Helsinki University of Technology Laboratory of Acoustics and Audio Signal Processing
Espoo 2006 Report 77

ANALYSIS, SYNTHESIS, AND PERCEPTION OF SPATIAL SOUND – BINAURAL LOCALIZATION MODELING AND MULTICHANNEL LOUDSPEAKER REPRODUCTION

Juha Merimaa

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 2006 Report 77

ANALYSIS, SYNTHESIS, AND PERCEPTION OF SPATIAL SOUND – BINAURAL LOCALIZATION MODELING AND MULTICHANNEL LOUDSPEAKER REPRODUCTION

Juha Merimaa

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 11th of August 2006, 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 951-22-8290-9
ISSN 1456-6303

Otamedia Oy
Espoo, Finland 2006

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

Author Juha Merimaa

Name of the dissertation

Analysis, Synthesis, and Perception of Spatial Sound –
Binaural Localization Modeling and Multichannel Loudspeaker Reproduction

Date of manuscript 12.7.2006	Date of the dissertation 11.8.2006
☑ 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/Audio technology
Opponent(s)	Prof. Armin Kohlrausch
Supervisor	Prof. Matti Karjalainen
(Instructor)	

Abstract

In everyday audio environments, sound from several sources arrives at a listening position both directly from the sources and as reflections from the acoustical environment. This thesis deals, within some limitations, with analysis of the resulting spatial sound field, reproduction of perceptually relevant features of the sound as measured in a chosen listening position, as well as with modeling of the related auditory localization.

For the localization, the auditory system needs to independently determine the direction of each source, while ignoring the reflections and superposition effects of any possible concurrently arriving sound. A modeling mechanism with these desired properties is proposed. Interaural time difference (ITD) and interaural level difference (ILD) cues are only considered at time instants when only the direct sound of a single source has non-negligible energy within a critical band and, thus, when the evoked ITD and ILD represent the direction of that source. It is shown how to identify such time instants as a function of the interaural coherence (IC). The source directions suggested by the selected ITD and ILD cues are also shown to imply the results of a number of published psychophysical studies.

Although the room reflections are usually suppressed in auditory localization, they contribute to the perception of the acoustical environment. The reviewed physical analysis techniques and psychoacoustical knowledge on spatial hearing are applied in development of the Spatial Impulse Response Rendering (SIRR) method. SIRR aims at recreating ITD, ILD, IC, and monaural localization cues by using a perceptually motivated analysis-synthesis method. The method is described in the context of multichannel loudspeaker reproduction of room responses with convolving reverberators. The analyzed quantities consist of the time- and frequency-dependent direction of arrival and diffuseness of sound. Based on the analysis data and a recorded omnidirectional signal, multichannel responses suitable for reproduction with any chosen surround loudspeaker setup are synthesized. In formal listening tests, it is shown that SIRR creates a more natural spatial impression than can be achieved with conventional techniques.

Keywords Auditory localization, binaural models, precedence effect, multichannel reproduction, room responses

ISBN (printed) 951-22-8290-9	ISSN (printed) 1456-6303
ISBN (pdf) 951-22-8291-7	ISSN (pdf)
ISBN (others)	Number of pages 191

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

Print distribution Report 77 / 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 Juha Merimaa

Väitöskirjan nimi

Tilaäänen analyysi, synteesi ja havaitseminen - binauraalinen paikannusmallinnus ja monikanavakaiutintoisto

Käsikirjoituksen jättämispäivämäärä 12.7.2006	Väitöstilaisuuden ajankohta 11.8.2006

☑ Monografia ☐ Yhdistelmäväitöskirja (yhteenveto + erillisartikkelit)

Osasto	Sähkö- ja tietoliikennetekniikan osasto
Laboratorio	Akustiikan ja äänenkäsittelytekniikan laboratorio
Tutkimusala	Akustiikka/audiotekniikka
Vastaväittäjä(t)	Prof. Armin Kohlrausch
Työn valvoja (Työn ohjaaja)	Prof. Matti Karjalainen

Tiivistelmä

Tyypillisissä ääniympäristöissä usean äänilähteen ääni saapuu kuuntelupaikkaan sekä suoraan lähteistä että heijastuksina akustisesta ympäristöstä. Tämä väitöskirja käsittelee tuloksena olevan äänikentän analyysiä valitussa kuuntelupisteessä, ihmiskuulijan kannalta oleellisten tilaominaisuuksien toistoa ja auditorisen paikannuksen mallinnusta.

Auditorisen paikannuksen tapauksessa kuulojärjestelmä pystyy yleensä toisistaan riippumatta määrittämään äänilähteiden suunnat jättäen huomiotta huoneheijastukset ja mahdollisen samanaikaisesti saapuvan äänen aiheuttamat summautumisilmiöt. Väitöskirjassa esitetään nämä ominaisuudet sisältävä suuntakuulon mallinnusmekanismi. Korvien väliset aikaero (ITD) ja tasoero (ILD) huomioidaan vain ajanhetkinä, jolloin ainoastaan yhden lähteen suoran äänen energia on merkityksellinen analysoidulla kriittisellä kaistalla, ja jolloin ITD ja ILD siis kuvaavat kyseisen äänilähteen suuntaa. Työssä osoitetaan, että tällaiset ajanhetket voidaan tunnistaa korvien välisen koherenssin (IC) avulla. Mallin avulla valittujen ITD- ja ILD-vihjeiden osoitetaan myös vastaavan useiden aikaisemmin julkaistujen psykofyysisten kokeiden tuloksia.

Vaikka huoneheijastuksia ei yleensä huomioida auditorisessa paikannuksessa, ne vaikuttavat akustisen ympäristön kuulohavaintoon. Väitöskirjassa kuvattuja analyysimenetelmiä ja tietoa tilakuulon toiminnasta hyödynnetään Spatial Impulse Response Rendering (SIRR) -menetelmän kehittämisessä. SIRR pyrkii toistamaan ITD-, ILD-, IC- ja monauraaliset paikannusvihjeet psykoakustisesti motivoidun analyysi-synteesimenetelmän avulla. Menetelmä esitetään sovellettuna konvoluoivien kaikulaitteiden avulla tapahtuvan huonevasteiden monikanavakaiutintoistoon. Analysoidut suureet ovat äänen tulosuunta ja diffuusisuus ajan ja taajuuden funktiona. Analyysitulosten ja pallosuuntakuvioisen mikrofonisignaalin avulla syntetisoidaan millä tahansa valitulla monikanavakaiutinjärjestelmällä tapahtuvaan toistoon soveltuvat monikanavavasteet. Formaaleissa kuuntelukokeissa osoitetaan myös, että SIRR tuottaa luonnollisemman tilavaikutelman kuin perinteiset tekniikat.

Asiasanat Auditorinen paikannus, binauraaliset mallit, presedenssiefekti, monikanavainen äänentoisto, huonevasteet

ISBN (painettu) 951-22-8290-9	ISSN (painettu) 1456-6303
ISBN (pdf) 951-22-8291-7	ISSN (pdf)
ISBN (muut)	Sivumäärä 191

Julkaisija Teknillinen korkeakoulu, Akustiikan ja äänenkäsittelytekniikan laboratorio

Painetun väitöskirjan jakelu Julkaisu 77 / TKK, Akustiikan ja äänenkäsittelytekniikan laboratorio

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

Preface

Since as early as I can remember, I have always been fascinated by sound. I like to hear it, I like to create it, and I even like to think about it. It was not long after starting my undergraduate studies at Helsinki University of Technology that it became clear to me that I also wanted to work with sound. So, here I am now, writing the preface for my dissertation, and what a journey it has been. Not only intellectually but also literally, the work on sound has taken me to quite a few places in the world. The work reported in this thesis has been carried out at Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Finland and, during the years 2003–2004, at Institute of Communication Acoustics, Ruhr-Universität Bochum, Germany. Although the manuscript was completed earlier, some minor editing was also done during my current visit to MARCS Auditory Laboratories, University of Western Sydney, Australia.

First and foremost, I am grateful to my supervisor Prof. Matti Karjalainen, who introduced me to the scientific side of sound. Matti's down-to-earth attitude combined with his innovativeness and will to keep on learning, have inspired me ever since. I am quite sure he did not anticipate the work that would follow from giving a young Ph.D. student a newly constructed microphone array.

Just when I was getting deeper into binaural psychoacoustics, I had the chance to join the team of Prof. Jens Blauert for what turned out to be two full years. Jens is not only a great music lover, but he also happens to be perhaps the world's leading expert on spatial hearing. I would like to call him the "unofficial co-supervisor" of this thesis.

I have had the opportunity to work with incredibly many talented people. The most important parts of the work reported in this thesis were done in collaboration with Drs. Christof Faller and Ville Pulkki. I would like to thank (mainly doctors or soon-to-be doctors) Jukka Ahonen, Jonas Braasch, Jörg Buchholz, Wolfgang Hess, Toni Hirvonen, Jyri Huopaniemi, Tapio Lokki, Pedro Novo, Kalle Palomäki, Timo Peltonen, Andreas Silzle, Miikka Tikander, and John Worley for further collaboration and/or numerous discussions related to my work. I also thank the previously mentioned people and all my other colleagues at TKK Acoustics lab, IKA, and the HOARSE project for numerous discussions completely unrelated to my work. There are too many of you to list all the names.

I doubt that I would have ever made it up to this point without the help of secretaries Lea Söderman and Edith Klaus with any practical problems I was able to come up with. There were more than a few. I am grateful to the pre-examiners of this thesis, Drs. Durand Begault and Nick Zacharov, for their helpful comments on

the manuscript, and to Catherine Kiwala for proofreading. I would also like to thank Prof. Rainer Martin for approving my extended stay in Bochum, Prof. John Mourjopoulos for his overwhelming Greek hospitality during a short visit to Audio Group, University of Patras, as well as Dr. Jörg Buchholz and Prof. Denis Burnham for the chance to visit MARCS.

Very special thanks for the crucial opportunities to momentarily forget the science and dive into the expressive side of sound go to Polirytmi, its more or less progressive side projects, and IKA Swing. Last but definitely not least, I am indebted to my parents and all my friends, for making me what I am and for being there for me.

The thesis was funded by the Graduate School on Electronics, Telecommunications and Automation (GETA), the research training network for Hearing Organization and Recognition of Speech in Europe (HOARSE, HPRN-CP-2002-00276), and partly supported by Tekes (VÄRE technology program, project TAKU), Academy of Finland (project 105780), Emil Aaltosen säätiö, and Kaupallisten ja teknillisten tieteiden edistämissäätiö.

Sydney, 30th June, 2006

Juha Merimaa

Contents

List of Symbols

Scalar quantities

a	Acceleration
a_{xy}	Cross-products for computing running cross-correlation
b	Bias
c	Speed of sound
c_0	Cue selection threshold
c_{12}	Maximum of coherence function
d	Distance between microphones
e	Directivity pattern of a microphone
f	Frequency
f_s	Sampling frequency
j	Imaginary unit
m	Mass (Chapter 2)
m, n	Sample indices as a function of time
p	Sound pressure as a function of time (Chapter 2); signal power (Chapter 4)
r	Radial distance
s	Distance
t	Time
u	Particle velocity as a function of time
w	Weighting function
x	Time-domain signal
x, y, z	Cartesian coordinates

A	Equivalent absorption area; amplitude
D	Distance between ears
E	Energy density
F	Force
I	Sound intensity
L	Level
ΔL	Level difference; ILD
N_f	Number of eigenfrequencies
N_r	Number of reflections
P	Pressure
P_0	Equilibrium pressure
R	Cross-correlation function
S	Surface area

T	Observation interval (Chapter 2); time constant (Chapter 4)
V	Volume
W	Work; omnidirectional B-format microphone signal
X, Y, Z	Figure-of-eight B-format microphone signals
Z_0	Characteristic acoustic impedance

α	Absorption coefficient (Chapter 2); forgetting factor (Chapter 4)
β	Directivity parameter of a gradient microphone
γ	Normalized cross-correlation, i.e., coherence function
ϕ	Elevation angle
φ	Phase variable
θ	Azimuth angle; angle of arrival
λ	Wavelength of sound
ρ	Density of a fluid
ρ_0	Mean density of a fluid
σ	Standard deviation
τ	Time lag; ITD
ψ	Diffuseness estimate
ω	Angular frequency

Complex notation of time-domain signals is indicated with $\hat{}$. Fourier transforms of the lower case symbols are denoted with respective upper case symbols with always explicitly indicating the dependence on frequency.

Vector quantities

$\mathbf{e}$	Unit vector
$\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z$	Unit vectors in the directions of the corresponding Cartesian coordinate axes
$\mathbf{n}$	Normal vector of a surface
$\mathbf{x}$	Position vector
$\mathbf{X}'$	Vector notation for B-format signals: $\mathbf{X}' = X\mathbf{e}_x + Y\mathbf{e}_y + Z\mathbf{e}_z$

Other vector quantities are denoted with bold-face symbols of the respective scalar quantities (magnitudes or one-dimensional components of the vectors). The square of a vector is defined as the square of its magnitude.

Operators

d	Differential operator
∂	Partial differential operator
∇	Vector operator $\mathbf{e}_x\partial/\partial x + \mathbf{e}_y\partial/\partial y + \mathbf{e}_z\partial/\partial z$
$\mathsf{E}\{\cdot\}$	Expectation
$\mathsf{Im}\{\cdot\}$	Imaginary part of a complex number
$\mathsf{Re}\{\cdot\}$	Real part of a complex number
$\langle\cdot\rangle$	Time average
$\lvert\cdot\rvert$	Absolute value
$\lVert\cdot\rVert$	Magnitude of a vector

List of Abbreviations

2-D	Two-dimensional
3-D	Three-dimensional
ANOVA	Analysis of variance
ASW	Auditory source width
BMLD	Binaural masking level difference
BRIR	Binaural room impulse response
EC	Equalization and cancellation
EE	Excitation–excitation
EI	Excitation–inhibition
ERB	Equivalent rectangular bandwidth
ERD	Equivalent rectangular duration
FFT	Fast Fourier transform
GCC	Generalized cross-correlation
GUI	Graphical user interface
HRTF	Head-related transfer function
IACC	Interaural cross-correlation
IC	Interaural coherence
ICI	Interclick interval
IFFT	Inverse fast Fourier transform
ILD	Interaural level difference
ITD	Interaural time difference
JND	Just noticeable difference
LEV	Listener envelopment
LF	Lateral energy fraction
LSO	Lateral superior olive
MAA	Minimum audible angle
MAMA	Minimum audible movement angle
MLS	Maximum length sequence
MOS	Mean opinion score
MSO	Medial superior olive
PDF	Probability density function
RIR	Room impulse response
RMS	Root mean square
RT	Reverberation time
SIRR	Spatial Impulse Response Rendering
SNR	Signal-to-noise ratio
SOC	Superior olivary complex

SPL	Sound pressure level
STFT	Short-time Fourier transform
TDR	Target-to-distracter ratio
TKK	Helsinki University of Technology (Teknillinen korkeakoulu)

Chapter 1

Introduction

Human beings have a remarkable ability to observe their surroundings through hearing. While vision is limited to only the frontal direction, hearing enables detection, identification, and localization of sound sources in any direction around the listener. Furthermore, human listeners are able to roughly estimate properties of the acoustical environment, such as the size of a room or a hall or the existence of a nearby wall, based on listening to the sound of a source present in the environment.

Apart from contributing to orientation, spatial features of sound can also add considerably to the pleasure of listening. Enthusiasts and performers of classical music have long been known to appreciate good concert halls and, indeed, some of the most highly regarded concert halls in the world were built more than a hundred years ago (Beranek, 1996). However, a deeper understanding of the underlying acoustical phenomena and especially of the related human perception is much more recent and a topic for ongoing research. Additionally, possibilities for creating and experiencing various spatial sound environments have multiplied with the development of sound recording and reproduction technology.

This thesis deals with both localization in everyday listening situations consisting of possibly concurrent sound sources and an acoustical environment, as well as with multichannel loudspeaker reproduction of acoustical environments. Both themes are inherently spatial, and for the most part, this thesis's treatment will be limited to spatial features. Nevertheless, three different fields of research need to be considered: 1) physical sound fields, 2) perception, and 3) reproduction techniques. Within the context of the thesis, the fields can be seen as different perspectives under the general theme of spatial sound. These three perspectives will be introduced in some more detail in Section 1.1, followed by a description of the coordinate systems used throughout the thesis in Section 1.2, and an overview of the thesis in Section 1.3.

1.1 The three perspectives

Although physical sound fields, perception, and reproduction are often studied separately, they are inseparably linked (see also Blesser, 2001). The perception by a human listener exposed to a spatial sound field is, of course, closely related to the physical phenomena of sound propagation. If this were not the case, the auditory perception

would not correspond to the actual audio environment. Furthermore, reproduction techniques can use either a physical or a perceptual approach in reconstructing an audio environment, and frequently both approaches are involved concurrently; since the accuracy of physical reconstruction is in most cases limited by technology, the necessary compromises are often at least implicitly based on perceptual considerations.

1.1.1 Physical sound fields

The physical phenomena form the basis of the issues brought up in the thesis in the sense that the physical features are what is being perceived, even if perception may also be affected by the internal state of an observer. Furthermore, if accurate physical reproduction of spatial sound were feasible, perceptual phenomena related to reproduction would not necessarily need to be addressed. From a physical point of view, the acoustical waves emitted by a sound source inside an enclosed space are reflected, diffracted, scattered and partly absorbed by different obstacles (including the walls of the enclosure) as the emitted sound energy spreads within the space. The resulting reverberant sound field is thus a superposition of waves travelling in different directions, and may include sound emitted by several sources at different locations. Unless otherwise noted, all sound, regardless of the actual form of interaction, that arrives at a position of interest due to interaction of sound with the acoustical environment, will be denoted as reflections later in this thesis.

Physical sound fields can be explored either mathematically or sampled with one or more microphones for subsequent analysis and/or reproduction. For analysis and reproduction purposes, a single static microphone is not sufficient for capturing the spatial properties of the sound field. However, with microphone arrays, the sound field can be studied either as a function of spatial position or directionally within a single position, yielding what can be called a listener-centered representation. In both cases, the resolution of the sampling of the space is necessarily finite and poses limits to the accuracy of the subsequent analysis or reproduction.

Any sound field naturally depends on its excitation. In reproduction, it is often not essential to separate the excitation (for example, a talker or a musical instrument) from the effect of the acoustical environment. However, the separation is imperative for a signal processing analysis of the acoustical environment itself. The effect of the environment can be described with a transfer functions, which can be estimated from a measured response to a known excitation. The transfer functions of acoustical environments are called *room impulse responses* (RIRs), and they will play an important role in this thesis. In general, physical analysis will be used mainly as a tool for understanding the related perception and for realizing spatial sound reproduction.

1.1.2 Auditory perception

From a human perspective, what matters in a physical sound field is that which can be perceived. Blauert (1997, p. 1) defines conscious perception as a "subject-object relationship" where the perceiver (the subject) becomes aware of what is being perceived (the object) (see also Griffiths and Warren, 2004). In case of spatial sound, what is perceived is an *auditory space* which may consist of multiple *auditory events*.

As already mentioned, the auditory space and the physical space consisting of *sound events* are related. However, there is no simple general relation. Not all sound events produce auditory events, and the resulting auditory events do not necessarily coincide with the sound events[1]. The relations between the physical and sensory events are the subject matter in the field of psychophysics, or related to audio in the field of psychoacoustics, where test subjects are asked to report some aspect(s) of perception for a carefully planned set of physical stimuli.

From the literature (see Section 3.3), it is known that in most natural listening situations, the perceived directions of auditory events indeed correspond well to the directions of the physical sound sources emitting the sound that is associated with each auditory event. As outlined earlier, in everyday *complex listening situations*, sound from multiple sources, as well as reflections from the physical surroundings, arrive concurrently from different directions at the ears of a listener. In order for the localization of the auditory events to correspond to the sound sources, the auditory system needs not only to be able to independently localize the concurrently active sources, but also to be able to suppress the sound events related to the reflections. However, the information of the reflections is not fully suppressed. Instead, it can contribute to the auditory events related to the sources as well as to the general perception of the auditory space. This contribution is what makes it possible to appreciate, for instance, good concert halls or spatial sound reproduction.

Another field of research related to psychoacoustics is auditory modeling. Auditory models aim at predicting chosen features of perception based on the stimulus signals, usually building on computational simulations of the physiologically known parts of the hearing system. The models can be applied to the evaluation of audio technology and, for instance, to audio coding to reduce sound signals to perceptually relevant components (e.g., Brandenburg and Stoll, 1994; Brandenburg and Bosi, 1997). Novel auditory models also serve as hypotheses for further psychoacoustical and physiological research. Numerous auditory models have been proposed in the literature. Nevertheless, existing models have difficulties in predicting the localization in complex listening situations.

1.1.3 Sound reproduction

The historical trend in sound recording and reproduction has been from monaural (single channel) towards increasingly elaborate spatial recording and reproduction systems. Conceptually, a straightforward method is to record the sound from the ears of a listener or an artificial head (also called dummy head) and reproduce it with headphones (for a recent overview, see Hammershøi and Møller, 2005). However, such binaural reproduction suffers from individual differences in the way that the sound is conveyed to the ears of an individual listener as well as from the fact that the sound field rotates with head movements. Although certainly not less problematic, reproduction with multichannel loudspeaker systems does not suffer from these specific

[1] In fact, auditory events can even occur without any sound events as is the case in certain disease conditions (for example tinnitus). However, this thesis concentrates on auditory perception as a response to external sound events.

problems. Multichannel sound has also been gaining more and more importance in cinemas, home theaters, and various sound installations.

As already mentioned, reproduction methods can be either physically or perceptually motivated. Another distinction can be made between the goals of authentic and plausible reproduction. *Authentic reproduction* tries to faithfully recreate a physical sound field or a perception as they would occur in a chosen audio environment. *Plausible reproduction*, on the other hand, aims at "a suitable reproduction of all required quality features for a given specific application" (Pellegrini, 2001b). Hence, a plausible reproduction may be distinguishable from the original sound field as long as it acceptably fulfills the set requirements. Note that, with the above definitions, the requirements for plausibility may, however, change as increasingly authentic reproduction becomes possible.

Ideally, an audio engineer striving for realism would record in a carefully selected performance venue capturing both the music (or any other source signal) and the acoustics at the same time. The most common systematic approach for such multichannel recording involves using one microphone per each loudspeaker used later in the reproduction. This approach already involves the perceptual assumption that, in the reproduction, the resulting auditory events do not correspond to the individual loudspeaker signals and directions, but the joint operation of the loudspeakers recreates a more or less accurate reproduction of the original physical sound field or the resulting auditory space. The most accurate reproduction is typically confined to a small area called the *sweet spot*. However, the attainable physical accuracy even in the sweet spot is severely limited by conventional microphone technology. For these reasons, the choice of the microphone configuration is often based on perceptual considerations with the goal of a plausible reproduction.

If a desired recording venue is not available or more control over the resulting reproduction is desired, it is necessary to use close microphone techniques. In such techniques, several spot microphones are placed close to sound sources to yield fairly "dry" source signals with ideally no audible room effect. An artificial scene is then constructed in the post-processing phase by positioning these signals in desired directions using, for instance, amplitude panning. Moreover, the spatial impression of a room or a hall is created with the help of reverberators or by adding the signals of additional microphones placed further away from the source(s) in the recording room. Most often, the signals are combined *ad hoc* according to the perception and artistic considerations of the recording engineer. This approach also enables creating, according to artistic considerations, soundscapes that do not exist.

A considerable amount of research has been done on plausible reverberation algorithms that do not model any specific real acoustical environment (e.g., Gardner, 1998; Väänänen, 2003; see also Pellegrini, 2001a). However, with convolving reverberators, it has also become possible to achieve similar realism with close microphone techniques as when recording directly in a chosen venue. If a set of RIRs measured in the venue with a desired microphone system is available, convolving them with a dry source signal yields the same result as recording the source in the venue, with the limitation that the actual source might have had a different directivity and thus excited the room a little differently compared to the sound source used in the measurement of the RIRs.

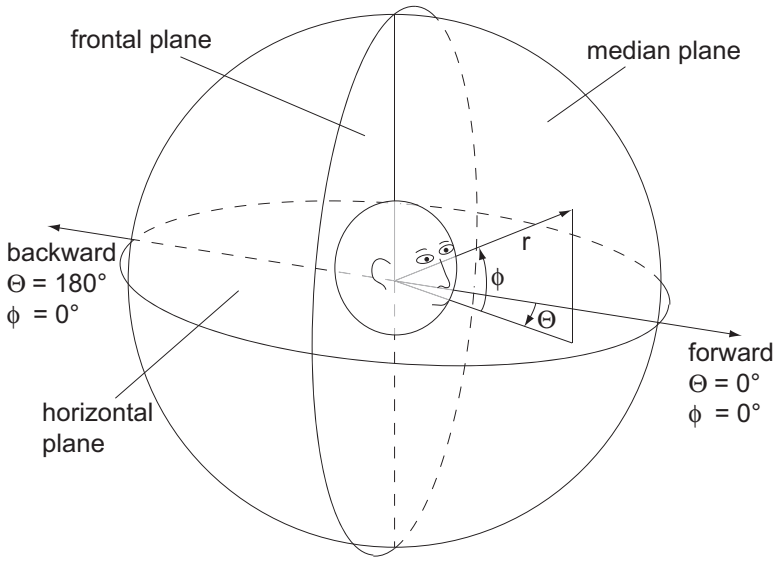

Figure 1.1: Head-related coordinate system; θ is the azimuth, ϕ is the elevation, and r is the distance (after Blauert, 1997, p. 14).

1.2 Coordinate systems

In a thesis concerned with spatial sound, it is frequently necessary to refer to a position in space. For a large part of the thesis, the *head-related coordinate system* will be used as the frame of reference. This coordinate system is illustrated in Figure 1.1, and as the name implies, the coordinates shift in conjunction with the movements of the head of a subject. The origin of the coordinate system lies halfway between the entrances of the two ear canals. The direction within the coordinate system is described with two angles: the *azimuth* θ and the *elevation* ϕ. Three important planes intersecting at the origin are also identified. The *horizontal plane* has a constant elevation $\phi = 0$ and, according to the adopted definition, the azimuth increases with movement to the left from the direction in front of the listener. The *median plane* (also called the median sagittal plane) lies at right angles to the horizontal plane and consists of positions equidistant to both ear canal entrances. Furthermore, the *frontal plane* is at right angles to both the horizontal and median planes and intersects both earcanal entrances. Positive elevation angles are used for directions above the horizontal plane.

The head-related coordinate system is well suited for describing features of spatial hearing. However, physical analysis and the related mathematical operations are often more conveniently described using the Cartesian coordinate system, as will be done. Directions in the Cartesian coordinate system can also be converted to azimuth and elevation. When describing phenomena where the presence of the head of a listener is not assumed, the conversion will be done such that the xy-plane corresponds to the horizontal plane. Furthermore, the positive x-axis is aligned with $\theta = 0$ (front), the positive y-axis with $\theta = 90°$ (left), and the positive z-axis with $\phi = 90°$ (up).

