaluate the performance of parametric spatial sound techniques and to aid in the development of such techniques.

Publication III: Evaluation of sound field synthesis techniques with a binaural auditory model

Wave field synthesis and Ambisonics use the interference of loudspeaker signals to reconstruct a sound field within the listening area. This common goal is approached differently in the two techniques. In Ambisonics, the inputs to the loudspeakers are extracted from signals captured with a coincident microphone array, whereas separate microphones are used for each loudspeaker in WFS. The effective listening area can be enlarged by increasing the number of loudspeakers, which reduces spatial aliasing in WFS. Similar improvements may be achieved in Ambisonics if also the ambisonic order is increased at the same time.

Traditionally, these techniques have been evaluated by inspecting reconstructed sound fields either visually or in terms of instrumental measures. An alternative approach is exploited in PIII. Several binaural listening scenarios were simulated using HRTFs in order to evaluate sound fields reconstructed with WFS and Ambisonics techniques employing circular loudspeaker arrays. The simulations resulted in binaural input signals that were processed with the binaural auditory model described in PI to obtain binaural activity maps for the different scenarios. The activity maps show artifacts in the reconstructed sound fields at off-sweet-spot-listening conditions, and these artifacts are shown to be in accordance with the results of a listening test evaluating the techniques in terms of spatial aspects. Additionally, the model is able to visualize how the individual loudspeaker signals result in audible coloration artifacts in WFS, although the first wavefront is reconstructed correctly. A previously conducted listening experiment verified the audibility of such artifacts. Consequently, the study presents a successful application of the model for the evaluation of sound field synthesis techniques.

Publication IV: "A Binaural Auditory Model for the Evaluation of Reproduced Stereophonic Sound"

Spatial sound reproduction capabilities of portable multimedia devices are limited due to the small size of these devices. The two loudspeakers in (some of) them cannot be positioned in a manner that enables the optimal stereophonic listening setup, nor can the small loudspeakers yield a flat magnitude response in the entire audible frequency range. Therefore, manufacturers are required to find alternative solutions to improve the spatial sound reproduction. A binaural auditory model provides an appealing research tool aiding developers to find the optimal solution for a given device. Such an auditory model should preferably also be computationally efficient so that the effects of different solutions on the reproduction can be evaluated in a productive manner.

With the aim to provide such a research tool, an application-motivated binaural auditory model is developed and presented in PIV. The model is constructed by refining elements from previously presented models of different auditory processing stages such that the model is able to evaluate both the direction(s) of sound event(s) and the specific loudness spectra from a binaural input signal. Considering the application, the sound reproduction of a device can be evaluated by recording the reproduction with a dummy head and by using the model to derive the above-mentioned metrics. In order to verify the applicability of the model, the performance of the model was evaluated using binaural recordings made in anechoic conditions. It was found that the the model is able to mimic the human localization performance and to estimate loudness in a manner that follows the theoretical loudness function. The latter aspect is considered important for reliable evaluation of distortion aspects from the specific loudness spectra (see Sec. 5.2.2) since the sound reproduction level is known to have an effect on the sound reproduction quality of portable multimedia devices. The model is also applied to estimate the stereo image width from binaural recordings of a music sample presented with different loudspeaker setups. The application proves that the model is able to detect differences in the perceived stereo image width and to demonstrate the functionality of stereo enhancement algorithms that are typically used in portable multimedia devices to create virtual sound sources outside the narrow loudspeaker span.

Publication V: "Fusion of spatially separated vowel formant cues"

Binaural auditory models may be used to visualize the auditory scene surrounding the listener as a binaural activity map. Typically, the map is visually inspected when information about the number of sound sources, their directions and ASWs, and the LEV is being extracted. The human auditory system extracts such information in auditory scene analysis where a separate stream is formed for each sound source. If such an analysis is to be emulated in auditory models, detailed knowledge about the effects of the monaural and binaural grouping cues on auditory scene analysis is needed. Here, the monaural grouping cues refer to the spectral contents of the different sound events and the binaural ones to the directional cues evoked by the events.

The relative impacts of the monaural and binaural grouping cues on perception of speech as a fusion of separate components were investigated in PV. There, a glottal inverse-filtering algorithm was first applied to extract the glottal source signals and vocal tract transfer functions from natural Finnish vowels. Subsequently, noise-excited counterparts for the eight different vowels were generated using filters derived from the extracted signals and transfer functions. The generated noise-excited vowels were then divided into their even and odd formant components, and a set of listening experiments were conducted where the two components of the vowel /æ/ were presented from different directions around the listener using a multichannel loudspeaker reproduction system in anechoic conditions. Both the amount of spatial separation between the two components and the directions of the components were varied.

It was found that the correct vowel is identified when the two components are presented simultaneously despite the fact that neither of the components was by itself sufficient for accurate identification of the vowel. Moreover, neither the spatial separation nor the directions of the components affected the vowel identification. Hence, the monaural grouping cues seem to be strong enough to maintain the perception of the vowel-identity despite the spatial separation between the components. Interestingly, a secondary auditory event was perceived at the same time when the vowel was correctly identified, but only when the even and odd formant components of the vowel /æ/ were spatially separated and presented symmetrically in front of the listener. This implies that the binaural grouping cues may have enough weight in auditory scene analysis to evoke the perception of the additional auditory event. Consequently, the findings bolster the idea that the processing streams of the auditory pathway are fused for the identification of the vowel, but two auditory images are perceived when the components evoke conflicting directional cues in *opposite* hemispheres in the *where* processing stream.

Publication VI: "Audibility of coloration artifacts in HRTF filter designs"

HRTF filters may be used to position virtual sound sources around the listener in binaural reproduction over headphones. Ideally, the perceived spatial impression is greatly improved compared to the unprocessed headphone reproduction, and no artifacts are introduced. Since the direct measurement of HRTFs and HpTFs at the eardrums of individual subjects suffers from technical difficulties, HRTF filters are typically generated from measurements made with the microphone positioned either at the blocked ear canal entrance or at the open ear canal entrance. Furthermore, non-individual HRTFs (and HpTFs) are often used in practical applications. However, the audio quality achieved with a binaural reproduction technique is affected by the choices made during the design of the HRTF filters. The aim of the study reported in PVI was to assess how the choice of the HRTF filter design method affects the amount of introduced coloration.

In order to achieve the goals of the study, a set of HRTF filters for headphone reproduction were first generated from the HRTFs and HpTFs measured at the blocked ear canal entrance and with a pressure-velocity sensor at the open ear canal entrance. Reference filters were also designed from individual probe microphone measurements at the eardrum. Thereafter, the HRTF filters were compared to the reference filters, and a set of FIR filters were constructed to describe the differences in the magnitude responses. Then, individual stimuli were generated for each participant by processing pink noise and instrumental music samples with the FIR filters. Upon acquisition of the stimuli, the perceived amount of coloration introduced by the different HRTF filter designs were assessed following the "double-blind triple stimulus with hidden reference" test paradigm [148] (see Sec. 3.2.2). The stimuli were reproduced with a monophonic loudspeaker setup, and the assessors were asked to rate the impairment introduced by the FIR filter processing. The obtained impairment rating was then used as a measure of the coloration introduced by the given HRTF filter design.

All design methods were found to introduce coloration. Moreover, methods using non-individual blocked ear canal measurements were perceived as most colored while the method using pressure-velocity measurements at the open ear canal entrance was found to introduce the least amount of coloration. It was also found that a significant amount of the coloration is introduced at high frequencies. Consequently, the results of the study support the idea that the perceived quality of binaural reproduction using HRTF filters may be improved if the HRTF filters are designed using careful measurements of the responses from the eardrum together with individual headphone compensation.

7. Concluding remarks

Spatial sound technologies aim for high quality reproduction of a sound scene that has been either recorded with microphones or generated artificially. There are also technologies that aim to extend an actual sound scene with additional sounds that are embedded there [261]. Most of the recently developed technologies aim for a reproduction where the perception of the listener is the same as if he or she was present in the original sound scene. As it is the listener who finally decides whether these technologies have succeeded in reproducing the sound with high quality or not, the use of human subjects in formal listening tests remains the only reliable method to assess the quality of these techniques. This aspect was also recognized in this thesis work, and therefore a discriminative sensory assessment was used in PVI to measure the audibility of coloration artifacts caused by different HRTF filter designs. In this study, the method used to design the HRTF filter was found to have a significant impact on the perceived amount of coloration, and consequently, the perceived quality of headphone reproduction may be improved by careful measurement of the HRTFs and HpTFs at the eardrum.

