Helsinki University of Technology Laboratory of Acoustics and Audio Signal Processing

Espoo 2007 Report 83

PERCEPTUAL AND MODELING STUDIES ON SPATIAL SOUND

Toni Hirvonen

TEKNILLINEN KORKEAKOULU
TEKNISKA HÖGSKOLAN
HELSINKI UNIVERSITY OF TECHNOLOGY
TECHNISCHE UNIVERSITÄT HELSINKI
UNIVERSITE DE TECHNOLOGIE D'HELSINKI

Helsinki University of Technology Laboratory of Acoustics and Audio Signal Processing

Espoo 2007 Report 83

PERCEPTUAL AND MODELING STUDIES ON SPATIAL SOUND

Toni Hirvonen

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission for public examination and debate in Auditorium S4, Department of Electrical and Communications Engineering, Helsinki University of Technology, Espoo, Finland, on the 7th of December 2007, at 12 o'clock noon.

Helsinki University of Technology
Department of Electrical and Communications Engineering
Laboratory of Acoustics and Audio Signal Processing

Teknillinen korkeakoulu
Sähkö- ja tietoliikennetekniikan osasto
Akustiikan ja äänenkäsittelytekniikan laboratorio

Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing
P.O. Box 3000
FIN-02015 TKK
Tel. +358 9 4511
Fax +358 9 460 224
E-mail lea.soderman@tkk.fi

ISBN 978-951-22-9050-5
ISSN 1456-6303

Multiprint Oy
Espoo, Finland 2007

HELSINKI UNIVERSITY OF TECHNOLOGY P. O. BOX 1000, FI-02015 TKK http://www.tkk.fi	ABSTRACT OF DOCTORAL DISSERTATION

Author Toni Hirvonen

Name of the dissertation

Perceptual and Modeling Studies on Spatial Sound

Date of manuscript 29.10.2007	Date of the dissertation 7.12.2007
□ Monograph	☒ Article dissertation (summary + original articles)

Department	Department of Electrical and Communications Engineering
Laboratory	Laboratory of Acoustics and Audio Signal Processing
Field of research	Acoustics, spatial hearing, multichannel reproduction, auditory modeling
Opponent(s)	Prof. Jens Blauert
Supervisor	Prof. Matti Karjalainen
(Instructor)	D. Sc. (Tech.) Ville Pulkki

Abstract

Humans have the ability to perceive various spatial auditory attributes, such as the localization and width of sound sources. The study of spatial hearing is important not only in terms of basic perceptual research, but also because ever more sophisticated audio reproduction algorithms and systems are introduced to consumers. From such systems, listeners regularly perceive complicated spatial auditory scenes involving several simultaneous sounds from different directions. These scenes can be thought as being complex ones, as opposed to perceiving a single, point-like source in an anechoic environment.

The first part of this thesis investigates the perceptual issues related to such complex sound scenes via subjective listening tests. A single anechoic source results in localization cues that most listeners unambiguously interpret as indicating the actual direction of the sound. In the case of several interfering sound sources, the cues may vary greatly as a function of frequency. As illustrated by the results presented here, this is a common occurrence in modern multichannel reproduction systems. To gain further insight on this little-researched phenomenon, specific test cases where localization cues were manipulated as a function of frequency in the horizontal plane were investigated. The subjects reported the localization and width of the complex sounds, and these responses revealed several interesting phenomena. Most importantly,the listeners always perceived a horizontally wide sound source as being much narrower than it's physical width. Strong perceptual contrasts were also found to be significant.

Another focus of this thesis is auditory modeling. The stimuli used in the previous experiments were simulated utilizing established auditory modeling techniques. The simulation results were not found to correspond entirely with the psychoacoustical results in all cases , prompting additional weighting of different frequencies in the modeling. This thesis also introduces a novel, general auditory model concept inspired by recent psychoacoustical results that partly contradict the previous modeling approaches. The model's capacity to account for common spatial hearing phenomena was examined. The initial simulation results validate the proposed concept. Quantitative comparisons with psychoacoustical results, including the data obtained from the listening tests performed in this thesis, are planned to be done in the future.

Keywords Spatial hearing, Auditory modeling, Virtual Sources

ISBN (printed) 978-951-22-9050-5	ISSN (printed) 1456-6303
ISBN (pdf) 978-951-22-9051-2	ISSN (pdf)
ISBN (others)	Number of pages 152

Publisher Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing

Print distribution Report 83 / TKK, Laboratory of Acoustics and Audio Signal Processing, Espoo, Finland

☒ The dissertation can be read at http://lib.tkk.fi/Diss/

TEKNILLINEN KORKEAKOULU PL 1000, 02015 TKK http://www.tkk.fi	VÄITÖSKIRJAN TIIVISTELMÄ

Tekijä Toni Hirvonen

Väitöskirjan nimi

Tutkimuksia tilaäänen havaitsemisesta ja mallinnuksesta

Käsikirjoituksen jättämispäivämäärä 29.10.2007	Väitöstilaisuuden ajankohta 7.12.2007
☐ Monografia	☒ Yhdistelmäväitöskirja (yhteenveto + erillisartikkelit)

Osasto	Sähkö- ja tietoliikennetekniikan osasto
Laboratorio	Akustiikan ja äänenkäsittelytekniikan laboratorio
Tutkimusala	Alustiikka, Tilakuuleminen
Vastaväittäjä(t)	Prof. Jens Blauert
Työn valvoja	Prof. Matti Karjalainen
(Työn ohjaaja)	TkT Ville Pulkki

Tiivistelmä

Ihmisillä on kyky havainnoida äänen tilaominaisuuksia, kuten äänilähteen lokalisaatiota ja leveyttä. Tilaäänen tutkimus on tärkeää sekä havainnoinnin perustutkimuksen, että myös kuluttajille tarjottavien hienostuneiden äänentoistojärjestelmien yleistymisen kannalta. Tälläiset järjestelmät tuottavat kuulijoille äänimaisemia, jotka koostuvat useista samanaikaisista ja erisuuntaisista lähteistä. Tälläiset äänimaisemat voidaan ajatella monimutkaisiksi verrattuna yhteen pistemäiseen lähteeseen heijastuksettomassa tilassa.

Tämän väitöskirjan ensimmäinen osa tutkii edellisten kaltaisiin monimutkaisiin äänimaisemiin liittyvää havainnointia subjektiivisilla kuuntelukokeilla. Yksittäinen äänilähde kaiuttomassa tilassa tuottaa lokalisaatiovihjeitä, jotka useimmat kuulijat assosioivat äänilähteen todelliseen suuntaan. Useiden lähteiden tapauksessa vihjeet taas voivat vahdella huomattavasti taajuuden funktiona. Kuten työn tuloksista nähdään, tälläiset tilanteet ovat yleisiä moderneissä äänentoistojärjestemissä. Ilmiön tutkimista varten konstruoitiin kuuntelukoenäytteitä, joissa horisontaalitason suuntavihjeitä manipuloitiin halutulla tavalla. Kuulijat ilmaisivat kompleksisten ääninaytteiden lokalisaation ja leveyden ja nämä tulokset paljastivat lukuisia mielenkiintoisia ilmiöitä. Kuulijat eivät koskaan havainneet fyysiesti leveän äänilähteen koko leveyttä, vaan arvoivat leveyden paljon vähäisemmäksi. Vahvojen perkeptuaalisten kontrastien vaikutus havaittiin myös merkittäväksi.

Väitöskirjan toinen tutkimuskohde on auditorinen mallinus. Edellämainittujen kokeiden koenäytteet simuloitiin käyttäen tunnettuja mallinnustekniikoita. Simulaatiotulokset eivät täysin vastanneet kuuntelukokeen tuloksia kaikissa tapauksissa, josta johtuen ehdotettiin eri taajuuskaistojen lisäpainotusta mallinnuksessa. Väitöskirja esittelee myös uudenlaisen yleisen auditorisen mallin konseptin, johon viimeaikaiset, vanhojen mallinnustekniikoiden kanssa ristiriitaiset, neurofysiologiset tulokset ovat vaikuttaneet. Mallin kykyä selittää yksinkertaisia psykoakustisia ilmiöitä tutkittiin. Alustavat simulaatiotulokset validoivat ehdotetun konseptin. Kvantitatiivisten vertailujen teko psykoakustisten tulosten, mukaalukien tämän väitöskirjan näytteiden, kanssa on suunnitteilla tulevaisuudessa.

Asiasanat Tilaääni, Kuulon mallinnus, Virtuaalilähteet

ISBN (painettu) 978-951-22-9050-5	ISSN (painettu) 1456-6303
ISBN (pdf) 978-951-22-9051-2	ISSN (pdf)
ISBN (muut)	Sivumäärä 152

Julkaisija Teknillinen korkeakoulu, Akustiikan ja äänenkäsittelytekniikan laboratorio

Painetun väitöskirjan jakelu Report 83 / TKK, Akustiikan ja äänenkäsittelytekniikan laboratorio, Espoo, Suomi

☒ Luettavissa verkossa osoitteessa http://lib.tkk.fi/Diss/

Preface

As with many people I have worked with, it was mainly the enjoyment I got from music that inspired me to study acoustics during my undergraduate years. After finishing my masters thesis in 2002, I became interested in doing some kind of acoustical research in an academic environment. I started working at the Helsinki University of Technology (TKK) Laboratory of Acoustics and Audio Signal Processing the same year, and continued to carry out the research presented here from 2003 to 2007. During these years, I feel like that I learned many, many things.

Here are some people I would like to acknowledge: My award-winning thesis instructor Dr. Ville Pulkki was the key person who enabled me to finish this book. Ville not only hired me but he also ended up being my sole co-author in the publications included in this thesis. My supervisor Prof. Matti Karjalainen is a legendary figure in the acoustics lab and always a reliable source of information. The expertise and innovative ideas of Ville and Matti continue to be an inspiration.

People who have had some work- or study-related dealings with me : The other three professors of the acoustics lab: Prof. Unto K. Laine, Prof. Paavo Alku, and Prof. Vesa Välimäki. My roommates Mr. Carlo Magi and Mr. Petri Korhonen. Colleagues with whom I have worked with: Dr. Juha Merimaa, Dr. Antti Kelloniemi, Dr. Henri Penttinen, Dr. Laura Lehto, Dr. Tuomas Paatero, Mr. Jukka Ahonen, Mr. Miikka Tikander, Mr. Jyri Pakarinen, Mr. Matti Airas, Ms. Heidi-Maria Lehtonen, and the rest of the aku-staff. Special appreciation goes to Mrs. Lea Söderman and Mr. Jussi Hynninen for handling many practical issues.

I have received funding from the Academy of Finland (projects no. 201050 and 105780) and from Emil Aaltonen Foundation. Finally, thanks to my mother and father, to the rest of my family, and to my friends (you know who you are).

Contents

List of Publications

This thesis consists of an overview and of the following publications:

I Pulkki, V. and Hirvonen, T., ”Localization of Virtual Sources in Multichannel Audio Reproduction”, IEEE Transactions on Speech and Audio Processing, Vol 13, No. 1, Jan 2005.

II Hirvonen, T. and Pulkki, V., ”Center and Spatial Extent of Auditory Events as Caused by Multiple Sound Sources in Frequency-Dependent Directions”, Acta Acustica united with Acustica, Vol 92, No. 2, Jan 2006.

III Hirvonen, T., ”Segregation of Two Simultaneously Arriving Narrowband Noise Signals as a Function of Spatial and Frequency Separation”, in Proceedings of 8th International Conference on Digital Audio Effects, (DAFx’05), Madrid, Spain, September 20-22, 2005.

IV Hirvonen, T. and Pulkki V., ”Perception and Analysis of Selected Auditory Events with Frequency-Dependent Directions”, Journal of the Audio Engineering Society , Vol 54, No. 9, Sep 2006.

V Hirvonen, T. and Pulkki, V., ”Interaural Coherence Estimation with Instantaneous ILD”, in Proceedings of 7th Nordic Signal Processing Symposiun (NORSIG 2006), Reykjavik, Iceland, June 7-9, 2006.

VI Hirvonen, T. and Pulkki, V., ”Predicting Binaural Masking Level Difference and Dichotic Pitch Using Instantaneous ILD Model”, AES 30th Int. Conference, 2007.

VII Pulkki, V. and Hirvonen, T., ”Computational Count-Comparison Models for ITD and ILD decoding”, ICA 19th Int. Congress on Acoustics, 2007.

Author's contribution

Publication I

The first author of the paper is mainly responsible for this research. The author of the thesis implemented and conducted the listening tests and reported them in Section 6, as well as contributed in revising the inital draft.

Publication II

Planning and revising of this paper was done in collaboration with the second author. The present author conducted the listening tests and wrote the initial draft of the paper.

Publication III

The present author is responsible for this research.