In some occasions related to discussion on auditory localization (Section 3.3), the direction of a sound source is also described using the *three-pole coordinate system*. Each of the these three coordinates defines a cone. The *left–right direction* is defined

as the angle between the median plane and the line connecting the origin of the head-related coordinate system to the sound source. Constant left–right directions thus form cones about the interaural axis, which will be later referred to as cones of confusion (see Section 3.2.1). Note that in the horizontal plane, the azimuths θ and $(180° - \theta)$ have the same left–right direction. The *up–down direction* is equal to the elevation angle and the *front–back direction* describes the angle between the frontal plane and the line connecting the origin of the head-related coordinate system to the sound source.

1.3 Outline and contributions of the author

This thesis considers the three interconnected perspectives of spatial sound in the same order as introduced earlier: physical analysis, perception, and reproduction. The order is motivated by the use of both physical analysis and perceptual knowledge in the development of a reproduction technique in Chapter 5. The thesis is by no means a complete description of all these fields of research, but rather a selection of interrelated subtopics. Nevertheless, due to the interdisciplinary nature of the work, the literature review is fairly extensive including some basics of each discipline in order to make the thesis more readable to professionals of one of the subtopics. Chapters 2 and 3 predominantly comprise background. The main scientific contributions of the author are presented in Chapters 4 and 5, followed by a summary and overall conclusions in Chapter 6.

More specifically, Chapter 2 deals with physical sound fields. Basic propagation of sound in enclosed spaces as well as microphone techniques and directional analysis of the sound propagation are described. The analysis methods provide the necessary tools for the Spatial Impulse Response Rendering (SIRR) method (Chapter 5). Furthermore, they can be used to visualize directional room responses as proposed by Merimaa *et al.* (2001). The chapter includes some mathematical derivations by the author, including a method for energetic analysis based on B-format microphones. Various microphone techniques were also discussed earlier by Merimaa (2002), and the author's publications related to measurements of spatial sound include Peltonen *et al.* (2001), Merimaa *et al.* (2005b), and Vassilantonopoulos *et al.* (2005).

Chapter 3 reviews essential features of human hearing, including physiology of the auditory periphery, psychoacoustics of auditory localization, time and frequency resolution of binaural hearing, and related auditory models. This information is later needed in both Chapters 4 and 5. Some perceptual properties of room responses are also briefly discussed as further background for Chapter 5. Scientific work related to the last mentioned topic (not reviewed in this thesis) was published by Merimaa and Hess (2004) and Merimaa *et al.* (2005a).

Chapter 4 describes a novel auditory modeling mechanism for describing human binaural localization in multi-source scenarios and reverberant environments. It is proposed that in such complex listening situations, the instantaneous interaural coherence is used to select or give more weight to cues that correspond to the actual localization. The method is examined with numerous model simulations published earlier by Faller and Merimaa (2004). Although the implications of the cue selection

model are used in the formulation of the SIRR method, the original order of the work was the opposite. Early work on SIRR (Merimaa and Pulkki, 2003) suggested that reproducing lowered coherence stabilizes the sound image by suppressing localization cues. This intuition was shared by Christof Faller based on his work on the Binaural Cue Coding method (Faller and Baumgarte, 2001, 2003; Baumgarte and Faller, 2003). All related work presented in this thesis was done in close collaboration with Faller. The ideas were jointly developed and both authors contributed equally to the writing of the original paper. Faller did most of the work on programming the simulations, whereas a majority of the simulated cases were proposed by the present author.

Finally, the SIRR method for multichannel reproduction of directional room responses is presented in Chapter 5. SIRR is a perceptually motivated analysis-synthesis method that will be shown to be able to create more natural reproductions than conventional techniques. The work was done in collaboration with Ville Pulkki, who suggested the initial idea. The present author added the analysis and synthesis of diffuseness and developed the first implementation and all applied analysis methods. Pulkki later contributed to improvements in the diffuse synthesis and did most of the work on the evaluation of SIRR. Different stages of development and evaluation of SIRR were published earlier by Merimaa and Pulkki (2003, 2004, 2005), Pulkki *et al.* (2004a,b), and Pulkki and Merimaa (2005, 2006), and in the related patent applications (Lokki *et al.*, 2003–2006). For freely available SIRR responses, see Merimaa *et al.* (2005b).

Chapter 2

Physical Analysis of Spatial Sound

2.1 Introduction

Sound can be defined as wave motion within matter. In fluids (gases and liquids), sound manifests itself as longitudinal waves involving propagating disturbances in the local pressure, density, and temperature of the fluid, together with motion of the fluid elements. In the case of typical sound events, none of these quantities is permanently changed, nor are the fluid elements permanently moved, but sound causes temporary changes in the local state of the fluid. Once the sound wave front has passed, the fluid returns to its equilibrium state. Two essential quantities in this chapter—the sound pressure and particle velocity—also describe changes from the equilibrium. The changes are typically very small. For instance, sound pressure level (SPL) is expressed relative to $2 \cdot 10^{-5}$ Pa (just audible with normal hearing), whereas the mean atmospheric pressure is roughly 10^5 Pa. It should also be emphasized that the particle velocity is the velocity related to the local oscillatory motion of the fluid elements and it is not the same as the speed of sound describing the speed of propagation of the disturbances (Cremer and Müller, 1982a; Fahy, 1989).

From the human perspective, which is prevalent in the later chapters, sound pressure is the most important quantity and what the ear is sensitive to. However, as a scalar quantity, sound pressure measured in a single position does not provide information on the spatial properties of the sound field. For spatial investigations it is necessary either to study the sound field as a function of position (e.g., Berkhout *et al.*, 1997) or the directional propagation of sound through a point in space. The discussion in this thesis is limited to the latter method, which can be considered *listener-centered* analysis. This choice is motivated by the later application of the analysis methods to the reproduction of sound as captured in a single listening position (see Chapter 5). In practice, however, the listener-centered analysis also requires sampling the sound field over a finite distance, area, or volume. In a source-free region, the directional and spatial properties of the sound field are also related such that knowing the directional properties allows constructing the spatial sound field around the measurement point and vice versa (Williams, 1999). With a limited number of spatial samples or, equivalently, limited directional resolution, the choice of the listener-centered analysis corresponds to focusing the accuracy around a chosen position.

The listener-centered directional analysis can be performed in several different ways. For the purposes of this thesis, we are interested in directional microphones and microphone systems commonly applied in spatial sound reproduction. The discussion on microphone technology illustrates the problems encountered in the reproduction and serves as background for Chapter 5. The related analytical methods, on the other hand, can be used to gain more understanding of the directional sound field and they provide necessary tools for later parts of this thesis. More specifically, it is necessary to establish methods for analyzing the time- and frequency-dependent direction of arrival and diffuseness of sound for the Spatial Impulse Response Rendering (SIRR) method (see Chapter 5). In accordance with the overall human perspective of the thesis, all discussion will be limited to audio frequencies, i.e., to the frequency range considered audible for a human listener (20 Hz to 20 kHz).

Related to reverberant acoustical environments, interaction of airborne sound with solid structures, such as different obstacles and the walls of a room, is of course also of interest. Indeed, such interaction is what creates the spatial sound field in an enclosed space. However, mathematical treatment of the related physical phenomena is beyond the scope of this thesis, and the interested reader is referred to acoustics textbooks (e.g., Cremer and Müller, 1982a,b; Pierce, 1989; Kuttruff, 2000). Sound reflection, scattering, and diffraction are described qualitatively throughout the thesis using the concepts of geometrical acoustics, rather than being treated in greater detail.

This chapter is organized as follows: Section 2.2 consists of background on sound propagation in general and especially in room environments. Section 2.3 reviews some directional microphone types and microphone array techniques. The chosen techniques will also be seen as related to the energetic analysis of sound fields described in detail in Section 2.4. The energetic analysis is applied to visualization of directional room responses in Section 2.5, and it will be used later in SIRR. However, SIRR is not limited to energetic analysis. Hence, some alternative analysis methods are briefly outlined in Section 2.6. Finally, the summary and conclusions of the chapter are given in Section 2.7.

2.2 Background

Sound propagation can be treated either as wave phenomena or approximated with geometrical acoustics using the concept of sound rays. Although it is also applicable for quantitative computations, geometrical acoustics is especially suitable for qualitatively describing sound propagation and interaction of sound with solid structures. Yet another possibility for dealing with sound propagation is the use of measurable transfer functions. As mentioned in Section 1.1.1, the transfer functions of acoustical environments are denoted as room impulse responses (RIRs), or simply room responses and they will be used throughout this thesis. The concepts of geometrical acoustics are first introduced in Section 2.2.1 and some basic properties of and measurement techniques for RIRs are described in Section 2.2.2.

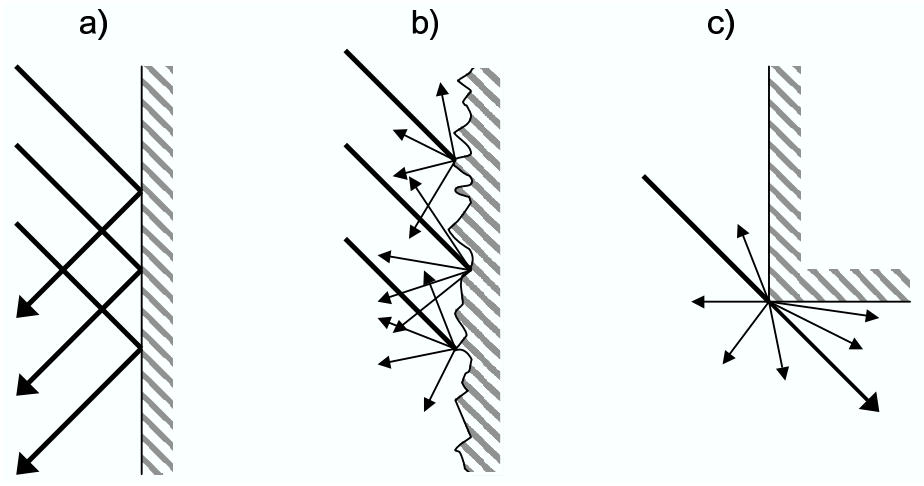

Figure 2.1: Illustration of a) specular reflection, b) scattering (diffuse reflection), and c) edge diffraction.

2.2.1 Geometrical acoustics

In geometrical acoustics (Cremer and Müller 1982a, Part I; Kuttruff 2000, Chapter 4), the sound power emitted by a source is divided into portions propagating in different directions according to the directivity of the source. For a point source yielding a spherical wave front (diverging rays), the power along each ray is attenuated as $1/r^2$, where r denotes the distance from the sound source. Furthermore, the power of the rays is reduced as a function of distance by the frequency-dependent absorption of the air. This absorption depends on the temperature and humidity and increases as a function of frequency (ISO 9613-1, 1993).

When the propagating sound reaches a rigid obstacle, the resulting behavior depends on the relation of the wavelength of the sound to the dimensions of the obstacle and its surface structure. The interaction with a large smooth surface creates a mirror-like *specular* reflection. On the other hand, if the surface is rough compared to the wavelength of the sound, a *diffuse* reflection takes place, scattering the sound rays in multiple directions away from the surface. Specular and diffuse reflections are illustrated in panels a) and b) of Figure 2.1, respectively. In both cases the surface may also absorb part of the sound energy in a frequency-dependent manner. Since the reflection behavior depends on the roughness of the surface relative to the wavelength of sound, it should be noted that low frequencies (long wavelength) may be reflected specularly and higher frequencies (shorter wavelength) increasingly diffusely from a single surface. Furthermore, at even higher frequencies, portions of the same surface may again act as specular reflectors.

Yet another phenomenon takes place when an obstacle is blocking the direct path of propagation. The shadowing effect of the obstacle depends again on its size compared to the wavelength of sound such that when the wavelength is large, the sound effectively "bends" around the obstacle. This phenomenon is called *diffraction*. In addition to the bending, a wedge of any angle apart from 90° (the inside of a rectangular corner) and 180° (which is actually not a wedge) also scatters sound in all directions resulting in edge diffraction as illustrated in panel c) of Figure 2.1 (Svensson *et al.*,

1999; Torres *et al.*, 2000; see also Pulkki and Lokki, 2003). The edge diffraction is also the microscopic mechanism responsible for diffuse reflection (Dalenbäck *et al.*, 1994).

2.2.2 Room responses

Room acoustical phenomena excluding (typically undesired) resonances of mechanical structures are approximately linear, which makes it possible to use the room responses to describe the transfer of sound from one point in space to another. Room responses can be applied both in objective and subjective analysis of room acoustics, as well as in auralization[1] and reproduction. Moreover, RIRs for any of the previously mentioned tasks can be either directly measured or computed using geometrical description and specifications of the surface materials of an acoustical space (e.g., Allen and Berkley, 1979; Lehnert and Blauert, 1992; Kleiner *et al.*, 1993; Savioja *et al.*, 1999; Blauert *et al.*, 2000; Lokki, 2002; Novo, 2005). The discussion and examples in this chapter will be limited to measured RIRs, thus avoiding the need to consider imperfections in the practical computation methods.

An impulse response of a room or a hall consists of a multitude of sound events produced by the reflections, scattering, and diffraction phenomena described in the previous section. RIRs are typically divided into three subsequent parts: the direct sound, early reflections, and late reverberation. The direct sound corresponds to the sound propagating via a straight line between a source and a receiver. It is always the first sound event in a response unless, of course, some obstacle is blocking the direct path and thus preventing the direct sound from occurring. However, even in such a case at least some low frequency sound is usually diffracted to the measurement position before any of the early reflections. Apart from the attenuation caused by the distance to the sound source and absorption of the air, the direct sound event does not in principle depend on the acoustical environment or source and receiver positions. However, the seats of a concert hall may form an acoustical resonator structure, causing some attenuation of sustained direct sound passing at near-grazing incidence across them, which is known as the seat dip effect (e.g., Takahashi, 1997; Davies and Cox, 2000)

The early reflections are mainly discrete sound events, whereas the late reverberation corresponds to diffuse sound arriving simultaneously from several directions. According to a standard definition for concert halls (ISO 3382, 1997), the sound arriving within 80 ms from the direct sound is considered as early reflections and the sound arriving after 80 ms as late reverberation. However, this division is mainly of a perceptual nature (see Section 3.5), and from a physical point of view there is no single time instant when the early reflections change to late reverberation. Instead, the density of reflections grows steadily as the emitted sound energy spreads in the room. Neglecting scattering and diffraction, the average density of reflections arriving to a receiver position can be shown to follow

$$\frac{\mathrm{d}N_r}{\mathrm{d}t} = 4\pi \frac{c^3 t^2}{V}\,, \tag{2.1}$$

[1] Auralization is an acoustic analog of visualization defined by Kleiner *et al.* (1993) as "the process of rendering audible, by physical or mathematical modeling, the sound field of a source in a space, in such a way as to simulate the binaural listening experience at a given position in the modeled space."

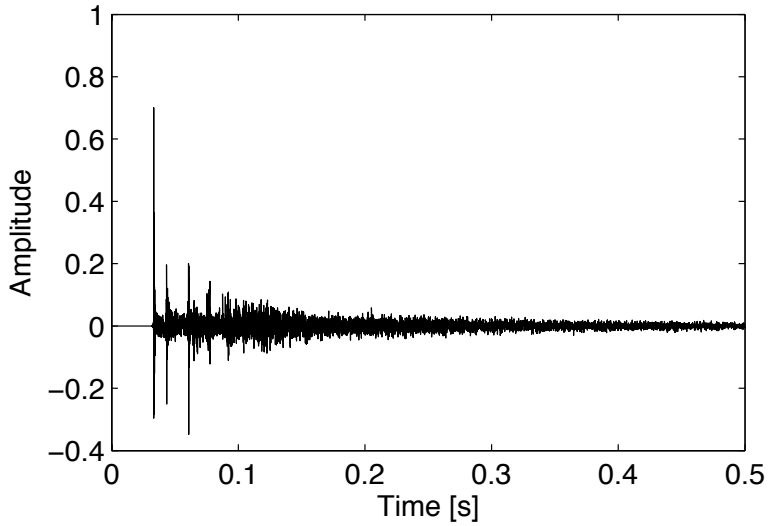

Figure 2.2: Example of a concert hall response: Time-domain plot for the first 0.5 s of response `s2_r2_o` from the database of Merimaa *et al.* (2005b).

where N_r is the number of reflections, t is the time from the direct sound, c is the speed of sound, and V is the volume of the room (Kuttruff, 2000, p. 98). The sound is also attenuated by the absorption of the reflecting surfaces, and if the absorption is uniform on all surfaces, the total sound energy follows an exponential decay curve. Figure 2.2 illustrates one RIR measured from a medium-sized concert hall. The relatively sparse peaks in the beginning are the direct sound followed by some early reflections. Furthermore, the diffusers on the walls of the hall result in a considerable amount of sound being directed to the measurement position also between the first specular reflections.

The decay of energy during the propagation of sound has been throughly studied in the literature and is most often characterized by reverberation time (RT), defined as the time that it takes for the sound inside a hall to decay 60 dB after a sound source is turned off (ISO 3382, 1997). Sabine (1900) derived the famous law for the RT, often quoted nowadays in the form

$$\mathrm{RT} = 0.163\frac{V}{A}\ \text{seconds}, \tag{2.2}$$

where $A = \alpha S$ is the equivalent absorption area of the room and α is the mean absorption coefficient and S is the surface area of the boundaries in the room (Kuttruff, 2000, p. 118–119). The equation is strictly valid only for decay starting from an ideally diffuse field in a room with uniformly absorbing boundaries and negligible air absorption. However, it still often serves as the first approximation of the reverberation time. Several correction terms for the equation have also been proposed later (e.g., Kuttruff 2000, Chapter 5; Barron and Lee 1988; Barron 1995a; see also Barron 1995b). When a steady sound is switched on in a room, the room also features a buildup complementary to the decay (Schroeder, 1966).

The frequency domain features of the steady state sound field in a room have also provoked interest. Rooms exhibit modes where some frequencies are amplified due to

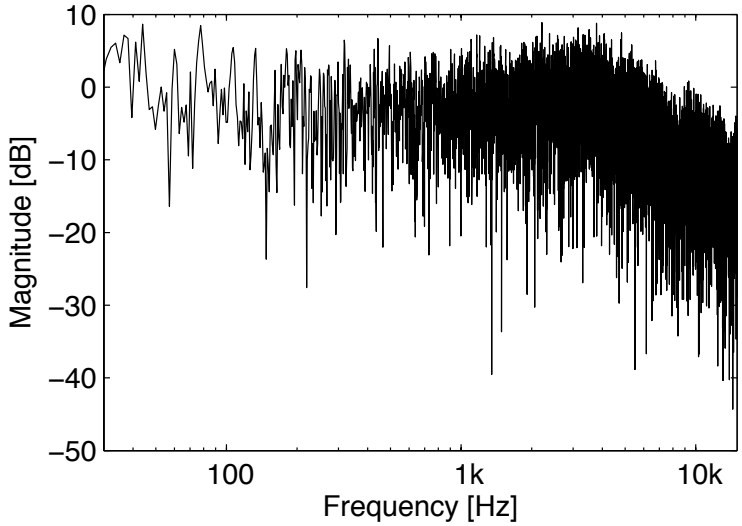

Figure 2.3: Example of a concert hall response: Frequency-domain plot of a 4 s long measurement of the same response shown in Figure 2.2. The decay of the level above about 4 kHz is due to absorption both by the air and by the surfaces of the hall.

interference of reflected sound traveling in different directions. From the wave field point of view, the interference creates more or less ideal standing waves. The modal frequencies corresponding to the standing waves (also called eigenfrequencies) depend on the dimensions and geometry of the room. Furthermore, the modes that are excited and can be observed depend on the source and measurement positions, respectively. However, indication for the general modal behavior can be obtained by considering the modes of a rectangular room. The average density of modes per Hz in a rectangular room can be approximated with

$$\frac{\mathrm{d}N_f}{\mathrm{d}f} = 4\pi V \frac{f^2}{c^3}\,, \tag{2.3}$$

where N_f is the number of modes and f is the frequency (Kuttruff, 2000, p. 70). The form of the Eq. (2.3) is similar to Eq. (2.1) and it can be seen that the density of the modes grows with the square of frequency. Consequently, apart from low frequencies, the modes usually cannot be observed individually. As an example of the frequency domain behavior, the Fourier transform of the RIR shown earlier in Figure 2.2 is illustrated in Figure 2.3. For further discussion on statistics of room responses, see e.g. Schroeder (1954a,b, 1962, 1965, 1996), Polack (1993), Jot *et al.* (1997), and Blesser (2001).

As already mentioned, RIRs can be readily measured. The measurement process consists of exciting a room with a chosen method, capturing the resulting sound with one or more microphones, and of potential post-processing involving deconvolution of the excitation signal and system compensation. The choice of sound sources and microphones naturally affects the measurement results. Nowadays loudspeakers are almost exclusively used as sources. ISO 3382 (1997) requires using a loudspeaker (or generally any source) as close to omnidirectional as possible in measurements pertaining to acoustical analysis. However, for auralization purposes it may be desirable to

use a more typical loudspeaker with closer resemblance to the directivity of natural sound sources, as done by e.g. Ben-Hador and Neoran (2004).

In acoustical measurements, there is always inherent background noise and consequently maximizing the signal-to-noise ratio (SNR) becomes a concern. The SNR is usually limited by the achievable sound levels of a loudspeaker and it can be improved by using prolonged excitation signals that can be deconvolved out of the measured responses (for an overview, see Müller and Massarani, 2001, 2005). The most popular current excitation signals are the maximum length sequences (MLS; Schroeder, 1979) and logarithmically swept sinusoids (Farina, 2000; Müller and Massarani, 2001, 2005; for other methods, see also Heyser, 1967; Berkhout *et al.*, 1980; Aoshima, 1981; Suzuki *et al.*, 1995). The optimal excitation signal depends on the measurement system. MLS has theoretically very attractive properties but it relies on the assumption of linear time-invariant systems. On the other hand, logarithmic sweeps are able to separate harmonic distortion produced, for instance, by the loudspeaker from the linear response, which makes sweeps in many cases superior compared to other excitations (for an experimental comparison, see Stan *et al.*, 2002). All RIRs presented in this thesis have been measured with logarithmic sweeps. In some cases the measurement process has also involved careful compensation of the magnitude response of the measurement system (for a case study, see Merimaa *et al.*, 2005b).

2.3 Directional microphones and microphone arrays

Microphones usually transduce the sound pressure into an electrical signal. An ideal omnidirectional pressure microphone senses the movement of a diaphragm as a response to the pressure variations on one side of the diaphragm. Although it is also possible to directly measure other direction-dependent quantities of the sound field (Olson, 1991), pressure microphones or microphones arrays can also be made sensitive to the direction of arrival of sound. The resulting directional signals can then be utilized both in analysis of the sound field (e.g. Okubo *et al.*, 2000; Gover, 2002) and in reproduction (see Section 5.2).

In directional microphone techniques, the signals from several sampling points are combined such that sound arriving from a desired direction is amplified and/or sound from undesired directions is attenuated. Single-diaphragm directional microphones commonly used in recording applications utilize the difference of sound pressure on both sides of the diaphragm. Equivalent processing can also be realized with two closely spaced omnidirectional microphones, as will be discussed in this section. Furthermore, various acoustical constructions can be used to impose direction-dependent delays, interference, and/or filtering within the spatial sampling process (for an overview, see Olson, 1991; Eargle, 2001).

More flexible directional processing than with single-diaphragm microphones can be achieved with larger microphone arrays. The construction of a directional signal from an array is usually denoted as *beamforming*. A general design procedure for array-based beamforming has a vast number of free parameters starting from the number

and layout of the microphones and extending to frequency-dependent weighting and delaying of each individual microphone signal. Additionally, the microphones in the array can have built-in directivity, and the weights and delays of the individual signals can be varied dynamically in order to realize beam steering. In general, the higher the number of microphones, the higher the directivity that can be achieved. The wide range of audio frequencies, however, poses special problems on the design of highly directive wideband microphone arrays because the (effective) dimensions of the array and the spacing of the microphones within the array need to be proportional to the wavelength of sound. This difficulty also motivates the search for alternative reproduction methods, such as the SIRR technique developed later in this thesis.

A comprehensive treatment of microphone array techniques is beyond the scope of this thesis. This section is limited to two specific microphone techniques commonly encountered in high-quality recording applications: The basic principles and problems of the gradient techniques are introduced in Section 2.3.1, and the B-format, based on spherical harmonic decomposition of the sound field, is described in Section 2.3.2. This section establishes the link between gradient techniques and first order B-format, which will be important for the derivation of the energetic analysis based on B-format signals. B-format signals will also be extensively used later in Chapter 5. For other microphone array techniques and implementations, see e.g. Broadhurst (1982), Flanagan (1985), Kummer (1992), Pumphrey (1993), Holm *et al.* (2001), Meyer (2001), and Merimaa (2002).

2.3.1 Gradient microphones

A basic gradient microphone senses the difference in sound pressure between two closely spaced points. As already mentioned, an equivalent system to a single-diaphragm microphone using the pressure difference between the front and back of the diaphragm can be constructed from two closely spaced omnidirectional microphones whose responses are subtracted from each other. The differentiation can be seen as a finite difference approximation of the gradient of the sound field under the assumption that the dimensions of the microphone system are small compared to the wavelength of sound, hence the name gradient techniques. Furthermore, the pressure gradient is proportional to the particle velocity, as will be shown in Section 2.4.3. For this reason, gradient microphones are often called *velocity microphones*.

In the following, it is assumed that the sound pressure is sampled ideally in the two measurement points and the microphone or microphones do not affect the sound field. For a more detailed analysis of practical gradient microphones, see Olson (1991, Sections 8.3–8.5), Eargle (2001, Chapters 4–5), and Olson (1973)[2]. With the previous assumptions, the directivity and frequency response of a gradient microphone in the far field of a source where the arriving wave fronts can be approximated as planar, can be derived as follows. Consider a plane wave with an angular frequency ω arriving to a microphone pair with an angle θ as depicted in Figure 2.4. Let the sound pressure halfway between the microphones be $\hat{p}_{\mathrm{mid}}(t,\omega) = Ae^{j\omega t}$, where A is the complex

[2]Olson (1973) actually discusses gradient loudspeakers, but the methods are analogous to gradient microphone techniques.

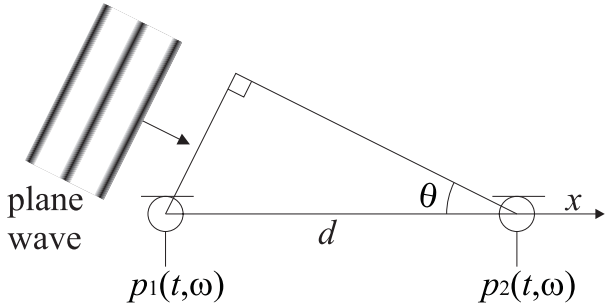

Figure 2.4: Plane wave arriving from angle θ to a microphone pair placed on the x-axis with a distance d between the microphones.

amplitude, j is the imaginary unit, and $\hat{}$ denotes a complex notation of a time-domain signal such that the instantaneous sound pressure is given as $p(t,\omega) = \mathsf{Re}\{\hat{p}(t,\omega)\}$. The sound pressure at the positions of the two microphones can be now written as

$$\hat{p}_1(\omega,\theta) = \hat{p}_{\mathrm{mid}}(\omega) e^{j\omega \frac{d\cos(\theta)}{2c}} \tag{2.4}$$

and

$$\hat{p}_2(\omega,\theta) = \hat{p}_{\mathrm{mid}}(\omega) e^{-j\omega \frac{d\cos(\theta)}{2c}}\,, \tag{2.5}$$

where d is the distance between the microphones, c is the speed of sound, and the exponential terms describe frequency-independent delays of $\Delta t = d\cos(\theta)/2c$. Note that the explicit dependence of the quantities on t has been dropped for typographic clarity.