However, conducting such a test properly is time consuming. As a consequence, listening tests are often impractical for evaluating whether a modification of a specific parameter has an effect on the quality of the sound reproduction using the method under development. Hence, auditory models that mimic the processing in the human auditory pathway with computational operations can provide an appealing alternative to the direct use of human listeners.

The idea of using a binaural auditory model for the evaluation of the reproduced sound is addressed in this thesis work. The work was initiated in the Master's Thesis of the author, where it was shown that a binaural auditory model developed by the author can be used to find spectral differences between the sound reproductions of mobile phones, differences which are audible to human listeners. The development of that model was continued, and in PIV it was shown that the model can also be applied to address the perceived stereo-image width of stereophonic loudspeaker setups and to show the functionality of stereo-widening algorithms that are often used in small portable multimedia devices to create virtual sound sources outside the narrow loudspeaker span of such devices.

However, it was later found that such an auditory model, which was designed for the evaluation of stereophonic sound reproduction of small multimedia devices, was not able to explain the results of psychoacoustical experiments in challenging sound scenarios, such as in the presence of multiple simultaneous talkers or in a diffuse field. Since such challenging sound scenarios provide also the biggest challenges for spatial sound reproduction techniques, it was decided that the processing in the human auditory pathway needs to be modeled more accurately. As a consequence, a new binaural auditory model was developed to emulate the functionality of the nuclei in the auditory pathway based on neurophysiological data and results from psychoacoustical experiments. The development resulted in a count-comparison-based model (PI) that visualizes the output as a binaural activity map that matches with human perception in several binaural listening scenarios.

The developed model was also successfully applied to evaluate spatial sound reproduction techniques in PII and PIII, both showing good agreement between model outputs and listening test results. The former study addressed spatial artifacts introduced by nonlinear time-frequency domain techniques in challenging sound scenarios, while sound fields reconstructed with wave field synthesis and Ambisonics were evaluated in the latter. However, these evaluations focused on impairments in the spatial impression, and coloration and non-linear distortion, among other important attributes affecting the overall quality perception, were, to a large extent, excluded. The ability to address such attributes requires more profound knowledge about the processing in the auditory pathway, especially about interactions between the directional and spectral information in auditory scene analysis. Such a requirement provided the motivation for the psychoacoustical experiment described in PV. The publication demonstrates that such interactions exist and that they have an influence on the auditory scene perception.

To summarize, this thesis work demonstrates that the same computational model can fulfill both requirements that were specified in Chapter 1. Firstly, the developed model with its transmission-line model of the cochlea, probabilistic model of the inner hair-cell, and functional models of MSO and LSO nuclei emulates the functionality of hearing in a sufficient detail that allows the model to account for several binaural listening phenomena such as echo suppression in the precedence effect, lateralization of band-limited noise, binaural interference, and the perception of widely distributed sound sources. Consequently, the model brings up possibilities to test new theories about hearing and to further emulate the perceptual mechanisms to combine the auditory, vestibular, and visual information. Secondly, the demonstrated ability of the model to visualize artifacts in various spatial sound reproductions opens up other application areas for the developed model. That is, this type of models may in the future replace application-specific models in the evaluation of spatial audio.

The main limitation of the current model is that visual inspection of the resulting binaural activity map is needed to evaluate whether the output of the model is in accordance with results from psychoacoustical experiments. Additionally, evaluation of the performance of a given spatial audio reproduction technique currently comprises a visual comparison of the activity maps obtained for the technique under inspection and for the target associated with the technique. Consequently, one evident area for future work is to extend the current model with high-level algorithms that would analyze the binaural activity map. For instance, artificial neural networks could be trained to provide measures about the number of auditory events, their directions and ASWs based on the binaural activity map. Another important topic for future development comprises pursuing the investigations concerning the interactions between the spectral and directional information so that the model could be extended to provide metrics related to binaural timbre perception. Also other directions for future work were identified in the publications.

Bibliography

[1] A. Kohlrausch, J. Braasch, D. Kolossa, and J. Blauert, “An Introduction to Binaural Processing,” in *The Technology of Binaural Listening* (J. Blauert, ed.), pp. 1–32, Springer-Verlag, Berlin, Germany, 2013.

[2] S. P. Thompson, “On the function of the two ears in the perception of space,” *Phil. Mag. Series 5*, vol. 13, no. 83, pp. 406–416, 1882.

[3] Lord Rayleigh, “On our perception of sound direction,” *Phil. Mag. Series 6*, vol. 13, no. 74, pp. 214–232, 1907.

[4] J. Blauert, *Spatial Hearing. The psychophysics of human sound localization*, pp. 37–50, 140–155, 164–176. Cambridge, MA, USA: MIT Press, 2nd ed., 1997.

[5] E. B. Goldstein, *Sensation and Perception*, pp. 339–357, 375–390. Wadsworth-Thomson Learning, sixth ed., 2002.

[6] G. von Békésy and E. G. Wever, *Experiments in hearing*. New York, NY, USA: McGraw-Hill, 1960. McGraw-Hill series in psychology.

[7] A. G. Møller, ed., *Hearing: Anatomy, Physiology, and Disorders of the Auditory System*, pp. 75–150. San Diego, CA: Academic Press, 2nd ed., 2006.

[8] J. K. Moore, “The human auditory brain stem as a generator of auditory evoked potentials,” *Hear. Res.*, vol. 29, pp. 33–43, 1987.

[9] D. H. Sanes, “An in vitro analysis of sound localization mechanisms in the gerbil lateral superior olive,” *J. Neuroscience*, vol. 10, pp. 3494–3506, Nov. 1990.

[10] N. B. Cant and R. L. Hyson, “Projections from the lateral nucleus of the trapezoid body to the medial superior olivary nucleus in the gerbil,” *Hear. Res.*, vol. 58, pp. 26–34, Feb. 1992.

[11] W. B. Warr, “Fiber degeneration following lesions in the anterior ventral cochlear nucleus of the cat,” *Exp. Neurol.*, vol. 14, pp. 453–474, 1966.

[12] N. L. Strominger and A. J. Strominger, “Ascending brainstem projections of the anteroventral cochlear nucleus in the rhesus monkey.,” *J. Comp. Neurol.*, vol. 143, pp. 217–242, 1971.

[13] J. Brunso-Bechtold, G. C. Thompson, and R. B. Masterton, “Study of the organization of auditory afferents ascending to the central nucleus of the inferior coiliculus in the cat,” *J. Comp. Neurol.*, vol. 197, pp. 705–722, 1981.

[14] B. Grothe, M. Pecka, and D. McAlpine, "Mechanisms of sound localization in mammals," *Physiol. Rev.*, vol. 90, pp. 983–1012, Jul. 2010.

[15] C. Tsuchitani and J. Bourdeau, "Encoding of stimulus frequency and intensity by cat superior love s-segment cells," *J. Acoust. Soc. Am.*, vol. 42, pp. 794–805, Feb. 1960.

[16] J. J. Guinan, B. E. Norris, and S. S. Guinan, "Single auditory units in the superior olivary complex. ii: Locations of unit categories and tonotopic organization," *Intl. J. Neurosci*, vol. 4, pp. 147–166, 1972.

[17] M. E. Scheibel and A. B. Scheibel, "Neurophil organization in the superior olive of the cat," *Exp. Neurol*, vol. 43, pp. 339–348, 1974.

[18] B. Grothe, "Sensory systems: New roles for synaptic inhibition in sound localization," *Nat. Rev. Neurosci.*, vol. 4, pp. 540–550, 2003.

[19] D. J. Tollin, K. Koka, and J. J. Tsai, "Interaural level difference discrimination thresholds for single neurons in the lateral superior olive," *J. Neuroscience*, vol. 28, no. 19, pp. 4848–4860, 2008.

[20] P. Joris, "Envelope coding in the lateral superior olive. II. Characteristic delays and comparison with responses in the medial superior olive," *J. Neurophysiol*, vol. 76, pp. 2137–2156, Oct. 1996.

[21] D. J. Tollin and T. C. T. Yin, "Interaural Phase and Level Difference Sensitivity in Low-Frequency Neurons in the Lateral Superior Olive," *J. Neurosci.*, vol. 25, pp. 10648–10657, Nov. 2005.