Publication IV

Planning and revising of this paper was done in collaboration with the second author. The present author conducted the listening tests and wrote the initial draft of the paper.

Publication V

Planning and revising of this paper was done in collaboration with the second author. The present author conducted the model simulations and wrote the initial draft of the paper.

Publication VI

Planning and revising of this paper was done in collaboration with the second author. The present author was responsible for the model simulations and wrote the initial draft of the paper.

Publication VII

Planning and revising of this paper was done in collaboration with the first author. The first author was responsible for the model simulations and wrote the initial draft of the paper.

List of Abbreviations

ANOVA	Analysis of Variance
ASW	Auditory Source Width
BEP	Binaural Edge Pitch
BMLD	Binaural Masking Level Difference
CB	Critical Band
CN	Cochlear Nucleus
CF	Characteristic Frequency
DeCo	Decorrelation signal output of auditory model
EC	Equalization-Cancellation
EE	Excitation-Excitation
ERB	Equivalent Rectangular Bandwidth
GTFB	Gammatone Filterbank
HP	Huggins Pitch
HRTF	Head-Related Transfer Function
IACC	Interaural Cross-Correlation
IE	Inhibition-Excitation
IHC	Inner Hair Cell
IIR	Infinite Impulse Response
ILD	Interaural Level Difference
ILDA	Interaural Level Difference Angle
ITD	Interaural Time Difference
ITDA	Interaural Time Difference Angle
K-S	Kolmogorov-Smirnov
LSO	Lateral Superior Olive
MAA	Minimum Audible Angle
MSO	Medial Superior Olive
MGB	Medial Geniculate Body
OHC	Outer Hair Cell
RMS	Root Mean Square
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level

List of Symbols

D	RMS deviation of reproduction system
$\langle \bar{D} \rangle$	run RMS displacement of reproduction system
$\langle \bar{E} \rangle$	mean displacement error of reproduction system
$f(\Theta)$	microphone polar pattern as a function of azimuth angle
g	loudspeaker gain factor
N_0	diotic noise signal
Θ	azimuth angle
τ	time constant
ϕ	elevation angle
p	p-value, probability of null hypothesis
ρ	interaural correlation
$\langle \bar{s} \rangle$	run standard deviation of reproduction system displacement
S_π	signal with a 180° interaural phase shift
S_m	monaural signal
w	perceptual weight factor

1 Introduction

According to an old legend, the famous Greek skeptic Pyrrho died when he walked over a cliff [1]. He supposedly demonstrated his extreme skepticism by walking blindfolded toward the edge and refusing to believe that the verbal warnings of his observing students were in fact real at all. True or not, the Pyrrho myth contains a couple of valuable lessons. Most importantly, the story seems to say that one should be observant of the surrounding world or at the very least accept that it is not an illusion. This was especially important in ancient and prehistorical times when the world presented serious threats to survival on daily basis.

So how does one go about observing the world? Senses are generally defined as faculties by which outside stimuli are perceived [2]. Furthermore, many philosophers agree that sensory input is the only way of attaining information from the outside world. A certain type of sensory cell responds to a certain type of stimulus and transmits signals to a region of the brain where the signals are interpreted. Although humans have at least seven senses, the most important ones are taught to be the classical five defined in ancient times: sight, hearing, touch, smell, and taste. However, there is no clear consensus on what constitutes as a sense, or how the signals from various sensory cells are mapped to the brain. This is one of the reasons why it is helpful to consider human perception as a whole, as is done in Gestalt psychology [3]. Its basic idea is that humans form their experiences based on holistic and parallel processing of different sensory inputs.

The opposite of holism is reductionism, the idea that a complex phenomenon or system can be described by the sum of its simpler parts [4]. Much of the everyday scientific research on perception is based on this idea, as scientists usually work on a narrow, specific area. At the same time however, the majority of scientific community is against extreme or ”greedy” reductionism where everything is eventually explained by particle physics. There are numerous examples that contradict reductionism, such as the fact that music cannot be experienced listening to individual notes, and simple molecules that constitute organisms are not considered living. Nevertheless, only during this century have holistic ideas, such as Gestalt psychology and systems theory [5], gained major scientific acceptance. This can be partly attributed to emergence of sophisticated experimental methods that allow the complex stimuli and analysis.

This thesis is concerned with one of the classical senses, hearing. If interpreted freely, the myth of Pyrrho suggests an interesting presumption about our auditory perception: only when the most important sense, sight, is impaired, does it become necessary to rely on hearing as the primary source of information. Visual dominance over hearing has been shown also in modern experiments, for example with the ventriloquism effect, where a voice is perceived to emerge from other than the actual sound source [6]. There visual domination is however not universal. In accordance to Gestalt principles, conflicting sensory information can result in a totally unexpected perception, as demonstrated by the McGurk effect and related experiments [7, 8]. In the original experiment, a video of one phoneme's production is dubbed with a recording of a different phoneme being spoken. Often, the perceived phoneme is a neither, but rather some intermediate phoneme. In any case, the holistic principles should be kept in mind in hearing research.

From a historical viewpoint, most people associate hearing and auditory phenomena with verbal communication and music. These enable efficient means of exchanging information and obtaining artistic pleasure. Somewhat less familiar attribute of hearing is the ability to spatially analyze the auditory scene, i.e., to distinguish different sounds and the directions from which they emanate. Evolutionarily speaking, the ability to pinpoint the direction of a threatening sound seems important. It can be however argued that the evolutionary role of hearing in localization is simply to give a rough direction of the sound so that the eyes can be turned towards the sound.

The modern era, on the other hand, has seen the development of telecommunication and entertainment systems that make complex auditory scenes an everyday occurrence. Nowadays, sound engineers can ever more accurately control the sounds arriving to the listener's ears and want to know how they are perceived. Consequently, the research fields of spatial hearing [9], auditory object analysis [10], auditory scene analysis [11], and ecological psychoacoustics [12] have been born.

1.1 Scope of the Thesis

Narrowing down the broad topic of auditory perception, it can be said that spatial hearing is the main focus of this thesis. As with all senses, there is a relation between the physical space (sound sources) and the perceived auditory space (auditory objects and events) [9]. Clarifying this relationship is the underlying motivation here, subjective listening tests being the basic method of investigation. Listening test data is obtained from both original experiments and previous research. The approach taken in here is similar to general subjective testing, where the goal is to use the listener as an objective measuring tool similar to, e.g., a microphone [13].

This thesis work can be thought as a continuation to the work of Pulkki on spatial hearing, virtual auditory sources, and auditory reproduction in general [14]. The work started with a research about commonly used multiple-loudspeaker systems. These systems produced interesting auditory scenes that were analyzed with measurements and tested with human listeners. The scenes contained sometimes uncommonly behaving spatial cues, which humans use to perceive the direction of the sound. The open questions of this research led to a series of listening tests where complex auditory scenes were examined more precisely. The tests utilized a fairly complex setup which allowed for multi-source auditory stimuli. The aim was to apply the holistic principles at least to a certain extent, and not to focus on simplest possible auditory phenomena or use a simulated reproduction of complex auditory scenes.

Nevertheless, the approach taken in the experiments presented in this thesis is ultimately, if not reductionist, at least generalizing, compared to complicated audiovisual scenes with music or speech stimuli and room reverberation. In order to keep the number of test cases and research questions within available resources, the stimuli were limited so that temporal effects were minimized and the test signals were random noise.

Auditory modeling research is the other focus of this thesis. Here, the term refers to the use of computational scientific modeling techniques to somehow simulate the human auditory system. Although the subjective investigations discussed above included some modeling of the results, there is an abrupt shift in focus from perceptual research, represented by the first four publications of this dissertation, to experimental auditory modeling in the last three publications. This was motivated by the modeling studies performed at the end of the performed perceptual studies. During the analysis of the auditory scenes used in listening tests, it seemed that the presently utilized modeling techniques did not describe well the listeners' responses. This along with the recent neurophysiological results inspired to examine a novel idea for auditory modeling.

However, the model was not intended to completely solve the issues that rose from the subjective results within the scope of the present dissertation. Rather, the purpose was to examine and model some simple spatial hearing phenomena and thus validate the novel modeling concept. It should also be kept in mind that the suggested model is intended to be a general auditory model and arguably more based on human physiology than other similar models. Much research is needed in order to fully develop such a model. However, in order to provide a fitting closure for the dissertation, the present state of the model development, as well as some informal experiments are discussed in Section 7.

In conclusion, the author carefully suggests that these studies aimed for a slightly more holistic approach than is generally utilized with similar topics. Alas, however, some generalization is necessary in scientific research: in order to make predictions, phenomena must be classified to a limited number of categories and assume that these behave similarly in the future as they do in the past. It is the wish of the author that the information discarded this way in this research is not as essential as in the legend of Pyrrho.

1.2 Organization of the Thesis

The thesis consists of this introduction and seven articles that detail the research done by the author. The relevant results from the previous studies are discussed in the following chapters of the introduction. First, the physiology of the hearing system is introduced. The physiological information is strongly in the background when discussing hearing as perception in a separate chapter. Finally, auditory modeling, as well as scientific modeling in general, are discussed. All chapters deal mainly with cues and phenomena related to spatial hearing.

2 Physiology of Hearing

This chapter aims to give a brief overview of the mechanisms of the auditory system as to where they are relevant to the perceptual issues discussed in Chapter 3. The anatomy of the human auditory system has been studied extensively over the years and not all its aspects are covered here. For a more profound analysis, see e.g., [15]. There is a consensus of the relevant parts that constitute the trail ascending from the ear, where the sound pressure is transformed into neural impulses, to the brain cortex, where perceptions are formed. Smaller parts and organs are here associated to either of the two larger wholes: the hearing organ and the auditory pathway. Although the examination is mostly for a single side, it should be kept in mind that many parts of the human hearing system are essentially symmetrical on both sides of the head; human hearing has two channels which interact with each other. This fact will play greater part in Section 3 but interaction between the channels is noted also in this section when appropriate.

2.1 Hearing Organ

The function of the hearing organ is to transform the variations in sound pressure to neural impulses. The hearing organ is constituted of the outer, the middle, and the inner ear, which can be seen in figure 2.1. When talking about the physiology and function of the hearing organ, one should also mention the rest of the human body. The reason for this is that the shape of the body notably affects the incoming sound waves. While pinnae and the rest of the head cause the most significant effects, reflections from shoulders and the torso should not be ignored [15]. The total effect of the body is referred to as the Head-Related Transfer Function, or HRTF [16].

The HRTF is also always a two-channel function, defined for both ears separately.

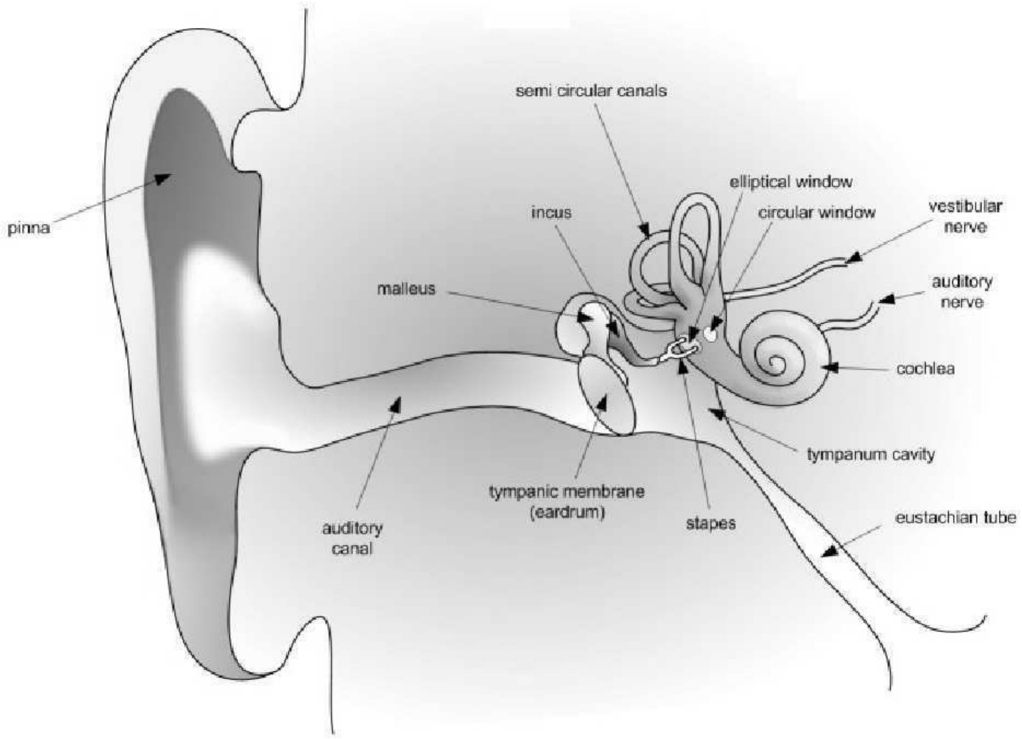

Figure 2.1: The anatomy of the human hearing organ. (Figure from [17])

If the sound emanates from either side of the head, the arrival times to the ears are different. The maximum time difference is approximately 0.7 ms for a normal human head The side which the sound reaches first is called ipsilateral, the other side being contralateral. Because of wave propagation around the head, the sound amplitude is attenuated at high frequencies approximately above 1.5 kHz on the contralateral side.