Using a trigonometric identity, the difference of the two microphone signals can be written as

$$\hat{p}_{\mathrm{dif}}(\omega,\theta) = \hat{p}_1(\omega,\theta) - \hat{p}_2(\omega,\theta) = \hat{p}_{\mathrm{mid}}(\omega) 2j \sin\left(\omega \frac{d\cos(\theta)}{2c}\right)\,. \tag{2.6}$$

Some magnitude responses according to Eq. (2.6) are shown in Figure 2.5 as a function of $d/\lambda = fd/c$, where $\lambda = 2\pi c/\omega$ is the wavelength of sound. At low frequencies, the illustrated directivity patterns have the shape of a figure-of-eight which starts to split into multiple beams above $d/\lambda = 0.5$. The splitting is the result of *spatial aliasing*, which usually sets the upper frequency limit for the operation of any directional microphone system. Another prominent feature is a considerable variation in the magnitude response. When

$$\omega \frac{d}{2c} \ll 1 \Leftrightarrow \frac{d}{\lambda} \ll \frac{1}{\pi}\,, \tag{2.7}$$

the small angle approximation $\sin(x) \approx x$ can be used to yield

$$\hat{p}_{\mathrm{dif}}(\omega,\theta) \approx \hat{p}_{\mathrm{mid}}(\omega) j\omega \frac{d\cos(\theta)}{c}\,. \tag{2.8}$$

This equation clearly displays a figure-of-eight (cosine) directivity, first-order highpass characteristics (multiplication with ω) and a 90° phase shift (j) which is an artifact of the differentiation.

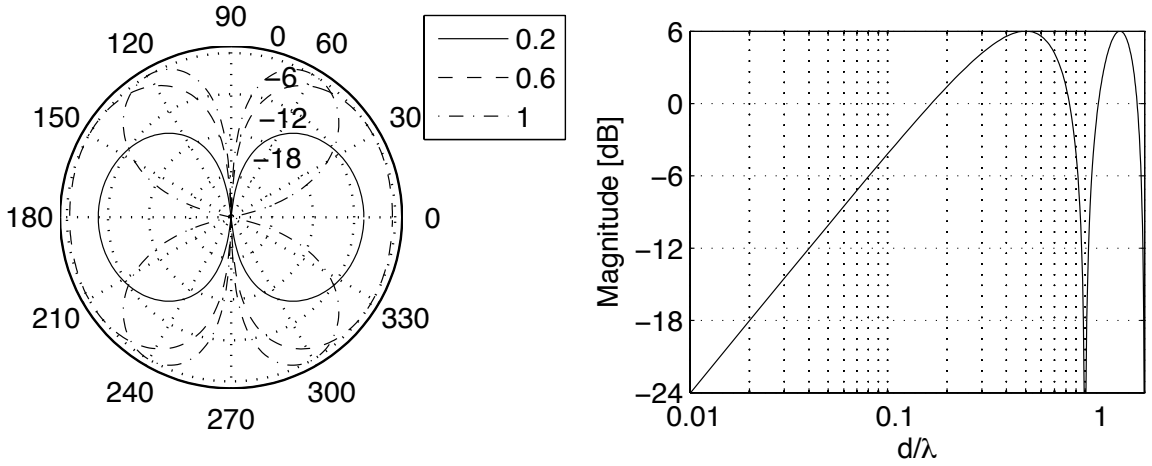

Figure 2.5: Logarithmic magnitude responses of an ideal gradient microphone according to Eq. (2.6). Left: Polar plots as a function of θ for $d/\lambda = 0.2,\ 0.6, 1$. Right: Magnitude response for $\theta = 0$ as a function of d/λ.

To derive a signal with a frequency dependent magnitude proportional to the sound pressure, a first-order lowpass filter needs to be applied to the response (for realization of the lowpass filtering in single-diaphragm microphones, see Olson, 1991, p. 275–279). In individually used gradient microphones, it may not be necessary to compensate for the phase shift or overall sensitivity of the gradient microphone. However, an ideal microphone sampling the sound field with a figure-of-eight directivity pattern at low frequencies can be theoretically implemented as

$$\hat{p}_8(\omega, \theta) = \hat{p}_{\mathrm{mid}}(\omega) \cos(\theta) = -\frac{jc}{\omega d} \left[\hat{p}_1(\omega, \theta) - \hat{p}_2(\omega, \theta)\right] . \tag{2.9}$$

The lower operational frequency limit of a gradient microphone is usually defined by measurement errors. With the pressure difference at a fixed distance decreasing towards low frequencies, measurement noise or phase errors in the sampling begin to dominate the low-frequency output unless the frequency range is limited or the microphone is designed to approach omnidirectional at low frequencies. The pressure difference at low frequencies can, of course, be increased by increasing the distance between the sampling points, but this will result in spatial aliasing starting at lower frequencies.

From Eqs. (2.8) and (2.9), it is obvious that sound events arriving from opposite directions are transduced with opposite phases. This property enables modification of the directivity pattern by adding an omnidirectional component to the gradient signal. The omnidirectional part then cancels out some of the negative lobe of the figure-of-eight pattern. General first order directivity patterns constructed this way are of the form

$$e(\theta) = \beta + (1 - \beta) \cos(\theta), \quad \beta \in [0, 1] , \tag{2.10}$$

where β is the directivity parameter describing the proportion of the omnidirectional component. The directional effect of β with some typical values is illustrated in Figure 2.6. It should be emphasized that in using the previous summation principle it is

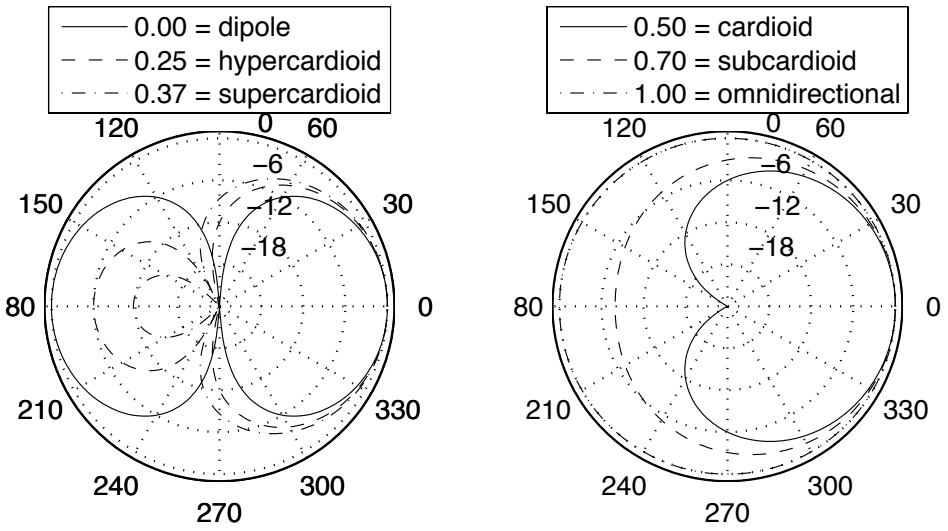

Figure 2.6: Polar plots of the logarithmic magnitude responses of first-order gradient directivity patterns according to Eq. (2.10) with some typical values of β.

important to fully equalize the figure-of-eight signal according to Eq. (2.9). However, the same directivity patterns can also be realized by introducing a delay to one of the omnidirectional signals before the differentiation. Specifically, for $\beta < 1$ it is easy to show that adding a delay of $\Delta t = (d/c)[\beta/(1-\beta)]$ to the signal $\hat{p}_2(\omega)$ in Eq. (2.6) yields a low-frequency directivity according to Eq. (2.10)[3].

So far, the discussion has only involved propagating plane waves. However, in general sound fields, the magnitude of the sound pressure may vary between the two microphones. For the purposes of the current discussion, consider a gradient microphone placed close to a sound source. The magnitude of the sound pressure resulting from a point source decreases as $1/r$ over the distance (corresponding to the attenuation of power as $1/r^2$, as stated earlier in Section 2.2.1). For small r, the difference in the pressure at the two sampling points can be substantial due to the different relative distances to the source. Furthermore, this pressure difference is constant as a function of frequency, whereas the difference due to the gradient of the sound field decreases towards low frequencies. Consequently, placement of the microphone close to a source creates a relative amplification of low frequencies, which is known as the *proximity effect* (Eargle, 2001, p. 77–79 and 91–93).

Gradient techniques can, of course, be extended to utilize sampling of the sound field at more than two positions. Second order gradient patterns can be formed by subtracting the signals of two closely spaced gradient microphones (e.g., Woszczyk 1984; Sessler *et al.* 1989; Olson 1991, p. 311–319; Eargle 2001, p. 116–123; see also Korenbaum 1992, Sessler and West 1992, and Elko *et al.* 1994), and the methods can be generalized to higher orders. However, the frequency range limitations grow more severe with increasing order of the gradient, thus limiting the achievable directivity. In practice, higher than first-order gradient microphones are rare in high-quality

[3]Elko (2000) also proposed an alternative technique for facilitating the equalization by constructing the general directivity patterns as a weighted sum of two back-to-back cardioid responses formed first with delays d/c (i.e., $\beta = 0.5$).

recording applications.

Microphones with steerable beams are another possibility for utilizing several sampling points. As shown earlier, various first-order directivity patterns can be formed as a sum of an omnidirectional and a figure-of-eight signal. Hence, all that is needed for a general steerable first-order gradient microphone is a method to steer the figure-of-eight pattern. A two-dimensional steering can be formulated based on the following trigonometric identity

$$\cos(\theta - \theta') = \cos(\theta)\cos(\theta') + \sin(\theta)\sin(\theta')\,, \tag{2.11}$$

where θ' is the desired change in the angle of orientation. The right-hand side of the equation shows that such a change can be realized by including the output of a second microphone with a directivity pattern $\sin(\theta)$, i.e., that of a concentric figure-of-eight microphone perpendicular to the first microphone. Similarly in three dimensions, a microphone signal $p_8(\theta', \phi')$ with a figure-of-eight pattern pointing at azimuth θ' and elevation ϕ' (see coordinate systems in Section 1.2) can be formed as a weighted combination of three figure-of-eight microphone signals aligned with the Cartesian coordinate axes with the following ideal direction-dependent relation to the sound pressure of a plane wave

$$X = p\cos(\theta)\cos(\phi)\,, \quad Y = p\sin(\theta)\cos(\phi)\,, \quad Z = p\sin(\phi)\,, \tag{2.12}$$

yielding

$$p_8(\theta', \phi') = \cos(\theta')\cos(\phi')X + \sin(\theta')\cos(\phi')Y + \sin(\phi')Z\,. \tag{2.13}$$

It happens that the signals X, Y, and Z defined above are also proportional to the first-order B-format signals discussed below.

2.3.2 B-format

As mentioned earlier, B-format signals are based on a spherical harmonic decomposition of the sound field. The sound field on an imaginary sphere around a measurement point can be expressed as a weighted sum of an infinite number of spherical harmonics, in a manner analogous to the representation of a periodic signal with its Fourier series (Williams, 1999). The spherical harmonics correspond to directivity patterns whose directivity increases with the order[4] of the harmonics. A truncation of the spherical harmonic series to a certain order allows a representation of the sound field with a limited directional resolution. In practice, the truncation is necessary due to limitations in the measurement technology and a limited number of available microphones for a recording array.

The zero-order B-format component W is an omnidirectional signal. Disregarding the sensitivity of the microphone, it can thus be defined as $W = p$. The first order signals X, Y, and Z correspond to the orthogonal gradient (figure-of-eight) signals

[4]The term order may create some confusion and it is used here according to a common convention in B-format and Ambisonics literature. What is actually meant with order n is spherical harmonics Y_n^m of degree $n \geq 0$ including orders $m \in [-n, n]$. For a description of the spherical harmonics, see Williams (1999, Section 6.3).

defined in Eq. (2.12) and multiplied by $\sqrt{2}$ (Farrar, 1979a; Jagger, 1984; see also Gerzon, 1973)[5]. Higher-order components, however, are not directly proportional to the gradient directivity patterns of the same order but correspond to linear combinations of the gradient patterns up to the same order (Gerzon, 1973; Cotterell, 2002).

It is obvious that first-order B-format recording can also be realized by four closely spaced, properly aligned, and calibrated discrete microphones: one omnidirectional and three with a figure-of-eight directivity pattern. However, array techniques make it possible to achieve better coincidence of the individual components. A suitable recording array can be easily constructed from microphone pairs aligned at each Cartesian coordinate axis, and array systems for first-order B-format recording are also commercially available. Practical implementations of higher-order recording are, on the other hand, fairly recent and require a relatively high number of microphones.

The spherical harmonic decomposition can be directly applied in the Ambisonics sound reproduction method (see Section 5.2.2). However, due to the possibility to easily form and steer any general first order gradient pattern, the applications of B-format are not limited to Ambisonics. Furthermore, the relation of the omnidirectional and gradient signals to sound pressure and particle velocity, respectively, makes it easy to use B-format in energetic analysis of sound, as will be shown in Section 2.4.8. This property and the commercial availability of first-order B-format microphones make them important for this thesis. In the following, some available first-order implementations are described, followed by a brief review of the recent higher-order techniques.

Soundfield microphones

Soundfield microphones are the best-known commercial devices capable of recording first-order B-format. Soundfield microphone systems consist of a multicapsule microphone unit connected to a preamplifier and control unit. The microphone unit includes four closely spaced subcardioid microphone capsules arranged at the corners of a regular tetrahedron. The four output signals of the capsules are denoted as A-format, which is then converted to B-format (and optionally other adjustable outputs) by the control unit. The conversion consists of simple summation, differentiation, and equalization operations (Gerzon, 1975; Craven and Gerzon, 1977; Farrar, 1979a,b; Jagger, 1984).

As the four capsules are not perfectly coincident, spatial aliasing is, of course, an issue at high frequencies. Gerzon (1975), Farrar (1979a), and Jagger (1984) all claim adequate directional operation up to approximately 10 kHz. Indeed, measurements by Farina (2001a,b) show that the directivity patterns are roughly similar from 125 Hz to 8 kHz, although there are some distortions and asymmetries at all frequencies indicating an imperfect estimation of the spatial gradients. Furthermore, the relative levels of the omnidirectional and figure-of-eight signals deviate considerably as a function of frequency, which makes the general directivity patterns somewhat

[5]The purpose of the 3 dB ($\sqrt{2}$) gain of X, Y, and Z relative to W is to make their energy closer to that of W in a diffuse sound field, which was important in oder to maximize the SNR in analog recordings. Gerzon (1973) actually uses a gain of $\sqrt{3}$ which equalizes the levels according to the normalized real spherical harmonics.

frequency-dependent, and will have some consequences for the energetic analysis.

Microflown technologies

Microflown (de Bree, 2003) is a recent advancement in direct measurement techniques for the particle velocity (for the exact relation between the sound pressure gradient and particle velocity, see Section 2.4.3). The Microflown sensor is made by microtechnology and consists of two heated platinum wires less than 1 μm thick. Particle velocity induces asymmetric changes in the temperature of the wires and consequently in their resistance, which can be easily measured. The sensors offer an extremely wide frequency range covering all audio frequencies with a figure-of-eight directivity pattern and good signal-to-noise ratio. Furthermore, a microscale array with three orthogonal velocity sensors and a pressure microphone is also available. Although the frequency response of the Microflown (de Bree, 2003, Figure 12) would seem to require some equalization for recording purposes, the technology would be an attractive alternative for future work related to the analysis and reproduction aspects of this thesis.

TKK 3-D array

The TKK array (Peltonen *et al.*, 2001; Merimaa, 2002) is a custom microphone system built by Goutarbès at the Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology (TKK). A picture of the array is shown in Figure 2.7. The array consists of twelve omnidirectional miniature electret microphone capsules arranged as six concentric pairs, two pairs on each of x-, y-, and z-coordinate axes. The inner pairs are set 10 mm apart from each other and the outer pairs have a spacing of 100 mm. The main motivation for the two spacings is to extend the bandwidth in first-order directional processing by cross-fading between the two pairs aligned in the same direction. The usable frequency range is roughly 100 Hz – 10 kHz. Furthermore, the orthogonal geometry of the array makes extracting B-format straightforward, and the array could also be used for forming some second-order gradient patterns at a more limited frequency range. For a more comprehensive description of the possible applications, see Merimaa (2002).

Higher-order B-format microphone arrays

Instead of gradient techniques, the recent higher-order B-format systems are all based on direct spherical harmonic sampling of the sound field. As already mentioned, the number of necessary microphones grows with increasing directivity, and the frequency range limitations are more severe than at first order. Abhayapala and Ward (2002) described a spherical shell array consisting of 25 omnidirectional microphones on a spherical surface with a radius of 4 cm. The array was capable of recording spherical harmonics up to third order at a frequency range of 340 Hz – 3.4 kHz. Furthermore, Meyer and Elko (2002) considered 32 microphones mounted on the surface of a rigid sphere (with a radius of 5 cm) providing some advantages to the frequency range and yielding third order spherical harmonics with the spatial aliasing staying at moderate levels up to 5 kHz.

Figure 2.7: TKK 3-D microphone array

The design principles of Laborie *et al.* (2003) appear very promising and allow using arbitrary array layouts and directivity patterns for the individual microphones. Furthermore, the compromise between high directivity and SNR at low frequencies can be adjusted using a single parameter. An illustrated B-format prototype consisted of a spherical array of 24 omnidirectional microphones where some of the microphones were placed inside the sphere, yielding third order spherical harmonics with reasonable directional accuracy even above 5 kHz. A theoretical framework for all the previously mentioned techniques was recently presented by Poletti (2005). For related theoretical analysis, see also Williams (1999), Poletti (2000), Cotterell (2002), and Rafaely (2005).

2.4 Energetic analysis

In the previous section, some possibilities for recording the sound pressure as a function of the direction of arrival of sound were described. Another possibility for directional analysis is to observe the transfer of energy. The transfer of energy indeed takes place even if in Section 2.1 it was stated that in case of typical sound events the movement of fluid elements and changes in local pressure are only temporary disturbances. The sound energy is created by the action of a sound source on the fluid immediately surrounding it, and it is transported with the propagating disturbances.

The energetic analysis presented in this section has a firm mathematical basis and for the treatment, some basic concepts and assumptions need to be defined more

precisely. The sound pressure p is defined as the change from the equilibrium pressure P_0 of the fluid, and the total pressure is thus $P = P_0 + p$. Correspondingly, two quantities are introduced for the density of the medium: ρ is the instantaneous density (analogous to P) and ρ_0 is the mean (equilibrium) density. As mentioned in Section 2.1, p/P_0 is typically very small and the same applies to the relation of $(\rho - \rho_0)/\rho_0$. Hence, the approximation $\rho \approx \rho_0$ will be used often. Throughout this section, it will also be assumed that the temperature of the fluid is constant apart from changes induced by the sound, and convection, i.e., net movement of the fluid (which cannot be created by the temporary disturbances related to sound propagation) is negligible within the volume of interest. These assumptions imply that P_0 and ρ_0 are constant and do not contribute to the transfer of energy.

Although parts of this section can be found in acoustics text books, the methodology is not commonly known to audio engineers. Since the derivation of the necessary concepts is fairly straightforward, it is worth presenting in this thesis. The section starts with the manifestation and propagation of sound energy in Sections 2.4.1 and 2.4.2, respectively. The relation between the sound pressure and the particle velocity is derived in Section 2.4.3. Sections 2.4.4 and 2.4.5 discuss the energy propagation in cases where part of the sound energy is oscillating locally. Section 2.4.6 describes a frequency domain energetic analysis, and the established measurement technique using microphone pairs is presented in Section 2.4.7. Finally, the measurement technique based on B-format signals is newly derived in Section 2.4.8.

2.4.1 Sound energy

Sound energy manifests itself both as kinetic energy of the fluid elements in motion and as potential energy of the regions where the pressure deviates from its equilibrium value. As can be expected based on elementary physics, the kinetic energy of a fluid element per unit volume is equal to

$$E_{\mathrm{kin}} = \frac{1}{2}\rho u^2 \,, \tag{2.14}$$

where u is the particle velocity (Fahy, 1989, p. 47). Furthermore, for practical considerations, ρ can be approximated with ρ_0.

A change in the potential energy equals the work done in changing the volume of a fluid element. Per unit volume the work is

$$\mathrm{d}W = -P\frac{\mathrm{d}V}{V} \,, \tag{2.15}$$

where V is the volume of the element. A major part of the work is done by P_0. However, this part of the work is reversed when returning to the equilibrium volume[6] and consequently it does not contribute to the sound energy (Cremer and Müller,

[6] Assuming that the acting forces are conservative, i.e., that the process is reversible. As a matter of fact, the changes in the pressure also lead to changes in the temperature which could result in a non-reversible process. However, the pressure changes happen so quickly and the heat conduction of the air is so small that the exchange of heat between neighboring elements can to first approximation be neglected and the process is effectively adiabatic (Cremer and Müller, 1982b, p. 7).

1982b, p. 14; Fahy, 1989, p. 47). Hence, the change in the potential energy can be written as

$$\mathrm{d}E_{\mathrm{pot}} = -p\frac{\mathrm{d}V}{V}\,. \tag{2.16}$$

The change of volume $\mathrm{d}V/V$ can also be related to sound pressure. The detailed derivation of this relation requires discussion of the kinetic gas theory and is beyond the scope of this thesis. For small pressure variations, such as typical sound levels, Fahy (1989, p. 48) gives the following approximation

$$\frac{\mathrm{d}V}{V} \approx \frac{\mathrm{d}p}{\rho_0 c^2}\,, \tag{2.17}$$

Substituting this into Eq. (2.16) and solving for E_{pot} yields

$$E_{\mathrm{pot}} = \frac{p^2}{2\rho_0 c^2}\,. \tag{2.18}$$

The speed of sound is actually affected by the fact that the process is adiabatic instead of isothermic, i.e., that temperature changes also occur. Thus, involving c in Eq. (2.18) accounts also for the thermal potential energy within the limits of the approximation in Eq. (2.17). The total energy per unit volume, denoted as *energy density* of the sound field, can now be written in the form

$$E = E_{\mathrm{pot}} + E_{\mathrm{kin}} = \frac{p^2}{2\rho_0 c^2} + \frac{1}{2}\rho_0 u^2\,. \tag{2.19}$$

This expression applies to any general sound field with small pressure changes and negligible convection (Fahy, 1989, p. 48). By introducing the characteristic acoustic impedance of the medium defined as $Z_0 = \rho_0 c$, Eq. (2.19) can be written in the form

$$E = \frac{1}{2}\rho_0 \left(Z_0^{-2} p^2 + u^2\right)\,. \tag{2.20}$$

2.4.2 Sound intensity

The sound intensity introduced in this subsection is a quantity describing the transport of the kinetic and potential energy in a sound field. Consider an imaginary surface embedded in a fluid. Let us assume further that the small dissipative forces present in the fluid can be neglected and that the surface is not in direct contact with any solid object. The rate at which work is done on one side of the surface by the fluid on the other side can be written as

$$\frac{\mathrm{d}W}{\mathrm{d}t} = \mathbf{F}\cdot\mathrm{d}\mathbf{s}/\mathrm{d}t = \mathbf{F}\cdot\mathbf{u} = p\mathrm{d}\mathbf{S}\cdot\mathbf{u}\,, \tag{2.21}$$

where $\mathbf{F}$ is the force vector acting on the surface, $\mathbf{s}$ is the movement of the surface, and $\mathbf{u}$ is the particle velocity vector. Furthermore, $\mathrm{d}\mathbf{S} = \mathrm{d}S\mathbf{n}$ is an elemental vector area where $\mathrm{d}S$ is the area of the surface and $\mathbf{n}$ is a unit vector normal to the surface

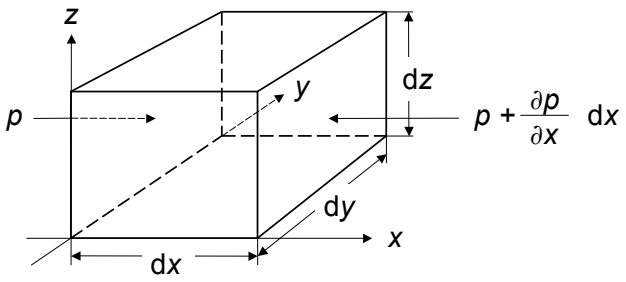

Figure 2.8: A rectangular fluid element with illustration of the forces acting on the x-axis (after Cremer and Müller, 1982b, p. 3).

and directed into the fluid receiving the work. This work is what transfers the energy in either direction across the surface. The work rate per unit area can be written as

$$\frac{\mathrm{d}W}{\mathrm{d}t\mathrm{d}S} = pu_n\,, \tag{2.22}$$

where $u_n = \mathbf{u}\cdot\mathbf{n}$. The quantity $I_n = pu_n$ is defined as the component of instantaneous sound intensity normal to the surface. Furthermore, the instantaneous intensity vector is defined as

$$\mathbf{I} = p\mathbf{u}\,, \tag{2.23}$$

describing thus the instantaneous flow of energy (Fahy, 1989, p. 49).

The opposite direction of the intensity vector can be used as an estimate for the direction of arrival of sound, as needed later in this thesis. The properties of the sound intensity will be discussed further in Section 2.4.4. However, it is necessary to first derive the relation between the sound pressure and particle velocity.

2.4.3 Relation between sound pressure and particle velocity

It is obvious that for the intensity analysis it is necessary to be able to measure the particle velocity. Although direct velocity transducers exist (Olson, 1991; see also Microflown technologies on p. 22), the particle velocity can also be estimated based on common sound pressure measurements. In Section 2.3.1, it was already mentioned that the velocity is related to the pressure gradient of the sound field. Formally, the relation can be derived as follows (Cremer and Müller, 1982b, p. 3–5):.

Consider Newton's dynamic law applied to a small rectangular fluid element of volume $\mathrm{d}V = \mathrm{d}x\mathrm{d}y\mathrm{d}z$, where $\mathrm{d}x$, $\mathrm{d}y$, and $\mathrm{d}z$ are the lengths of the edges. Such an element is illustrated in Figure 2.8. The forces acting on $\mathrm{d}V$ in the positive and negative x-direction are

$$F_{x+} = p\,\mathrm{d}z\mathrm{d}y \tag{2.24}$$

and

$$F_{x-} = \left(p + \frac{\partial p}{\partial x}\mathrm{d}x\right)\mathrm{d}z\mathrm{d}y\,, \tag{2.25}$$

respectively. The difference in the forces results in an acceleration $a_x = \mathrm{d}u_x/\mathrm{d}t$ of the mass $m = \rho_0\,\mathrm{d}V$, where u_x is the component of the particle velocity in the x-coordinate direction and the density has been approximated with ρ_0. From Newton's

law we now have for $F_{x+} - F_{x-}$

$$-\frac{\partial p}{\partial x}\mathrm{d}V = \rho_0 \mathrm{d}V \frac{\mathrm{d}u_x}{\mathrm{d}t}\,. \tag{2.26}$$

Eliminating $\mathrm{d}V$ and considering that corresponding equations hold for the y- and z-directions, we can write

$$-\nabla p = \rho_0 \frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}\,, \tag{2.27}$$

where ∇ is the vector operator denoting gradient.