[22] P. X. Joris, C. E. Schneider, and A. Rees, "Neural processing of amplitude-modulated sounds," *Physiol. Rev.*, vol. 84, pp. 541–577, Apr. 2004.

[23] D. Irvine, "Physiology of the auditory brainstem," in *The Mammalian Auditory Pathway: Neurophysiology* (A. N. Popper and R. R. Fay, eds.), pp. 157–231, New York, NY, USA: Springer-Verlag, 1992.

[24] B. Gordon, "Receptive fields in deep layers of cat superior colliculus," *J. Neurophysiol.*, vol. 36, pp. 157–178, Mar. 1973.

[25] A. R. Palmer and A. J. King, "The representation of auditory space in the mammalian superior colliculus," *Nature*, vol. 299, pp. 248–249, Sep. 1982.

[26] B. E. Stein and M. A. Meredith, *The Merging of the Senses*. Cambridge, MA, USA: MIT Press, 1993.

[27] G. A. Calvert, "Crossmodal Processing in the Human Brain: Insights from Functional Neuroimaging Studies," *Cereb. Cortex*, vol. 11, pp. 1110–1123, Dec. 2001.

[28] C. K. Peck, "Visual-auditory interactions in cat superior colliculus: their role in the control of gaze," *Brain Res.*, vol. 420, pp. 162–166, Sep. 1987.

[29] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. Academic Press, fourth ed., 1997.

[30] B. Scharft, "Critical bands," in *Foundations of modern auditory theory* (J. V. Tobias, ed.), vol. 1, pp. 157–202, New York: Academic Press, 1970.

[31] B. Scharft, M. Florentine, and C. H. Meiselman, “Critical band in auditory lateralization,” *Percept. Psychophys.*, no. 42, pp. 215–223, 1964.

[32] G. K. Yates, “Cochlear structure and function,” in *Hearing* (B. C. J. Moore, ed.), San Diego, CA: Academic Press, 1995.

[33] L. Robles, M. A. Ruggero, and N. C. Rich, “Basilar membrane mechanics at the base of the chinchilla cochlea. I. Input–output functions, tuning curves, and response phases,” *J. Acoust. Soc. Am.*, vol. 80, pp. 1364–1374, Nov. 1986.

[34] R. D. Patterson, “Auditory filter shapes derived with noise stimuli,” *J. Acoust. Soc. Am.*, vol. 59, pp. 640–654, 1976.

[35] B. R. Glasberg and B. C. J. Moore, “Derivation of auditory filter shapes from notched-noise data,” *Hear. Res.*, vol. 47, pp. 103–138, Aug. 1990.

[36] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and models*, pp. 28–29, 61–93, 223–226. Springer, second updated ed., 1999.

[37] B. C. J. Moore and B. R. Glasberg, “Suggested formulae for calculating auditory-filter bandwidths and excitation patterns,” *J. Acoust. Soc. Am.*, vol. 74, no. 3, pp. 750–753, 1983.

[38] K. Keen, “Preservation of Constant Loudness with Interaural Amplitude Asymmetry,” *J. Acoust. Soc. Am.*, vol. 52, no. 4, pp. 1193–1196, 1972.

[39] V. P. Sivonen and W. Ellermeier, “Directional loudness in an anechoic sound field, head-related transfer functions, and binaural summation,” *J. Acoust. Soc. Am.*, vol. 119, pp. 2965–2980, May 2006.

[40] V. P. Sivonen, “Directional loudness and the underlying binaural summation for wideband and reverberant sounds,” *J. Acoust. Soc. Am.*, vol. 121, pp. 2852–2861, May 2007.

[41] G. von Békésy, “Zur Theorie des Hörens. Über das Richtungshören bei einer Zeitdifferenz oder Lautstärkeungleighheit der beiderseitigen Schalleinwirkungen,” *Physik. Zeitschr.*, pp. 824–835, 857–868, 1930.

[42] W. A. Yost, “Lateral position of sinusoids presented with interaural intensive and temporal differences,” *J. Acoust. Soc. Am.*, vol. 70, pp. 397–409, Aug. 1981.

[43] F. L. Wightman and D. J. Kistler, “The dominant role of low-frequency interaural time differences in sound localization,” *J. Acoust. Soc. Am.*, vol. 91, pp. 1648–1661, Mar. 1992.

[44] 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.*, vol. 111, pp. 2219–2236, May 2002.

[45] C. Trahiotis and R. M. Stern, “Lateralization of bands of noise: Effects of bandwidth and differences of interaural time and phase,” *J. Acoust. Soc. Am.*, vol. 86, pp. 1285–1293, Oct. 1989.

[46] S. S. Stevens and E. B. Newman, “The Localization of Actual Sources of Sound,” *Am. J. Psychol.*, vol. 48, pp. 297–306, Apr. 1936.

[47] G. Boerger, *Die Lokalisation von Gausstönen*. PhD thesis, Technische Universität, Berlin, Germany, 1965.

[48] M. B. Gardner, "Lateral localization of 0° or near-0° oriented speech signals in anechoic conditions," *J. Acoust. Soc. Am.*, vol. 44, no. 3, pp. 797–802, 1968.

[49] E. Shotter, *Absolute Auditory Object Localization*. PhD thesis, Loughborough University, Leicestershire, UK, Jun. 1997.

[50] A. W. Mills, "On the minimum audible angle," *J. Acoust. Soc. Am.*, vol. 30, no. 4, pp. 237–246, 1958.

[51] H. Wallach, E. B. Newman, and M. R. Rosenzweig, "The precedence effect in sound localization," *Am. J. Psychol.*, vol. 42, pp. 315–326, 1949.

[52] R. Litovsky, S. Colburn, W. A. Yost, and S. Guzman, "The precedence effect," *J. Acoust. Soc. Am.*, vol. 106, pp. 1633–1654, Oct. 1999.

[53] J. L. Flanagan and B. J. Watson, "Binaural unmasking of complex signals," *J. Acoust. Soc. Am.*, vol. 40, no. 2, pp. 546–468, 1966.

[54] R. T. Carhart, T. W. Tillman, and E. S. Greetis, "Perceptual masking in multiple sound backgrounds," *J. Acoust. Soc. Am.*, vol. 45, no. 3, pp. 694–703, 1969.

[55] D. McFadden and E. G. Pasanen, "Lateralization at high frequencies based on interaural time differences," *J. Acoust. Soc. Am.*, vol. 59, pp. 634–639, Mar. 1976.

[56] A. Kohlrausch, "The influence of signal duration, signal frequency and masker duration on binaural masking level differences," *Hear. Res.*, vol. 23, pp. 267–273, Feb. 1986.

[57] B. Kollmeier and R. H. Gilkey, "Binaural forward and backward masking: Evidence for sluggishness in binaural detection," *J. Acoust. Soc. Am.*, vol. 87, pp. 1709–1719, Apr. 1990.

[58] V. Best, F. J. Gallun, S. Carlile, and B. G. Shinn-Cunningham, "Binaural interference and auditory grouping," *J. Acoust. Soc. Am.*, vol. 121, pp. 1070–1076, Feb. 2007.

[59] T. Hirvonen and V. Pulkki, "Perceived distribution of horizontal ensemble of independent noise signals as function of sample length," in *Proc. AES 124th Convention*, (Amsterdam, the Netherlands), May 17-20 2008. Paper No. 7408.

[60] O. Santala and V. Pulkki, "Directional perception of distributed sound sources," *J. Acoust. Soc. Am.*, vol. 129, pp. 1522–1530, Mar. 2011.

[61] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of sound*, pp. 47–394. Cambridge, MA, USA: MIT Press, 1994.

[62] L. P. A. S. van Noorden, "Minimum differences of level and frequency for perceptual fission of tone sequences ABAB," *J. Acoust. Soc. Am.*, vol. 61, no. 4, pp. 1041–1045, 1977.

[63] R. A. Rasch, “The perception of simultaneous notes such as in polyphonic music,” *Acustica*, vol. 40, pp. 21–33, 1978.

[64] C. J. Darwin and N. S. Sutherland, “Grouping frequency components of vowels: When is a harmonic not a harmonic?,” *Q. J. Exp. Psychol.-A.*, vol. 36, pp. 193–208, 1984.