The visible part of the outer ear is called the pinna. Its folds act as direction-dependent filters at frequencies above approximately a few kHz, and it amplifies the sound to the inside of the ear. The pinna surrounds the ear canal, a tube about 26 mm in length and 7 mm in diameter. The ear canal acts as a linear resonator for the incoming sound waves around the 3.3 kHz frequency. The ear canal also protects the ear drum, where the outer ear terminates. The total effect of the outer ear is comparable to a direction-dependent linear filter that generally amplifies frequencies in the 2-4 kHz range up to 15 dB.

The middle ear consists of a structure of three bones (malleus, incus, and stapes) known as ossciles. Their function is to couple the vibrations of the ear drum to the oval, i.e. elliptical, window, the ”input” of the cochlea. As the cochlea is filled with fluid, the impedance differences would cause a 30 dB reduction in acoustic power between the eardrum and the cochlea if there were not such a coupling device [18].

From the auditory viewpoint, the inner ear consists of the cochlea. It is the crucial site where pressure vibrations are encoded to neural signals that then ascend up the auditory pathway. The cochlea is essentially a curved tube whose volume is divided into several membrane-separated, fluid-filled compartments along its length. In the middle of the tube, on top of the basilar membrane, is the organ of Corti, a structure lined with inner and outer hair cells (IHCs and OHCs, respectively).

According to the classical functional view, the incoming vibrations from the elliptical window propagate in the fluids and the membranes so the basilar membrane vibrates, causing the hair cells move against the tectorial membrane and send out neural spikes in the auditory nerve, to which the hair cells are connected [19]. The organ of Corti has about 20000-30000 nerve receptors distributed along the length of the cochlea. The basilar membrane has a maximum displacement depending on the frequency of the vibration so that low- and high-frequency sounds cause a maxima at the end of the cochlea and near the elliptical window, respectively. It is commonly thought that the cochlea acts as a spectrum analyzer or similarly to a band-pass filter bank.

More profound investigations have revealed some additional facts about the cochlea. The number of OHCs exceeds the number of IHCs over three times. OHCs are connected to the efferent neurons, which carry signals back from the central nervous system. Although OHCs are generally thought to implement mainly dynamic compression and suppression of the sidebands [20], it is interesting to find feedback this powerful already in the peripheral neural stages. It has also been suggested that the position of the hair cell on the basilar membrane is not the only factor to the cochlear frequency selectivity, but that some inner and outer hair cells themselves could respond to certain frequencies better than to others [21]. Although the spectrum analyzer analogy seems to hold well in most situations, the function of the cochlea is not precisely known at present.

2.2 Auditory Pathway

The neural pathway from the sensory cells to the brain cortex is the most complex among senses for hearing. This notion is based on the number of nuclei involved, the numerous connections between them, and the apparent complexity of the neural processing, especially between the two ear signals. Figure 2.2 illustrates the relevant nuclei and neural paths and will be used as a basis for this section. This physiological overview is mostly based on firmly established facts given in, e.g., [22]. More recent research and functional speculations are discussed in the next chapter.

The cochlear nerve is the first site of neuronal processing of the transformed data from the inner ear. Several interesting observations have been made already on this early stage. Firstly, when recording responses of cochlear nerve fibers, it can be seen that different fibers have different tuning curves, i.e., produce most activation with different frequencies [22]. The ”resonant” frequency of a given fiber is called its characteristic frequency (CF). However, the activation is not independent of level; as the intensity of a pure tone increases, the more it activates fibers whose CF is not the tone frequency [23]. This is somewhat unintuitive considering the human frequency resolution (see Section 3.1). Furthermore, the fibers vary greatly in rate of spontaneous activity, activation threshold, and post-stimulus behavior.

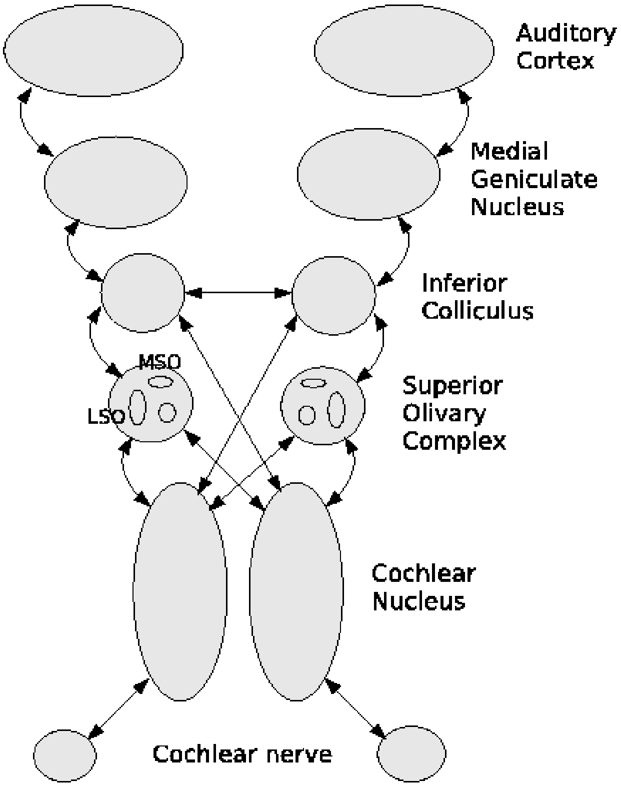

Figure 2.2: A schematic view of the nuclei in the left and right sides of the human auditory pathway and the connections between them.

One of the most interesting properties of auditory neural processing is the loss of phase-locking to the temporal structure of stimulus at high frequencies. This phenomenon can already be seen at the level of the cochlear nerve. As indicated by experiments with other mammals [24, 25], phase-locking starts to decline already below 1 kHz and is no longer detectable at frequencies above a few kHz. The loss of phase-locking generally worsens when moving to higher processing stages, which indicates that processes that depend on the temporal precision of the neural impulses must be achieved already at the lower levels.

The Cochlear Nucleus (CN) is the first relay station in the auditory system. At the level of the CN, the neural signals from each ear remain separated. A prominent feature of the CN that can be detected in all major nuclei of the auditory system is tonotopic organization: the CFs of auditory neurons in the CN are organized so that the neurons in the same area respond to same frequencies. However, tonotopic organization becomes more complex in the higher processing stages. The neuron tuning curves are wider and more complicated already at the level of CN than at the cochlea [26].

Ascending to Superior Olivary Complex (SOC), the auditory system begins to process the neural information binaurally, as both SOCs receive inputs from the ipsilateral and the contralateral CN. Two important ”sub-nuclei” inside the SOC are the low-frequency-oriented Medial Superior Olive and the high-frequency oriented Lateral Superior Olive (MSO and LSO, respectively). The neurons in these nuclei display the two basic processes in binaural processing: excitation and inhibition. Most LSO neurons are generally thought as the previous type, with the contralateral signal providing inhibition of the ipsilateral signal excitation (EI-type). Neurons in the MSO mainly receive excitatory input from both ears (EE-type), although important inhibitionary mechanisms have also been discovered in MSO [27].

Inferior Colliculus (IC) and the Medial Geniculate Nucleus (MGN) are the next nuclei in the pathway. From here upwards, any description of the neural mechanisms includes much speculation. Interaction between the left and right channels at the IC level indicates further interaural processing, possibly combining horizontal and elevation information. There is general evidence of tonotopic organization and both IE- and EE-type neurons in both IC and MGN.

Through the mid-brain, the neural signals travel to the Primary Auditory Cortex, which is the most highly organized processing unit of sound in the brain. The tonotopic organization is ever more complex here. Investigations of mammalian cortex lateral belts with bandpass noise stimuli have revealed that frequency is represented in several areas simultaneously [28–30]. The multiple representations could possibly be used to analyze the spectral components with multiple bandwidths and frequency resolutions simultaneously [31]. Organization of cortex neurons based on stimulus amplitude modulation, as well as the spatial direction of the stimulus, in addition to tonotopic organization have been speculated but not confirmed. One difficulty of determining the higher neural mechanisms is their strong dependence on the type of stimulus (e.g., tone vs. speech). Finally, it should be remembered that the cortex also provides input to the lower processing stages.

In summary, the auditory pathway most probably includes numerous mechanisms that can be adaptively used to determine the precept of a given stimulus. Also, research on the plasticity of the neural system indicates that these mechanisms can change over time [32].

3 Perception of Spatial Sound and Related Phenomena

This section discusses the auditory perception limited to spatial hearing and other issues related to this thesis. Many of these phenomena are commonly thought to stem mainly from the mechanisms of the peripheral or lower levels of the auditory system, whose physiology is better known than that of the higher levels. For this reason, some functional hypotheses are also discussed here. It is of course clear that the percept itself is not formed until at the cortex, and so all perception in principle involves every level of the given sensory system. In any case, even if the assumption of the low-level mechanism dominance is accepted, perception itself cannot be seen physiologically and must be measured with subjective tests.

3.1 Frequency Resolution

Depending on the person, humans can hear frequencies approximately between 20 Hz and 20 kHz (lower frequencies can be sensed via tactile input). Outside this range, the basilar membrane does not vibrate, hence there is no produced activity in the auditory system. When the basilar membrane does vibrate at some audible frequency, it often causes frequency masking: if sounds significantly different in level are close each other in frequency, the vibrations caused by the louder sound overpower those caused by the quieter sound at the same place of the membrane and make it inaudible.

The term ”frequency resolution” refers to the notion that spectral components within a certain frequency band are processed together as in the masking example above. Ever since the experiments by Fletcher [33], it has been the assumption that the peripheral auditory system functions as an array of overlapping passband filters. The effective bandwidth of each filter is called the critical band (CB). Glasberg and Moore used a method derived from Fletcher to determine the properties of the auditory filters; the detection threshold of a signal tone, masked by a noise notched around the signal frequency, is measured as a function of the notch bandwidth [34]. The experiments resulted in the Equivalent Rectangular Bandwidth (ERB), which approximates the CB as a function of its center frequency by the equation:

$$\mathrm{ERB}(f_c) = 24.7(4.37 * f_c + 1), \tag{3.1}$$

where f_c is the center frequency in kHz.

In addition to psychoacoustical measurements, the ERB scale corresponds closely to physiological results [35, 36], and is therefore widely used. For a more complex approximation, additional facts revealed by the masking experiments can be taken into account. The auditory filters are not rectangular but shaped more like asymmetrical cones, and their shape depends on the level of the sound. Also, it has been noted that the CB varies for different stimulus types, e.g., by being wider for speech [18], and narrower for forward-masking stimuli [37]. Several implementations have been proposed to account for the nonlinear nature of frequency selectivity, one example being [38].

3.2 Interaural Difference Cues

Even though the visual sense is dominant in perception, human hearing is capable of many extraordinary feats, which may be attributed to the complexity of the auditory system. Listeners are able to localize sounds with accuracy; at best, the minimum audible angle (MAA) is about 1° azimuth at the frontal direction [9]. Additionally, a complex scene that results from multiple sound sources and reflections can be effectively analyzed, as illustrated by the cocktail-party effect. Researchers of spatial hearing have identified numerous cues that humans use for these tasks. This section is focused on two such prominent cues: interaural time difference (ITD) and interaural level difference (ILD).

The modern view of these interaural cues dates back to Lord Rayleigh's duplex theorem [39]. With the assumption of a spherical head without pinnae, he theorized the existence of the two (hence 'duplex') cues induced by the distance between the ears and level difference caused by head shadowing. It has since been recognized that ITD and ILD are mainly responsible for the sound localization in the horizontal plane [9, 40].

Maintaining the head symmetry assumption, the same interaural cues can be produced from several locations as long as the sound incidence angle to the ear remains unchanged. The spherical cone-shaped area where this happens is called the cone of confusion (see figure 3.1). Even though spectral HRTF information can be used to resolve the direction of sound within the cone, listeners may have difficulties doing so. The sound direction implied a certain interaural cue value may also vary between individuals for physiological reasons. The azimuth angles implied by ITD and ILD cues are in this work referred as ITDA and ILDA, respectively.