The total acceleration in Eq. (2.27) is

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} = \frac{\partial \mathbf{u}}{\partial t} + u_x\frac{\partial \mathbf{u}}{\partial x} + u_y\frac{\partial \mathbf{u}}{\partial y} + u_z\frac{\partial \mathbf{u}}{\partial z}\,, \tag{2.28}$$

where u_x, u_y, and u_z are the corresponding Cartesian components of the particle velocity vector. The first term describes acceleration of the fluid particles due to time variation and the other three express acceleration due to convection. Neglecting the convective terms (according to an earlier assumption) Eq. (2.27) can be written in the form

$$-\nabla p = \rho_0 \frac{\partial \mathbf{u}}{\partial t}\,, \tag{2.29}$$

known as the *linearized fluid momentum equation*. This equation will be needed both in the description of the active and reactive intensity in the following as well as in the derivation of the measurement techniques for sound intensity in Sections 2.4.7 and 2.4.8.

2.4.4 Active and reactive intensity

So far all quantities have been discussed as instantaneous. In sound fields such as in rooms, the direction and magnitude of the intensity will, of course, vary as a function of time. Furthermore, the intensity will be later analyzed as a function of frequency, which necessarily involves averaging over a finite time either according to the impulse responses of a filterbank or within a time window applied in Fourier analysis. In this section, some properties of the average intensity over time are discussed.

In a general sound field involving sound propagating in several directions, not all local movement of sound energy corresponds to a net transport. The part of the instantaneous intensity that does contribute to a net transport is denoted *active intensity* and can be calculated simply as

$$\mathbf{I}_a = \langle \mathbf{I}(t) \rangle\,, \tag{2.30}$$

where $\langle\cdot\rangle$ denotes time averaging.

In a monochromatic (single frequency) stationary sound field, the active intensity can be shown to depend on the phase relations of the harmonically varying sound pressure and the particle velocity vector. Conceptually the dependence can be explained as follows. In an ideal traveling plane wave, the direction of $\mathbf{u}(t)$ is constant and its magnitude varies in the same or opposite phase with $p(t)$ (whichever is the

case is a matter of choice of coordinate directions). As a product of $p(t)$ and $\mathbf{u}(t)$, the resulting intensity thus has a constant sign and direction, indicating that all energy propagation takes place in the same direction. However, a phase difference between $p(t)$ and any directional component of $\mathbf{u}(t)$ results in an oscillation of the direction of the intensity vector.

Some more insight into the issue as well as into the origin of the phase differences can be obtained by considering a general stationary monochromatic sound field (Mann *et al.*, 1987; Fahy, 1989, pp. 54–55; Mann and Tichy, 1991; Schiffrer and Stanzial, 1994). Let us again represent the sound pressure using the complex notation. Specifically, let

$$\hat{p}(\mathbf{x}, t) = A(\mathbf{x}) e^{j[\omega t + \varphi(\mathbf{x})]} , \tag{2.31}$$

where $\mathbf{x}$ is the position vector, $A(\mathbf{x})$ is the (real) space-dependent amplitude, and $\varphi(\mathbf{x})$ is the space-dependent phase. The linearized fluid momentum equation (2.29) relates the particle velocity to the pressure gradient of the sound field. Now, the pressure gradient of Eq. (2.31) is

$$\nabla \hat{p} = (\nabla A + jA\nabla\varphi)\, e^{j(\omega t + \varphi)} , \tag{2.32}$$

where the explicit indication of the dependence of the quantities on t and $\mathbf{x}$ has been dropped for typographic clarity. In a monochromatic field, the particle velocity can be solved from Eq. (2.29) by integration over time yielding

$$\hat{\mathbf{u}} = \frac{j}{\omega\rho_0}\nabla\hat{p} = \frac{1}{\omega\rho_0}\,(-A\nabla\varphi + j\nabla A)\, e^{j(\omega t + \varphi)} . \tag{2.33}$$

The intensity can now be computed as a product of Eqs. (2.31) and (2.33). Considering first the component of intensity associated with $\hat{\mathbf{u}}$ in quadrature (having a 90° phase shift) with $\hat{p}$ yields the following real representation

$$\mathbf{I}_r(t) = -\frac{\sin 2(\omega t + \varphi)}{4\omega\rho_0}\nabla A^2 . \tag{2.34}$$

As can be seen, the time average of Eq. (2.34) is zero, indicating no contribution to the net transport of energy. This component of the instantaneous intensity is denoted as *reactive intensity*. Furthermore, anticipating the outcome, the component corresponding to $\mathbf{u}$ and p in phase is defined as the active component of the instantaneous intensity

$$\mathbf{I}_a(t) = -\frac{A^2\cos^2(\omega t + \varphi)}{\omega\rho_0}\nabla\varphi . \tag{2.35}$$

Indeed, $\mathbf{I}_a$ has a constant sign independent of t, signifying an exclusive contribution to the net transport of energy. Hence, it can be confirmed that the time-averaged active intensity depends solely on the in-phase components of the sound pressure and particle velocity (Fahy, 1989, pp. 54–55).

Eq. (2.35) shows that the active component of instantaneous intensity is proportional to the spatial gradient of phase, indicating that the active intensity is perpendicular to surfaces of uniform phase. The reactive component (Eq. 2.34), on the other hand, depends on the spatial gradient of the amplitude of the sound pressure.

Such gradients can be produced as a result of interference, for instance, in standing waves. However, also in propagating spherical waves or in any other wave front whose pressure decreases as a function of distance from the source, there is a reactive (oscillatory) intensity component due to the spatial gradient. For spherical waves the relative strength of the reactive component decreases as a function of distance as the wave front approaches an approximation of a plane wave (note the relation between the reactive intensity and the proximity effect of directional microphones). By noting that $\cos^2(\omega t + \varphi) = [1 + \cos 2(\omega t + \varphi)]/2$, it can also be seen that, at a fixed position, the flow of energy shifts periodically between the active and reactive intensity components.

For non-stationary or non-monochromatic sound fields, the particle velocity cannot be calculated according to Eq. (2.33). Schiffrer and Stanzial (1994) proposed an alternative general decomposition of $\mathbf{u} = \mathbf{u}_p + \mathbf{u}_q$ (using real notation), where $\mathbf{u}_p$ is in phase and $\mathbf{u}_q$ in quadrature with p. The component $\mathbf{u}_p$ is defined as

$$\mathbf{u}_p = \frac{\langle p\mathbf{u}\rangle p}{\langle p^2\rangle}\,, \tag{2.36}$$

having the same time dependence as p, and average magnitude and direction of the component of $\mathbf{u}$ contributing to the active intensity. Furthermore, for $\mathbf{u}_q$ we have

$$\mathbf{u}_q = \mathbf{u} - \mathbf{u}_p = \frac{\langle p^2\rangle\mathbf{u} - \langle p\mathbf{u}\rangle p}{\langle p^2\rangle}\,. \tag{2.37}$$

Now, alternative active and reactive components of instantaneous intensity can be defined as

$$\mathbf{I}'_a(t) = p\mathbf{u}_p = \frac{\langle p\mathbf{u}\rangle p^2}{\langle p^2\rangle} \tag{2.38}$$

and

$$\mathbf{I}'_r(t) = p\mathbf{u}_q = \frac{\langle p^2\rangle p\mathbf{u} - \langle p\mathbf{u}\rangle p^2}{\langle p^2\rangle}\,, \tag{2.39}$$

respectively[7].

It can be shown that the definitions in Eqs. (2.38) and (2.39) coincide with the earlier definitions for monochromatic sound fields (Schiffrer and Stanzial, 1994; Stanzial *et al.*, 1996; see also Stanzial and Prodi, 1997). However, in general sound fields, the root mean square (RMS) value of $\mathbf{I}'_r(t)$ will be different compared to the RMS of a sum of the reactive intensity components $\mathbf{I}_r(t)$ (Stanzial and Prodi, 1997). Nevertheless, it is easy to show that also $\langle \mathbf{I}'_r(t)\rangle = 0$. The amplitude of the oscillations of the newly defined reactive intensity in different directions could also be analyzed yielding ellipsoids describing the *energy polarization* in the sound field (Stanzial *et al.*, 1996; Stanzial and Prodi, 1997; for figures of the three dimensional ellipsoids, see Stanzial *et al.*, 2000). However, the validity of the time-dependence of the presented decomposition is not fully clear. Stanzial and Prodi (1997) also discuss another alternative definition where, instead of decomposing $\mathbf{u}$, p is decomposed into a component in phase and in quadrature with $\mathbf{u}$, yielding different instantaneous values for the radiating and oscillating intensity but the same time averages as using a decomposition of $\mathbf{u}$.

[7]Stanzial and Prodi (1997) have later used the terms *radiating* and *oscillating* intensity to describe the newly defined active and reactive intensity components, respectively.

2.4.5 Diffuseness

Energy transport in a sound field can also be discussed relative to the diffuseness of the sound field. According to Nélisse and Nicolas (1997), a generally accepted definition of diffuseness states that in a perfectly diffuse field, the energy density is constant in all points within the volume of interest. Furthermore, a typical definition requires that all directions of propagation of sound be equally probable (Nélisse and Nicolas, 1997; Kuttruff, 2000, Section 5.1). These definitions together imply no net transport of energy, i.e., $\langle \mathbf{I} \rangle = 0$.

Based on the previous discussion on active and reactive intensity, the first thought for a measure of the degree of diffuseness might be to use the ratio of the reactive and active intensity components. However, it was already pointed out that different definitions of reactive intensity yield different values in general sound fields. Furthermore, the reactive intensity can also vanish in points where the particle velocity is zero. According to the linearized fluid momentum equation this corresponds to the vanishing of the spatial gradient of the sound pressure, which could theoretically happen, for instance, in the antinodes (pressure maxima) of an ideal standing wave, from where the energy is transferred symmetrically to different directions.

Another possibility is to compare the active intensity to the energy density. The quantity $\langle \mathbf{I} \rangle / E$ describes the velocity of sound energy propagation and it can be shown to be bound between $[0, c]$ (Schiffrer and Stanzial, 1994; Stanzial *et al.*, 1996). The value of c indicates that all energy is being transferred, whereas a smaller value implies that part of the energy is locally confined. The fraction of the propagating sound energy is thus given by $\|\langle \mathbf{I} \rangle\| / \langle cE \rangle$, where $\| \cdot \|$ denotes the norm of a vector. This quantity has also been discussed as a field indicator by Stanzial *et al.* (1996) and applied to acoustical analysis of an opera house by Stanzial *et al.* (2000). However, we define the *diffuseness estimate* ψ as the proportion of the locally confined energy. From Eqs. (2.20) and (2.23) we now have

$$\psi = 1 - \frac{\|\langle \mathbf{I} \rangle\|}{\langle cE \rangle} = 1 - \frac{2Z_0 \, \|\langle p\mathbf{u} \rangle\|}{\langle p^2 \rangle + Z_0^2 \langle \mathbf{u}^2 \rangle} \,. \tag{2.40}$$

A value of $\psi = 0$ signifies the absence of any local oscillation of energy. Furthermore, for an ideally diffuse sound field, $\psi = 1$. However, the converse is not necessarily true, i.e., the value $\psi = 1$ does not guarantee an ideally diffuse sound field in the sense that all directions of propagation would be equally probable. Generally, based on energetic analysis in a single point, there is no possibility to analyze the exact directional distribution of oscillating intensity, because symmetric oscillations yield the same zero intensity (due to zero particle velocity) as no oscillations in the same direction. Such symmetric oscillations are fairly theoretical, and in practical broadband and/or time-variant sound fields the energy polarization described earlier could provide useful information in addition to ψ. However, it cannot serve as a substitute for ψ. Finally, note that an instantaneous value $\psi(t)$ could also be defined but it will not be needed in this thesis.

2.4.6 Frequency distributions

For derivation of the frequency distribution of sound intensity, consider first a general sound field with multiple harmonic components at different frequencies. As a product of two quantities including multiple harmonic components, intantaneous intensity will include frequency components at the sum and difference frequencies of the original harmonic components. Thus, the direct frequency analysis of the instantaneous intensity does not reflect the harmonic structure of the sound field being analyzed, and as such is of little practical significance. What is of interest is the contribution of sound pressure and particle velocity at the same frequency or frequency range to the total intensity.

The Fourier transform effectively decomposes an analyzed quantity into stationary monochromatic components. As shown in Section 2.4.4, in stationary monochromatic sound fields the active intensity is related to the in-phase components of the sound pressure and particle velocity. Hence

$$\mathbf{I}_a(\omega) = \mathsf{Re}\left\{P^*(\omega)\mathbf{U}(\omega)\right\} , \tag{2.41}$$

where $P(\omega)$ and $\mathbf{U}(\omega)$ are the Fourier transforms of the sound pressure and particle velocity over a finite time window[8], respectively, and * denotes complex conjugation (Fahy, 1977, 1989, Section 5.3). Since Fourier transforms are inherently complex-valued, the hat signifying a complex signal is omitted from the notation. Eq. (2.41) represents the real part of the cross-spectrum[9] of the sound pressure and particle velocity. Furthermore, the imaginary part $\mathsf{Im}\{P^*(\omega)\mathbf{U}(\omega)\}$ describes the mean amplitude of the oscillating intensity within each Cartesian coordinate direction. However, this kind of an analysis of the reactive intensity does not provide information on the polarization of the oscillations which may yield higher reactivity in between the Cartesian coordinate axes. As such, the frequency distribution of the reactive intensity is not very useful for later applications. Consequently, henceforth the frequency distributions will only be derived for the active intensity.

The mutual phases of the sound pressure and particle velocity do not affect the energy density, whose frequency distribution is simply the magnitude of the Fourier transform of Eq. (2.20). Hence

$$E(\omega) = \frac{1}{2}\rho_0 \left[Z_0^{-2}\left|P(\omega)\right|^2 + \left|\mathbf{U}(\omega)\right|^2\right] , \tag{2.42}$$

where $|\cdot|$ denotes the absolute value of a complex number and the square of a vector quantity is defined as the square of its norm. The frequency distribution of the diffuseness estimate is thus

$$\psi(\omega) = 1 - \frac{\left\|\mathbf{I}_a(\omega)\right\|}{cE(\omega)} = 1 - \frac{2Z_0\left\|\mathsf{Re}\left\{P^*(\omega)\mathbf{U}(\omega)\right\}\right\|}{\left|P(\omega)\right|^2 + Z_0^2\left|\mathbf{U}(\omega)\right|^2} . \tag{2.43}$$

[8]In order to maintain the intensity as the *rate* of work per unit area, the Fourier transform needs to be defined such that it is normalized by time. However, later in this thesis we will not be interested in the absolute magnitude of $\mathbf{I}_a(\omega)$ but its direction and magnitude related to other Fourier transformed quantities, so the normalization will not be necessary as long as all quantities are computed over the same time.

[9]The cross-spectrum of two signals x_1 and x_2 is typically expressed as $G_{x_1x_2} = X_1(\omega)X_2^*(\omega)$ instead of $X_1^*(\omega)X_2(\omega)$ (see Section 2.6.1). However for intensity we maintain the above definition according to Fahy (1989, p. 97).

It was already mentioned that frequency analysis can also be performed using a filterbank, i.e., a number of bandpass filters with different center frequencies. In the filterbank approach, the sound pressure and each particle velocity component can be simply fed through the filterbank prior to the energetic analysis, and the equations from the previous sections can be directly applied to compute both instantaneous and time-averaged intensity, energy density, and diffuseness at each frequency band. The Fourier analysis presented in this section can also be applied to short consecutive (or overlapping) time windows yielding a *short-time Fourier transform* (STFT). Note that the STFT components with the same frequency from subsequent time windows can also be seen as down-sampled filterbank signals.

2.4.7 Microphone pair measurements

After describing the theory of energetic analysis, we now turn to practical measurements. Computing the desired quantities from a probe yielding directly sound pressure and particle velocity is trivial apart from measurement errors (see Fahy, 1989, Section 5.2) and will not be discussed further. Another common measurement technique uses a pair of omnidirectional microphones (Fahy, 1989, p. 92). Consider a pair of microphones yielding the signals $p_1(t)$ and $p_2(t)$ placed on the x-axis with a spacing of d as shown earlier in Figure 2.4. From the linearized fluid momentum equation (2.29) we have

$$u_x(t) = -\frac{1}{\rho_0} \int_{-\infty}^{t} \frac{\partial p(\tau)}{\partial x} \mathrm{d}\tau \,. \tag{2.44}$$

Using a finite difference approximation yields

$$u_x(t) \approx \frac{1}{\rho_0 d} \int_{-\infty}^{t} [p_1(\tau) - p_2(\tau)] \,\mathrm{d}\tau \,, \tag{2.45}$$

which approximates the particle velocity at the point halfway between the microphones under the assumption that d is small compared to the wavelength of sound. The sound pressure at the same point is estimated by the average of the two microphone signals

$$p(t) \approx \frac{1}{2} \left[p_1(t) + p_2(t) \right] \,. \tag{2.46}$$

Eqs. (2.45) and (2.46) can be directly substituted into earlier time domain formulae to compute the intensity, energy density, and diffuseness. In a discrete time implementation, the integration in Eq. (2.45) is simply replaced by a summation over time. Similar equations, of course, hold for the y- and z-coordinate directions and computing the three dimensional particle velocity can be realized with three pairs of microphones.

Frequency distributions can be computed directly according to Eqs. (2.41), (2.42), and (2.43). However, for the active intensity it is possible to derive a simplified formula as follows (Chung, 1978; Fahy, 1989, Section 5.3): The Fourier transform of the estimated particle velocity from Eq. (2.45) is

$$U_x(\omega) \approx -\frac{j}{\omega \rho_0 d} \left[P_1(\omega) - P_2(\omega) \right] \,. \tag{2.47}$$

Note that apart from the constant $1/(\rho_0 c) = 1/Z_0$ the equation is similar to Eq. (2.9) describing the response of an ideal gradient microphone in a monochromatic sound field. Furthermore, the Fourier transform of the sound pressure is

$$P(\omega) \approx \frac{1}{2}\left[P_1(\omega) + P_2(\omega)\right] . \tag{2.48}$$

Now, substituting Eqs. (2.47) and (2.48) into Eq. (2.41) yields

$$\begin{aligned} I_{ax}(\omega) &\approx -\mathsf{Re}\left\{\frac{j}{2\rho_0\omega d}[P_1^*(\omega) + P_2^*(\omega)][P_1(\omega) - P_2(\omega)]\right\} \\ &= \frac{j}{2\rho_0\omega d}\mathsf{Im}\left\{|P_1(\omega)|^2 - |P_2(\omega)|^2 - P_1^*(\omega)P_2(\omega) + P_1(\omega)P_2^*(\omega)\right\} \\ &= -\frac{j}{\rho_0\omega d}\mathsf{Im}\left\{P_1^*(\omega)P_2(\omega)\right\} , \end{aligned} \tag{2.49}$$

where for the last equality the fact that the absolute value of any complex variable is real and the relation $\mathsf{Im}\{P_1(\omega)P_2^*(\omega)\} = -\mathsf{Im}\{P_1^*(\omega)P_2(\omega)\}$ have been used. It can thus be seen that the active intensity is directly related to the cross spectrum of the two microphone signals.

As already mentioned, the approximations of both the sound pressure and particle velocity are valid only when the distance between the microphones is small compared to the wavelength of sound. The error in the intensity estimate at high frequencies depends on the sound field being measured and the intensity tends to be estimated too low (for a derivation of the error for some ideal sound fields, see Fahy, 1989, Section 5.5.1). Furthermore, the estimate of u exhibits spatial aliasing similar to the gradient microphones (see Figure 2.5), which is reflected in the intensity estimate. For a practical upper limit of the validity of the approximations, Chung (1978) proposed an angular frequency of $\omega \approx c/d$. The lower frequency limit, on the other hand, is determined as in gradient microphones by measurement noise and mismatch between the two microphones (for related analysis, see Jacobsen, 2002). The frequency range problems can be alleviated by using a solid spacer between the microphone pairs (Jacobsen *et al.*, 1998; Juhl and Jacobsen, 2004) or by using microphone pairs at different distances, as done in the TKK 3-D Array (see p. 22). Another alternative method involving averaging over a spherical microphone array was also proposed by Kuttruff and Schmitz (1994).

2.4.8 B-format measurements

Due to the relation of the zero- and first-order B-format signals (see Section 2.3.2) to the sound pressure and pressure gradient, respectively, they can be readily used for energetic analysis. According to an earlier definition (disregarding the sensitivity of the actual microphone system),

$$p = W . \tag{2.50}$$

As differential signals, X, Y, and Z are related to the finite difference approximation of the particle velocity. However, they are equalized such that for an ideal plane wave they yield signals relative to the sound pressure according to Eq. (2.12) multiplied by

$\sqrt{2}$. For a plane wave propagating in the direction of the unit vector $\mathbf{e}_p$, the signals can be thus expressed, using a vector notation, as

$$\mathbf{X}' = X\mathbf{e}_x + Y\mathbf{e}_y + Z\mathbf{e}_z = \sqrt{2}p\mathbf{e}_p \,, \tag{2.51}$$

where $\mathbf{e}_x$, $\mathbf{e}_y$, and $\mathbf{e}_z$ are unit vectors in the directions of the corresponding Cartesian coordinate axes. Introducing again the complex notation for time domain signals, the sound pressure of the plane wave can be written as

$$\hat{p}(\mathbf{x}, t) = Ae^{j\omega(t-\mathbf{x}\cdot\mathbf{e}_p/c)} \,. \tag{2.52}$$

Now, from the linearized fluid momentum equation (2.29)

$$\hat{\mathbf{u}} = -\frac{1}{\rho_0}\int_{-\infty}^{t} \nabla\hat{p}\mathrm{d}\tau = \frac{1}{\rho_0 c}\hat{p}\mathbf{e}_p = \frac{1}{\sqrt{2}Z_0}\hat{\mathbf{X}}' \,, \tag{2.53}$$

where the explicit indication of the dependence of the quantities on $\mathbf{x}$ and t has again been dropped for typographic clarity.

As for the microphone pair techniques, Eqs. (2.50) and (2.53) can be directly substituted to earlier formulae to compute the intensity, energy density, and diffuseness. For the frequency distributions we have according to Eqs. (2.41), (2.42), and (2.43)

$$\mathbf{I}_a(\omega) = \frac{1}{\sqrt{2}Z_0}\mathsf{Re}\left\{W^*(\omega)\mathbf{X}'(\omega)\right\} \,, \tag{2.54}$$

$$E(\omega) = \frac{1}{2}\rho_0 Z_0^{-2}\left[|W(\omega)|^2 + |\mathbf{X}'(\omega)|^2/2\right] \,, \tag{2.55}$$

and

$$\psi(\omega) = 1 - \frac{\sqrt{2}\,\|\mathsf{Re}\left\{W^*(\omega)\mathbf{X}'(\omega)\right\}\|}{|W(\omega)|^2 + |\mathbf{X}'(\omega)|^2/2} \,. \tag{2.56}$$

In practical implementations, the measurement accuracy of the B-format signals, of course, needs to be considered. B-format is most often recorded with Soundfield microphones (see p. 21) and the involved analog signal processing makes the error analysis even more difficult than with a microphone pair. Based on the measurements of Farina (2001a,b), a Soundfield microphone will introduce some frequency-dependent bias in the direction of the intensity estimate due to distortion and asymmetries of the directivity patterns. Furthermore, the frequency-dependent deviations in the gain of W relative to X, Y, and Z will increase the diffuseness estimate. Some more insight into the accuracy will be gained in the next section where measurements of the same room response are analyzed both using the TKK 3-D array and a Soundfield microphone.

2.5 Visualization of directional room responses

The energetic analysis methods presented in the previous section can be directly applied to visualization of measured room responses as proposed by Merimaa *et al.* (2001). Contrary to common use of active intensity averaged over a considerable

time, the analysis presented here will be performed using STFT[10]. This method yields a uniform time-frequency resolution, but corresponding analysis with a non-uniform resolution proportional to that of the human peripheral hearing has also been presented by Merimaa *et al.* (2001). Here, the analysis will be limited to sound pressure and active intensity. However, the analysis data are of exactly the same form as that used later in the SIRR method (see Chapter 5) where also the diffuseness estimate will be employed and illustrated.

Figure 2.9 presents visualizations of the analysis results for two responses measured with the same source and receiver positions and two different microphone systems in the Promenadikeskus concert hall located in Pori, Finland (Merimaa *et al.*, 2005b). The data in the two upper panels were computed with microphone pair techniques from measurements with the TKK 3-D array, and the data in the lower panels from B-format measurements with a Soundfield MKV microphone system (response `s2_r2_sf` from the database of Merimaa *et al.*, 2005b; for similar visualizations of other acoustical environments, see also Merimaa *et al.*, 2001 and Vassilantonopoulos *et al.*, 2005).

The plots illustrate the direction-dependent arrival of sound to a measurement position during a time period of 100 ms starting from slightly before the arrival of the direct sound. The figures consist of active sound intensity vectors plotted on top of an omnidirectional sound pressure spectrogram. For the visualization, the three-dimensional (3-D) vectors have been projected into the horizontal and the median plane. In both planes, vectors pointing to the right represent sound emanating from the frontal direction, defined as the direction towards the stage from the measurement position in the audience area and parallel to the side walls of the hall. In the horizontal plane, a vector pointing down signifies sound arriving from the right side, and in the median plane a vector pointing down represents sound arriving from above.

Both the active intensity and the spectrograms were computed using 128 sample long Hann windowed time frames with a sampling frequency of 48 kHz. The time-frequency representations have been adjusted for illustrational purposes. Zero-padding and largely overlapping windows were used to smooth the spectrogram, and the intensity vectors are plotted only for positions of local maxima of the magnitude of the active intensity as a function of time. Furthermore, the data are thresholded to the levels shown in the color bars of the figures in order to highlight the strongest reflections. The lengths of the vectors are proportional to the component of the logarithmic magnitude of the active intensity in the plane in question, and the underlying spectrograms are also displayed on a logarithmic (dB) scale.

The figure shows the direct sound arriving from slightly to the right at about 30 ms, followed by a reflection from the right side wall at 40 ms and from the front wall at 60 ms. The measured hall has a large number of diffusers and it can be seen that the diffuseness of the sound field increases quickly, resulting in fairly random directions of the intensity vectors in the late response. However, a strong discrete reflection from the left can still be identified at about 90 ms.

The TKK-array measurements involved a careful compensation for the magni-

[10]Note that Begault and Abel (1998) have also performed analysis of direction of arrival of room responses within short time windows. They used the zero-lag cross-coherences between omnidirectional and figure-of-eight responses to estimate the direction of arrival which is, in fact, equivalent to computing the broadband active intensity vector normalized by signal power.