[65] D. Deutsch, “Two-channel listening to musical scales,” *J. Acoust. Soc. Am.*, vol. 57, pp. 1156–1160, 1975.

[66] V. Best, B. G. Shinn-Cunningham, E. J. Ozmeral, and N. Kopco, “Exploring the benefit of auditory spatial continuity,” *J. Acoust. Soc. Am.*, vol. 127, no. 6, pp. 258–264, 2010.

[67] D. E. Broadbent, “A note on binaural fusion,” *Q. J. Exp. Psychol.*, vol. 7, pp. 46–47, 1955.

[68] T. C. Rand, “Dichotic release from masking for speech,” *J. Acoust. Soc. Am.*, vol. 55, pp. 678–680, 1974.

[69] A. H. Schwartz and B. G. Shinn-Cunningham, “Dissociation of perceptual judgments of ’what’ and ’where’ in an ambiguous auditory scene,” *J. Acoust. Soc. Am.*, vol. 128, pp. 3041–3051, Nov. 2010.

[70] B. G. Shinn-Cunningham, A. K. C. Lee, and A. J. Oxenham, “A sound element gets lost in perceptual competition,” *PNAS*, vol. 104, pp. 12223–12227, May 2007.

[71] J. S. Snyder, M. K. Gregg, and C. Alain, “Attention, awareness, and the perception of auditory scenes,” *Front. Pscyhol.*, vol. 3, pp. 1–15, Feb. 2012.

[72] B. G. Shinn-Cunningham, “Object-based auditory and visual attention,” *Trends Cogn. Sci.*, vol. 12, pp. 182–186, May 2008.

[73] H. A. Witkin, S. Wapner, and T. Leventhal, “Sound localization with conflicting visual and auditory cues,” *J. Exp. Psychol.*, vol. 43, pp. 58–67, Jan. 1952.

[74] C. V. Jackson, “Visual factors in auditory localization,” *Q. J. Exp. Psychol*, vol. 5, no. 2, pp. 52–65, 1953.

[75] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” *Nature*, vol. 264, pp. 746–748, Dec. 1976.

[76] M. Barron and A. H. Marshall, “Spatial impression due to early lateral reflections in concert halls: the deviation of a physical measure,” *J. Sound and Vibration*, vol. 77, no. 2, pp. 211–232, 1981.

[77] J. Blauert and W. Lindeman, “Auditory spaciousness: some further psychoacoustic analyses,” *J. Acoust. Soc. Am.*, vol. 80, pp. 533–542, Aug. 1986.

[78] J. S. Bradley and G. A. Souldore, “The influence of late arriving energy on spatial impression,” *J. Acoust. Soc. Am.*, vol. 94, no. 4, pp. 2263–2271, 1995.

[79] D. Griesinger, “The psychoacoustics of apparent source width, spaciousness and envelopment in performance spaces,” *Acta Acoustica United with Acoustica*, vol. 83, no. 4, pp. 721–734, 1997.

[80] R. Mason and F. Rumsey, “Interaural time difference fluctuations: their measurement, subjective perceptual effect, and application in sound reproduction,” in *Proc. AES 19th Intl. Conf.*, (Schloss Elmau, Germany), Jun. 2001.

[81] Y. Ando, “Subjective preference in relation to objective parameters of music sound fields with a single echo,” *J. Acoust. Soc. Am.*, vol. 62, pp. 1436–1441, Dec. 1977.

[82] T. Lokki, J. Pätynen, A. Kuusinen, and S. Tervo, “Disentangling preference ratings of concert hall acoustics using subjective sensory profiles,” *J. Acoust. Soc. Am.*, vol. 132, pp. 3148–3161, Nov. 2012.

[83] M. Ashby and J. Maidment, *Introducing Phonetic Science*. Cambridge University Press, illustrated ed., 2005.

[84] M. Cooke and D. P. W. Ellis, “The auditory organization of speech and other sources in listeners and computational models,” *Speech Communication*, vol. 35, pp. 141–177, Oct. 2001.

[85] J. Lyzenga and B. C. J. Moore, “Effect of frequency-modulation coherence for inharmonic stimuli: Frequency-modulation phase discrimination and identification of artificial double vowels,” *J. Acoust. Soc. Am.*, vol. 117, pp. 1314–1325, Mar. 2005.

[86] R. E. Remez, P. E. Rubin, D. B. Pisoni, and T. Carrell, “Speech perception without traditional speech cues,” *Science*, vol. 212, pp. 947–950, May 1981.

[87] D. E. Broadbent and P. Ladefoged, “On the fusion of sounds reaching different sense organs,” *J. Acoust. Soc. Am.*, vol. 29, pp. 708–710, 1957.

[88] A. M. Liberman and I. G. Mattingly, “A Specialization for Speech Perception,” *Science*, vol. 243, pp. 489–494, Jan. 1989.

[89] D. H. Whalen and A. M. Liberman, “Speech perception takes precedence over non-speech perception,” *Science*, vol. 237, pp. 169–171, Jul. 1987.

[90] S. Bentin and V. Mann, “Masking and stimulus intensity effects on duplex perception: A Confirmation of the dissociation between speech and nonspeech modes,” *J. Acoust. Soc. Am.*, vol. 88, pp. 64–74, Jul. 1990.

[91] E. C. Cherry, “Some experiments on the recognition of speech, with one and with two ears,” *J. Acoust. Soc. Am.*, vol. 25, no. 5, pp. 975–979, 1953.

[92] J. C. R. Licklider, “The influence of interaural phase relations upon the masking of speech by white noise,” *J. Acoust. Soc. Am.*, vol. 20, no. 2, pp. 150–159, 1948.

[93] M. T. M. Scheffers, *Sifting vowels. Auditory pitch analysis and sound segregation*. PhD thesis, University of Groningen, 1983. Summary.

[94] W. A. Yost, R. H. Dye, and S. Sheft, “A Simulated ¨Cocktail Party¨ with Up to Three Sound Sources,” *Percept. Psychophys*, vol. 58, no. 7, pp. 1026–1036, 1996.

[95] M. L. Hawley, R. Y. Litovsky, and H. S. Colburn, “Speech intelligibility and localization in a multi-source environment,” *J. Acoust. Soc. Am.*, vol. 105, pp. 3436–3448, Jun. 1999.

[96] G. Lorho, *Perceived Quality Evaluation: An application to sound reproduction over headphones*. PhD thesis, Aalto University, 2010.

[97] F. Rumsey, *Spatial Audio*. Oxford, England: Focal Press, 2001.

[98] J. Eargle, *The Microphone Book*. Woburn, MA: Focal Press, 2001.

[99] A. D. Blumlein, "U.K. Patent 394,325, 1931." Reprinted in Stereophonic Techniques, Audio Eng. Soc., NY, USA, 1986.

[100] H. Wierstorf and S. Spors, "Sound field synthesis toolbox," in *Proc. AES 132nd Convention*, (Budapest, Hungary), Apr. 26-29 2012. eBrief No. 50.

[101] B. B. Bauer, "Phasor Analysis of Some Stereophonic Phenomena," *J, Acoust. Soc. Am.*, vol. 33, pp. 1536–1539, Nov. 1961.

[102] A. Wilska, *Untersuchungen über das Richtungshören*. PhD thesis, University of Helsinki, 1938.

[103] H. Møller, "Fundamentals of Binaural Technology," *Appl. Acoust.*, vol. 36, no. 3/4, pp. 171–218, 1992.

[104] F. L. Wightman, D. J. Kistler, and M. Arruda, "Perceptual consequences of engineering compromises in synthesis of virtual auditory objects (A)," *J. Acoust. Soc. Am.*, vol. 92, no. 4, p. 2332, 1992.

[105] N. I. Durlach, A. Rigopulos, X. D. Pang, W. S. Woods, A. Kulkarni, H. S. Colburn, and E. M. Wenzel, "On the externalization of auditory images," *Presence*, vol. 1, no. 2, pp. 251–257, 1992.

[106] F. L. Wightman and D. J. Kistler, "Resolution of front–back ambiguity in spatial hearing by listener and source movement," *J. Acoust. Soc. Am.*, vol. 105, pp. 2841–2853, May. 1999.

[107] A. Silzle, "Selection and tuning of HRTFs," in *Proc. AES 122nd Convention*, (Munich, Germany), May 2002.