Since the temporal fine structure of the high-frequency signal is not accurately encoded, and the shadowing effect occurs mostly at high frequencies, the traditional frequency division between the ITD (prominent below 1.5 kHz) and ILD (prominent above 1.5 kHz) seems natural. Therefore, the classical view of ILD is a long-time averaged level difference with a maximum resolution of approximately 1 dB SPL, as opposed to the temporally accurate ITD mechanism that has approximately 10 μs resolution. However, the ITDs of the high-frequency signal envelope can also contribute to lateralization in absence of low frequencies [41, 42], and ILDs can be effective also at low frequencies. Aside from static resolution, the speed at which the rapid changes in direction are tracked is an important factor for spatial cues. A

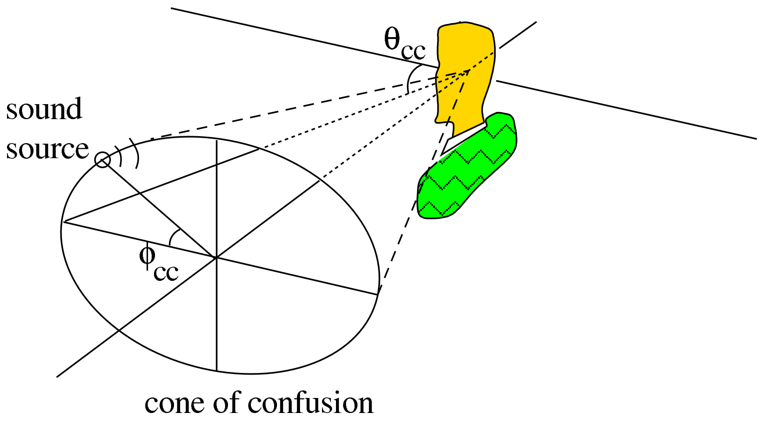

Figure 3.1: Cone of confusion. θ_{cc} indicates the azimuth and ϕ_{cc} the elevation angle of the sound source within the cone of confusion.

number of studies have shown that the ITD mechanism is quite sluggish, with an integration time constant of 45-250 ms [43], and that the interaural level difference (ILD) decoding is at least equally fast if not faster [42, 44].

Assuming a free-field, point-like sound source, ITDs and ILDs occur in certain 'natural' combinations depending on the azimuth direction of the source [45]. The auditory system learns to identify these cues as indicating the azimuth direction in question [46]. A spatial cue is said to be consistent if it indicates a single direction over a wide frequency range. By examining the interaural cues separately or in unnatural combinations (usually with headphones), their relative salience can be approximated. It has been found that when both ITD and ILD are consistent over a wide frequency range, but implied source directions within θ_{cc} (ITDA and ILDA) conflict, the low-frequency ITD cues dominate [47,48]. In the case where either ITD or ILD is set to be inconsistent as a function of frequency, thus giving conflicting directional information, the consistent cue is more prominent [40].

Related to cue salience, Raatgever and Bilsen [49] have performed their so called 'dominant region' experiments. The purpose of their studies was to extract a perceptual weighting function for the salience of binaural components in localization as a function of frequency. The tests were performed so that three frequency bands were presented to subjects using headphones. Initially, the middle frequency band had a larger ITD value than the other two bands, and the subjects adjusted the amplitude of this band until the subjective lateral direction of the auditory event was the same as when the two other bands had the larger ITD value. The tests were performed below 1200 Hz and a dominant region was found, centered around 600 Hz. Thus the results imply that some frequencies might contribute to the localization of complex sounds more prominently than others when the cues vary as a function of frequency.

The interaural cues do not always interact totally; with conflicting cues it is likely that the subject does not perceive a natural, point-like source [45, 50]. Investigators have for a large part dismissed the notion of total trading between the interaural cues that was suggested by early research, although these experiments established that trading can occur to some degree [51]. Significant argument for the dismissal of total trading comes from the mammalian physiological studies. It is generally thought that the ITD and ILD cues are initially encoded with separate mechanisms in different nuclei, namely in MSO and LSO, respectively [22]. Conflicting cues thus lead to complex listening situations which are discussed in Section 3.4.

There is good evidence that neurons implement coincidence detection between the ear signals as a function of ITD via neural excitation in MSO [52, 53]. MSO neurons respond best to low-frequency sounds, and the neurons with the same CFs have been show to respond to same ITDs. Perhaps conterintuitively, most of these ITD values exceed the physiological range implied by the size of the human head [54, 55]. Also, recent research emphasizes the role of neural inhibition in the MSO as improving the time resolution of ITD encoding [56]. This mechanism is in great demand, as the ITD processing requires far better resolution (10-20 μs) than any other neural process [57].

LSO neurons have generally higher best frequencies than those in MSO and predominantly show inhibitory processes: the ear signals are commonly thought to be effectively subtracted producing the ILD cue. As opposed to the traditional view of ILD being a long-time average, recent studies have shown that the LSO inhibition is very fast [58, 59]. This coincides with subjective tests suggesting that humans are able to detect fast ILD changes [42, 44]. The processing of the envelopes of high-frequency signals has also been speculated to take place in LSO [59].

Other spatial cues besides the interaural difference cues exist as well, such as the direction dependent filtering caused by the pinna. The pinna itself is thought to resolve whether the sound is coming from the front or the back of the listener. and the notches of the pinna are important in determining the elevation of the sound [9]. Also important in localization are the many temporal issues of spatial hearing, such as the precedence effect [60]. However, these are outside the focus of this thesis.

3.3 Binaural Signal Detection and Dichotic Pitch

Binaural hearing brings about many perceptual phenomena that are not present with monaural hearing. Many of the related studies focus on binaural signal detection. Starting with the research by Hirsh [61], it has been established that if different interaural manipulations are applied to a signal and a masker, the detection threshold of the signal is likely to be reduced. The threshold reduction induced by binaural hearing is referred to as binaural masking level difference (BMLD). As BMLD tests are easy to perform with simple equipment, extensive data exists on the subject.

The much used classic test paradigm is to present a tone signal that is phase-shifted between the ears by 180° among a wideband noise masker (N_0S_π), and compare the obtained threshold to the N_mS_m (monaural signal and masker) or to the N_0S_0 (diotic signal and masker) reference. In general, the classic BMLD is prominent (up to 15 dB) at low signal frequencies, but also present (up to 5 dB) at the higher frequency range above approximately 2 kHz. When presenting the signal monaurally ($N_0 S_m$), BMLD is approximately 6 dB smaller than in the $N_0 S_\pi$ configuration.

However, BMLD may occur in any situation where the signal and the masker have different interaural parameters. Many different test paradigms on the topic have been conceived, see e.g., [62] for a review on earlier research and [63] for more recent experiments. Some of these results are not as straightforward to interpret as the classic case. For example, it has been found that the detection performance varies between individual masker waveforms [64, 65].

In an attempt to generalize the BMLD phenomenon, researchers have linked binaural detection to the discrimination of coherence, or correlation between the ear signals [66]. Interaural correlation is an important perceptual attribute of spatial hearing and can be detected by humans very accurately [67]. Durlach *et al.* derived a mathematical expression to the interaural correlation ρ of a classic BMLD stimulus [68]:

$$\rho = (1 - \mathrm{SNR})/(1 + \mathrm{SNR}), \tag{3.2}$$

where SNR indicates the signal-to-noise power ratio of the BMLD stimulus. However, other results indicate that a simple generalization of BMLD as coherence detection mechanism is not sufficient [69].

Dichotic pitch occurs when two different broadband ear signals induce a pitch perception with simultaneous presentation but fail to do so when the two signals are presented monaurally. In a sense, dichotic pitches fall into the category of auditory illusions as there is no real signal to detect. Huggins pitch is the first and the most famous of the dichotic pitches [70]. The corresponding stimulus is implemented by creating a phase transition of 360° during a narrow frequency band of a broadband noise. The percept that occurs with this stimulus clearly has a specific frequency similarly as the percepts created by a sinusoid or narrowband noise. Other types of binaural pitch stimuli also exist, some of which create a perception that is less like a single tone, but more akin to a harmonic cluster [71, 72].

3.4 Perceptual Analysis of Complex Auditory Situations

As seen from the previous sections, the potential of spatial hearing goes beyond mere localization of a point-like source in free-field. This is good, since every-day life presents us with numerous complex auditory situations. Common examples are perceiving speech among noise or an instrument among an orchestra. Also, modern audio reproduction systems typically consist of several loudspeakers positioned in a moderately reverberant listening room, with sophisticated algorithms determining the sound produced by each speaker (e.g [73, 74]).

It seems baffling that humans are able to identify and associate the parts (e.g., the talker or the instrument) of such complex scenes based on the waveforms of the two ear signals alone. However, this viewpoint does not take into account the holistic principles of perception; the sensory input from all modalities contributes to the process. Also, there is rarely a situation where we do not have prior information and expectations about the scene.

How to go about analyzing such a scene scientifically? The whole complex percept can be described with abstract attributes, such as: ”envelopment,” ”broadness,” ”sense of space,” ”balance,” ”timbre,” and ”spaciousness,” [75–77]. However, the meaning of these words may vary between different listeners unless listener training is applied.

In addition to using abstract descriptions, the approach commonly taken is that ”objects,” [10] ”events,” [9] and ”streams,” [11] can be perceived among the massive information flow. Perceptual objects emerge from the background in a ”figure-ground” manner, with only a part of the available information constituting the significant ”figure” [78]. A typical example of the figure-ground mechanism is given by the Rubin’s face-vase illustration [79], where both figures cannot be perceived simultaneously. Thus, the role of perceptual ”edges” of sharp contrasts in some perceptual dimension are emphasized. Some theories also speculate that extensive unconscious segregation analysis takes place prior to conscious grouping [80].

This approach gives rise to certain philosophical and semantic questions: what do we really mean when we label physical objects with words of natural languages and identify them with certain auditory perceptions? Is this strategy more useful than harmful? Notably, Wittgenstein has explored the limitations of language as a way of relaying information [81].

Leaving the semantic questions aside, traditional Gestalt theory, as well as the cross-modality considerations by Gibson [82], provide principles according to which objects are grouped; these include the familiar cues of common fate, proximity, closure, continuation etc. Bregman lists grouping cues specific to audition, e.g., harmonicity, phase, arrival time/onset (precedence effect [60]), interaural cues, modulation, frequency proximity and overall spectral similarity [11]. Frequency proximity is often thought as auditory equivalent to spatial proximity in the visual domain, and it is tempting to associate auditory objecthood to mere comparison between spectral templates. However, the perceptual mechanisms related to tonotopic maps in the auditory cortex are not well known at present.

The formation of auditory virtual sources is an interesting phenomenon that coincides with the Gestalt principles. A typical example occurs with pairwise panning: if equidistant loudspeakers in a space with little reflections emit the same signal, the listener perceives the direction of the sound as in between the speakers. The perceived direction can be manipulated by adjusting signal amplitudes [83]. If the loudspeaker signals have structural or temporal differences in the listening position, a number of interesting things can happen to the percept.

Gardner [84] describes several listening conditions that may result in fusion, displacement, or spatial broadening of a virtual source created by two or more real sound sources. Image fusion, ”the governing phenomenon”, is mediated by commonality of auditory features in accordance with Gestalt theory. Dissimilarity between features causes the fused sound to break up into separate sources. Several other studies support the notion of commonality in image fusion, for example [85] and [86]. Sound displacement, on the other hand, is strongly tied to sound localization. Several studies have determined that simultaneous distracting signals strongly affect localization of the target signal [87–91]. Localization and general directional quality of virtual sources have been studied before with amplitude panning [14] and Ambisonics [92].

If the similarities in some respect can cause two or more sounds to fuse, then it is natural to assume that differences in the sounds mediate the breakup of fusion or broadening of a virtual source. The broadening phenomenon is usually described with auditory source width (ASW). ASW has long been recognized as a seminal feature in spatial auditory perception, e.g., in describing concert hall acoustics [93, 94]. In this work, ASW mainly refers to the perceived width of a specific auditory event or object.

In a normal listening room, ASW is often linked with attributes such as the amount of early reflections [95]. Early research on source width, or tonal volume, showed it to be a function of loudness, duration of the sound, interaural characteristics, and frequency [96, 97]. Potard and Burnett state [98]: ”Low pitched sound sources need a greater distance for one wavelength to unfold and tend to have a larger apparent width than high pitched sound sources.” An important factor affecting the perceived width is the interaural correlation [9]. Recently, Mason et al. have derived a model for this dependence with headphone experiments where the subjects describe the width of a sound localized inside the head [99].

4 Modeling Auditory Spatial Perception

This section discusses computational scientific models from the auditory point of view. For a general introduction to scientific modeling, see [100]. There exist numerous models in hearing research ranging from theories of perceptual grouping of objects [78] to signal algorithms representing the function of hair cells and the auditory nerve [101]. The focus in this section, as in the previous ones, is however on spatial hearing. Prior to introducing some well-known auditory models, a general (and accepted) limitation of scientific modeling is discussed: the model may make explicit assumptions that are known to be false, or incomplete, in some detail.