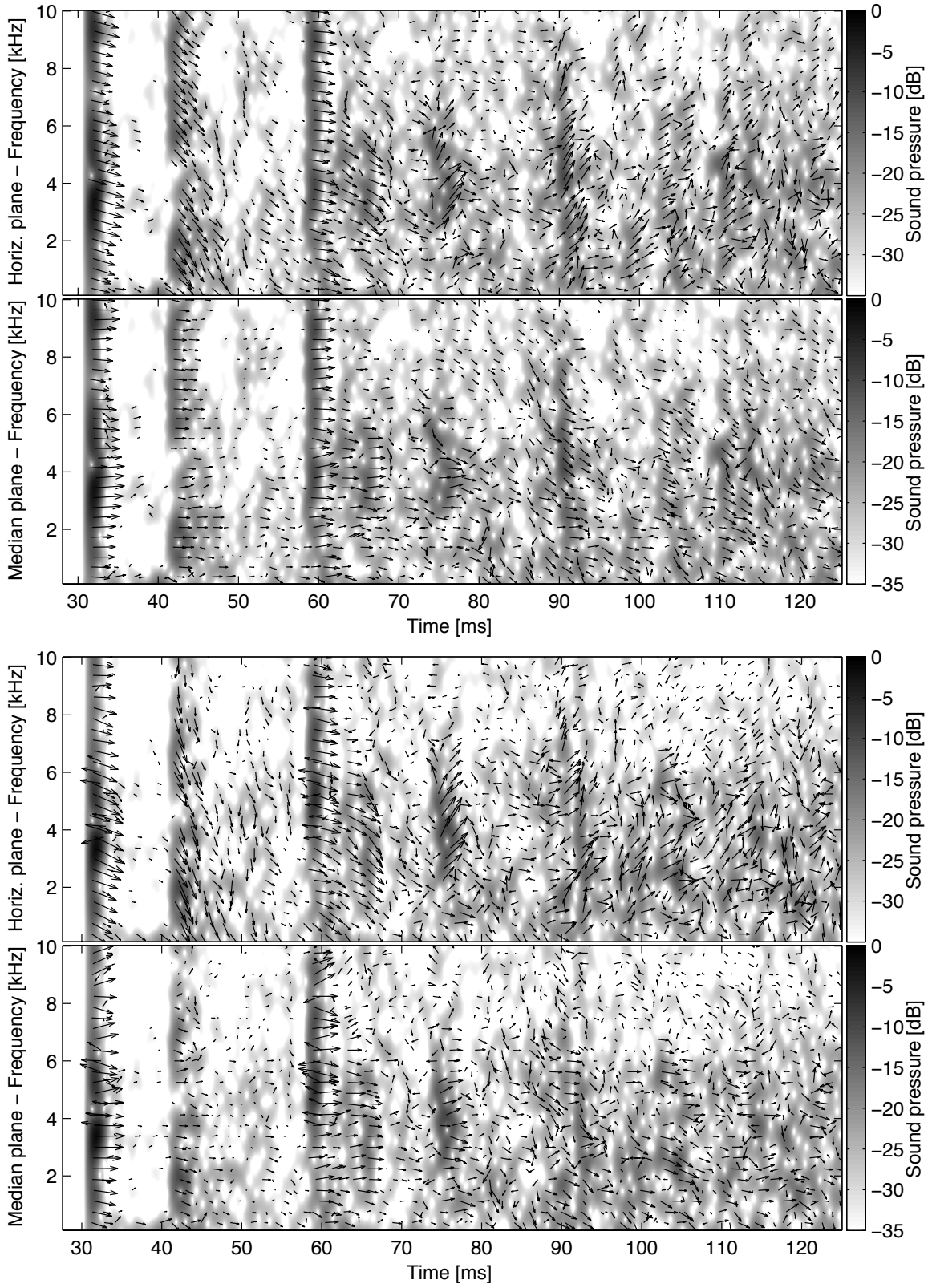

Figure 2.9: Visualization of a directional room responses measured with the TKK 3-D array (two upper panels) and a Soundfield MKV B-format microphone (two lower panels). See text for details.

tude responses of each capsule and can be considered fairly accurate up to the limit $f = c/(2\pi d) \approx 5.4$ kHz recommended by Chung (1978). Taking into account that the sound sources and microphone systems were repositioned between the two measurements, the correspondence between the two measurements is reasonably good, suggesting that the Soundfield microphone can also yield reliable results below about 5 kHz.

Some more insight into the measurement errors at higher frequencies can be gained by observing the directions of the intensity vectors corresponding to the first sound events across frequencies. For the direct sound, it is known that all frequencies should arrive from the same direction. The same can also be assumed to hold approximately for the first discrete reflections. It is obvious that the Soundfield microphone measurement exhibits some errors. Figure 2.9 suggests that above about 5 kHz the intensity vectors derived from the Soundfield microphone are somewhat biased in a direction dependent way. There are also some vectors indicating propagation of energy to the opposite direction compared to the actual wave front. The reason for these apparently erratic components is unknown but might be related to reflections from the casing of the microphone system or from the microphone stand. Based on a visual inspection, the TKK array, however, does not seem to produce large directional errors even somewhat above the suggested high frequency limit.

2.6 Alternative analysis methods

The energetic analysis described in Section 2.4 already provides the necessary tools for analyzing the direction of arrival and diffuseness of sound as needed in SIRR (see Chapter 5). However, energetic analysis is not the only possible way to estimate these quantities. For the sake of completeness, some alternative microphone array techniques suitable for the same purposes are briefly reviewed in this section. Estimation of the direction of arrival will be described first in Section 2.6.1 followed by a discussion on alternative measures of diffuseness in Section 2.6.2.

2.6.1 Direction of arrival

Numerous methods for estimating the direction of arrival of sound or the location of a sound source based on microphone array signals have been presented in the literature. A simple method suitable for steerable beamformers is based on searching for the direction of the beam that maximizes the output (e.g., Flanagan *et al.*, 1985). Instead of scanning the space with a beam, modern techniques, however, often consider jointly what can be seen as the output of a bank of beamformers. Within certain limits, this approach enables simultaneous localization of multiple concurrent sources (e.g., Wax and Kailath, 1983; see also Hahn and Tretter, 1973). Abel and Begault (1998) have also presented a related maximum likelihood estimator for concurrent localization of multiple sources based on B-format signals.

Another common method uses estimation of time delays between a set of microphones to derive the position of a sound source. The time delays are typically estimated as the lag value maximizing the cross-correlation estimate between two micro-

phone signals

$$R_{x_1x_2}(\tau) = \frac{1}{T-\tau}\int_{\tau}^{T} x_1(t)x_2(t-\tau)\mathrm{d}t\,, \tag{2.57}$$

where x_1 and x_2 are the signals of the two microphones, τ is the time lag, and T is the observation interval. A constant time delay corresponds to a hyperbolic surface of rotation, and by using several microphone pairs with different orientations, it is possible to estimate the position of the source as the intersection of such hyperbolic surfaces (e.g., Fang, 1990; Chan and Ho, 1994; Brandstein *et al.*, 1997; Brandstein and Silverman, 1997; Berdugo *et al.*, 1999; for alternative methods related to the time delay estimation, see also Jacovitti and Scarano, 1993; Zhang and Er, 1996; Stuller and Hubing, 1997; Brandstein and Silverman, 1997; Benesty, 2000).

The cross-correlation can also be computed as the inverse Fourier transform of the cross-spectral density $G_{x_1x_2}$ of the two microphone signals. The cross-spectral density is defined as

$$G_{x_1x_2}(f) = X_1(f)X_2^*(f)\,, \tag{2.58}$$

where $X_1(f)$ and $X_2(f)$ are the Fourier transforms of x_1 and x_2, respectively. By introducing a weighting function $W(f)$ for the cross-spectral density, we can write the *generalized cross-correlation* (GCC) function as

$$R_{\mathrm{GCC}}(\tau) = \int_{-\infty}^{\infty} W(f)G_{x_1x_2}(f)e^{j2\pi f\tau}\mathrm{d}f\,. \tag{2.59}$$

Several weightings have been described by Knapp and Carter (1976). Interestingly, many of the weightings aim at suppressing distracting sound such as room reflections or background noise, whereas the focus in this chapter has been on specifically analyzing the reflections, as will also be done later in the SIRR method. However, we will take the opposite point of view in Chapter 4, where cross-correlation will be used to estimate time delay in a model of the human hearing (see also Section 3.6.1). For now, let us simply state that the maximum likelihood estimator for the time delay has a weighting function (Knapp and Carter, 1976; Carter, 1987)

$$W_{\mathrm{ML}}(f) = \frac{1}{|G_{x_1x_2}(f)|}\cdot\frac{|\gamma_{12}(f)|^2}{[1-|\gamma_{12}(f)|^2]}\,, \tag{2.60}$$

where $\gamma_{12}(f)$ is the cross-coherence of x_1 and x_2 given by

$$\gamma_{12}(f) = \frac{G_{x_1x_2}(f)}{\sqrt{G_{x_1x_1}(f)G_{x_2x_2}(f)}}\,. \tag{2.61}$$

The first term in Eq. (2.60) effectively flattens the amplitude of the cross-spectral density, and the second term gives more weight to frequencies with a high coherence.

A third category of localization methods unrelated to any previous discussion includes the so-called signal-subspace techniques, which relate the covariance matrix of the signals of a microphone array to the locations of the sources (e.g., Wang and Kaveh, 1985; Schmidt, 1986). Using higher-order signal statistics within the same context has also been proposed by Bourennane and Bendjama (2002).

2.6.2 Diffuseness

As mentioned in Section 2.4.5, in an ideal diffuse sound field, the energy density is constant over the sound field. Hence, a natural alternative to the energetic diffuseness estimate is to measure the spatial uniformity of the sound field (see Nélisse and Nicolas, 1997). However, studying the sound field over considerable distances does not fit in the adopted listener-centered directional analysis and will not be discussed further. Another obvious possibility related to the equal probability of all directions of propagation is to use highly directional beamforming technique to analyze the distribution of sound arriving from different directions (e.g., Gover, 2002).

Yet another alternative is again based on coherence between two microphones (e.g., Cook *et al.*, 1955; Bodlund, 1976; Chu, 1981; Nélisse and Nicolas, 1997). For a single plane wave the peak of the coherence function always has a value of one. However, in the presence of sound waves propagating in several directions, they are summed with different phase relations at the positions of the microphones thus yielding partially incoherent signals. For a diffuse field, Cook *et al.* (1955) derived and experimentally verified the relation

$$\gamma_{12} = \frac{\sin(2\pi d/\lambda)}{2\pi d/\lambda}\,, \tag{2.62}$$

where d is the distance between the microphones (see also Chu, 1981). For a small distance compared to the wavelength of sound, the coherence thus exhibits high values even in a diffuse field. The coherence of partly diffuse sound fields is typically between one and the value for an ideal diffuse field, as will be discussed later in more detail in Sections 3.5.2 and 5.4.2.

2.7 Summary and conclusions

After background discussion on sound propagation in rooms, a substantial part of this chapter concentrated on different directional analysis techniques. The described directional microphones and microphone array techniques can be directly applied to both directional analysis of sound fields and to sound reproduction (for further discussion on the reproduction aspects, see Section 5.2.2). However, it is difficult to realize high directivity over a wide frequency range with a reasonable number of microphones, which limits the achievable directional resolution.

The presented analytic methods for estimating the direction of arrival and diffuseness of sound will be needed later in the development of the Spatial Impulse Response Rendering (SIRR) method which does not rely on highly directional microphones (see Chapter 5). The most comprehensively described method suitable for such analysis consists of analysis of the directional propagation of sound energy. Two important quantities were derived: The active intensity describes the direction and magnitude of the net transport of energy and the diffuseness estimate measures the fraction of locally confined sound energy. The energetic analysis was also applied to visualization of directional room responses. The briefly reviewed alternative analysis methods included, among others, techniques based on computing the cross-correlation and/or coherence between microphone signals.

In general, which one of the presented analysis methods is most appropriate for a certain application depends on the objectives of the analysis, on the sound field being analyzed, and on the available microphones and/or microphone arrays. Especially in analysis pertaining to sound reproduction, the availability aspect should not be underestimated. For this reason, the commerically available B-format microphones and the related energetic techniques will be used as the main analysis tools in the SIRR method in Chapter 5, although it will also be seen that SIRR could easily utilize other directional analysis techniques.

Chapter 3

Spatial Hearing

3.1 Introduction

The concepts of auditory space and auditory events were already introduced in Section 1.1.2. This chapter reviews the auditory spatial perception and related auditory models in more detail. The main emphasis is on *localization*, defined as "the law or rule by which the location of an auditory event (e.g., its direction or distance) is related to a specific attribute or attributes of a sound event" (Blauert, 1997, p. 37)[1]. Some localization results reviewed in this chapter will be studied later with a novel auditory modeling mechanism in Chapter 4. Furthermore, within the context of this thesis, the discussion of auditory perception has the function of telling what is important for a human listener in the spatial sound field, i.e., what needs to be reproduced (see the SIRR method in Chapter 5). For this purpose, the treatment will be extended to some other perceptual attributes affected by the directional properties of a physical sound field. However, these attributes will not be directly investigated later in this thesis and, consequently, the review will be briefer than that of localization.

As will be seen, the dominant cues for localization are the differences between the two ear input signals. These differences also help in detecting a signal in the presence of a masker with different interaural attributes. One way to describe the advantage given by a specific interaural stimulus configuration is the *binaural masking level difference* (BMLD), defined as the difference in the minimum level of the signal enabling auditory detection, compared to a similar situation where the signal has the same interaural attributes as the masker. General discussion on detection will be fairly limited. However, detection studies have provided important information on the resolution of both the monaural and binaural auditory system, which can be employed in auditory models as well as in the development of SIRR. For the later discussion, it is important to distinguish between three types of masking experiments: In simultaneous masking, the signal does not extend beyond the temporal limits of the masker, whereas in forward masking, the signal follows the masker. Backward

[1]The full definition by Blauert (1997) also includes the possibility that the localization of an auditory event may involve information from beyond the auditory modality (see e.g. Wallach, 1940; Blauert, 1997, Section 2.5.2; Zwiers *et al.*, 2003; Kohlrausch and van de Par, 2005). However, in this thesis, the discussion on localization will be limited to the auditory system.

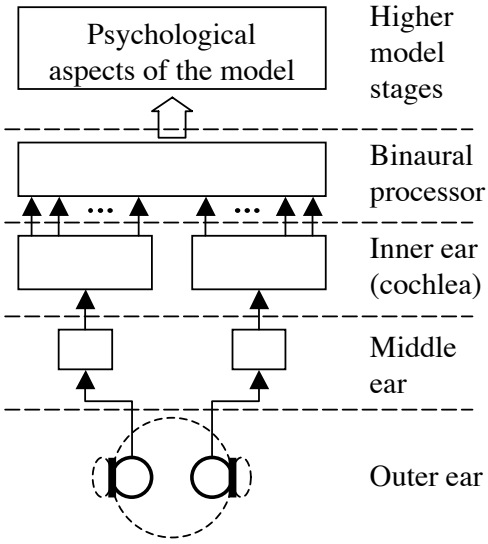

Figure 3.1: A generic model of spatial hearing.

masking, with the signal preceding the masker, also occurs to some extent but will not be discussed in this thesis.

The role of physiological processes in hearing was already mentioned in Section 1.1.2. The human hearing process features a number of stages involved in the auditory perception. The structure of a generic model of spatial hearing is illustrated in Figure 3.1. There is little doubt about the first stages of the auditory system; i.e., the physical and physiological functioning of the outer, middle, and inner ear are known and understood to a high degree. Although the physiology of these first stages will also be briefly described, the main point of view in this chapter is functional. That is, rather than trying to incorporate every anatomical detail into the discussion, we are concerned with what the ear does to a sound signal, how this processing affects the perception, and how the processing can be simulated. The stage of the binaural processor is already less well known, and with current knowledge, the interaction between the binaural processor and the higher level cognitive processes needs to be addressed through indirect psychophysical evidence. In the lack of conclusive knowledge, the operation of the stages above the auditory periphery is, for the most part, treated psychophysically.

This chapter is organized as follows: The physiology and modeling of the auditory periphery are described in Section 3.2. Localization is discussed in Section 3.3. The time and frequency resolution of spatial hearing is reviewed in Section 3.4. Section 3.5 briefly describes other features of spatial perception apart from localization. Binaural models are introduced in Section 3.6, followed by the summary of the chapter in Section 3.7.

3.2 Auditory periphery

The main task of the auditory periphery is to sample the sound field and to convert the vibrations of the air into neural signals which are then processed at the higher

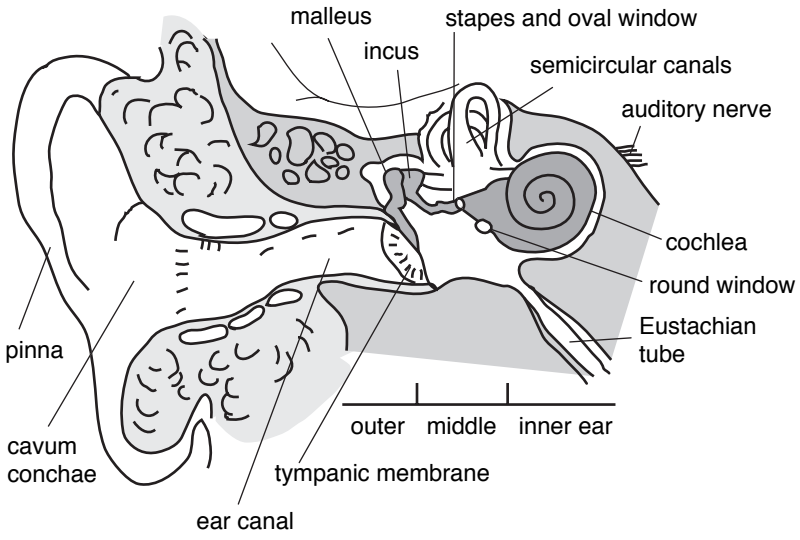

Figure 3.2: Cross-section of the ear.

levels of the brain. However, the peripheral stages also have a considerable effect on the information available to the higher levels. Furthermore, the head and torso play an important role in spatial hearing. Although they are sometimes not considered part of the auditory periphery, the treatment here is extended to include the physical propagation of sound into the ears. The cross-section of the ear is illustrated in Figure 3.2, and the operation of the parts with auditory functions will be described in this section in the order of propagation of sound through the auditory periphery, starting from the effects of the head, torso, and outer ears in Section 3.2.1, continuing to the middle ear in Section 3.2.2, and to the inner ear in Section 3.2.3.

3.2.1 Head, torso, and outer ears

The physical sound field is coupled to the auditory system predominantly through the outer ears, which consist of the pinnae and the ear canals. Some sound reaches the auditory system also through bone conduction, i.e., via sound-induced motion of the skull that is directly coupled to the middle and inner ear. However, bone conduction yields levels in the inner ear that are at least 40 dB lower than those excited by the air conduction through the outer ears (Hudde, 2005). With such a low relative level, bone conduction is usually perceptually irrelevant if the ear canals are not blocked, as will be assumed throughout this thesis.

Neglecting bone conduction, the two ear input signals include all the information about the external sound field that is available to the listener. The combined effect of the head, torso, and pinnae on sound can be measured, analyzed, and simulated using *head-related transfer functions* (HRTFs), which characterize the transmission of sound from a specific point in free field to the ears of a listener (e.g., Wightman and Kistler, 1989a; Møller *et al.*, 1995; Blauert, 1997; Hammershøi and Møller, 2005; Riederer, 2005; for freely available HRTFs, see Gardner and Martin, 1994, 1995; Algazi *et al.*, 2001). Some HRTFs for different directions are illustrated in Figure 3.3, and in the following, the direction-dependent features of the HRTFs will be discussed in more

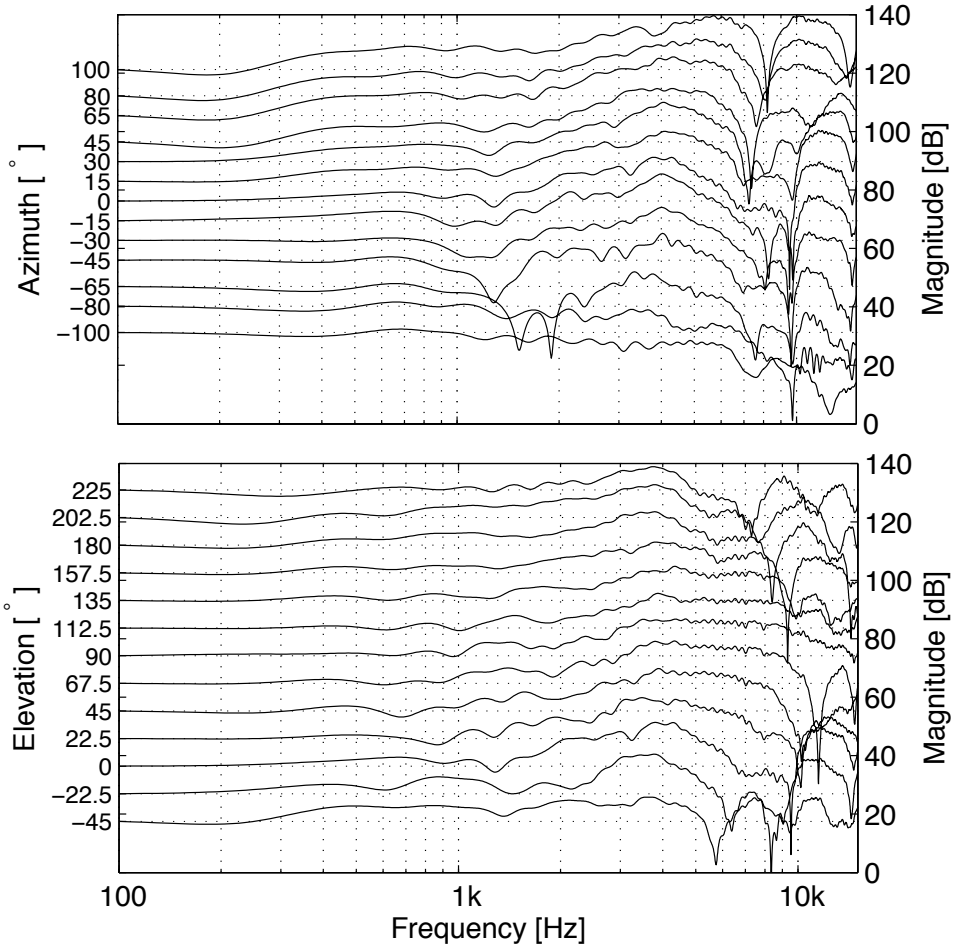

Figure 3.3: Magnitude responses for the left ear of a listener (subject 3 from the CIPIC HRTF database of Algazi *et al.*, 2001) for sound sources with different angles in the horizontal plane (upper panel) and in the median plane (lower panel). The magnitude scale is shown on the right side of the panels and different HRTFs have been offset according to the angles of arrival as shown on the left side of the panels. Note that the measurements do not include the resonance of the ear canal.

detail, establishing a physical basis for what will be denoted as *localization cues*. For later discussion, note that the transfer functions to the ears of listener can also be measured in a chosen acoustical environment yielding what is called *binaural room impulse responses* (BRIRs) (for freely available BRIRs, see Merimaa *et al.*, 2005b).

To begin with, the head has a substantial effect on the ear input signals. Consider, for instance, a situation with a sound source on one side of the head. The interaction of the head with the sound field increases the sound pressure level in the *ipsilateral* ear on the side of the sound source, whereas the sound reaching the *contralateral* ear on the other side needs to propagate (diffract) around the head. As discussed in Section 2.2.1, the shadowing effect of an obstacle depends on its size relative to the wavelength of sound. Hence, the existence of the head induces a frequency-dependent *interaural level difference* (ILD) that increases towards high frequencies. ILDs can be seen as higher levels for positive azimuths in Figure 3.3. Furthermore, the longer propagation

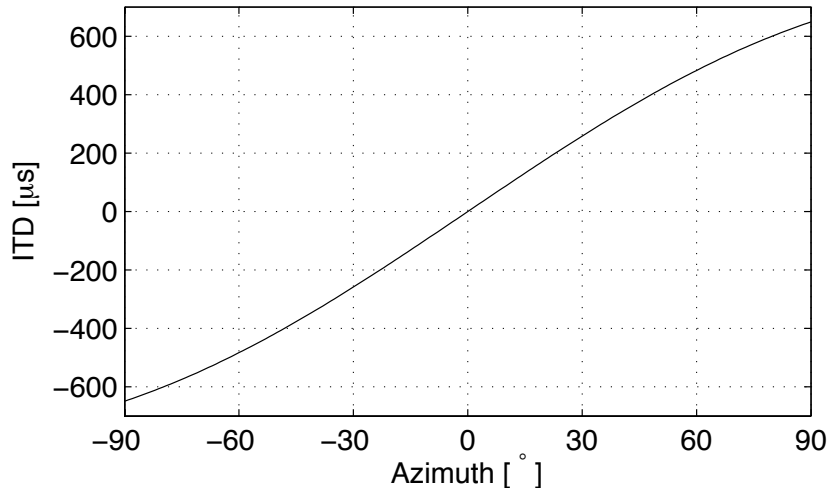

Figure 3.4: ITD as a function of azimuth angle according to Eq. (3.1) with $D \approx 17$ cm (mean of head widths and depths in the database of Algazi *et al.*, 2001).

path to the contralateral ear gives rise to an *interaural time difference* (ITD). Using a spherical model of the head, ITDs for a sound source in the far field of the listener can be approximated with

$$\tau = \frac{D}{2c}\left[\theta + \sin(\theta)\right] , \tag{3.1}$$

where D is the distance between the ears, c is the speed of sound, and θ is the azimuth of the source (Blauert, 1997, p. 75). ITDs according to Eq. (3.1) are illustrated in Figure 3.4.

For a first approximation, ITDs and ILDs have constant values for all source positions with a constant difference in the path length to the ears. Similar to earlier discussion on two microphones (see Section 2.6.1), such positions form hyperbolic surfaces of rotation about the interaural axis. Moreover, in the far field, the surfaces can be approximated with the so-called *cones of confusion* (e.g., Blauert, 1997, p. 179). However, for sources close to the head, the attenuation of sound as a function of distance makes the ILDs distance-dependent[2]. The ITDs are also slightly affected by distance very close to the head (Blauert, 1997, p. 76; Duda and Martens, 1998), but the effect is weaker. Using again a spherical model of the head, Shinn-Cunningham *et al.* (2000) showed that the positions with equivalent ITDs and ILDs create tori of confusion. Nevertheless, at distances greater than 2 m from the head, the dependence of both the ITDs and ILDs on the distance is perceptually negligible, and the approximation of the cones of confusion holds.