[108] N. Zacharov and G. Lorho, "Subjective Evaluation of Virtual Home Theatre Sound Systems for Loudspeakers and Headphones," in *Proc. AES 116th Convention*, (Berlin, Germany), May 2004.

[109] F. E. Toole, "The acoustics and psychoacoustics of headphones," in *Proc. AES 2nd Intl. Conf.*, (Anaheim, CA, USA), May 1984.

[110] R. C. Maher, E. Lindemann, and J. Barish, "Old and new techniques for artificial stereophonic image enhancement," in *Proc. AES 101st Convention*, (Los Angeles, CA, USA), Nov. 1996. Paper No. 4371.

[111] B. S. Atal and M. R. Schroeder, "Apparent sound source translator." US Patent no. 3,236,949, Feb. 1966.

[112] S. E. Olive, "Evaluation of five commercial stereo enhancement 3d audio software plug-ins," in *Proc. AES 110th Convention*, (Amsterdam, the Netherlands), May 2001. Paper No. 5386.

[113] ITU, *Multichannel stereophonic sound system with and without accompanying picture ITU-R Recommendation BS.775-1*, 1997.

[114] G. Barry, B. Lawlor, and E. Coyle, “Sound source separation: Azimuth discrimination and resynthesis,” in *Proc. 7th Intl. Conf. on Digital Audio Effects*, (Naples, Italy), pp. 240–244, Oct. 5–8 2004.

[115] C. Faller, “Multiple-loudspeaker playback of stereo signals,” *J. Audio Eng. Soc.*, vol. 54, pp. 1051–1064, Nov. 2006.

[116] G. Barry and D. Kearney, “Localization quality assessment in source separation-based upmixing algorithms,” in *Proc. AES 35th Intl. Conf.*, (London, UK), Feb. 11–13 2009. Paper No. 33.

[117] G. Theile and H. Wittek, “Principles in surround recordings with height,” in *Proc. AES 130th Convention*, (London, England), p. Paper No. 8403, May 13–16 2011.

[118] K. Hamasaki, K. Hiyama, and R. Okumura, “The 22.2 Multichannel Sound System and Its Application,” in *Proc. AES 118th Convention*, (Barcelona, Spain), May 28-31 2005. Paper No. 6406.

[119] V. Pulkki, “Virtual Sound Source Positioning Using Vector Base Amplitude Panning,” *J. Audio Eng. Soc.*, vol. 45, pp. 456–466, Jun. 1997.

[120] A. J. Berkhout, “A Holographic Approach to Acoustic Control,” *J. Audio Eng. Soc.*, vol. 36, pp. 977–995, Dec. 1988.

[121] G. Theile, “Wave field synthesis– a promising spatial audio rendering concept,” in *Proc. 7th Intl. Conf. on Digital Audio Effects*, (Naples, Italy), pp. 125–132, Oct. 2004.

[122] A. J. Berkhout, D. de Vries, and P. Vogel, “Acoustic control by wave field synthesis,” *J. Acoust. Soc. Am.*, vol. 93, pp. 2764–2778, 1993.

[123] S. Spors and J. Ahrens, “Spatial sampling artifacts of wave field synthesis for the reproduction of virtual point sources,” in *Proc. AES 126th Convention*, (Munich, Germany), May 2009. Paper No. 7744.

[124] F. Volk, J. Konradl, and H. Fastl, “Simulation of wave field synthesis,” in *Proc. Acoustics’08*, (Paris, France), Jun. 29–Jul. 4 2008.

[125] H. Wierstorf, A. Raake, and S. Spors, “Binaural assessment of multichannel reproduction,” in *The Technology of Binaural Listening* (J. Blauert, ed.), pp. 255–278, Berlin-Heidelberg, Germany: Springer-Verlag, 2013.

[126] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, “Creating interactive virtual acoustic environments,” *J. Audio Eng. Soc.*, vol. 47, no. 9, pp. 675–705, 1999.

[127] M. A. Gerzon, “Periphony: With-height sound reproduction,” *J Audio Eng. Soc.*, vol. 21, pp. 2–10, Feb. 1973.

[128] M. A. Gerzon, “Criteria for evaluating surround-sound systems,” *J Audio Eng. Soc.*, vol. 25, pp. 400–408, Jun. 1977.

[129] K. Farrar, “Soundfield microphone,” *Wireless World*, vol. 85, pp. 48–50, Oct. 1979.

[130] A. Solvang, “Spectral Impairment of Two-Dimensional Higher Order Ambisonics,” *J. Audio Eng. Soc.*, vol. 56, pp. 267–279, Apr. 2008.

[131] S. Bertet, J. Daniel, E. Parizet, and O. Warusfel, “Influence of Microphone and Loudspeaker Setup on Perceived Higher Order Ambisonics Reproduced Sound Field,” in *Proc. Ambisonics Symposium*, (Graz, Austria), Jun. 25-27 2009.

[132] V. Pulkki, J. Merimaa, and T. Lokki, “Reproduction of Reverberation with Spatial Impulse Response Rendering,” in *Proc. AES 116th Convention*, (Berlin, Germany), May 8-11 2004. Paper No. 6057.

[133] J. Vilkamo, T. Lokki, and V. Pulkki, “Directional Audio Coding: Virtual Microphone-Based Synthesis and Subjective Evaluation,” *J. Audio Eng. Soc.*, vol. 57, pp. 709–724, Sept. 2009.

[134] V. Pulkki, “Spatial sound reproduction with Directional Audio Coding,” *J Audio Eng. Soc.*, vol. 55, pp. 503–516, Jun. 2007.

[135] S. Berge and N. Barrett, “High Angular Resolution Planewave Expansion,” in *Proc. 2nd Intl. Symposium on Ambisonics and Spherical Acoustics*, (Paris, France), May 6-7 2010.

[136] S. Berge and N. Barrett, “A new method for B-format to binaural transcoding,” in *Proc. AES 40th Intl. Conf.*, (Tokyo, Japan), Oct. 8–10 2010. Paper No. 6-5.

[137] M.-V. Laitinen and V. Pulkki, “Binaural Reproduction For Directional Audio Coding,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, (New Paltz, NY), Oct. 2009.

[138] J. Ahonen, “Microphone Configurations for Teleconference Application of Directional Audio Coding and Subjective Evaluation,” in *Proc. AES 40th Intl. Conf.*, (Tokyo, Japan), Oct. 8–10 2010. Paper No. 5.

[139] S. Bech and N. Zacharov, *Perceptual Audio Evaluation – Theory, Method and Application*. Chichester, England: John Wiley & Sons, Ltd., 2006.

[140] S. E. Olive and F. E. Toole, “The detection of reflections in typical rooms,” *J. Audio Eng. Soc.*, vol. 37, pp. 539–553, Jul. 1989.

[141] F. E. Toole, “Subjective evaluation: Identifying and controlling the variables,” in *Proc. AES 8th Intl. Conf.*, (Washington D.C., USA), pp. 95–100, May 1990.

[142] S. Bech, “Timbral aspects of reproduced sound in small rooms. I,” *J. Acoust. Soc. Am.*, vol. 97, pp. 1717–1726, Mar. 1995.

[143] S. Bech, “Timbral aspects of reproduced sound in small rooms. II,” *J. Acoust. Soc. Am.*, vol. 99, pp. 3539–3549, Jun. 1996.

[144] ITU, *Methods for Subjective Determination of Transmission Quality ITU-T Recommendation P.800*, 1990.

[145] H. Fastl, “Psycho-acoustics and sound quality,” in *Communication Acoustics* (J. Blauert, ed.), pp. 139–162, Berlin-Heidelberg, Germany: Springer-Verlag, 2005.

[146] L. L. Thurstone, “A law of comparative judgment,” *Psychol. Rev.*, vol. 34, pp. 273–286, Jul. 1927.

[147] ITU, *Methods for the Subjective Assessment of Intermediate Quality Level of Coding Systems ITU-R Recommendation BS.1534-1*, 2003.

[148] ITU, *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems ITU-R Recommendation BS.1116-1*, 1997.

[149] ITU, *Methods for Subjective Assessment of Sound Quality – General Requirements ITU-R Recommendation BS.1284*, 1998.

[150] T. Nakayama, T. Miura, O. Kosaka, M. Okamoto, and T. Shiga, "Subjective Assessment of Multichannel Reproduction," *J. Audio Eng. Soc.*, vol. 19, pp. 744–751, Oct. 1971.