4.1 General Considerations

While the modern computational modeling concept owes much to the reductionist theories of Descartes, it was Turing that took the approach to the ultimate [102]; he contemplated the possibility of computers so sophisticated that their responses to stimuli would be indistinguishable from that of humans. Turing had showed that it is possible to compute any algorithm with a simple theoretical computer (the Turing machine) [103], and his hypothesis was that all of brain activity could be represented by a limited number of algorithms in some language of symbols. There are several counterarguments to this theory, namely the computer does not ”understand” the symbols if it lacks the cultural knowledge of humans, and the familiar doubt of how well can the symbols of the language represent reality.

If we however consider Turing’s theory in the sense of perceptual modeling, we arrive at the question whether the algorithms should mimic the mechanisms of the sensory pathway, i.e., should the model be physiologically accurate. The most important limitation to this approach is the lack of physiological knowledge. Already at the lower processing levels there are serious unanswered physiological questions, as discussed in Section 2. Furthermore, each individual has in principle a unique neural physiology, and small variations may result in unexpected changes in a dynamic nonlinear system like the brain over the course of time [104]. Understandably, the exact physiological modeling of the entire brain is not possible.

It may be argued that the aim of scientific modeling is not a perfect artificial intelligence as such, but rather to describe some specific phenomena separately, e.g some low-level neural mechanisms. In auditory modeling, however, the usual goal is to get a representation of the entire percept and this cannot be done by modeling the better known lower parts of the processing path alone. For example, the frequency resolution measured from the auditory nerve contrasts the corresponding human performance at high sound levels [23].

Should perceptual models then mimic the physiology at all or just try to produce as similar output as possible compared to human observers? A neural network, for example, would be a suitable tool for learning measured and scaled human responses to some specific situation if the investigator has enough subjective data of the phenomenon in question. A problem arises of how can such a model be applied to other phenomena if it is conceived as a ”black box” for a specific situation. The only solution to this problem seems to be to mimic the actual mechanisms of human hearing. Applicability is also the reason that overly complex models should be avoided. If the number of free parameters is large, the model simply implements a transformation of coordinates or curve-fitting, as discussed by Colburn and Durlach [105].

In spatial auditory modeling, the restrictions of physiological knowledge and the requirement of applicability has resulted in an emphasis on the peripheral neural mechanisms and low-level binaural interactions, combined with some kind of abstract parts describing high-level neural processing. Hence the extensive use of the term ”binaural model”. At the top of the processing chain is often a decision-making device, which is where the computational algorithm makes ecological, human-like, verdicts about the stimulus.

4.2 Modeling of Auditory Periphery

Commonly, the monaural periphery of binaural models consists of possible HRTF-filtering of the two ear signals, an ERB-filterbank, and simulation of neural transformation. The ERB-filterbank approximates the human frequency resolution by dividing the signal into separate critical bands, as shown in figure 4.1. The illustrated responses are derived from Slaney’s approximation [106] of the Gammatone filterbank [107]. The shape and bandwidth of these filters is based on psychoacoustical and physiological measurements.

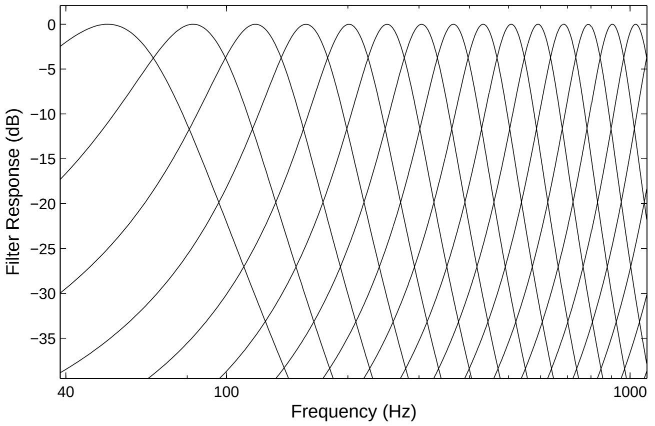

Figure 4.1: An example ERB-filterbank magnitude responses in the range of 40-1000 Hz. Human frequency resolution is approximated by filtering the signal into separate critical bands. Here the filters are spaced 1 ERB apart in frequency.

Another part of peripheral models is commonly neural transduction of the incoming signal. While complex neuron models have been developed, the neural transformation can be most simply characterized as half-wave rectification of the input signal, as neural impulses cannot be negative. This also corresponds roughly to the release rate of the transmitter substance of the IHCs [101,108]. After this, low-pass-filtering is usually applied to the signal in order to simulate the loss of phase-locking, as discussed in Section 2.2.

4.3 Popular Binaural Modeling Concepts

In 1948, Jeffress proposed an elegant binaural model that has remained highly influential [109]. The model presents a way of obtaining the ITD by comparing the inputs of the two ear signals, and thus inferring the lateral location of the sound source. As seen from figure 4.2, both ear signals travel trough separate delay lines, in between which lie the coincidence detection neurons (EE-type). Effectively this system approximates cross-correlation calculation. The output of the coincidence neurons creates an accurate representation of the azimuth space that can be analyzed by the upper levels, provided that the inputs are synchronized and the unit delays are sufficiently small. If the coincidence detection is implemented separately for each critical band, the three-dimensional output (frequency-time-neural activity) can be interpreted as a topographical map of the azimuth space.

The inspiration for the Jeffress model probably rose from the fact that low-frequency ITD is the dominant localization cue and its encoding requires the most temporally accurate processing in the nervous system. The early trading experiments between ITD and ILD also implied that ILD could also be somehow represented as ITD. Even though the two cues are mainly processed at different nuclei and total time-intensity trading was dismissed, the Jeffress delay-line concept soon became the standard view of human ITD processing. Some neurophysiological evidence of Jeffress-type mechanisms was also found in single cases [110].

Evidence against the physiological validity of the Jeffress model has since surfaced,

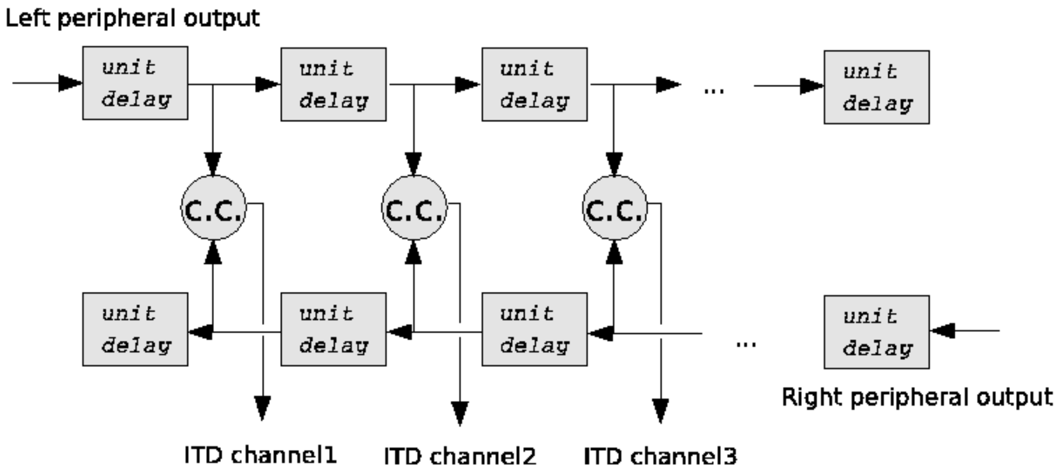

Figure 4.2: The Jeffress delay line concept. Coincidence counting (C.C.) units approximate the cross correlation between the ear signals.

as summarized by Grothe [56]. He argues that the avian ITD processing may resemble the Jeffress delay-line, but that the mammalian system is quite different. An azimuthal space map resulting from the Jeffress model has not been found in the auditory system. As discussed in Section 3.2, the majority of the gerbil and guinea-pig MSO cells respond maximally to ITDs outside the physiological range [54, 111]; The cell's best frequency and best ITD are also correlated and the slopes of the ITD functions are steepest at zero ITD, where localization is most accurate. Grothe therefore suggests that the relative activity of the entire population of MSO cells, rather than the distribution within the MSO, represents the azimuth of the sound.

Since the emergence of the previous results, Joris and Yin have defended the physiological validity of the Jeffress model and pointed out problems in the modeling approach based on relative activity of the two hemispheres [55]. They argue that while the capacity of such mechanisms may be questioned, there is sufficient evidence to support the existence of mammalian delay lines. However, their conclusions are mainly uncertain; no recent physiological study has validated the accurate topographical map of characteristic delays predicted by the Jeffress model. Currently, there is no clear consensus on these physiological issues.

Another important binaural model is the equalization-cancellation (EC) model proposed by Durlach [112]. The EC-model was mainly intended for explaining binaural detection and not localization. More so than the Jeffress model, the EC-model is also an abstract concept not based on physiology. The two ear signals are first band-pass filtered and random gain and time jitter is applied to them, which limits the model's performance. The basic mechanism of the model is to equalize the gains and the phases of the ear signals prior from subtracting, or canceling the similar parts. The cancellation can be thought as an EI-mechanism, in contrast to EE-coincidence detection in the Jeffress model. This process effectively eliminates a diotic masker. The remaining dichotic signal can then be analyzed with Durlach's original SNR-approach or with an optimal detector based on the signal detection theory [113].

4.4 Developments on Spatial Auditory Modeling

Jeffress's original model has been complemented with numerous extensions. As mentioned, the standard peripheral part of contemporary binaural models consists of the critical-band filtering and a neural transform. In addition, a decision device is needed to analyze the output activity pattern. These important additions to the Jeffress concept were initiated by Colburn [114, 115]. Stern et al. emphasized the importance of consistent near-zero ITD cues [116]. Lindemann extended the capacity of the Jeffress model to process ILD and precedence effect cues [117, 118], as well as to emphasize the natural combinations of ITDs and ILDs [45].

A recent model proposed by Breebaart et al. is based on both the Jeffress delay lines and the EC-concept [119–121]. The ITD information is obtained with the familiar delay line, which is complemented by an additional attenuation line, originated by

Reed and Blum [122], at each delay value. The activity is however determined by EI-type neurons instead of EE-type neurons in the Jeffress model. The activity pattern thus produced is affected by both ITD and ILD cues, and is analyzed by an optimal detector. The decision device stores pattern templates that can then be compared to the stimulus in forced-choice detection experiments similarly as in real listening tests. Even though such maps have not been specifically found in the auditory pathway, the model by Breebaart et al. is notable in that it has only one parameter to be adjusted and it has been meticulously compared with psychoacoustical results of various phenomena.

4.5 The Performance of Auditory Models

The present view among researchers is that binaural models are able to account for a large amount of results from simple subjective experiments. One of the most comprehensive comparisons to date have been done by Breebaart et al. [119–121]. Generally, the delay and attenuation line mechanisms give ITD and ILD cues that correspond to the azimuth position of a broadband, point-like source. In addition to localization other phenomena can also be considered. These include temporal adaption and sluggishness, as well as the precedence effect, which are however not essential for the research presented in this thesis.

Coincidence-counting auditory models have been widely utilized to predict psychoacoustical data such as classic BMLD experiments [116]. The activity patterns show ”dimples” caused by the S_π signal among the masker. Colburn was able to describe much of the classical BMLD results by forming a decision variable based on the coincidence output [115]. In addition to coincidence-counting models, the EC-model can successfully account for BMLD results, as that is what the model was designed for [112].

Dichotic pitch is very similar to the BMLD case in a binaural modeling sense [116, 123]. Cross-correlation models produce notable effects at the pitch frequency, whereas EC models produce a decorrelation peak. Culling et al. have compared the two approaches for various dichotic pitches [71, 72]. They concluded that a modified EC model provides the most coherent prediction results in the tested cases.

One can always ask to what extent the previous results prove the worth of binaural models. The stimuli used in the binaural listening experiments are predominantly not very ecological, i.e., similar to ”real-life” sounds. ”Traditional” listening tests have favored simple tones and noises, partly due to the limitations of early auditory equipment. Even in the context of ”laboratory stimuli” the binaural models often need adjusting when going beyond the basic case, for example in a BMLD situation [64]. Scientists often successfully use several incompatible models of one and the same target system for predictive purposes, which is considered bad practice [100]. For these reasons, the applicability of binaural models to predict the results for some future test cases is uncertain. Even if predictions with binaural models can be made, it is difficult to determine which model works best overall; in addition to its structure, the performance of the model depends on the task and the decision process that is used.