From Figure 3.3, it is obvious that there are also differences in the HRTFs in the median plane, which is a special cone of confusion yielding no interaural differences for a perfectly symmetric head. The pinna flange attenuates sound from behind the listener, and the more detailed structures of the pinnae serve as direction-dependent acoustic filters giving rise to *monaural localization cues*. However, due to the size of

[2]Although the effect is slightly different, the cause is the same as in the proximity effect of directional microphones (see p. 19).

the pinnae, their effect is noticeable only at frequencies above 2–3 kHz. Reflections from the torso, on the other hand, have been shown to have a notable elevation-dependent effect on the ear input signals at frequencies below 3 kHz (Gardner, 1973a; Avendano *et al.*, 1999). The discussed direction-dependent filtering effects also affect the ILDs and especially the pattern of ILDs across frequencies as a function of source position (see Duda, 1997). Furthermore, the pinnae have some effect on the ITDs at high frequencies. However, studying the effects of the pinnae is complicated by considerable individual differences, and a more detailed discussion is beyond the scope of this thesis. For further spectral considerations, see Blauert (1969/70, 1997, Section 2.2.3), Shaw (1974, 1997), Hebrank and Wright (1974), Mehrgardt and Mellert (1977), Middlebrooks *et al.* (1989), and Middlebrooks (1992, 1997, 1999a), and for the frequency dependence of ITDs, Mehrgardt and Mellert (1977), Blauert (1997, Section 2.2.3), Middlebrooks and Green (1990), and Gaik (1993).

Apart from creating the direction-dependent cues, the resonances of the pinnae also increase the efficiency of transmission of sound into the ear canal (Pickles, 1988, Section 2.B) and thus contribute to the sensitivity of the ear. The most prominent effect of the ear canal itself (effectively prolonged by the cavum conchae, see Figure 3.2) is a resonance at around 3 kHz. However, due to its small lateral dimensions, at audio frequencies, only one-dimensional sound propagation takes place within the ear canal (Møller, 1992; Hammershøi and Møller, 1996; Blauert, 1997, Sections 2.2, 4.2; Hudde and Engel, 1998c; Hudde, 2005). This observation is important because it implies that the ear canal itself does not have a direction-dependent effect on sound. In other words, HRTFs measured at any point in the ear canal contain the same spatial information, although different equalizations of the HRTFs may be needed depending on the application and the measurement position (Møller, 1992; Hammershøi and Møller, 2005). Note that the HRTFs shown in Figure 3.3 were measured with a blocked ear canal; hence, they do not show the resonance described above.

3.2.2 Middle ear

The middle ear (Pickles, 1988, Section 2.B; Hudde and Engel, 1998a,b,c; Hudde, 2005) consists of the eardrum and a chain of three ossicles: malleus, incus, and stapes, as illustrated earlier in Figure 3.2. The eardrum converts the sound pressure at the end of the ear canal into mechanical vibrations of the ossicles. The last ossicle, the stapes, then transmits the vibrations into the fluids of the inner ear through the oval window. The main function of the middle ear is to implement an efficient transmission of sound energy from the air to the fluids. In engineering terms, the middle ear performs impedance matching, which is best at a frequency range of 800 Hz – 4 kHz (Hudde, 2005; Puria *et al.*, 1997 measured efficient transmission to slightly lower frequencies), thus giving the transfer through the middle ear a bandpass character. The combined effect of the ear canal and the middle ear also shifts the overall resonance peak of the ear from the 3 kHz of the ear canal alone to 4 kHz (Hudde, 2005).

In binaural auditory models, the effect of the middle ear has often been neglected or taken implicitly into account by weighting functions describing the salience of different frequencies. However, some models for the middle ear have also been proposed. Electrical circuit analogies have been presented by Zwislocki (1962) and Hudde and Engel

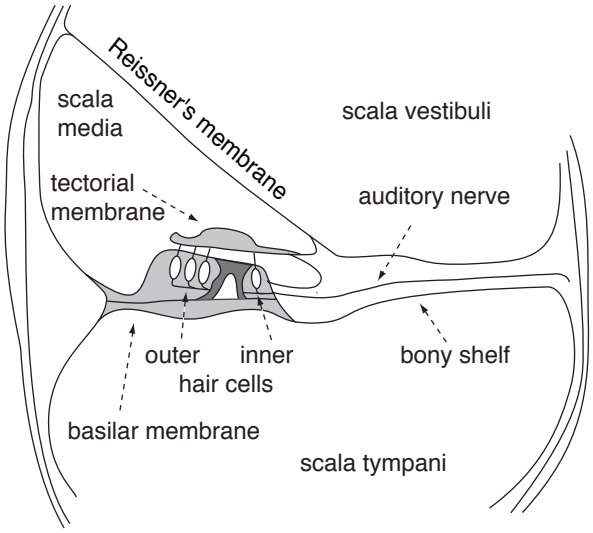

Figure 3.5: Cross-section of the cochlea.

(1998a,b,c). Furthermore, in their binaural auditory model, Breebaart *et al.* (2001a) used a bandpass filter with a rolloff of 6 dB/octave below 1 kHz and -6 dB/octave above 4 kHz.

3.2.3 Inner ear

The inner ear is where the conversion of mechanical vibrations to neural signals takes place. The inner ear actually consists of two organs coupled by common fluids: the vestibular organ (including the semicircular canals shown in Figure 3.2) and the cochlea. However, only the cochlea has an auditory function and will be discussed here. The cochlea has a coiled shape with a cross-section as illustrated in Figure 3.5. The cross-section displays three fluid-filled ducts: scala vestibuli, scala media, and scala tympani. A movement of the stapes in the middle ear is transmitted into a displacement of the fluid of the scala vestibuli and further to the scala media through the Reissner's membrane. The movement of the fluid in the scala media causes a wave-like displacement in the basilar membrane relative to the tectorial membrane, which bends the stereocilia (actin filaments) on top of the inner hair cells. The bending further elicits spikes in the fibers of the auditory nerve (Pickles, 1988, Chapter 3; Hudde, 2005). In addition to realizing the essential neural transduction, the cochlea also has an important function in performing a spectral analysis of sound. In the following, the spectral analysis and the operation of the basilar membrane will be discussed first, followed by an introduction to the actual neural transduction.

Basilar membrane and frequency analysis

The frequency analysis capability of the cochlea is closely related to its anatomy. The stiffness of the basilar membrane is highest at the *base* of the cochlea near the oval window and decreases towards the other end, the *apex*. According to the traveling wave theory, a sound of certain frequency creates a resonance at a specific position

along the basilar membrane with waves of lower frequencies propagating deeper towards the apex. The inner hair cells around the resonance position then receive highest excitation. However, the organ of Corti also includes the active outer hair cells whose existence is known to increase the sensitivity and to sharpen the tuning of the cochlea (Hudde, 2005). The exact nature of the active operation is not known. It has traditionally been assumed that the vibrations of the outer hair cells are fed back to the basilar membrane. Nevertheless, Bell and Fletcher (2004) recently proposed an alternative mechanism in which the outer hair cells might form a secondary resonance structure that directly stimulates the inner hair cells through a jetting fluid.

According to Hudde (2005), no published model so far is able to fully simulate the physiological operation of the cochlea without either making some generalized assumptions or involving unexplained feedback loops. Nevertheless, the functionality of the cochlea can be simulated to a certain degree. In auditory modeling, it is common to use a bandpass filterbank to simulate the spectral analysis. In order to develop the shape of the filters, it is possible to directly measure the stimulus-dependent mechanical motion of the basilar membrane, or the frequency-dependent responses of the afferent auditory nerve fibers, in a laboratory animal. Close to the *characteristic frequency* (the frequency yielding maximum response for a certain position on the basilar membrane or for a certain auditory nerve fiber), the frequency selectivity of both measurements appears similar, suggesting that the frequency selectivity is derived from the basilar membrane (Ruggero *et al.*, 1997; see also Pickles, 1988; Geisler and Cai, 1996).

In addition to physiological measurements, it is also possible to exploit the indisputable psychoacoustic evidence of human auditory processing within the so-called *critical bands* in the derivation of the shapes of the auditory filters. Critical band processing can be observed, for instance, in loudness (Zwicker *et al.*, 1957) and consonance perception (Plomp and Levelt, 1965; Plomp and Steeneken, 1968), absolute threshold of hearing, masking phenomena, and perception of phase (Zwicker, 1961; Zwicker and Fastl, 1990, Section 6; see also Zwicker and Terhardt, 1980). Most comprehensive knowledge on the auditory filters has been gained from simultaneous notched-noise masking experiments, where the detection threshold of a signal as a function of spectral distance to masking noise above and/or below the signal frequency is studied (e.g., Patterson, 1974, 1976; Patterson and Nimmo-Smith, 1980; Moore, 1987; Glasberg and Moore, 1990; Moore *et al.*, 1990; Shailer *et al.*, 1990; Zhou, 1995; Sommers and Gehr, 1998; Yasin and Plack, 2005; see also Moore and Glasberg, 1983; Glasberg *et al.*, 1984; Moore *et al.*, 1995). In their landmark paper, Glasberg and Moore (1990) included compensation for the transmission of sound through the outer and middle ear in the fitting process. Using a gammatone filter model, Slaney (1993) further developed a widely used auditory filterbank implementation corresponding to the equivalent rectangular bandwidth (ERB) scale of Glasberg and Moore (1990). The resulting filters for a range of center frequencies are illustrated in Figure 3.6. Furthermore, the ERB scale has been shown to correspond closely to Greenwood's (1961; 1990) frequency-position function of the physiological mapping of characteristic frequencies to different positions in the cochlea (for a comparison plot, see Slaney, 1993, p. 3).

The active role of the outer hair cells was already mentioned, and it can be seen as two nonlinear phenomena: compression and suppression. In addition to having

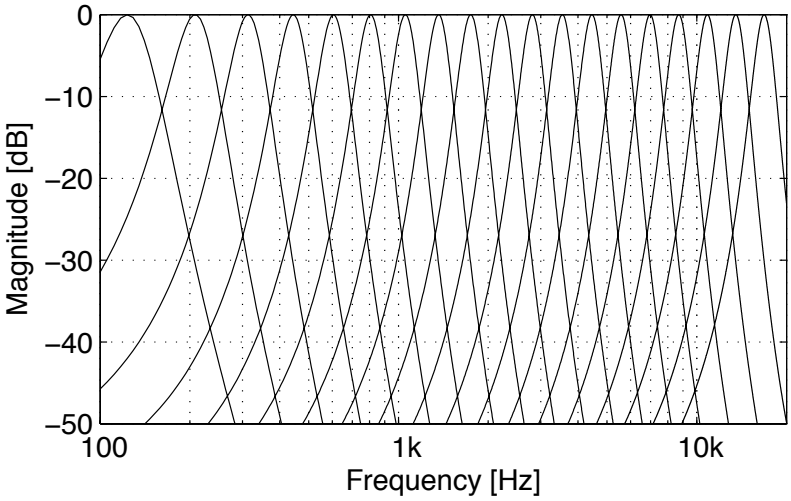

Figure 3.6: Frequency responses of a gammatone filterbank (Slaney, 1998) with a spacing of two ERBs between center frequencies.

direct consequences, for instance, in forward masking, both the compression and suppression also affect the auditory filters. As the SPL of a specific narrowband sound increases, the compression reduces the gain at the position of the basilar membrane being maximally excited at low levels. However, the gain at the skirts of the auditory filter is less reduced, which makes the characteristic frequency to shift slightly and the filter to appear broadening with increasing SPL[3] (e.g., Weber, 1977; Glasberg and Moore, 1990, 2000; Rosen *et al.*, 1998). Ruggero *et al.* (1997) measured physiological cochlear responses growing linearly up to about 20 dB SPL and with a slope of 0.2 dB/dB in the intensity range 40–80 dB SPL at the characteristic frequency (see also Rhode and Recio, 2000). Furthermore, the compression is almost immediate, starting within 100 μs from stimulus onset (Recio *et al.*, 1998). Comparable compression rates have also been measured psychoacoustically for human listeners (e.g., Oxenham and Plack, 1997; Plack and Oxenham, 1998, 2000; Moore *et al.*, 1999; Nelson *et al.*, 2001; Lopez-Poveda *et al.*, 2003; Plack and Drga, 2003; Nelson and Schroder, 2004).

Contrary to compression, the suppression is a reduction of gain at a certain position in the cochlea by sound at frequencies outside the corresponding auditory filter. The suppression decreases with increasing frequency separation of the suppressor and the suppressed signal. The well-known two-tone suppression (suppression of a tonal response by another tone) has been measured physiologically (e.g., Nuttall and Dolan, 1993) and cochlear responses to more complex stimuli were investigated recently by Recio and Rhode (2000), and Rhode and Recio (2001a,b) (for related psychoacoustical studies, see also Oxenham and Plack, 1997; Hicks and Bacon, 1999; Yasin and Plack, 2003; Plack and Drga, 2003). The effect of the suppression on frequency selectivity is similar to compression in the sense that it also causes broadening of the auditory filters. Heinz *et al.* (2002) have argued that due to the fast dynamic changes

[3]Actually, it is the active operation of the outer hair cells that makes the filters narrow at low levels, so in fact the broader responses at higher levels are closer to the natural response of the basilar membrane itself (Ruggero *et al.*, 1997; Moore *et al.*, 1999).

in the tuning due to the nonlinear phenomena, it is actually improper to discuss the shape of the auditory filters without specifying the stimulus configuration, and that the notched-noise masking experiments overestimate the width of the filters because of suppression. Indeed, sharper tuning has been found in forward masking compared to simultaneous masking experiments (e.g., Moore and Glasberg, 1981; Glasberg *et al.*, 1984; Sommers and Gehr, 1998).

From the above discussion, it is obvious that the linear gammatone filterbank lacks generality. Dynamic extensions realizing a signal-dependent bandwidth have been recently proposed by several authors (e.g., Robert and Eriksson, 1999; Meddis *et al.*, 2001; Lopez-Poveda and Meddis, 2001; Irino and Patterson, 2001; Zhang *et al.*, 2001; Heinz *et al.*, 2001; Patterson *et al.*, 2003; see also Shamma *et al.*, 1986; Shamma and Morrish, 1987; Giguère and Woodland, 1994a,b; Glasberg and Moore, 2000). However, the static implementation by Slaney (1993) is still the dominant approach used successfully in a number of recent auditory models and modeling studies (e.g., Dau *et al.*, 1997a,b; Pulkki *et al.*, 1999a; Breebaart *et al.*, 2001a,b,c; Braasch, 2002; Braasch and Blauert, 2003; Buchholz and Mourjopoulos, 2004a,b; Braasch, 2005), proving that it can also produce valid results. As a special case of frequency selectivity, the linear gammatone filterbank should be able to provide sufficient approximation of the cochlear frequency selectivity at least for phenomena which are not notably level dependent and involve broadband sound, since for such stimuli the gammatone filters are bound to have close to the correct shape at some level.

Neural transduction

It was already established that the frequency selectivity observed in the auditory nerve fibers initiates from the level of the basilar membrane. However, the neural transduction has other properties important for understanding and modeling the auditory perception. Anatomically, the auditory nerve consists of approximately 30,000 nerve fibers leading from the organ of Corti to the cochlear nucleus in the brainstem. Despite a larger number of outer hair cells compared to inner hair cells in the cochlea, 90–95% of the nerve fibers are connected to the inner hair cells with approximately 20 fibers innervating each inner hair cell (Pickles, 1988, pp. 78-79). The active role of the outer hair cells was also already discussed, and the following concentrates on the auditory nerve fibers connected to the inner hair cells and thus conveying neural information towards the higher levels of the auditory system.

The firing of each auditory nerve fiber as a response to the excitation of the inner hair cell is statistical, and the firing rate grows with increasing excitation up to a certain saturation level. Furthermore, the nerve fibers exhibit spontaneous firing even without external auditory stimulation. Liberman (1978) divided the auditory-nerve units into three classes based on their spontaneous firing rate. High-spontaneous-rate units have low thresholds, i.e., they start responding to sound at their characteristic frequency at barely audible levels. Medium- and low-spontaneous-rate units, on the other hand, have larger thresholds, and the thresholds of the low-rate units may even exceed the saturation levels of the high-rate units. Thus, not only the rate of firing of single auditory nerve fibers grows as a function of SPL, but also the number of neural units firing.

Detailed physiology of the neural transduction has been reviewed by Pickles (1988, Section 5.C), and the following description is mainly based on the physiologically motivated neural model of Meddis (Meddis, 1986, 1988; Meddis *et al.*, 1990; see also Hewitt and Meddis, 1991) which was recently revised by Sumner *et al.* (2002, 2003b). In this model, the auditory nerve fibers are stimulated by a release of a transmitter substance from the inner hair cells, according to physiology. The transmitter is released through a permeable membrane of the inner hair cells, and the permeability is increased by bending of the stereocilia in one direction only. Hence, the release rate of the transmitter resembles a half-wave rectified output of the auditory filters (Meddis, 1986; Sumner *et al.*, 2002).

The probability that a spike process will be triggered is determined by the instantaneous amount of the neurotransmitter in the cleft between the inner hair cell and the auditory nerve fibers. However, the inner hair cells store a limited amount of the transmitter. Following a release, part of the transmitter is returned to the hair cell through a reprocessing stage, causing a small delay. Furthermore, a certain amount of the transmitter is irretrievably lost and replaced by new transmitter manufactured in the hair cell. After the onset of a strong stimulus, the amount of free transmitter in the hair cell decreases, causing a gradual decrease in the release rate until it is counterbalanced by the reuptake and manufacturing processes (Meddis, 1986). This causes an adaptation behavior measured from the auditory nerve fibers as a sharp onset response with the spike rate dropping rapidly over the first 10–20 ms and more slowly over the next several minutes (Pickles, 1988, pp. 80–81; see also Sumner *et al.*, 2003a).

The ability of the neural signals to follow the fine structure of the stimulus is limited both by the rate at which the transmitter can be cleared from the cleft (Meddis, 1988) and by inertia of the release process that can be characterized as low-pass filtering by the inner hair cell membrane (Sumner *et al.*, 2002). At high frequencies, the nerve fibers fire with equal probability in every part of the cycle of a sinusoidal stimulus; i.e., the firing rate is determined by the envelope of the signal. However, at low frequencies, the responses of the nerve fibers are increasingly phase-locked to the stimulus, and the synchronization increases slightly with stimulus amplitude (Sumner *et al.*, 2002; for measurements see Johnson, 1980).

The essential functional components of the neural transduction described above are half-wave rectification, low-pass filtering, adaptation, and a certain level of spontaneous activity. A detailed functional model consisting of physiologically possible elements was proposed by Ross (1996). In less detailed peripheral models, the functional components of the neural transduction have often been implemented as cascaded signal processing operations yielding an output signal that reflects some kind of average firing rate statistics of the auditory nerve. The spontaneous activity is typically simulated by adding low-level random noise to the output signals of an auditory filterbank before the main neural transduction model. For the transduction, half-wave rectification followed by lowpass filtering at 1 kHz and some kind of signal-dependent compression modeling the adaptation has been used by several authors (e.g., Dau *et al.*, 1996; Kohlrausch and Fassel, 1997; Buchholz and Mourjopoulos, 2004a; see also Karjalainen, 1985, 1996).

Another neural transduction model with square-law half-wave rectification followed

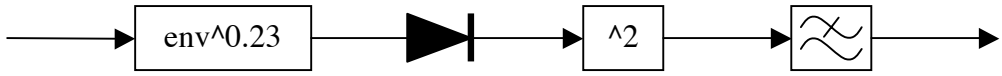

Figure 3.7: Bernstein *et al.*'s (1999) neural transduction model, consisting of envelope compression, square-law half-wave rectification, and low-pass filtering.

by a fourth-order low-pass filter with a cutoff frequency of 425 Hz has been proposed by Bernstein and Trahiotis (1996). Their model is motivated by physiological studies and specifically fitted to predict frequency-dependent binaural detection performance with a cross-correlation model (see Section 3.6.1). In order to account for a larger set of stimuli, the model was also later extended with an envelope compression stage prior to the half-wave rectification, simulating roughly the compression of the basilar membrane at medium (40–70 dB) SPL (Bernstein *et al.*, 1999). The structure of the extended model is illustrated in Figure 3.7.

3.3 Localization

After describing what kind of information is extracted by the peripheral auditory system, we now turn to how the information is utilized in auditory localization. As mentioned in Section 3.1, the physiological mechanisms processing the information coded into the signals of the auditory nerve are less well-known than those of the auditory periphery. Nevertheless, based on psychophysical evidence, there is no doubt that the direction-dependent cues produced by the head, torso, and pinnae (see Section 3.2.1) contribute to localization. Furthermore, the localization is affected by the operation of the inner ear, as will be seen in this section.

There are several ways to psychophysically measure the human localization performance. Localization accuracy can be studied using the absolute difference between the location of a physical sound event and the reported location of an auditory event. This measure includes any systematic bias and/or possible large differences due to stimulus-specific confusions, which will be discussed later in this section. Another possibility not concerned with whether the locations of the auditory event and the sound event coincide is to study the *just noticeable differences* (JNDs), i.e., the smallest changes in the localization cues or the location of a source leading to a change in the position of the resulting auditory event. The JNDs thus reflect the accuracy of the directional processing in the auditory system. The JND of the direction of a source is also denoted as *minimum audible angle* (MAA), and Blauert (1997, p. 21) uses the term *localization blur*. The MAA has been found to be both stimulus- and direction-dependent. For azimuthal localization in the horizontal plane, the MAA in front of a listener is at best about 1° and at the sides between 3–10 times its value for the forward direction. On the other hand, the MAA of elevation localization is in front of a listener at best about 4° and increases for other directions within the median plane (Blauert, 1997, Section 2.1; see also Perrott and Saberi, 1990).

Although some systematic differences between the locations of the auditory events and sound events may occur, auditory localization of a single source in free-field is

conceptually simple because all localization cues suggest the same direction. However, in the complex listening situations, i.e., in the presence of several sound sources and/or room reflections, oftentimes sound from several different directions concurrently reaches the position of the listener. Furthermore, the superposition of sound emanating from several directions may result in instantaneous localization cues that often do not correspond to any of the source directions. Nevertheless, humans have a remarkable ability to resolve also such complex composites of sound into separate localizable auditory events at directions corresponding to the sound sources.

Several studies discussed in this section have presented conflicting or nonrealistic localization cues to listeners through headphones. In such cases, the sound is typically localized inside the head somewhere on the interaural axis, which will be denoted as *lateralization*. Nevertheless, with a realistic simulation of the transduction of sound into the ears of the listener (using HRTFs or BRIRs), externalization can be achieved in headphone studies (Hartmann and Wittenberg, 1996; Kulkarni and Colburn, 1998). A number of investigators have also shown that localization with individualized HRTFs is, at least in the left–right direction, comparable to localization of corresponding real sources (e.g., Wightman and Kistler, 1989b; Bronkhorst, 1995; Møller *et al.*, 1996; Kulkarni and Colburn, 1998; Middlebrooks, 1999b; Langendijk *et al.*, 2001; see also Wenzel *et al.*, 1993). Hence, unless there is a reason to believe that the reproduction method has affected the results, localization studies with HRTF stimulation will be treated equivalently to localization with real sources.

Localization as such also involves distance of sound sources. However, the localization cues discussed so far do not provide reliable information on distance except for sources close to the listener (the tori of confusion). Hence, this section will be limited to directional localization; perception of distance will be briefly discussed in Section 3.5.3. This section is organized as follows: The effect of the individual cues on localization is discussed in Section 3.3.1, and Sections 3.3.2 and 3.3.3 review localization results for two different complex listening situations.

3.3.1 Effect of individual localization cues

The introduced localization cues can be divided into two categories: binaural and monaural (spectral) cues. The binaural cues manifest themselves as differences between the ear input signals, whereas evaluation of the spectral cues depends on the ability of a listener to distinguish between the spectrum of the source signal and the effect of the source position on the spectrum. This makes the binaural cues more robust and generally more important for localization, although, as discussed earlier, at a first approximation, they can only determine the cone (or torus) of confusion that a source is located in. Indeed, a considerable amount of front–back and up–down confusions has been reported in localization experiments (e.g., Wightman and Kistler, 1989b; Makous and Middlebrooks, 1990; Wenzel *et al.*, 1993; Møller *et al.*, 1996; Middlebrooks, 1997, 1999b). However, Perrett and Noble (1997) and Wightman and Kistler (1999) reported almost complete disappearance of the front–back confusions when their listeners were instructed to move their heads (for the effect of head movements, see also Wallach, 1939, 1940; Begault *et al.*, 2003). Furthermore, in the study of Perrett and Noble (1997), localization within cones of confusion remained

possible also with distorted monaural cues, i.e., with the help of the changes induced only in the binaural cues by head movements.

For the rest of this section, the discussion will be limited to the effect of static localization cues. Starting from the binaural cues, according to the classic duplex theory of binaural localization (Strutt [Lord Rayleigh], 1907), the ITDs are used for localization at low frequencies and ILDs at high frequencies. However, the theory is not fully correct even if it does reflect the frequency-dependent importance of ITDs and ILDs (see Section 3.3.2). For sinusoidal signals, the effectiveness and detectability of the ITDs does indeed decrease at high frequencies. This happens for two physiological reasons: (1) above approximately 800 Hz, the period of a sinusoid is shorter than range of naturally occurring delays, making determination of the actual ITD from the steady state phase difference of the sinusoidal signal ambiguous, and (2) the decline in phase locking of the auditory nerve at high frequencies makes estimation of the ITD from the phase difference increasingly difficult (Blauert, 1997, Section 2.4.1). Nevertheless, interaural envelope delays contribute to lateralization at considerably higher frequencies (e.g., Henning, 1974, 1980; McFadden and Pasanen, 1976; Nuetzel and Hafter, 1976; Dye *et al.*, 1994a; Blauert, 1997, pp. 150–154, 316–318). Using the so-called "transposed stimuli" where the temporal information available in low-frequency waveforms is imposed on the envelopes of high-frequency stimuli (van de Par and Kohlrausch, 1997), Bernstein and Trahiotis (2002, 2003) even showed that the transposed high-frequency ITDs can have an equal effect as the same ITDs at low frequencies. The JNDs of ITDs are, at best, in the order of 10 μs and increase with an increasing displacement of the lateralized auditory event from the midline (Blauert, 1997, p. 153, 316).

ILDs have also been found effective throughout the audible frequency range with JNDs at best slightly below 1 dB. As in the case of ITDs, the JNDs of ILDs increase somewhat with the displacement of the auditory event from the midline. Furthermore, the sensitivity to ILDs seems to be lower at around 1 kHz than at other frequencies (Grantham, 1984b; Yost and Dye, 1988; Blauert, 1997, pp. 160–165, 316). However, the duplex theory is correct in the sense that naturally occurring ILDs for low-frequency sources farther than about 2 m in free field may be small enough to be negligible, especially compared to the ITDs.

Localization based on strictly monaural cues can only be studied with headphones or, assuming perfect symmetry of the head, within the median plane in free field. Outside the median plane, the filtering by the pinnae also produces frequency-dependent ILDs that change as a function of position within a cone of confusion (Duda, 1997) and have been shown to contribute to resolving the direction of a source (Wightman and Kistler, 1997a; see also Searle *et al.*, 1975; Morimoto, 2001; Jin *et al.*, 2004). Nevertheless, similarities between localization in the median plane and other cones of confusion have been found. Both in the median plane (Blauert, 1969/70, 1997, pp. 107–113), as well as in other cones of confusion (Middlebrooks, 1992, 1997), the localization judgements of narrowband stimuli tend to cluster around certain ranges, depending on their frequency but independent of the actual direction of a stimulus. Furthermore, the directions of the clustering are related to elevations from which a broadband stimulus would produce high levels at the frequency in question due to HRTF filtering (see the median plane HRTFs in Figure 3.3).