[151] A. Gabrielsson and H. Sjögren, "Perceived sound quality of sound-reproducing systems," *J. Acoust. Soc. Am.*, vol. 66, pp. 1019–1033, Apr. 1979.

[152] N. Zacharov and K. Koivuniemi, "Unravelling the perception of spatial sound reproduction: Analysis & external preference mapping," in *Proc. AES 111st Convention*, (New York, NY), Nov. 30–Dec. 3, 2001.

[153] C. Guastavino and B. F. G. Katz, "Perceptual evaluation of multi-dimensional spatial audio reproduction," *J. Acoust. Soc. Am.*, vol. 116, pp. 1105–1115, Aug. 2004.

[154] G. Lorho, "Perceptual evaluation of mobile multimedia loudspeakers," in *Proc. AES 122nd Convention*, (Vienna, Austria), May 2007.

[155] T. Hirvonen and V. Pulkki, "A Listening Test System for Automotive Audio – Part 3: Comparison of Attribute Ratings Made in a Vehicle with Those Made Using an Auralization System," in *Proc. AES 123rd Convention*, (New York, NY), Oct. 5–8 2007. Paper No. 7224.

[156] H. Møller and M. F. Sørensen and D. Hammershøi and C. B. Jensen, "Head-Related Transfer Functions of Human Subjects," *J. Audio Eng. Soc.*, vol. 43, pp. 300–321, May 1995.

[157] H. Møller and M. F. Sørensen and C. B. Jensen and D. Hammershøi, "Binaural Technique: Do We Need Individual Recordings?," *J. Audio Eng. Soc*, vol. 44, pp. 451–469, Jun. 1996.

[158] ITU, *Head and torso simulator for telephonometry ITU-T Recommendation P.58*, 1996.

[159] R. L. Goode, M. Killion, K. Nakamura, and S. Nishihara, "New knowledge about the function of the human middle ear: development of an improved analog model," *Am. J. Otol.*, vol. 15, pp. 145–154, Mar. 1994.

[160] B. C. J. 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. 224–237, 1997.

[161] E. A. Lopez-Poveda and R. Meddis, "A human nonlinear cochlear filterbank," *J. Acoust. Soc. Am.*, vol. 110, pp. 3107–3118, Dec. 2001.

[162] P. Søndegaard and P. Majdak, "The Auditory Modeling Toolbox," in *The Technology of Binaural Listening* (J. Blauert, ed.), pp. 33–56, Springer-Verlag, Berlin, Germany, 2013.

[163] A. Härmä, "HUTEar Matlab Toolbox version 2.0." `http://www.acoustics.hut.fi/software/HUTear/`, 2000. Accessed: Mar. 3, 2014.

[164] R. D. Patterson, I. Nimmo?Smith, D. L. Weber, and R. Milroy, "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," *J. Acoust. Soc. Am.*, vol. 72, pp. 1788–1803, Dec. 2003.

[165] R. D. Patterson and M. H. Allerhand, "Extending the domain of center frequencies for the compressive gammachirp auditory filter," *J. Acoust. Soc. Am.*, vol. 98, pp. 1892–1894, Oct. 1995.

[166] R. A. Lufti and R. D. Patterson, "On the growth of masking asymmetry with stimulus intensity," *J. Acoust. Soc. Am.*, vol. 76, pp. 739–745, Sep. 1984.

[167] L. Carney, "A model for the responses of low-frequency auditory-nerve fibers in cat," *J. Acoust. Soc. Am.*, vol. 93, pp. 401–417, Jan. 1993.

[168] R. Meddis, L. P. O'Mard, and E. A. Lopez-Poveda, "A computational algorithm for computing nonlinear auditory frequency selectivity," *J. Acoust. Soc. Am.*, vol. 109, pp. 2852–2861, Jun. 2001.

[169] T. Irino and R. Patterson, "A time-domain, level-dependent auditory filter: The gammachirp," *J. Audio Eng. Soc. Am.*, vol. 101, pp. 412–419, Jan. 1997.

[170] R. D. Patterson, M. Unoki, and T. Irino, "Extending the domain of center frequencies for the compressive gammachirp auditory filter," *J. Acoust. Soc. Am.*, vol. 114, pp. 1529–1542, Sep. 2003.

[171] A. R. D. Thornton, K. Shin, E. Gottesman, and J. Hine, "Temporal non-linearities of the cochlear amplifier revealed by maximum length sequence stimulation," *Clin. Neurophys.*, vol. 112, pp. 768–777, May. 2001.

[172] C. A. Shera, J. J. Guinan, and A. J. Oxenham, "Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements," *PNAS*, vol. 99, pp. 3318–3323, Mar. 2002.

[173] G. Zweig, "Finding the impedance of the organ of Corti," *J. Acoust. Soc.Am.*, vol. 89, pp. 1229–1254, Mar. 1991.

[174] D. T. Kemp, "Stimulated acoustic emissions from within the human auditory system," *J. Acoust. Soc. Am.*, vol. 64, pp. 1386–1391, Nov. 1978.

[175] E. de Boer, "Auditory physics. Physical principles in hearing theory. I," *Phys. Rep.*, vol. 62, pp. 87–174, Jun. 1980.

[176] D. O. Kim, C. E. Molnar, and J. W. Matthews, "An active cochlear model with negative damping in the partition: comparisons with rhode's ante- and postmortem observations," in *Psychophysical, physiological and behavioral studies in hearing* (G. van den Brink and F. A. Bilsen, eds.), pp. 7–14, Delft university press, Delft, 1980.

[177] S. T. Neely and D. O. Kim, "An active cochlear model showing sharp tuning and high sensitivity," *Hear. Res.*, vol. 9, pp. 123–130, 1983.

[178] S. Verhulst, T. Dau, and C. A. Shera, "Nonlinear time-domain cochlear model for transient stimulation and human otoacoustic emission," *J. Acoust. Soc. Am.*, vol. 132, pp. 3842–3848, Dec. 2012.

[179] H. Hudde and S. Becker, "A physiology-based auditory model elucidating the function of the cochlear amplifier and related phenomena. Part I: Model structure and computational method," in *Proc. Meetings on Acoustics*, vol. 19, (Montreal, Canada), Jun. 2-8, 2013. Paper No. 3aPP1.

[180] R. Meddis, "Simulation of mechanical to neural transduction in the auditory receptor," *J. Acoust. Soc. Am.*, vol. 79, pp. 702–711, Mar. 1986.

[181] G. K. Yates, I. M. Winter, and D. Robertson, "Basilar membrane nonlinearity determines auditory nerve rate-intensity functions and cochlear dynamic range," *Hear. Res.*, vol. 45, pp. 203–219, May. 1990.

[182] C. J. Sumner, E. A. Lopez-Poveda, L. P. O'Mard, and R. Meddis, "A revised model of the inner-hair cell and auditory-nerve complex," *J. Acoust. Soc. Am.*, vol. 111, pp. 2178–2188, May 2002.

[183] S. Ross, "A model of the hair cell?primary fiber complex," *J. Acoust. Soc. Am.*, vol. 71, pp. 926–941, Apr. 1982.

[184] R. Meddis, "Simulation of auditory-neural transduction: Further studies," *J. Acoust. Soc. Am.*, vol. 83, pp. 1056–1063, Mar. 1988.

[185] R. E. Wickesberg and D. Oertel, "Delayed, frequency-specific inhibition in the cochlear nuclei of mice: a mechanism for monaural echo suppression," *J. Neurosci.*, vol. 10, pp. 1762–1768, Jun. 1990.

[186] M. Bürk and L. van Hemmen, "Modeling the cochlear nucleus: A site for monaural echo suppression?," *J. Acoust. Soc. Am.*, vol. 122, pp. 2226–2235, Oct. 2007.

[187] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. I. Model structure," *J. Audio Eng. Soc. Am.*, vol. 99, no. 6, pp. 3615–2622, 1996.

[188] J. Buchholz and J. Mourjopoulos, "A computational auditory masking model based on signal-dependent compression. i. model description and performance analysis," *Acta Acustica united with Acustica*, vol. 90, pp. 873–886, Sep./Oct. 2004.

[189] M. L. Jepsen, S. D. Ewert, and T. Dau, "A computational model of human auditory signal processing and perception," *J. Acoust. Soc. Am.*, vol. 124, pp. 422–438, Jul. 2008.