One possible approach to these issues could be to focus more on creating tools and applications for specific tasks instead of models of human physiology or cognition. One example of such tool is the cocktail-party processor by Bodden [124]. Such tools need not to be limited by human performance. For example, a localization performed by a neural network can sometime exceed the human performance [125]. In this thesis however, the limitations of physiologically oriented auditory modeling are acknowledged and models are used mainly to test and better understand hypotheses about perceptual mechanisms.

5 Goal of the Thesis and Summary of Results

With the scientific background established in the previous sections, specific goals for the thesis are stated below, followed by summaries of the publications. As mentioned in Section 1.1, there are two main research questions:

First, how are complex auditory scenes, where different frequencies arrive from different azimuth directions, perceived in terms of localization, width, and fusion/ segregation? In order to gain insight, previous results by other authors can be extrapolated to a certain degree, but no study has specifically targeted the hypotheses investigated in Publications I-IV.

Similarly, the second research question of the thesis is based on utilizing recent physiological results to spatial auditory modeling in a novel way. The aim is to present a combination of ITD and ILD models on a conceptual level and to investigate whether the proposed models can account for common binaural phenomena. Specifically, the idea of fast ILD encoding is utilized in Publications V-VI. This is the main difference compared to other models utilizing EI-type neurons. Publication VII combines a novel ITD model with the ILD model and summarizes the presented concept.

Publication I

This paper represents the first attempt to analyze the directional perception of noise virtual sources created with multichannel reproduction techniques in an anechoic chamber using a binaural auditory model and listening tests. The investigated loudspeaker setups were typical five- and eight-channel systems. Different microphone techniques and pair-wise panning were also simulated in the reproduction. The auditory model was utilized to compute the frequency-dependent ITDA and

ILDA that predict the cone of confusion in which a sound source lies. These values were compared with the auditory pointer localization data from listening tests.

The listening test results matched the simulation results generally well, although there were some systematic deviations. The model gave the most reliable predictions with virtual sources near the median plane, and at low frequencies. Farther from the median plane, the output of the model was in general hard to interpret, and it suggested directions nearer the median plane than those that were actually perceived. Both the simulation results and the listening tests suggest that with the 5.1 reproduction setup it is impossible to create virtual sources in directions farther than 70° from the median plane with the tested reproduction systems. With the eight-channel setup, the bias toward the median plane was prominently smaller. Some of the test configurations resulted in frequency-dependent perception of the virtual source. It was mainly in these cases that the model predictions failed. Generally, it seems that although ITDA and ILDA behave differently at low frequencies, the listeners relied on the ITD cue only.

Publication II

The previous paper established that sound reproduction techniques sometimes produce virtual sources whose directional cues propose multiple directions at a time. Also, many natural sound sources are not point-like, but spatially widened. This paper investigates how sound events whose localization cues indicate different azimuth direction as a function of frequency are perceived. A horizontally wide (45°) sound source was created by presenting spectrally consecutive, non-overlapping, bandlimited noise samples simultaneously from the different loudspeakers of a loudspeaker grid in an anechoic environment. The narrowband samples together formed a broadband stimulus. The order of the narrowband noise samples in the loudspeakers, as well as the total frequency range of the samples, was varied from case to case.

In each test case, the subjects were asked to indicate the perceived center of gravity of the sound image, as well as all the loudspeakers that they perceived to radiate sound. Generally, the perceived center could not be predicted merely by a simple model using a previously established frequency weighting function for binaural salience. Alternative frequency weights were calculated analytically from the listening test results. These calculations indicate that it is difficult to establish any general dominant frequencies, but that the lowest and highest bands were perceptually important in all examined frequency regions. The results also indicated that the perceived width of the sound sources produced by the nine-loudspeaker setup was, in all cases, less than half of the actual width of the source. This implies that some frequency bands from different loudspeakers fused together spatially. The main effects on perceived width were caused by the utilized frequency range in each test case.

Publication III

As a related investigation to Publication II, the perceptual segregation and fusion of two simultaneously arriving ERB-bandwidth noise components was studied as a function of their frequency and azimuth separation in an anechoic chamber. The subjects adjusted the frequency gap between the components until they heard two separate auditory objects. The results indicate that when both components are above 1.5 kHz in frequency, i.e., in the range where fine structure information of a complex stimulus is more difficult to determine, the frequency gap threshold is notably increased. The present hypothesis is that at high frequencies, where only the signal envelope can be analyzed accurately, the components are more difficult to segregate. The effect of azimuth separation between the ERB-bands was not prominent compared to the frequency dependency. Although the main trends can be observed easily from the results, the inter-subject deviations were relatively large. It is therefore hypothesized that determining the segregation of two frequency components utilizes higher-level processing prominently.

Publication IV

The purpose of this study was to gain further insight into the perception of virtual sources with varying directional cues by using 12 separate test cases designed to test specific hypotheses. More profound statistical analysis tools compared to the previous study were also experimented with. All cases utilized the frequency range between 200-1179 Hz. When comparing the perceptual salience of the middle versus the high/low frequencies, it was found that the middle range had a notably smaller effect on the perceived directional distribution of sound. Also, the spatial order of the frequency band signals was found to have a small but significant effect on the perceived width of the sound event. The cases with larger abrupt changes in the localization cues were perceived slightly wider than those in which the cues changed more moderately. Three cases were tested with both continuous noise and click train stimuli. The click cases were perceived notably narrower, mostly as radiating from one or two speakers. Simulations of the test cases implemented with a cross-correlation based auditory model were also examined and compared to the subjective results. The simulation results were not found entirely satisfactory, mainly because the spatial distributions obtained from the simulations were too similar in different test cases. Thus, additional weighting of different CBs in the cross-correlation model was suggested.

Publication V

As discussed in Section 1.1, the focus of the dissertation at this point shifts from perceptual research to auditory modeling. This paper presents a novel computational auditory model inspired by recent neurophysiological findings, as well as the inability of the common models to predict all psychoacoustical results from the previous experiments. The model utilizes the instantaneous ILD to predict spatial hearing cues that humans perceive. Based on previous listening tests, it is hypothesized that ILD decoding is fast and that the instantaneous difference signal can be utilized effectively by the auditory system. However, the ILD model is to be complemented by a separate ITD model analogous to the physiology of MSO and LSO. The main focus in this paper is on the interaural coherence cue, i.e. the perceived similarity between the waveforms of the ear signals. Simulations show that the proposed model concept is suitable predicting known psychoacoustical results.

Publication VI

This paper also investigates the previously implemented auditory model, this time using two common binaural detection cases: BMLD and binaural pitch. The model output produced pronounced peaks at the signal frequencies with BMLD stimulus. The dichotic pitch cases, Huggins pitch and binaural edge pitch stimuli produced a notable peak at the pitch frequency. The difference between the monaural and dichotic signal as well as individual masker effects were also studied in BMLD simulations. Again, the model concept was deemed suitable to account for detection data. No direct comparison with psychoacoustical detection data is given is this paper. Rather, the model output is thought to be processed by upper-level pattern recognition, or similar systems.

Publication VII

In this paper, the previously introduced ILD model is combined with a model for the ITD. The combined model is designed to implement the recently-discovered neurophysiological mechanisms of MSO and LSO relevant to coding the ITD and

ILD cues. The ILD model is in this paper is an updated version of the previous model. Some additional processing stages are now omitted or implemented in the other parts of the model. Initial model simulations with simple ITD and ILD cases show strong correspondence with physiological measurements.

6 Conclusions

The novel research of this thesis is not so much focused on a certain hypothesis or problem, but rather touching several topics. Human auditory spatial perception is the key theme that links together the research papers presented here. Nevertheless, these studies form a somewhat logical continuum.

Investigation on sound reproduction systems revealed that sometimes the produced spatial cues were not consistent over frequency. The perception of sound events of this nature was further examined with a horizontal loudspeaker grid that emitted simultaneous noise band signals, thus creating effectively a spatially wide sound source. Specifically localization (dominant perceived direction), perceived width, spatial distribution, and segregation were considered in the analysis. Numerous perceptual phenomena were discovered in the tests.

Auditory modeling techniques were employed throughout the research by suggesting models for the psychoacoustical results. Novel techniques were employed and improvements suggested for the commonly used models. This aspect of the thesis work culminated in the most recent papers, where novel auditory models are developed, investigated in common psychoacoustical cases, and found to be capable of accounting for psychoacoustical phenomena.

Summary of main results:

- Commonly used sound reproduction methods were investigated and their capability to produce virtual sources in desired target directions was reported.
- Most algorithms produce auditory events, whose spatial cues vary undesirably as a function of frequency.

- Different frequency bands fused notably when presented simultaneously from different directions with the resulting auditory event being perceived as less than half of the actual source width.

- The spatial order of the frequency bands was found to be significant in width perception.

- Strong perceptual contrasts and the loss of phase-locking were deemed important in the segregation of different frequency bands.

- Although commonly used auditory models could be used to account for some of the subjective results, certain cases proved problematic for them.

- A novel LSO model based on the concept of instantaneous ILD has been developed and simulated with simple spatial stimuli.

- The LSO model was combined with a simple MSO model in order to form a novel, physiologically-based auditory model.

7 Future Directions

As with all scientific research, these studies gave rise to both open questions and inspiration for future investigations. The psychoacoustical research on wide sound sources did not include temporal phenomena, such as the precedence effect, whose inclusion would be the next logical step. A hypotheis was presented [126] that fusion of different frequency bands in the subjective listening of spatially wide sound sources could be decreased if the subjects were allowed to move their heads. This is certainly possible at least to some extent, as head movements are sometimes utilized by humans to solve confusions in real-life listening situations. However, the authors of the research at the time performed unofficial tests utilizing head movements. The unofficial listening indicated that head movements did not help segregate the different bands much, and certainly not all loudspeakers were perceived to emit sound. The possible cause for this is discussed in the last paragraph of this section. A separate experiment would be required to investigate head movements, as well as wide sounds from the side of the head.

In an attempt to draw the present dissertation together thematically, the newly implemented auditory model presented in Publication VII was informally tested using HRTF-simulated spatially wide test cases from Publications II and IV. Although the physiologically-based separation of MSO and LSO models, as well as the fast ILD concept may in the future be helpful also in these cases, some additional research is necessary prior to such applications. More specifically: 1) HRTF-data should be used to carefully adjust the output of the model. Thus the comparison and combination between the resulting cues of two model parts and separate frequency bands is not fully possible at this point. 2) Time constants of the model would require fine-tuning. 3) It has been shown that the nonlinearity of the cochlea explains some of the MSO output. This is presently modeled in a very heuristic manner. Another question is whether the nonlinear cochlear output should also be used with the LSO model. 4) MSO is modeled using guinea-pig data. Human physiology is possibly significantly different in terms of head size and contralateral input. 5) The suggested auditory model could be further tested against psychoacoustical data by implementing a decision device according to signal detection theory.

A general feature of the suggested auditory model is that, unlike the traditional cross-correlation model, it does not suggest that separate sounds from several directions are detected simultaneously. However, the detected direction may change rapidly. The authors hypothesize that is also the case in the actual hearing system, as it might help explain the subjective results where different frequency bands were partially fused, as well as the decrease of perceived width with short sounds that were examined in Publication IV. In order to perceive separate sounds from, for example, all loudspeakers of a horizontal speaker ensemble, the ITD and ILD cues created by the different speaker signals should vary very fast. Needles to say, further research is required to support this hypothesis.

References

[1] Wikipedia article, “Pyrrho,” URL: http://en.wikipedia.org/w/index.php?title=-Pyrrho&oldid=100954504, December 2006.

[2] Wikipedia article, “Sense,” URL: http://en.wikipedia.org/w/index.php?title=-Sense&oldid=101881557, January 2007.

[3] M. Werheimer, “Gestalt theory,” in *Source Book of Gestalt Psychology*, W. D. Ellis, Ed., pp. 1–11. Harcourt, Brace and Co, New York, 1997, Available also at: http://gestalttheory.net/archive/wert1.html.

[4] R. Descartes, *Discourses Part V*, 1637.

[5] L. von Bertalanffy, *General System theory: Foundations, Development, Applications*, George Braziller, New York, revised edition, 1976.

[6] I. P. Howard and W. B. Templeton, *Human spatial orientation*, Wiley, London, 1966.

[7] H. McGurk and J. McDonald, “Hearing lips and seeing voices,” *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[8] D. Wright and G. Wareham, “Mixing sound and vision: The interaction of auditory and visual information for earwitnesses of a crime scene,” *Legal and Criminological Psychology*, vol. 10, no. 1, pp. 103–108, 2005.

[9] J. Blauert, *Spatial Hearing*, The MIT Press, Cambridge, MA, USA, revised edition, 1997.

[10] T. D. Griffths and J. W. Warren, “What is an auditory object?,” *Nature Neuroscience*, vol. 5, no. 11, pp. 887–892, 2004.