For broadband stimuli, utilization of the monaural localization cues is possible using either *a priori* knowledge or some assumptions about the stimuli. Blauert (1997, pp.103–105) reviews studies where familiarity with test signals has improved monaural localization. Furthermore, Wightman and Kistler (1997a,b) showed that considerable trial-to-trial fluctuations in the noise spectrum severely disrupt monaural localization. Zakarauskas and Cynader (1993) proposed that the spectrum of sound needs to be locally smooth for the extraction of the monaural cues to be possible. It is, however, not fully clear which specific peaks and/or notches in the HRTFs actually carry the monaural localization information, or whether it is the spectral shape across relatively broad bands that contributes to localization (for related research and discussion, see Roffler and Butler, 1967; Blauert, 1969/70; Searle *et al.*, 1975; Humanski and Butler, 1988; Kistler and Wightman, 1992; Middlebrooks, 1999b). Langendijk and Bronkhorst (2002) showed that the most important up–down cues are located in the 6–12 kHz and front–back cues in the 8–16 kHz frequency region. Nevertheless, Asano *et al.* (1990) found that frequencies below 2 kHz also have an effect on front–back localization.

Individual differences in the HRTFs, and especially in the pinna cues, were also mentioned in Section 3.2.1. Indeed, studies simulating free-field listening using nonindividualized HRTFs (i.e., HRTFs measured from some other listener or from a dummy head) have shown that localization in the left–right direction is fairly robust to individual differences, whereas localization performance in the front–back and up–down directions is generally degraded by the incorrect cues from another individual (e.g., Wenzel *et al.*, 1993; Bronkhorst, 1995; Møller *et al.*, 1996; see also Middlebrooks, 1999b; Begault *et al.*, 2003). The remarkable results of Hofman *et al.* (1998) showed that it is possible to learn to use modified spectral cues for up–down localization with some weeks of constant exposure to them. However, despite the capability for adaptation, these results reinforce the robustness of the binaural cues.

3.3.2 Conflicting cues and concurrent sound sources

More insight into the relative importance of different localization cues can be gained by investigating stimuli with conflicting cues. Classic trading experiments (Blauert, 1997, Section 2.4.3) have searched for opposing ITDs and ILDs that counterbalance each other resulting in a centralized lateralization[4]. The reported trading ratios appear to depend on stimulus type and level, as well as on individual listeners. Furthermore, the investigators have pointed out that although the trading is possible, it does not yield the same auditory event as presentation of identical signals to both ears. Instead, the resulting event may appear broader or even split into separate "time image" and "intensity image" (see especially Hafter and Carrier, 1972).

In general, localization depends on whether sound (consisting of one or more sound events) is localized as a single auditory event or multiple segregated auditory events. However, there seems to be no simple answer to when the fusion happens and how the resulting cues are combined (e.g., Gardner, 1973b; Buell and Hafter, 1991; Buell

[4]Also, lateralization studies with only ITD or ILD cues use, in fact conflicting cues in the sense that neither of these cues occurs alone in natural listening situations. Hence, the earlier results on the effects of a single binaural cue could also be seen as trading between the investigated cue and a central lateralization suggested by the other cue.

et al., 1994; Buell and Trahiotis, 1994, 1997; Dye *et al.*, 1994b, 1996; Dye, 1997; Hill and Darwin, 1996; Stellmack and Lufti, 1996; Bernstein and Trahiotis, 2004, 2005). For fused auditory events with conflicting cues, Wightman and Kistler (1992, 1997b) and Macpherson and Middlebrooks (2002) have shown that as long as low frequency energy is present, ITDs dominate the localization of broadband sound, despite ILDs and spectral cues suggesting other directions. However, in accordance with the classic duplex theory, ILDs are the dominant cues for highpass filtered stimuli. Furthermore, the reliability of the cues across frequencies also affects their salience such that inconsistency of ITDs may increase the weight of the ILDs (Buell *et al.*, 1994; Buell and Trahiotis, 1997; Wightman and Kistler, 1997a). Blauert (1997, p. 372) has proposed that the higher stages of the auditory system, involving at least partly cognitive processes, set up and test hypotheses as to what an appropriate perception would be. Such an approach would be consistent with the fact that reliable cues are more heavily weighted and it could also conceptually explain whether the fusion happens or not.

Also, in the case of segregated auditory events, sound sources with overlapping spectral content create at least instantaneous conflicting or erroneous localization cues. Several related localization and lateralization studies have been reviewed by Blauert (1997, Sections 3.2, 3.3, and 4.4). More recently, the effect of independent distracters on the localization of a target sound has been investigated by Good and Gilkey (1996), Good *et al.* (1997), Lorenzi *et al.* (1999), Hawley *et al.* (1999), Drullman and Bronkhorst (2000), Langendijk *et al.* (2001), Braasch and Hartung (2002), and Braasch (2002, 2003). Although some of the studies have shown statistically significant biases or degradation in the localization accuracy, the effects of introducing one or two distracters on localization are generally small. Only when the number of distracters is increased or the target-to-distracter ratio (TDR) is reduced does the localization performance begin to degrade notably. For most configurations of a target and a single distracter in the frontal horizontal plane, the accuracy stays very good down to a target level only a few dB above the threshold of detection (Good and Gilkey, 1996; Good *et al.*, 1997; Lorenzi *et al.*, 1999). However, the presence of concurrent sound can change the weighting of the ITD and ILD cues (Braasch, 2003). Some of the studies cited above will be discussed in more detail in connection with related auditory model simulations in Section 4.3.1.

3.3.3 Precedence effect and localization in rooms

Another complex listening situation occurs in the presence of room reflections. The physical propagation of sound in rooms was already discussed in Section 2.2.2. Despite the resulting numerous consecutive sound events, only a single auditory event typically occurs in the direction of the sound source. The phenomenon (or phenomena, see below) is nowadays usually called the precedence effect (originating from Wallach *et al.*, 1949), but some papers also refer to it as the Haas effect (according to Haas, 1949) or the law of the first wavefront (e.g., Blauert, 1997, Section 3.1.2). Before further discussion, it should be emphasized that the precedence effect is not suppression of all perceptual effects of, for instance, the room reflections, but only related to their fusion and individual localization.

Most precedence effect studies up to date have been conducted using simplified stimuli consisting of a lead sound from one direction (or with a certain lateralization) followed by an identical lag sound some milliseconds later from another direction (or lateralization). Extensive reviews have been given by Zurek (1987), Blauert (1997, Sections 3.1, 4.4.2, and 5.4), and Litovsky *et al.* (1999). Litovsky *et al.* (1999) divide the precedence effect further into three phenomena: the already mentioned fusion, localization dominance, and discrimination suppression. The fusion of a lead and a lag into a single auditory event breaks down when the delay between the lead and the lag is increased beyond the *echo threshold*. The echo threshold is strongly stimulus-dependent, and quantitative estimates from the literature vary between 2–50 ms (Blauert, 1997, Section 3.1.2; Litovsky *et al.*, 1999).

The directional perception of a fused pair of stimuli with an interstimulus delay shorter than 1 ms is called *summing localization*. For two identical stimuli with no delay between them, both sound events contribute equally to the localization (see also amplitude panning in Section 5.2). When a delay is introduced, the effect of the lag decreases up to a delay of approximately 1 ms, and for delays greater than that, *localization dominance* by the lead occurs, although the lag might still contribute slightly to the localization of the fused auditory event. Furthermore, increasing the level of the lag helps in shifting the auditory event back towards it (Litovsky *et al.*, 1999).

Experiments on the *discrimination suppression* aspect of the precedence effect study the ability of listeners to extract information about changes in the direction of the lead and/or the lag. Shinn-Cunningham *et al.* (1993) showed that the discrimination suppression is directly related to the amount of localization dominance. Small changes in the direction of a lag as part of a fused auditory event appear very difficult to notice, but large enough changes can be discriminated. The presence of the lag also degrades somewhat the discrimination of the lead (e.g., Litovsky and Macmillan, 1994; Litovsky *et al.*, 1999; Stellmack *et al.*, 1999). While fusion and localization dominance have been found to operate both in the horizontal and median planes (Blauert, 1971; Litovsky *et al.*, 1997; Rakerd *et al.*, 2000; Dizon and Litovsky, 2004), discrimination suppression has only been reported in the horizontal plane. However, in the median plane, the discrimination of changes in the direction of the lag appears to be mediated by perceptual changes in the (monaural) pitch of the fused auditory event, whereas the corresponding cue in the horizontal plane is a change in spatial perception (Litovsky *et al.*, 1999). Hence, it might be that the observed lag discrimination ability in the median plane does not reflect the properties of the spatial hearing.

The precedence effect has been traditionally considered to persist up to the echo threshold. However, both localization dominance (Litovsky and Shinn-Cunningham, 2001) and discrimination suppression (Tollin and Henning, 1998; Litovsky and Shinn-Cunningham, 2001) have recently been found to remain effective for delays even somewhat beyond the breakdown of the fusion. Furthermore, the precedence effect depends on previous stimulation. A *buildup* of precedence occurs when a lead/lag stimulus with a delay slightly above the echo threshold is repeated several times. During the first few stimulus pairs, the precedence effect is not active and two auditory events are independently perceived, but after the buildup, the clicks merge to a single auditory event in the direction of the lead (e.g., Clifton and Freyman, 1989, 1997; Freyman *et*

al., 1991; Grantham, 1996; Krumbholz and Nobbe, 2002). The buildup also increases lag discrimination suppression, although according to the experiments of Yang and Grantham (1997b), to a smaller extent than fusion.

The precedence effect literature also discusses a *breakdown* of precedence (also known as the Clifton effect) when, for instance, the directions of the lead and lag are suddenly swapped (Clifton, 1987; Clifton and Freyman, 1989, 1997; for more breakdown conditions, see Clifton *et al.*, 1994; McCall *et al.*, 1998). However, the results of Djelani and Blauert (2001, 2002) suggest that the buildup is actually direction-specific, and that what has been earlier reported as breakdown of precedence is actually a consequence of precedence not being built up for a new lag direction. With such an interpretation, the results of a breakdown following a spectral change (McCall *et al.*, 1998) would lead to the conclusion that the buildup of precedence is also frequency-specific. Djelani and Blauert (2002) also showed that without stimulus activity, the effect of the buildup decays slowly by itself.

Generally, for the operation of the precedence effect, it does not seem necessary for the lead and lag to be identical, although the results in the literature are somewhat contradictory. The studies of Blauert and Divenyi (1988) and Yang and Grantham (1997a) indicate that spectral overlap is needed and that the discrimination suppression operates within frequency channels. Nevertheless, lag discrimination suppression (Divenyi, 1992) and localization dominance Shinn-Cunningham *et al.* (1995) have been found effective even for leads and lags with nonoverlapping frequency content. Furthermore, Perrott *et al.* (1987) reported considerable reduction in fusion of due to using uncorrelated broadband lead and lag samples. However, Yang and Grantham (1997a) found that with brief 1-octave noise bursts using uncorrelated or temporally different lead and lag had little effect on discrimination suppression compared to correlated samples.

As mentioned earlier, the precedence effect is not full suppression of lagging sound events. Hence, studies of detectability of lagging sound events by any means possible do not belong under the precedence effect and will be briefly discussed in Section 3.5. With this limitation, few authors have investigated the different aspects of precedence effect with several lags or in the presence of concurrent sound. Ebata *et al.* (1968) showed that adding another click or continuous sound between the lead and the lag can extend the temporal scale of the precedence effect. Tollin and Henning (1999) also discovered in experiments with three consecutive fused clicks that the third click can change the spatial perception of the second click. In accordance with single lead-lag pairs, Djelani and Blauert (2001) found a breakdown of fusion when the pattern of multiple reflections was changed after a buildup phase. Furthermore, Chiang and Freyman (1998) showed that the presence of concurrent background noise decreases echo threshold and in some cases reduces localization dominance. Other related studies have mainly concentrated on localization accuracy in rooms, which can be seen as research on localization dominance with multiple lags (early reflections and late reverberation, denoted here together as reverberation). Overall, the localization accuracy seems to be slightly degraded by the reverberation (e.g. Hartmann, 1983; Begault, 1992; Giguère and Abel, 1993; Braasch and Hartung, 2002). However, Begault *et al.* (2003) found the opposite effect for azimuthal localization. Furthermore, Shinn-Cunningham (2000) showed that localization performance increases with practice in

a chosen room.

Both the bandwidth and onset of a stimulus appear to play a significant role in localization both in rooms as well as in simplified precedence effect conditions. Anomalies in the localization dominance for narrowband stimuli were first reported by Blauert and Cobben (1978). Hartmann (1983) found that for the precedence effect to work in rooms, the stimulus needs to be either broadband or have a strong attack transient (see also Hartmann and Rakerd, 1989; Hartmann, 1997). The localization of sinusoidal signals in the presence of a single reflection was investigated further by Rakerd and Hartmann (1985, 1986) with the result that the relative contributions of the direct sound and the (usually) incorrect steady state binaural cues on localization depend on the onset rate of the tones. Furthermore, it was verified that the suppression of the ongoing cues may last for several seconds. Giguère and Abel (1993) reported similar findings in reverberant environments for low-frequency noise with a bandwidth of one-third octave. However, rise time had little effect on the localization performance above 500 Hz, whereas increasing the reverberation time decreased the localization accuracy at all center frequencies. The bandwidth dependence of the precedence effect was investigated further by Braasch *et al.* (2003), who found that the localization dominance started to fail when the bandwidth of noise centered at 500 Hz was reduced to 100 Hz.

3.4 Frequency and time resolution of binaural hearing

The auditory system analyzes spatial sound with a limited time and frequency resolution. Knowing this resolution is important both for modeling studies and for reproduction purposes. In Section 3.2.3, it was established that the cochlea performs a frequency analysis which is reflected in monaural auditory processing. For the binaural auditory system, the question is thus whether the full frequency resolution is utilized. This will be discussed in Section 3.4.1. So far, few remarks have been made about the time resolution of the human hearing. The auditory system appears to be able to combine information from different time ranges depending on the task. In the context of this thesis, we are only interested in the time resolution of spatial perception often characterized with the term *binaural sluggishness*. The binaural sluggishness will be discussed in section 3.4.2. For temporal processing of monaural sound, see e.g. Moore *et al.* (1988) and Plack and Moore (1990).

3.4.1 Frequency resolution

Most comprehensive knowledge of the frequency resolution of the binaural hearing comes from detection studies, as was also the case for monaural hearing. By observing the dependence of BMLD on spectral configuration of binaural stimuli, the frequency resolution can be derived analogously to monaural hearing (see Section 3.2.3). Most related investigations agree in that the frequency resolution of binaural hearing is comparable to, or for certain stimulus configurations slightly lower (yielding larger critical bands) than, that of monaural hearing (Yama and Small, 1983; Hall *et al.*,

1983; Kohlrausch, 1988; Kollmeier and Holube, 1992; Langhans and Kohlrausch, 1992; van der Hejden and Trahiotis, 1998; Holube *et al.*, 1998; Breebaart *et al.*, 1999). However, masker energy from a considerably larger bandwidth has been found to affect the BMLD in some experiments with no spectrally flanking noise (Sever and Small, 1979; Hall *et al.*, 1983; Cokely and Hall, 1991). Several possible reasons for this discrepancy, including the dependence of the temporal properties of the stimuli on the bandwidth and asymmetry of the left and right auditory filters, have been discussed by van der Hejden and Trahiotis (1998). In conclusion, it appears that the frequency resolution of the binaural hearing reflects the resolution of the auditory periphery, although in some cases there may be interference from a somewhat larger frequency range.

3.4.2 Time resolution

Determining the time resolution of the binaural hearing is a little more complicated than that of the frequency resolution. Blauert (1972) found that a human listener is capable of tracking in detail the spatial movements of sound sources corresponding to sinusoidal fluctuations of the ITD and ILD cues up to only 2.4 and 3.1 Hz, respectively. It has also been shown that increasing the duration of a signal (Tobias and Zerlin, 1959; Nuetzel and Hafter, 1976; Buell and Hafter, 1988) or the temporal separation between two signals at different directions (Perrott and Pacheco, 1989) up to a length of at most a few hundred milliseconds results in a improvement in discrimination of interaural differences[5]. Nevertheless, Grantham and Wightman (1978) observed that their listeners were able to detect ITD fluctuations up to 500 Hz, not based on movement but on perceptual widening of the sound sources (see also Pollack, 1978). Furthermore, the experiments of Grantham (1984a) provided some evidence that fluctuations in ILDs are detectable up to even higher rates than those of ITDs.

The detectability of high frequency fluctuations in the binaural cues does not necessarily mean that the analysis of the cues is done at an equally high resolution. It has been proposed that the temporal processing of the binaural cues resembles integration with a sliding window, which effectively indicates lowpass filtering of the fluctuations. Indeed, studies of the *minimum audible movement angle* (MAMA), defined as the smallest detectable angular movement of a sound source, have shown that the MAMA grows with increasing angular velocity (e.g., Perrott and Musicant, 1977; Perrott and Pacheco, 1989; Saberi and Perrott, 1990; Chandler and Grantham, 1992; Saberi and Hafter, 1997; see also Grantham, 1986). The time required for the detection at each angular velocity is determined by the ratio of the MAMA to the angular velocity[6]. By observing the MAMAs at small angular velocities, Chandler and Grantham (1992) found that the minimum time to achieve an optimal detection is approximately 300–400 ms for horizontal movement in front of the listener and increases on the sides. Furthermore, Saberi *et al.* (2003) investigated dynamic changes in ITD and reported

[5]Increasing the signal duration also facilitates binaural detection, but the effect appears similar also in monaural conditions (Green, 1966; Kohlrausch, 1990) and hence does not specifically reflect the binaural processing.

[6]This follows from the elementary physical relation $s = vt$, where s is the (angular) movement, v is the (angular) velocity, and t is the time.

corresponding integration times between 100–400 ms, with the detection performance decreasing again for very slow movements lasting several seconds.

Another class of related studies has investigated the effect of time-varying interaural configurations of a masker on BMLD (Grantham and Wightman, 1979; Yost, 1985; Kollmeier and Gilkey, 1990; Holube *et al.*, 1998; Culling and Summerfield, 1998; see also Culling and Colburn, 2000). Some of the applied methods have also enabled determination of the shape of the integration window. For the best fitting shape, Kollmeier and Gilkey (1990) proposed a double-sided, nearly symmetric exponential window, and Culling and Summerfield (1998) proposed a Gaussian window. The estimated lengths of the window in the studies cited above vary considerably between individual listeners and stimuli. However, Breebaart *et al.* (2002) showed that some of the observed stimulus-dependence can be accounted for with a detection model that allows off-time listening, i.e., detection at time instants when the peak of the integration window is not centered on the signal (see also Culling and Summerfield, 1998). The best predictions were achieved with an equivalent rectangular duration (ERD) of 120 ms. Although lower, this value does not contradict the previous results for time integration related to the MAMA, since the skirts of a double-sided exponential or Gaussian integration window extend considerably farther than the ERD, contributing thus to *optimal* performance.

In a third group of related studies, the detectability of brief changes in ITD, ILD, or IC has been directly investigated. For the detection of a burst of interaurally uncorrelated noise between interaurally correlated noise, Akeroyd and Summerfield (1999) observed a mean ERD of 140 ms. Furthermore, using a cross-correlation detection model (see Section 3.6.1), they showed that the ERD did not depend significantly on frequency. Nevertheless, Bernstein *et al.* (2001) reported considerably shorter integration times for brief changes in ITD and ILD. Detection of both the ITD and ILD was described well with a symmetric double-exponential integration window with average time constants of 0.09 and 13.8 ms. Furthermore, this integration window was able to account for some of the results of Grantham and Wightman (1978). In a related study (Akeroyd and Bernstein, 2001), the best predictions for sensitivity to changes in either ITD or ILD were achieved with an ERD of 10 ms combined with a weighting function representing a brief loss of binaural sensitivity just after the onset of a sound.

As an explanation for the observed differences in the integration times, it has been proposed that the different tasks may tap different properties of the binaural auditory system (e.g., Kollmeier and Gilkey, 1990; Akeroyd and Bernstein, 2001; Bernstein *et al.*, 2001). An alternative explanation could be based on the multiple looks model proposed to explain some temporal properties of monaural detection (Viemeister and Wakefield, 1991[7]; see also Dai and Wright, 1995, 1999; Buus, 1999). The idea is that the auditory system has a short term memory of "looks" at the signal, which can be accessed and processed selectively. Hence, information could be combined over different time segments according to the current task. Hofman and van Opstal (1998) also found evidence that localization in the elevation direction is based on spectral integration in short time windows with a duration of a few ms, and the obtained

[7]For another alternative explanation for some of the results of Viemeister and Wakefield (1991), see Dau *et al.* (1997b).

information can then be combined by another mechanism over a larger time range.

To the knowledge of the author, possible multiple looks processing of the binaural cues has not been studied. However, it should be noted that a long double-sided integration window is inconsistent with the precedence effect, where the localization of a lead sound is practically unaffected by a lag some milliseconds later. Consequently, binaural information from shorter time windows in the order of some milliseconds must be available for the auditory system.

3.5 Spatial perception of room responses

So far, the perceptual discussion has been limited to localization and binaural detection. In this section, the treatment is extended to spatial perception related to acoustical environments. As mentioned earlier, the precedence effect is not a full suppression of lagging sound. Even if the room reflections are typically not individually localizable, they affect (among a number of monaural attributes, see e.g., Beranek, 1996) the spatial and timbral perception as well as distance localization. This section starts with a discussion of timbre and detectability of single reflections in Section 3.5.1. Both topics will be important later in the development of the SIRR method. The review of spatial impression in 3.5.2 gives some further background. Furthermore, in order to complete the earlier description of localization, perception of distance is briefly discussed in Section 3.5.3.

3.5.1 Timbre and detectability of single reflections

Timbre in general is a complex quantity depending on spectral, temporal and spatial properties of sound (for spectral and temporal aspects, see e.g., Risset and Wessel, 1999). Related to room responses, the summation of the direct sound and each reflection can be considered as a comb filter resulting in perceptual coloration of the sound, i.e., in a change in the timbre. Considering the monaural coloration due to a single reflection, Kates (1985) was able to predict the detectability of the reflection with the help of an auditory model, effectively realizing a short-time spectral analysis. However, the perceived coloration due to a reflection and the detectability of the reflection are also affected by its direction such that a laterally displaced reflection relative to the direct sound yields less coloration and a higher detection threshold (Zurek, 1979). The apparent contradiction of the decreased detectability of a laterally displaced reflection with the earlier discussion on binaural masking level difference has been shown to be due to higher detectability of a single reflection based on its timbral than spatial effects (Olive and Toole, 1989; Bech, 1998).

Related to the detection of reflections in the presence of other reflections, the pioneering work by Seraphim (1961) showed that the detectability of an individual reflection depends on the levels and directions of the other reflections both before and after the reflection in question. Olive and Toole (1989) found that the detectability strongly depends on a source signal convolved with a room response. Furthermore, adding independent reverberation reduced the detectability of single reflections as well as their effect on detectability of other single reflections. Nevertheless, some first re-

flections in a room may at their natural levels individually contribute both to the overall timbre (Bech, 1995, 1996) and to spatial perception (Bech, 1998; see also Cox *et al.*, 1993; Okano, 2002). Due to the increasing density of the reflections with time (see Section 2.2.2), the late reflections, on the other hand, are no longer individually perceptible.

The binaural decoloration aspect has also been investigated for multiple reflections. Using simulated room environments, Brüggen (2001) showed that one factor of the multidimensional coloration was related to the magnitude spectrum of sound. Furthermore, for binaural stimuli, the perceived coloration was often equal to the spectrally measured coloration in the less colored ear signal. In concert halls, the timbre has also been found to depend on the relative reverberation times at different frequencies (e.g., Gade, 1989; Beranek, 1996).

3.5.2 Spatial impression

Although only some first reflections in a room are individually detectable, the later reflections have a considerable collective effect on the spatial perception. Historically, the reverberation time has been considered as the most important characteristic of a room or a concert hall. Marshall (1967) was the first to explicitly point out that the directional distribution of room reflections affects what he called "spatial responsiveness" and will be denoted as spatial impression in this thesis. Blauert and Lindemann (1986a) later showed that spatial impression is a multidimensional perceptual attribute affected differently by early reflections and late reverberation. Nowadays, spatial impression in concert halls is usually divided into the *auditory source width* (ASW)[8] and the *listener envelopment* (LEV). The ASW describes the spatial effect of the acoustical environment that is associated with the auditory event corresponding to the sound source. On the other hand, the LEV is related to the perception of the environment itself (Beranek, 1996; Marshall and Barron, 2001; Rumsey, 2002).

Related perceptual attributes have also been searched for in studies on reproduction of spatial sound (e.g., Berg and Rumsey, 1999a,b, 2000a,b, 2001, 2002; Koivuniemi and Zacharov, 2001; Zacharov and Koivuniemi, 2001a,b,c). Berg and Rumsey (1999a,b, 2000b) clearly found the attributes ASW and LEV. Koivuniemi and Zacharov (2001), on the other hand, reported the attributes "sense of direction," "broadness," and "sense of space." A number of other spatial attributes, such as "sense of depth" and "room perception," were also identified. However, to the knowledge of the author, the dependence of the other attributes on properties of room responses or interaural cues has not been studied. Furthermore, the individual dimensions of spatial perception will not be investigated later in this thesis. Hence, the current discussion will be limited to the established ASW and LEV, which can be described in the context of room responses. Some of the other attributes will be briefly discussed in Section 5.2.2.

Auditory source width

The early work on spatial impression concentrated mainly on the effect of early lateral reflections on what has been later identified as ASW (e.g. Barron, 1971; Barron and

[8]The term *apparent source width* is also commonly used for ASW.

Marshall, 1981). Barron (1971) already discussed both later established measures of spatial impression: The *lateral energy fraction* (LF) is defined as the ratio of lateral energy to the total energy in a RIR. Furthermore, the *interaural cross-correlation* (IACC) is the IC computed from a BRIR[9]. For the evaluation of ASW, both the LF and IACC are computed over the early reflections, which, according to a standard definition, consist of the first 80 ms of a RIR starting from the direct sound (ISO 3382, 1997; for studies on the time limit, see Hidaka *et al.*, 1995). Although lateral energy naturally reduces the IACC, the two measures may yield somewhat different results. Especially at high frequencies, the IACC is more sensitive than LF to reflections that are not fully lateral but deviate somewhat from the median plane. Furthermore, this property appears to correlate with perception of ASW (Ando and Kurihara, 1986; Singh *et al.*, 1994; Okano *et al.*, 1998; see also Morimoto *et al.*, 1993, 1994; Bradley, 1994).