[190] R. L. Smith and J. J. Zwislocki, "Short-term adaptation and incremental responses of single auditory-nerve fibers," *Biol. Cybernetics*, vol. 17, pp. 169–182, 1975.

[191] L. A. Westerman and R. L. Smith, "Rapid and short-term adaptation in auditory nerve responses," *Hear. Res.*, vol. 15, pp. 249–260, Sep. 1984.

[192] M. Karjalainen, "A Binaural Auditory Model for Sound Quality Measurements and Spatial Hearing Studies," in *IEEE on Acoust., Speech and Sig. Proc.*, vol. 2, pp. 985–988, 1996.

[193] L. A. Jeffress, "A place theory of sound localization," *J. Comp. Physiol. Psychol.*, vol. 41, pp. 35–39, 1948.

[194] B. M. Sayers and E. C. Cherry, "Mechanism of binaural fusion in the hearing of speech," *J. Acoust. Soc. Am.*, vol. 29, no. 9, pp. 973–987, 1957.

[195] T. M. Shackleton, R. Meddis, and M. J. Hewitt, "Across frequency integration in a model of lateralization," *J. Acoust. Soc. Am.*, vol. 91, pp. 2276–2279, Apr. 1992.

[196] R. Stern and H. Colburn, "Theory of binaural interaction based on auditory-nerve data. IV. A model for subjective lateral position.," *J. Acoust. Soc. Am.*, vol. 64, no. 1, pp. 127–140, 1978.

[197] J. Blauert and W. Cobben, "Some consideration of binaural cross correlation analysis," *Acta Acoustica united with Acoustica*, vol. 39, pp. 96–104, Jan. 1978.

[198] L. A. Jeffress, "Binaural Signal Detection: Vector Theory," in *Foundations of Modern Auditory Theory* (J. V. Tobias, ed.), vol. II, pp. 349–368, New York: Academic Press, 1972.

[199] H. S. Colburn, "Theory of binaural interaction based on auditory-nerve data. II. Detection of tones in noise," *J. Acoust. Soc. Am.*, vol. 61, pp. 525–533, Feb. 1977.

[200] R. M. Stern and G. D. Shear, "Lateralization and detection of low-frequency binaural stimuli: Effects of distribution of internal delay," *J. Acoust. Soc. Am.*, vol. 100, pp. 2278–2288, Oct. 1996.

[201] W. Lindemann, "Extension of a binaural cross-correlation model by means of contralateral inhibition. I. Simulation of lateralization for stationary signals," *J. Acoust. Soc. Am.*, vol. 80, no. 6, pp. 1608–1622, 1986.

[202] W. Lindemann, "Extension of a binaural cross-correlation model by means of contralateral inhibition. II The law of the first wavefront," *J. Acoust. Soc. Am.*, vol. 80, no. 6, pp. 1623–1630, 1986.

[203] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Am.*, vol. 116, pp. 3075–3089, Nov. 2004.

[204] J. Breebaart, S. van de Par, and A. Kohlrausch, "Binaural processing model based on contralateral inhibition. I. Model structure," *J. Acoust. Soc. Am.*, vol. 110, pp. 1074–1088, Aug. 2001.

[205] J. Breebaart, S. van de Par, and A. Kohlrausch, "Binaural processing model based on contralateral inhibition. I. Dependence on spectral parameters," *J. Acoust. Soc. Am.*, vol. 110, pp. 1089–1104, Aug. 2001.

[206] J. Breebaart, S. van de Par, and A. Kohlrausch, "Binaural processing model based on contralateral inhibition. II. Dependence on temporal parameters," *J. Acoust. Soc. Am.*, vol. 110, pp. 1105–1117, Aug. 2001.

[207] J. Braasch and J. Blauert, “The precedence effect for noise bursts of different bandwidths. II. Comparison of model algorithms,” *Acoust. Sci. Tech.*, vol. 24, pp. 293–303, Jul. 2003.

[208] J. Braasch, “A precedence effect model to simulate localization dominance using an adaptive, stimulus parameter-based inhibition process,” *J. Acoust. Soc. Am.*, vol. 134, pp. 420–435, Jul. 2013.

[209] N. I. Durlach, “Equalization and cancellation theory of binaural masking-level differences,” *J. Acoust. Soc. Am.*, vol. 35, pp. 1206–1218, Aug. 1963.

[210] M. Dietz, S. D. Ewert, V. Hohmann, and B. Kollmeier, “Coding of temporally fluctuating interaural timing disparities in a binaural processing model based on phase differences,” *Brain Res.*, vol. 1220, pp. 234–245, Mar. 2008.

[211] M. Dietz, S. D. Ewert, and V. Hohmann, “Lateralization of stimuli with independent fine-structure and envelope-based temporal disparities,” *J. Acoust. Soc. Am.*, vol. 125, pp. 1622–1635, Mar. 2009.

[212] M. Dietz, S. D. Ewert, and V. Hohmann, “Auditory model based direction estimation of concurrent speakers from binaural signals,” *Speech Comm.*, vol. 53, pp. 592–605, May 2011.

[213] W. A. van Bergeijk, “Variation on a Theme of Békésy: A Model of Binaural Interaction,” *J. Acoust. Soc. Am.*, vol. 34, pp. 1431–1437, Sept. 1962.

[214] G. C. Stecker, I. A. Harrington, and J. C. Middlebrooks, “Location Coding by Opponent Neural Populations in the Auditory Cortex,” *PLoS Biol*, vol. 3, pp. 520–528, Mar. 2005.

[215] T. J. Park, A. Klug, M. Holinstat, and B. Grothe, “Interaural Level Difference Processing in the Lateral Superior Olive and the Inferior Colliculus,” *J. Neurophys.*, vol. 92, pp. 289–301, Jul. 2004.

[216] T. C. Yin and J. C. K. Chan, “Interaural time sensitivity in medial superior olive of cat,” *J. Neurophysiol.*, vol. 64, pp. 465–488, 1990.

[217] D. McAlpine and B. Grothe, “Sound localization and delay lines - do mammals fit the model?,” *Trends Neurosci.*, vol. 26, pp. 347–350, May 2003.

[218] M. Pecka, A. Brand, O. Behrend, and B. Grothe, “Interaural time difference processing in the mammalian medial superior olive: the role of glycinergic inhibition,” *J. Neurosci.*, vol. 28, pp. 6914–6925, Jul. 2008.

[219] N. Salminen, H. Tiitinen, S. Yrttiaho, and P. J. C. May, “The neural code for interaural time difference in human auditory cortex,” *J. Acoust. Soc. Am. EL.*, vol. 127, pp. 60–65, Feb. 2010.

[220] T. S. Jennings and H. S. Colburn, “Models of the superior olivary complex,” in *Computational Models of the Auditory System* (R. Meddis, E. A. Lopez-Poveda, A. Popper, and R. R. Fay, eds.), Springer Handbook of Auditory Research, ch. 4, pp. 65–96, New York, NY: Springer-Verlag, 2010.

[221] H. S. Colburn, “Theory of binaural interaction based on auditory-nerve data. i. general strategy and preliminary results on interaural discrimination,” *J. Acoust. Soc. Am.*, vol. 54, no. 6, pp. 1458–1470, 1973.

[222] H. S. Colburn, Y. Han, and C. P. Culotta, “Coincidence model of mso responses,” *Hear. Res.*, vol. 49, pp. 335–346, Nov. 1990.

[223] A. R. Brughera, E. S. Stutman, L. H. Carney, and H. S. Colburn, “A model with excitation and inhibition for cellsin the medial superior olive,” *Audit. Neurosci.*, vol. 2, pp. 219–233, 1996.

[224] A. Brand, O. Behrend, T. Marquard, D. McAlpine, and B. Grothe, “Precise inhibition is essential for microsecond interaural time difference coding,” *Nature.*, vol. 417, pp. 543–547, May. 2002.

[225] K. E. Hancock and B. Delgutte, “A physiologically based model of interaural time difference discrimination,” *J. Neurosci.*, vol. 24, pp. 7110–7117, Aug. 2004.

[226] V. Pulkki and T. Hirvonen, “Functional count-comparison model for binaural decoding,” *Acta Acustica united with Acustica*, vol. 95, pp. 883–900, 2009.