[11] A. S. Bregman, *Auditory Scene Analysis: The perceptual organization of sound*, MIT Press, Cambridge, MA, 1990.

[12] J. G. Neuhoff (ed.), *Ecological Psychoacoustics*, Elsevier, San Diego, CA, USA, 2004.

[13] S. Bech and N. Zacharov, *Perceptual Audio Evaluation - Theory, Method and Application*, John Wiley and Sons Ltd, West Sussex, England, 2006.

[14] V. Pulkki, *Spatial Sound Generation and Perception by Amplitude Panning Techniques*, Ph.D. thesis, Helsinki Univ. Tech., 2001, Available also at: http://lib.hut.fi/Diss/2001/isbn9512255324/.

[15] J. O. Pickles, *An Introduction to the Physiology of Hearing*, Academic Press, 1988.

[16] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The CIPIC HRTF Database, ," in *Proc. 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics*, New Paltz, New York, Oct. 2001, pp. 99–102, IEEE.

[17] Dan Pickard, "Human ear," URL: http://en.wikipedia.org/wiki/Image:HumanEar.jpg, 2006.

[18] H. Hudde, "A functional view on the peripheral human hearing organ," in *Communication Acoustics*, J. Blauert, Ed., pp. 43–74. Berlin, Berlin, 2005.

[19] G. von Bekesy, *Experiments in Hearing*, McGraw, New York, 1960.

[20] G. K. Yates, "Cochlear structure and function," in *Hearing*, B. J. C. Moore, Ed. Academic, San Diego, 1995.

[21] A. Bell and N. H. Fetcher, "The cochlear amplifier as a standing wave: 'squirting' waves between rows of outer hair cells?," *J. Acoust. Soc. Am.*, vol. 116, no. 2, pp. 1016–1024, 2001.

[22] A. R. Palmer, "Neural signal processing," in *Hearing*, B. J. C. Moore, Ed., pp. 75–121. Academic, San Diego, 1995.

[23] D. O. Kim and C. E. Molnar, "A population study of cochlear nerve fibers: comparison of spatial distributions of average-rate and phase-locking measures of responses to single tones," *J. Neurophysiol*, vol. 42, no. 1, pp. 16–30, 1979.

[24] D. H. Johnson, "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones," *J. Acoust. Soc. Am.*, vol. 68, no. 4, pp. 1115–1122, 1980.

[25] A. R. Palmer and I.J. Russel, "Phase-locking in the cochlear nerve of the guinea pig and its relation to the receptor potential of the inner hair cells," *Hear. Res.*, vol. 24, no. 1, pp. 1–15, 1986.

[26] J. E. Rose, R. Calambos, and J. R. Hughes, "Microelectrode studies of the cochlear nuclei of the cat," *Bull. Johns Hopkins Hosp.*, vol. 104, no. 5, pp. 211–251, 1959.

[27] B. Grothe and D. H. Sanes, "Synaptic inhibition influences the temporal coding properties of medial superior olivary neurons: an *in vitro* study," *J. Neurosci.*, vol. 14, pp. 1701–1709, 1994.

[28] J. P. Rauschecker and B. Tian, "Mechanisms and streams for processing of 'what' and 'where' in auditory cortex," in *Proc. Natl. Acad. Sci.*, USA, 2000, pp. 11800–11806, 97 (22),.

[29] J. P. Rauschecker and B. Tian, "Processing of band-passed noise in the lateral auditory belt cortex of the rhesus monkey," *J. Neurophysiol.*, vol. 91, no. 6, pp. 2578–2589, 2004.

[30] B. Tian and J. P. Rauschecker, "Processing of frequency-modulated sounds in the lateral auditory belt cortex of the rhesus monkey," *J. Neurophysiol.*, vol. 92, no. 5, pp. 2993–3013, 2004.

[31] C. E. Schreiner and J. R. Mendelson, “Functional topography of cat primary auditory cortex: distribution of integrated excitation,” *J. Neurophysiol..*, vol. 64, no. 5, pp. 1442–1459, 1990.

[32] M. P. Zwiers, A. J. Van Opstal, and G. D. Paige, “Plasticity in human sound localization induced by compressed spatial vision,” *Nat. Neurosci.*, vol. 6, no. 2, pp. 175–181, 2003.

[33] H. Fletcher, “Auditory patterns,” *Reviews of Modern Physics*, vol. 12, pp. 47–65, 1940.

[34] B. R. Glasberg and B. C. J. Moore, “Derivation of auditory filter shapes from notched-noise data,” *Hear. Res.*, vol. 47, no. 1-2, pp. 103–138, 1990.

[35] D. D. Greenwood, “Auditory masking and the critical band,” *J. Acoust. Soc. Am.*, vol. 33, no. 4, pp. 484–502, 1961.

[36] D. D. Greenwood, “A cochlear frequency-position function for several species - 29 years later,” *J. Acoust. Soc. Am.*, vol. 87, no. 6, pp. 2592–2605, 1990.

[37] B. C. J. Moore and B. R. Glasberg, “Auditory filter shapes derived in simultaneous and forward masking,” *J. Acoust. Soc. Am.*, vol. 70, no. 4, pp. 1003–1014, 1981.

[38] 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, no. 6, pp. 2852–2861, 2001.

[39] Lord Rayleigh (a.k.a. J. W. Strutt 3rd Baron of Rayleigh), “On our perception of sound direction,” *Phil. Mag.*, vol. 13, pp. 214–232, 1907.

[40] F. L. Wightman and D. J. Kistler, “Factors affecting the relative salience of sound localization cues,” in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H. Gilkey and T. R. Anderson, Eds. Lawrence Erlbaum Assoc., 1997.

[41] G. B. Henning, "Detectability of interaural delay in high-frequency complex waveforms," *J. Acoust. Soc. Am.*, vol. 55, no. 1, pp. 84–90, 1974.

[42] T. N. Buell and E. R. Hafter, "Discrimination of interaural differences of time in the envelopes of high-frequency signals: Integration times," *J. Acoust. Soc. Am.*, vol. 84, pp. 2063–2066, 1988.

[43] L. R. Bernstein, "Detection and discrimination of interaural disparities: Modern earphone-based studies," in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H. Gilkey and T. R. Anderson, Eds., pp. 117–138. Lawrence Erlbaum assoc., Mahwah, New Jersey, 1997.

[44] D. W. Grantham, "Discrimination of dynamic interaural intensity differences," *J. Acoust. Soc. Am.*, vol. 76, no. 1, pp. 71–76, 1984.

[45] W. Gaik, "Combined evaluation of intearaural time and intensity differences: Psychoacoustic results and computer modeling," *J. Acoust. Soc. Am.*, vol. 94, no. 1, pp. 98–110, 1993.

[46] R. Y. Litovsky and D. H. Ashmead, "Development of binaural and spatial hearing in infants and children," in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H. Gilkey and T. R. Anderson, Eds., pp. 571–589. Lawrence Erlbaum assoc., Mahwah, New Jersey, 1997.

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

[48] E. A. Macpherson and J. C. Middlebrooks, "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," *J. Audio Eng. Soc.*, vol. 111, no. 5, pp. 2219–2236, 2002.

[49] J. Raatgever, *On the Binaural Processing of Stimuli with Different Interaural Phase Relations*, Ph.D. thesis, Technische Högeschool Delft, 1980.

[50] E. R. Hafter and C. Carrier, “Binaural interaction in low-frequency stimuli: the inability to trade time and intensity completely,” *J. Acoust. Soc. Am.*, vol. 51, no. 6, pp. 1852–1862, 1972.

[51] G. G. Harris, “Binaural interactions of impulsive stimuli and pure tones,” *J. Acoust. Soc. Am.*, vol. 32, pp. 685–692, 1960.

[52] J. M. Goldberg and P. B. Brown, “Response of binaural neurons of dog superior olivary complex to dichotic tonal stimuli: Some physiological mechanisms of sound localization,” *J. Neurophysiol.*, vol. 32, pp. 613–636, 1969.

[53] T. C. T. Yin, P. X Joris, P. H. Smith, and J. C. K. Chan, “Neuronal processing for coding interaural time disparities,” in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H. Gilkey and T. R. Anderson, Eds., pp. 399–425. Lawrence Erlbaum Associates, Mahwah, New Jersey, 1997.

[54] D. McAlpine, D. Jiang, and A. R. Palmer, “A neural code for low-frequency sound localization in mammals,” *Nature Neurosci.*, vol. 4, no. 4, pp. 396–401, 2001.

[55] P. X. Joris and T. C. Yin, “A matter of time: internal delays in binaural processing,” *Trends Neurosci.*, vol. 30, no. 2, pp. 70–78, 2007.

[56] B. Grothe, “Sensory systems: New roles for synaptic inhibition in sound localization,” *Nature Neurosci.*, vol. 4, no. 7, pp. 540–550, 2003.

[57] R. G. Klumpp and H. R. Eady, “Some measurements of interaural time difference thresholds,” *J. Acoust. Soc. Am.*, vol. 28, no. 5, pp. 859–860, 1956.

[58] R. Batra, S. Kuwada, and D. C. Fitzpatrick, “Sensitivity to interaural temporal disparities of low- and high-frequency neurons in the superior olivary complex. I. heterogeneity of responses,” *J. Neurophysiol.*, vol. 78, no. 3, pp. 1222–1236, 1997.

[59] P. X. Joris and T. C. Yin, “Envelope coding in the lateral superior olive. I. sensitivity to interaural time differences,” *J. Neurophysiol.*, vol. 73, no. 3, pp. 1043–1062, 1995.

[60] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Gutman, “The precedence effect,” *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 1633–1654, October 1999.

[61] I. J. Hirsh, “The influence of interaural phase on interaural summation and inhibition,” *J. Acoust. Soc. Am.*, vol. 20, no. 4, pp. 536–544, 1948.

[62] N. I. Durlach and H. S. Colburn, “Binaural phenomenna,” in *Handbook of Perception*, E. C. Carterette and M. P. Friedman, Eds., pp. 365–466. Academic Press, 1978.

[63] R. H. Gilkey and T. R. Anderson, Eds., *Binaural and Spatial Hearing in Real and Virtual Environments*, Lawrence Erlbaum Assoc., Mahwah, NJ, US, 1997.

[64] H. S. Colburn, S. K. Isabelle, and D. J. Tollin, “Modeling binaural detection performance for individual waveforms,” in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H. Gilkey and T. R. Anderson, Eds., pp. 533–555. Lawrence Erlbaum assoc., Mahwah, New Jersey, 1997.

[65] S. K. Isabelle and H. S. Colburn, “Detection of tones in reproducible narrowband noise,” *J. Acoust. Soc. Am.*, vol. 89, no. 1, pp. 352–359, 1991.

[66] J. Koehnke, H. S. Colburn, and N. I. Durlach, “Performance in several binaural-interaction experiments,” *J. Acoust. Soc. Am.*, vol. 79, no. 5, pp. 1558–1562, 1986.

[67] J. F. Culling, H. S. Colburn, and M. Spurchise, “Interaural correlation sensitivy,” *J. Acoust. Soc. Am.*, vol. 110, no. 2, pp. 1020–1029, 2001.

[68] N. I. Durlach, K. J. Gabriel, H. S. Colburn, and C. Trahiotis, “Interaural correlation discrimination: Ii. relation to binaural unmasking,” *J. Acoust. Soc. Am.*, vol. 79, no. 5, pp. 1548–1557, 1986.

[69] L. R. Bernstein and C. Trahiotis, “Discrimination of interaural envelope correlation and its relation to binaural unmasking at high frequencies,” *J. Acoust. Soc. Am*, vol. 91, no. 1, pp. 306–316, 1992.

[70] E. M. Cramer and W. H. Huggins, “Creation of pitch through binaural interaction,” *J. Acoust. Soc. Am.*, vol. 30, no. 5, pp. 413–417, 1958.

[71] J. F. Culling, A. Q. Summerfield, and D. H. Marshall, “Dichotic pitches as illusions of binaural unmasking. I. huggins pitch and the binaural edge pitch,” 1998.

[72] J. F. Culling, D. H. Marshall, and A. Q. Summerfield, “Dichotic pitches as illusions of binaural unmasking. II. the fourcin pitch and the dichotic repetition pitch,” *J. Acoust. Soc. Am.*, vol. 103, no. 6, pp. 3527–3539, 1998.

[73] D. G. Malham and A. Myatt, “3-d sound spatialization using ambisonic techniques,” *Comp. Music J.*, vol. 19, no. 4, pp. 58–70, 1995.

[74] J. Merimaa and V. Pulkki, “Spatial impulse response rendering I: Analysis and synthesis,” *J. Audio Eng. Soc.*, vol. 53, no. 12, pp. 1115–1127, 2005.

[75] J. Blauert and W. Lindemann, “Auditory spaciousness: Some further psychoacoustic analyses,” *J. Acoust. Soc. Am.*, vol. 80, no. 2, pp. 533–542, 1986.

[76] F. Rumsey, “Spatial quality evaluation for reproduced sound: terminology, meaning and a scene-based paradigm,” *J. Audio Eng. Soc.*, vol. 50, no. 9, pp. 651–666, 2002.

[77] K. Koivuniemi and N. Zacharov, “The perception of spatial sound reproduction: Language development, verbal protocol analysis and listener training,”

in *Proc 111th Audio Eng. Soc. Convention*, New York, USA, 2001, Preprint # 5424.

[78] D. VanValkenburg and M. Kubovy, “From gibson’s fire to gestalts: A bridge-building theory of perceptual objecthood,” in *Ecological Psychoacoustics*, J. G. Neuhoff, Ed., pp. 113–147. Elseiver, San Diego, 2004.

[79] E. Rubin, *Synsoplevede Figurer*, Gyldendalske, Copenhagen, Denmark, 1921.

[80] E. S. Sussman, “Integration and segragation in auditory scene analysis,” *J. Acoust. Soc. Am*, vol. 117, pp. 1285–1292, 2005.

[81] L. Wittgenstein, *Philosophical Investigations*, Malden: Blackwell, 2001.

[82] J. J. Gibson, *The senses considered as perceptual systems*, Houghton Mifflin, Boston, 1966.

[83] V. Pulkki, “Virtual source positioning using vector base amplitude panning,” *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, June 1997.

[84] M. B. Gardner, “Image fusion, broadening, and displacement in sound localization,” *J. Acoust. Soc. Am.*, vol. 46, no. 2, pp. 339–349, 1969.

[85] W. M. Hartmann, “Pitch perception and the segregation and integration of auditory entities,” in *Auditory function: neurobiological bases of hearing*, Edelman et.al., Ed. Wiley, New York, 1988.

[86] S. McAdams, “Spectral fusion and the creation of auditory images,” in *Music, mind and brain: the neuropsychology of music*, M. Clynes, Ed., pp. 279–299. Plenum, New York, 1982.

[87] M. D. Good and R. H. Gilkey, “Sound localization in noise: The effect of signal-to-noise ratio,” *J. Acoust. Soc. Am.*, vol. 99, no. 2, pp. 1108–1117, 1996.

[88] E. H. A. Langendijk, D. J. Kistler, and F. L. Wightman, “Sound localization in the presence of one or two distracters,” *J. Acoust. Soc. Am.*, vol. 109, no. 5, pp. 2123–2134, 2001.

[89] J. Braasch and K. Hartung, “Localization in presence of a distracter and reverberation in the frontal horizontal plane. i. psychoacoustical data,” *ACUSTICA/acta acustica*, vol. 88, no. 6, pp. 942–955, 2002.

[90] J. Braasch, “Localization in presence of a distracter and reverberation in the frontal horizontal plane. ii. model algorithms,” *ACUSTICA/acta acustica*, vol. 88, no. 6, pp. 956–969, 2002.

[91] J. Braasch, “Localization in presence of a distracter and reverberation in the frontal horizontal plane. iii. the role of interaural level differences,” *ACUSTICA/acta acustica*, vol. 89, no. 4, pp. 674–692, 2003.

[92] M. J. Evans, A. I. Tew, and J. A. S. Angus, “Perceived performance of loudspeaker-spatialized speech for teleconferencing,” *J. Audio Eng. Soc.*, vol. 48, no. 9, pp. 771–785, September 2000.

[93] M. Barron, “The subjective effects of first reflections in concert halls - the need for lateral reflections,” *J. Sound Vib.*, vol. 15, no. 4, pp. 475–494, 1971.

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

[95] D. Griesinger, “The psychoacoustics of apparent source width, spaciousness and envelopment in performance spaces,” *Acustica*, vol. 83, pp. 721–731, 1996.

[96] E. G. Boring, “Auditory theory with special reference to intensity, volume, and localization,” *Am. J. Psych.*, vol. 37, no. 2, pp. 157–188, 1926.

[97] D. Perrot and T. Buell, “Judgments of sound volume: effects of signal duration, level, and interaural characteristics on the perceived extensity of broadband noise,” *J. Acoust. Soc. Am.*, vol. 72, no. 5, pp. 1413–1417, 1981.

[98] G. Potard and I. Burnett, “A study on sound source apparent shape and wideness,” in *Proceedings of Int. Conf. Auditory Display.*, 2003, pp. 25–28.

[99] R. Mason, T. Brookes, and F. Rumsey, “Frequency dependency of the relationship between perceived auditory source width and the interaural cross-correlation coefficient for time-invariant stimuli,” *J. Acoust. Soc. Am.*, vol. 117, no. 3, pp. 1337–1350, 2005.

[100] R. Frigg and S. Hartmann, “Models in science,” http://plato.stanford.edu/entries/models-science/, 2006.

[101] 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, no. 5, pp. 2178–2188, 2002.

[102] A. M. Turing, “Computing machinery and intelligence,” *Mind*, vol. 59, pp. 433–460, 1950, Available also at: http://loebner.net/Prizef/TuringArticle.html.

[103] A. M. Turing, “On computable numbers, with an application to the entscheidungsproblem,” in *Proc. London Mathematical Society*, 1936, vol. 42 of *2*, pp. 230–265.

[104] J. Gleick, *Chaos: Making a New Science*, Viking Penguin, New York, 1987.

[105] H. S. Colburn and N. I. Durlach, “Models of binaural interaction,” in *Handbook of Perception*, E. C. Carterette and M. P. Friedman, Eds., pp. 467–518. Academic Press, New York, 1978.

[106] M. Slaney, “An efficient implementation of the Patterson-Holdsworth filter bank,” Tech. Rep. 35, Apple Computer, Inc., 1993.

[107] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, D. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory Physiology and Perception*, Y. Cazals, L. Demany, and K. Horner, Eds., pp. 429–446. Pergamon, 1992.

[108] R. Meddis, "Simulation of mechanical to neural transduction in the auditory receptor," *J. Acoust. Soc. Am.*, vol. 79, no. 3, pp. 702–711, 1986.

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

[110] P. H. Smith, P. X. Joris, and T. C. Yin, "Projections of physiologically characterized spherical bushy cell axons from the cochlear nucleus of the cat: evidence for delay lines to the medial superior olive," *J. Comp. Neurol.*, vol. 331, no. 2, pp. 245–260, 1993.

[111] A. Brand1, O. Behrend1, T. Marquardt, D. McAlpine, and B. Grothe, "A neural code for low-frequency sound localization in mammals," *Nature*, vol. 417, no. 6888, pp. 543–547, 2002.

[112] N.I. Durlach, "Equalization and cancellation theory of binaural masking-level differences," *J. Acoust. Soc. Am.*, vol. 35, no. 8, pp. 1206–1218, 1963.

[113] D. M. Green, "Signal-detection analysis of equalization and cancellation model," *J. Acoust. Soc. Am.*, vol. 40, no. 4, pp. 833–838, 1966.

[114] H. S. Colburn, "Theory of binaural interaction based on auditory-nerve data. I. general strategy and preliminary results on interaural discrimination," *J. Audio Eng. Soc.*, vol. 54, no. 6, pp. 1458–1470, 1973.

[115] H. S. Colburn, "Theory of binaural interaction based on auditory-nerve data. II. detection of tones in noise," *J. Audio Eng. Soc.*, vol. 61, no. 2, pp. 525–533, 1977.

[116] R. M. Stern and C. Trahiotis, “Models of binaural perception,” in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H. Gilkey and T. R. Anderson, Eds., pp. 499–531. Lawrence Erlbaum assoc., Mahwah, New Jersey, 1997.

[117] W. Lindemann, “Extensions of a binaural cross-correlation model by contralateral inhibition. I. simulation of lateralization for stationary signals,” *J. Acoust. Soc. Am*, vol. 80, no. 6, pp. 1608–1622, 1986.

[118] W. Lindemann, “Extensions of a binaural cross-correlation model by contralateral inhibition. II. the law of the first wave front,” *J. Acoust. Soc. Am*, vol. 80, no. 6, pp. 1623–1630, 1986.

[119] J. Breebaart, S. van de Par, and A. Kohlrauch, “Binaural processing model based on contralateral inhibition. I. model structure,” *J. Acoust. Soc. Am.*, vol. 110, no. 2, pp. 1074–1088, 2001.

[120] J. Breebaart, S. van de Par, and A. Kohlrauch, “Binaural processing model based on contralateral inhibition. II. dependence on spectral parameters,” *J. Acoust. Soc. Am.*, vol. 110, no. 2, pp. 1089–1104, 2001.

[121] J. Breebaart, S. van de Par, and A. Kohlrauch, “Binaural processing model based on contralateral inhibition. III. dependence on temporal parameters,” *J. Acoust. Soc. Am.*, vol. 110, no. 2, pp. 1105–1117, 2001.

[122] 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, no. 3, pp. 1442–1453, 1990.

[123] P. Jeffrey Bloom, “Pitch of noise signals: Evidence for a ”central spectrum”,” *J. Audio Eng. Soc.*, vol. 61, no. 1, pp. 150–161, 1977.

[124] M. Bodden, “Modeling human sound-source localization and the cocktail-party effect,” *Acta Acustica*, vol. 1, pp. 43–55, 1993.

[125] J. Backman and M. Karjalainen, “Modelling of human directional and spatial hearing using neural networks,” in *Proc. ICASSP-93*, Minneapolis, 1993, pp. I–125 – I–128.

[126] J. Braasch, “Pre-examination comments for dissertation,” personal communication, 2007.

Report 83 Toni Hirvonen: PERCEPTUAL AND MODELING STUDIES ON SPATIAL SOUND 2007

HELSINKI UNIVERSITY OF TECHNOLOGY
LABORATORY OF ACOUSTICS AND AUDIO SIGNAL PROCESSING

61 A. Härmä: Frequency-Warped Autoregressive Modeling And Filtering. 2001

62 V. Pulkki: Spatial Sound Generation and Perception by Amplitude Panning Technique. 2001

63 T. Altosaar: Object-Based Modelling For Representation Processing Speech Corpora. 2001

64 M. Karjalainen (ed.): Electroacoustic Transducers and DSP. 2002

65 M. Karjalainen (ed.): Measurements and Modeling in Acoustics and Audio. Seminar in acoustics, spring 2002

66 C. Erkut: Aspects in analysis and model-based sound synthesis of plucked string instruments. 2002

67 P. Korhonen (ed.): Fonetiikan päivät 2002 – The Phonetics Symposium 2002. 2002

68 H. Järveläinen: Perception of attributes in real and synthetic string instrument sounds. 2003

69 M. Karjalainen (ed.): Spatial Sound Perception and Reproduction. CD-ROM. 2003

70 R. Väänänen: Parametrization, auralization, and authoring of room acoustics for virtual reality applications. 2003

71 Tom Bäckström: Linear predictive modelling of speech - constraints and line spectrum pair decomposition. 2004

72 Paulo A. A. Esquef: Interpolation of Long Gaps in Audio Signals Using Line Spectrum Pair Polynomials

73 Paulo A. A. Esquef: Model-Based Analysis of Noisy Musical Recordings with Application to Audio Retoration. 2004

74 Kalle Palomäki: Studies On Auditory Processing Of Spatial Sound And Speech By Neuromagnetic Measurements And Computational Modeling. 2005

75 Tuomas Paatero: Generalized Linear-In-Parameter Models — Theory And Audio Signal Processing Applications. 2005

76 Klaus Riederer: HRTFAnalysis: Objective And Subjective Evaluation Of Measured Head-Related Transfer Functions. 2005

77 Juha Merimaa: Analysis, Synthesis, And Perception Of Spatial Sound – Binaural Localization Modeling And Multichannel Loudspeaker Reproduction. 2006

78 Henri Penttinen, Jyri Pakarinen, Vesa Välimäki, Mikael Laurson, Henbing Li, and Marc Leman: Model-Based Sound Synthesis Of The Guqin. 2006

79 Henri Penttinen: Loudness And Timbre Issues In Plucked Stringed Instruments - Analysis, Synthesis, And Design. 2006

80 Antti Kelloniemi, Patty Huang, Vesa Välimäki, And Lauri Savioja: Spatial Audio And Reverberation Modeling Using Hyperdimensional Digital Waveguide Meshes. 2006

81 Antti Kelloniemi, Room Acoustics Modeling with the Digital Waveguide Mesh – Boundary Structures and Approximation. 2006

82 Laura Lehto: Occupational Voice - Studying Voice Production And Preventing Voice Problems With Special Emphasis On Call-Centre Employees. 2007

ISBN- 978-951-22-9050-5
ISSN 1456-6303