[227] D. McAlpine, D. Jiang, and A. R. Palmer, “A neural code for low-frequency sound localization in mammals,” *Nat. Neurosci.*, vol. 4, pp. 396–401, Apr. 2001.

[228] M. C. Reed and J. J. Blum, “A model for the computation and encoding of azimuthal information by the lateral superior olive,” *J. Acoust. Soc. Am.*, vol. 88, pp. 1442–1453, Sep. 1990.

[229] L. Yue and D. H. Johnson, “Optimal binaural processing based on point process models of preprocessed cues,” *J. Acoust. Soc.Am.*, vol. 101, pp. 982–992, Feb. 1997.

[230] T. Hirvonen and V. Pulkki, “Interaural Coherence Estimation with Instantaneous ILD,” in *Proc. 7th Nordic Signal Processing Symposium (NORSIG 2006)*, (Reykjavik, Iceland), pp. 122–125, Jun. 7-9 2006.

[231] ITU, *Method for objective measurements of perceived audio quality ITU-R Recommendation BS.1387-1*, 1998.

[232] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Spoer, J. G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feitten, “PEAQ–The ITU Standard for Objective Measurement of Perceived Audio Quality,” *J. Audio Eng. Soc.*, vol. 48, pp. 3–29, Jan./Feb. 2000.

[233] C. Faller, “Binaural cue coding-Part I: psychoacoustic fundamentals and design principles,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 509–519, Nov. 2003.

[234] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. D. H. Purnhagen, J. K. J. Hilpert, W. J. Rödén W. Oomen, K. Linzmeier, and K. S. Chong, “MPEG surround-the ISO/MPEG standard for efficient and compatible multichannel audio coding,” *J. Audio Eng. Soc.*, vol. 56, pp. 932–955, Nov. 2008.

[235] T. Spoer, R. Bitto, and K. Brandenburg, “System and method for evaluating the quality of multi-channel audio signals.” US Patent 7,024,259, Apr. 2006.

[236] I. Choi, B. G. Shinn-Cunningham, S. B. Chon, and K.-M. Sung, “Objective measurement of perceived auditory quality in multichannel audio compression coding systems,” *J. Audio Eng. Soc.*, vol. 56, pp. 3–17, Jan./Feb. 2008.

[237] J. Liebertrau, T. Spoer, S. Kämpf, and S. Scnneider, “Standardization of PEAQ-MC: Extension of ITU-R BS.1387-1 to Multichannel Audio,” in *Proc. AES 40th Intl. Conf.*, (Tokyo, Japan), Oct. 8-10 2010. Paper No. 3.

[238] F. Rumsey, S. Zielinski, P. Jackson, M. Dewhirst, R. Conetta, S. Geogre, S. Bech, , and D. Meares, “QESTRAL (Part 1): Quality Evaluation of Spatial Transmission and Reproduction using an Artificial Listener,” in *Proc. AES 125th Convention*, (San Francisco, CA, USA), Oct. 2-5 2008.

[239] P. J. B. Jackson, M. Dewhirst, R. Conetta, S. Zielinski, F. Rumsey, D. Meares, S. Bech, and S. Geogre, “QESTRAL (Part 3): system and metrics for spatial quality prediction,” in *Proc. AES 125th Convention*, (San Francisco, CA, USA), Oct. 2-5 2008.

[240] A. Härmä, T. Lokki, and V. Pulkki, “Drawing quality maps of the sweet spot and its surroundings in multichannel reproduction and coding,” in *Proc. AES 21st Intl. Conf.*, (St. Petersburg, Russia), Jun. 1-3 2002. Paper No. 64.

[241] R. Conetta, F. Rumsey, S. Zielinski, P. Jackson, M. Dewhirst, S. Bech, D. Meares, and S. Geogre, “QESTRAL (Part 2): Calibrating the QESTRAL model using listening test data,” in *Proc. AES 125th Convention.*, (San Francisco, CA, USA), Oct. 2-5 2008.

[242] P. Jackson, M. Dewhirst, R. Conetta, and S. Zielinski, “Estimates of perceived spatial quality across the listening area,” in *Proc. AES 38th Intl. Conf.*, (Piteå, Sweden), Jun. 13-15 2010.

[243] H. Fletcher and W. A. Munson, “Loudness, its definition, measurement and calculation,” *J. Acoust. Soc. Am.*, vol. 9, pp. 1–10, Jul. 1937.

[244] E. Zwicker and B. Scharf, “A model of loudness summation,” *Psych. Rev.*, vol. 72, pp. 3–26, Jan. 1965.

[245] ISO, *Acoustics – Method for Calculating Loudness Level ISO Recommendation 532*, 1975.

[246] E. Zwicker, “Procedure for calculating loudness of temporally variable sounds,” *J. Acoust. Soc. Am.*, vol. 62, pp. 675–682, Sep. 1977.

[247] B. R. Glasberg and B. C. J. Moore, “A model of loudness applicable to time-varying sounds,” *J. Audio Eng. Soc.*, vol. 50, pp. 331–342, May 2002.

[248] B. C. J. Moore and B. Glasberg, “Modeling binaural loudness,” *J. Acoust. Soc. Am.*, vol. 121, pp. 1604–1612, Mar. 2007.

[249] B. C. J. Moore and C.-T. Tan, “Development and validation of a method for predicting the perceived naturalness of sounds subjected to spectral distortion,” *J. Audio Eng. Soc.*, vol. 114, pp. 408–419, Jul. 2003.

[250] C.-T. Tan, B. C. J. Moore, N. Zacharov, and V.-V. Mattila, “Predicting the perceived quality of nonlinearly distorted music and speech signals,” *J. Audio Eng. Soc.*, vol. 52, pp. 699–711, Jul./Aug. 2004.

[251] B. C. J. Moore and C.-T. Tan, "Perceived naturalness of spectrally distorted speech and music," *J. Acoust. Soc. Am.*, vol. 52, pp. 900–914, Sep. 2004.

[252] V. Pulkki, M. Karjalainen, and J. Huopaniemi, "Analyzing Virtual Sound Source Attributes Using a Binaural Auditory Model," *J. Audio Eng. Soc.*, vol. 47, pp. 203–217, Apr. 1999.

[253] J. Braasch, "A binaural model to predict position and extension of spatial images created with standard sound recording techniques," in *Proc. AES 119th Convention.*, (New York, NY, USA), Oct. 7-10 2005.

[254] R. Baumgartner, P. Majdak, and B. Laback, "Assessment of Sagittal-Plane Sound Localization Performance in Spatial-Audio Applications," in *The Technology of Binaural Listening* (J. Blauert, ed.), pp. 93–119, Springer-Verlag, Berlin, Germany, 2013.

[255] 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 *Proc. AES 129th Convention*, (San Francisco, CA, USA), p. Paper No. 8296, Nov. 4-7 2010.

[256] D. W. Grantham and F. L. Wightman, "Detectability of varying interaural temporal differences," *J. Acoust. Soc. Am.*, vol. 63, pp. 511–523, Feb. 1978.

[257] J. Becker, "Spectral and Temporal Contribution of Different Signals to ASW Analyzed with Binaural Hearing Models," in *Proc. of the Forum Acousticum*, (Sevilla, Spain), Sep. 16-20 2002.

[258] W. Hess and J. Blauert, "Evaluation of auditory spatial impression in performance places," in *Proc. Forum Acousticum*, (Budapest, Hungary), Aug. 29- Sep. 2 2005.

[259] J. van Dorp Schuitman, D. de Vries, and A. Linday, "Deriving content-specific measures of room acoustic perception using a binaural, nonlinear auditory model," *J. Acoust. Soc. Am.*, vol. 133, pp. 1572–1585, Mar. 2013.

[260] S. George, S. Zielinski, F. Rumsey, P. Jackson, R. Conetta, M. Dewhirst, D. Meares, and S. Bech, "Development and validation of an unintrusive model for predicting the sensation of envelopment arising from surround sound recordings," *J Audio Eng. Soc.*, vol. 58, pp. 1013–1031, Dec. 2010.

[261] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, and G. Lorho, "Augmented reality audio for mobile and wearable appliances," *J. Audio Eng. Soc.*, vol. 52, pp. 618–639, June 2004.

Aalto-DD 109/2014

9 789526 057828

ISBN 978-952-60-5782-8
ISBN 978-952-60-5783-5 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

Aalto University
School of Electrical Engineering
Department of Signal Processing and Acoustics
www.aalto.fi