The frequency dependence of the ASW either related to early reflections or to the IACC of noise bursts has been investigated by several authors (e.g. Barron and Marshall, 1981; Blauert and Lindemann, 1986a,b; Blauert *et al.*, 1986; Morimoto and Maekawa, 1988; Hidaka *et al.*, 1995; Morimoto *et al.*, 1995; Potter *et al.*, 1995; Ueda and Morimoto, 1995; Okano *et al.*, 1998; Mason *et al.*, 2005). It has been shown that low-frequency (< 500 Hz) decorrelation between the ear input signals creates a high ASW. However, in natural sound fields, the IACC will always be high at low frequencies due to the small distance between the ears related to the wavelength of sound (Lindevald and Benade, 1986; see also the coherence between two microphones with a fixed distance in Section 2.6.2). Moreover, consistent with earlier discussion, Mason *et al.* (2005) showed that at high frequencies, the ASW is related to decorrelation of the envelopes instead of the waveforms of the ear input signals. As will be argued in Section 4.2.2, the envelopes are typically more correlated than waveforms, which reduces the range of IACC at high frequencies. Based on these considerations, it is not surprising that Hidaka *et al.* (1995) and Okano *et al.* (1998) found the ASW in natural sound fields to correlate best with the average of early IACCs measured at the octave bands centered at 500, 1000, and 2000 Hz.

Listener envelopment

The listener envelopment has for some reason attracted less attention than the ASW. As opposed to the ASW, the LEV is affected mainly by the late reverberation part of a RIR (Morimoto and Maekawa, 1989; Bradley and Soulodre, 1995a; Morimoto *et al.*, 2001; Soulodre *et al.*, 2003). Morimoto and Maekawa (1989) showed that the late IACC computed over the time from 80 ms to the end of a RIR[10] corresponds well to the perceived LEV. However, Bradley and Soulodre (1995a) found that rather than the late lateral energy fraction (which is related to the late IACC), the level of the late lateral energy proportional to the direct sound determined the LEV (see also Bradley

[9]Here the abbreviation IACC will be used to indicate that the IC is considered in the context of spatial impression. Note that a high degree of spatial impression correspond to high values of LF but low values of IACC.

[10]In practice, the end of a RIR is determined as the time instant where the background noise begins to dominate the measurement. The computation of the late IACC is also often limited to 1 s.

and Soulodre, 1995b). Soulodre *et al.* (2003) recently proposed a revised predictor for the LEV, including a higher relative weighting of the directional properties of the late reverberation than in the late lateral energy measure and frequency-dependent time integration limits.

Some studies have also found evidence that not only lateral reflections but also reflections from behind and above the listener may affect the LEV (Morimoto *et al.*, 2001; Furuya *et al.*, 2001; Wakuda *et al.*, 2003). However, Evjen *et al.* (2001) were not able to verify the effect of non-lateral sound energy on LEV (see also Barron, 2001; Hanyu and Kimura, 2001). Regarding the frequency dependence of LEV, Bradley and Soulodre (1995b) found the perceived envelopment to correspond best to the late lateral energy averaged over the 250, 500, and 1000 Hz octave bands (see also Soulodre *et al.*, 2003).

Discussion

Although successful in predicting the ASW and LEV in many listening tests, the current measures have several problems. Based on the earlier discussion on the binaural hearing, it is evident that the averaging times typically used to compute the LF and IACC are higher than the averaging done by the auditory system. Furthermore, when listening to a sound source in a room, the sound reaching the ears of a listener is, of course, the convolution of the source signal and the room response. Indeed, considerable source signal dependent differences in the spatial impression within a single room have been reported by Becker and Sapp (2001), Merimaa and Hess (2004), Merimaa *et al.* (2005a), and Lokki (2005). Moreover, the ASW and LEV have been found to to depend on the size of a room or a hall in a manner not predicted by the IACC (Merimaa and Hess, 2004; Merimaa *et al.*, 2005a).

The measures of ASW have also been shown to vary considerably between individual measurement positions in a hall even if the perceived ASW does not vary correspondingly (de Vries *et al.*, 2001; see also Pelorson *et al.*, 1992). As alternative measures, Griesinger (1992, 1997, 1999) and Mason *et al.* (2001a,b,c,d) have proposed measurement of the fluctuations of the ITD and/or ILD cues produced by the room reflections. However, in a comparison of a number of measures related to stimulus-specific ASW, the best predictions were achieved using IACC computed over segments of source activity (Mason and Rumsey, 2002). Note that these two measures are also related, since computing IACC over sound with fluctuating binaural cues yields low values.

3.5.3 Distance

In Section 3.2.1, it was already mentioned that in free-field the changes in the localization cues as a function of distance for sound sources beyond approximately 2–3 m are perceptually negligible. The perception of intermediate distances (3–15 m) is mainly based on loudness. At distances greater than that, the frequency-dependent absorption of the air (see Section 2.2.1) also creates spectral distance cues. However, the utilization of both loudness and spectral cues requires familiarity with or making some assumptions about the source signal. Consequently, both at intermediate and

large distances, large discrepancies between the distance of the auditory event and the sound source may occur for unfamiliar or amplified sounds (e.g. Coleman, 1962, 1968; Strybel and Perrott, 1984; Blauert, 1997, pp. 45–47, 118–127).

In room environments, the reverberation provides another distance cue. Since the direct sound is attenuated proportionally to the square of distance but the amount of reverberant energy varies much less within a room (see Barron and Lee, 1988; Barron, 1995a), the ratio of the direct to reverberant energy can be used for evaluating the distance. Indeed, both of the direct-to-reverberant ratio and loudness seem to be used in auditory distance localization within rooms (Zahorik, 2002a,b; see also Nielsen, 1993; Mershon, 1997; Bronkhorst and Houtgast, 1999). To the knowledge of the author, the possible effect of the directionality of room responses on the perception of distance has not been studied in detail. In an anechoic enviroment, Kurozumi and Ohgushi (1983) found that the cross-correlation of two noise signals emitted from two loudspeakers affected distance judgment. However, the utilized stimuli were highly unnatural and some of them may have resulted in in-head localization, which could explain the considerable changes in distance.

3.6 Binaural models

Several auditory models have been proposed to explain various aspects of human spatial perception. Modeling the localization based on the monaural cues is complicated by the individual differences, although some models have been proposed in the literature (e.g., Middlebrooks, 1992; Zakarauskas and Cynader, 1993; Hofman and van Opstal, 1998; Langendijk and Bronkhorst, 2002). This section is limited to the processing of binaural cues. Reviews of binaural auditory models have been given by Colburn and Durlach (1978); Colburn (1996), Blauert (1997, Sections 4.4.4 and 5.3), Stern and Trahiotis (1997), and Braasch (2005). Current models usually include a simulation of the auditory periphery as described in Section 3.2, and the processing within the models described in this section is assumed to follow such peripheral simulation.

Binaural auditory models are often divided into physiologically- or psychologically-oriented approaches, with the former aiming at simulating the behavior of neural units measured at the higher stages of the auditory system, whereas the latter work on a more abstract phenomenological basis. Anatomically, the auditory nerve fibers are connected to the left and right cochlear nuclei, which further project (among other nuclei) to the superior olivary complexes (SOCs). The lateral superior olive (LSO) and the medial superior olive (MSO) in the left and right SOCs are the first stages receiving input from both ears and thus capable of dealing with binaural information. The MSO receives excitatory input from both the ipsilateral and contralateral side, whereas the main contralateral input to the LSO is inhibitory. A disproportionately large area in the MSO is devoted to low frequencies, whereas the LSO deals predominantly with high frequencies. For this reason (according to the duplex theory of binaural localization), it has been traditionally considered that the MSO is a center for processing ITDs and LSO for ILDs (Pickles, 1988, Chapter 6; Kuwada *et al.*, 1997). Nevertheless, recent studies have found elements sensitive to ITDs also in the LSO

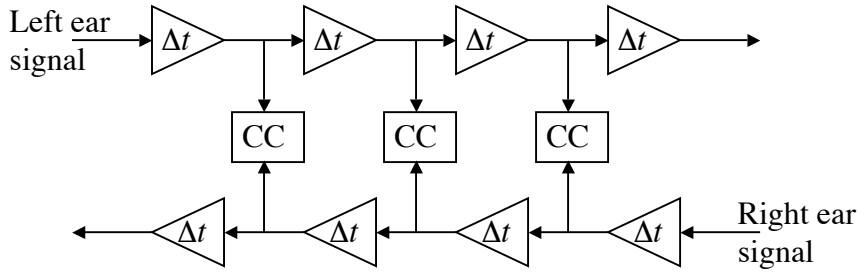

Figure 3.8: Illustration of a Jeffress (1948) coincidence structure for ITD extraction. The $\Delta\tau$ blocks are delays and the CC blocks depict coincidence detectors. In a cross-correlation model, the coincidence detectors would consist of multiplication and time integration.

(Joris and Yin, 1995) and inhibitory inputs to MSO (e.g., Brand *et al.*, 2002). Hence, the mutual roles of the LSO and MSO are not fully clear.

Most binaural models compute ITDs and ILDs separately, even if the recent results suggest that the ITD and ILD pathways are not fully separated. ILDs are often computed directly from the ratio of the signal powers of the left and right neural channels (Braasch, 2005), whereas extraction of ITD cues requires more involved processing. In the following, two types of binaural models will be discussed: The cross-correlation models (Section 3.6.1) are traditionally considered to resemble processing in the MSO and the excitation–inhibition (EI) models (Section 3.6.2) in the LSO. Pure cross-correlation models can be seen as excitation–excitation (EE) type, but the discussion will also include a model involving inhibition. The performance of the reviewed models in predicting localization in complex listening situations is discussed in Section 3.6.3.

3.6.1 Cross-correlation models

The majority of ITD extraction models are based on a coincidence structure proposed by Jeffress (1948) (see also Colburn, 1977; Stern and Colburn, 1978), or a cross-correlation implementation (e.g., Cherry and Sayers, 1956; Sayers and Cherry, 1957; Blauert and Cobben, 1978; Lindemann, 1986a,b; Stern *et al.*, 1988; Shackleton *et al.*, 1992; Stern and Shear, 1996; Trahiotis *et al.*, 2001) that can be seen as a special case of the coincidence structure. Such models include delay lines for the left and right neural signals, as shown in Figure 3.8, and the coincidence detectors produce a strong response at the delays where the phases of the left and right input signals match. Both inputs are thus excitatory (EE type processing), and the ITD is converted into a position of a peak along the delay line. Depending on the purposes of the modeling, the response may be averaged over the whole duration of the stimulus or within time windows corresponding to the time resolution of the binaural hearing (see Section 3.4.2). A detailed mathematical implementation of a cross-correlation model will be described later in Section 4.2.2.

In localization or lateralization studies of wideband stimuli, cross-correlation models typically compute a cross-correlogram displaying the cross-correlation functions at multiple critical bands. The information from several bands is most often combined

by integrating the cross-correlogram over frequency or by averaging the peak locations at different frequency bands. Stern *et al.* (1988) also used a psychoacoustically derived weighting function with a dominant region around 600 Hz to describe the salience of different frequency bands (for plots of other proposed frequency weighting functions, see Braasch, 2005, p. 92). Delay-weighting has also been proposed to take into account the assumed distribution of the coincidence detectors at different interaural delays (Colburn, 1977; Shackleton *et al.*, 1992; Stern and Shear, 1996).

Due to the periodic nature of the cross-correlation (or coincidence) operation, the peak structure reflects the period of narrowband input signals (see simulation results in Chapter 4). In order to resolve the ambiguity due to the periodicity, Stern *et al.* (1988) proposed weighting the cross-correlation peaks according to their straightness across frequency bands (for related perceptual studies, see Trahiotis and Stern, 1989; Buell *et al.*, 1994; Trahiotis *et al.*, 2001). Somewhat similar results were also achieved with the model of Shamma *et al.* (1989), which is not a cross-correlation model but yields patterns resembling cross-correlograms. Instead of using neural delays, the model relies on the frequency-dependent delays caused by the cochlear transduction and coincidence computation across frequency bands. The model of Shamma *et al.* (1989) was also shown to be sensitive to ILDs.

Some cross-correlation models have also combined evaluation of ILDs into the same model structure. For this purpose, Stern and Colburn (1978) used yet another delay weighting function with a Gaussian shape and the center of the weighting function on the ITD axis determined by the ILD according to time-intensity trading ratios. Taking a different approach, Lindemann (1986a) introduced contralateral inhibition into a cross-correlation model. The stationary inhibition component of the model makes the peak of the cross-correlation function shift towards the input with higher level. Additionally, Lindemann's model includes monaural processors that are involved in ILD processing, as well as a dynamic inhibition that will be discussed in Section 3.6.3. aqNevertheless, neither the model of Stern and Colburn (1978) nor the model of Lindemann (1986a) can explain the stimulus-dependence in the weighting of conflicting ITDs and ILDs (see Section 3.3.1). Gaik (1993) took a first step towards involving information on possible cue conflicts in an extension the model of Lindemann such that each of the inhibited coincidence detectors is tuned to a natural combination of ITDs and ILDs. With this extension, the model yields one peak for non-conflicting binaural cues and two peaks for an unnatural combination of the ITD and ILD. However, to the knowledge of the author, the relation of the strength of such two peaks to the psychophysical weighting of conflicting ITDs and ILDs has not been investigated.

Apart from localization and lateralization, cross-correlation models have also been used to study spatial impression based on temporal fluctuations of the binaural cues (e.g., Blauert, 1997; Bodden, 1998; Hess, 2006), and for predicting binaural detection. In detection studies, the presence of a signal with a different ITD compared to the masker lowers the peak of the cross-correlation function (unless the signal and the masker are both sinusoidals with the same frequency). It has been suggested that this is what the auditory system detects first. In cross-correlation-based detection models, the cross-correlation function is usually normalized by the signal power such that fully correlated signals in both ears give a maximum peak of one irrespective of their ITD and ILD (e.g., Osman, 1971, 1973; Bernstein and Trahiotis, 1996; Bernstein

et al., 1999; Akeroyd and Summerfield, 1999). In this thesis, the maximum value of the normalized cross-correlation function is denoted as *interaural coherence* (IC), and it will play an important role in the modeling studies in Chapter 4. Recall also that introducing physiologically motivated basilar membrane compression and neural transduction stages (see p. 52) in the model of Bernstein *et al.* (1999) enabled the use of similar IC variation thresholds to predict detection for various stimulus configurations[11].

Although evidence for processing that resembles cross-correlation has been found in the MSO (including mammalian species, see Yin and Chan, 1990; Yin *et al.*, 1997; Kuwada *et al.*, 1997), the physiological feasibility of the cross-correlation models has been recently questioned (McAlpine *et al.*, 2001; Brand *et al.*, 2002). Based on the new evidence, Grothe (2003) argued that most projection patterns in the mammalian MSO do not fit the concept of delay lines. Furthermore, van de Par *et al.* (2001) have reasoned that the precision needed for normalization of the cross-correlation function is so high that it is unlikely that the auditory system is performing the normalization *per se*. Nevertheless, cross-correlation models have been successful in extracting information that corresponds to human binaural perception and can thus serve at least as psychologically-oriented approaches.

3.6.2 Excitation–inhibition models

An alternative (or physiologically, perhaps additional) mechanism for extraction of binaural cues is based on an EI scheme. For ILD estimation, Reed and Blum (1990) proposed a model motivated by the physiological operation of the LSO. In the proposed model, the strengths of ipsilateral excitation and contralateral inhibition vary in opposite ways along an isofrequency slab of LSO, and the ILD is converted to a position where the excitation is canceled by the inhibition. Motivated by the finding of ITD sensitive elements in the LSO, Breebaart *et al.* (2001a) combined the ILD model of Reed and Blum and the delay line approach of Jeffress such that each tap of the delay lines is connected to a chain of attenuators. The resulting ILD processing is illustrated in Figure 3.9, and in the model of Breebaart *et al.*, this ILD structure replaces the coincidence detectors shown earlier in Figure 3.8.

Instead of the one-dimensional cross-correlation function per each critical band, the model of Breebaart *et al.* (2001a) yields a two-dimensional representation with a sound event inducing a local minimum at the positions of the corresponding ITD and ILD. Furthermore, reduction of IC results in an increase in the activity at the positions of the minima. Although the model was designed for and tested in binaural detection (Breebaart *et al.*, 2001a,b,c), it can be readily used for localization (see Section 3.6.3). The model also resembles the equalization and cancellation (EC) theory of BMLD (Durlach, 1963; see also Durlach, 1966). In EC, the signal in one ear is first equalized

[11] Specifically, the compression was necessary to predict detection in the presence of masking noises with different envelope structures (Bernstein *et al.*, 1999). Interestingly, Breebaart *et al.* (1999) were also unable to model the detection of a sinusoidal signal in the presence of different multiplied noise maskers with normalized cross-correlation. Since the variations in the masker induced considerable changes in the masker envelope, it is possible that introducing the peripheral stages of Bernstein *et al.* would have also enabled modeling the stimuli of Breebaart *et al.* (1999).

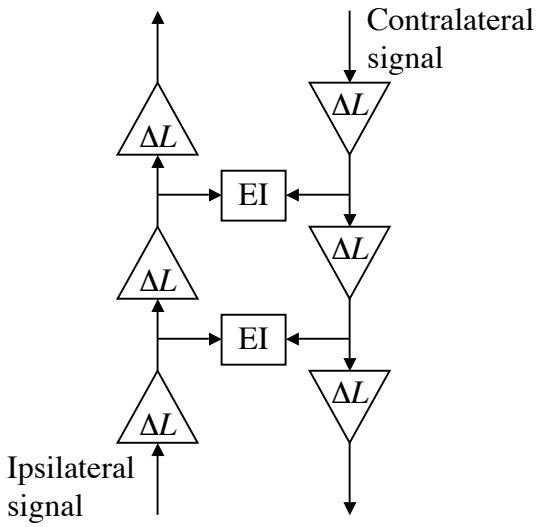

Figure 3.9: Functional view of the ILD extraction model of Reed and Blum (1990) as implemented by Breebaart *et al.* (2001a). The ΔL blocks are attenuators and the EI blocks perform subtraction of the contralateral signal from the ipsilateral signal followed by half-wave rectification.

relative to the other ear using a delay and a gain such that the masking components are exactly the same in both signals. Subsequently, the masker is cancelled by a subtraction. The model of Breebaart *et al.* (2001a) effectively performs the same process for all possible delays and level differences within the range being considered. In both models, internal noise is also added to the signals in order to limit the accuracy of the equalization process.

3.6.3 Localization in complex listening situations

Few binaural modeling studies have specifically considered localization in complex listening situations. To begin with, Blauert and Cobben (1978) concluded that for a precedence effect scenario, the correct cross-correlation peaks were available but the basic cross-correlation model could not explain how to identify them. However, the model of Lindemann (1986a) was shown to be able to simulate several precedence effect phenomena (Lindemann, 1986b) with the help of its dynamic inhibition. Effectively, the dynamic inhibition tends to hold a cross-correlation peak (and suppress new peaks) as long as the binaural cues of the input signals contribute to the current peak. When the contribution ends or is weakened, the inhibition fades away with a time constant of 10 ms.

A different phenomenological model for the precedence effect was proposed by Zurek (1987). The model uses a temporary localization inhibition triggered by an onset detector[12]. A cross-correlation implementation of the model was later developed by Martin (1997). Furthermore, Djelani and Blauert (2002) demonstrated promising results for modeling the direction-specific buildup of precedence by using onset detectors

[12]Note the similarity of this inhibition to the loss of binaural sensitivity after the onset of sound in the time integration window of Akeroyd and Bernstein (2001) (see Section 3.4.2).

to control the strength of the dynamic inhibition in the model of Lindemann (1986a).

Recently, Hartung and Trahiotis (2001) were also able to simulate the precedence effect for pairs of clicks without any inhibition, just taking into account the properties of the peripheral hearing. However, their model was not able to predict the localization of continuous narrowband noises in a comparison of several models by Braasch and Blauert (2003). The best results were achieved with a combined analysis of ITD cues with the model of Lindemann (1986a) and ILD cues using the model of Breebaart *et al.* (2001a) extended with temporal inhibition. For independent localization of concurrent sources with non-simultaneous onsets, Braasch (2002) has also proposed a cross-correlation difference model.

3.7 Summary

In this chapter, the operation of the auditory system and the perception of spatial sound, as well as related auditory models, were described. It was established that the signals arriving at the ears of a listener are analyzed within frequency bands. Furthermore, the two most important determinants of spatial auditory perception are the interaural time difference (ITD) and interaural level difference (ILD) cues. These cues are extracted at low frequencies from the waveforms and at high frequencies from the envelopes of the ear input signals and analyzed with a limited temporal resolution.

Human localization utilizes the ITDs, ILDs, and spectral localization cues. For a first approximation, the ITDs and ILDs determine a cone of confusion, and the spectral cues are used for localization within the cones of confusion. The relative perceptual weighting of the individual localization cues as a function of frequency depends on the stimuli. However, ITDs are typically more important at low frequencies and ILDs at high frequencies. The human localization accuracy is also fairly good even in the presence of concurrent sound and room reflections. Nevertheless, models of binaural localization have difficulties in predicting human localization in such complex listening situations.

Even if room reflections are suppressed in localization of sound sources, they have an important effect on the spatial and timbral perception. Some of the first reflections may individually contribute both to the timbre and the spatial impression. However, most reflections are not individually detectable but instead collectively determine the perception. Related to the spatial impression, the early reflections contribute mainly to the auditory source width and late reflections to the listener envelopment. Both attributes have been traditionally measured either by the fraction of lateral energy or using the interaural cross-correlation of the early and late parts of a room response.

Chapter 4

Binaural Cue Selection Model

4.1 Introduction

As discussed in Section 1.1.2, auditory models are a part of basic research on auditory perception, and they can also be applied in audio technology. Although the current modeling study was initially inspired by the experiences with the Spatial Impulse Response Rendering method (see Chapter 5), it also serves as further proof for the need to reproduce the chosen binaural cues. As also described earlier, in everyday complex listening situations, sound from multiple sources, as well as reflected, scattered, diffracted sound from the physical surroundings, arrives subsequently and/or concurrently from different directions at the ears of a listener. Based on the literature reviewed in Section 3.3, the human auditory system is very good at resolving such composites of sound into separate localizable auditory events at directions corresponding to the sources, while suppressing the localization of reflections. However, existing binaural localization models have difficulties in predicting the localization in complex listening situations (see Section 3.6.3).

In this chapter, a single modeling mechanism to explain various aspects of auditory localization in complex listening situations is proposed. The basic approach is very straightforward: Only ITD and ILD cues occurring at time instants when they represent the direction of one of the sources are selected, while other cues are ignored. It will be shown that the interaural coherence (IC) can be used as an indicator for these time instants. More specifically, by selecting ITD and ILD cues coinciding with IC cues larger than a certain threshold, one can, in many cases, obtain a subset of ITD and ILD cues similar to the corresponding cues of each source presented separately in free-field.

The proposed cue selection method is implemented within the framework of a model with a physiologically motivated peripheral stage, whereas the remaining parts are analytically motivated. As discussed in Section 3.3, the frequency-dependent weighting of the binaural cues is stimulus-dependent and may change in the presence of concurrent sound. Furthermore, a general model for predicting these changes does not exist. For these reasons, the treatment in this chapter will be limited, apart from some discussion, to the ITDs and ILDs occurring at individual critical bands without considering their exact role in the final localization judgement. The presented simu-

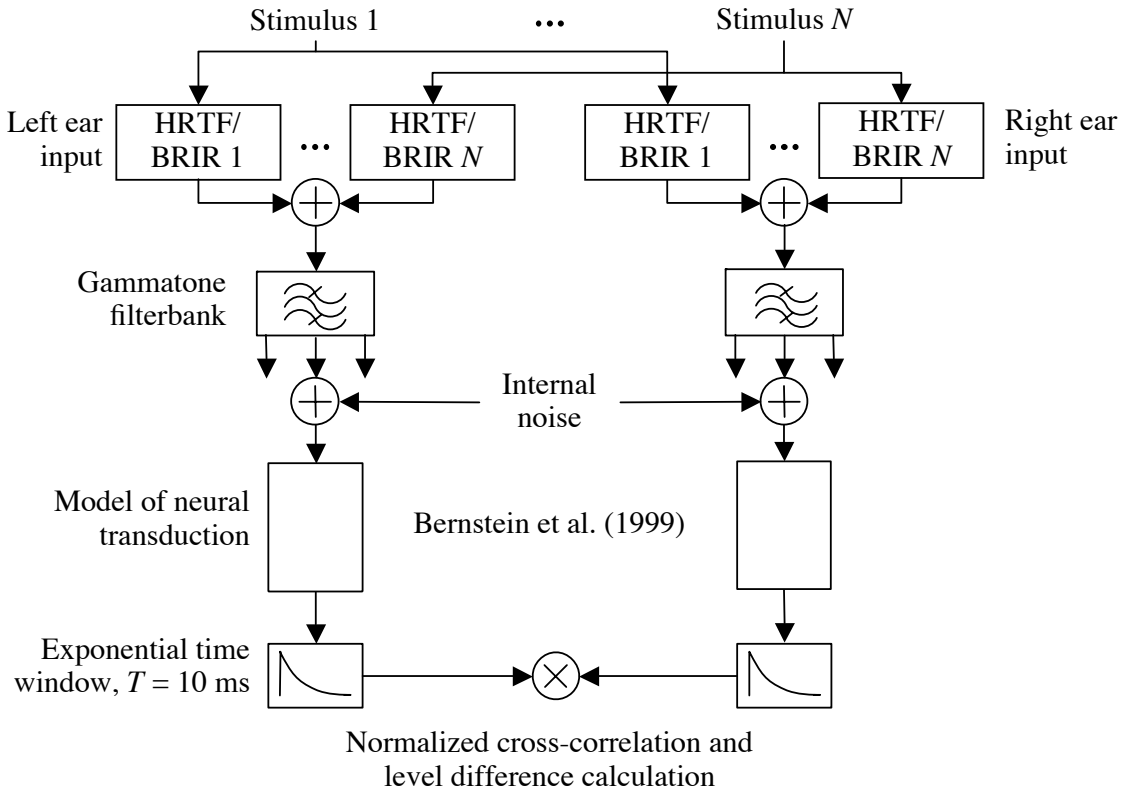

Figure 4.1: Structure of the peripheral part and binaural processor of the cue selection model. For an illustration of the neural transduction part, see Figure 3.7.

lation results, nevertheless, reflect psychophysical data from a number of localization experiments reviewed earlier, involving both independent distracters and precedence effect conditions.

This chapter follows closely the paper of Faller and Merimaa (2004) with some extended discussion. The chapter is organized as follows: The binaural model, including the proposed cue selection mechanism, is described in Section 4.2. The simulation results are presented in Section 4.3, with a short discussion of each case related to similar psychophysical studies. Section 4.4 includes a general discussion of the model and the presented results, followed by conclusions in Section 4.5.

4.2 Model description

The model builds upon existing auditory models, adding a cue selection mechanism for explaining the localization in complex listening situations. The overall structure of the model is similar to the organization of a generic binaural model shown earlier in Figure 3.1. The peripheral part and the binaural processor are illustrated in more detail in Figure 4.1, and will be described in Sections 4.2.1 and 4.2.2. The novel cue selection mechanism will be introduced in Section 4.2.3, followed by a discussion of the features of the model in Section 4.2.4.

4.2.1 Auditory periphery

The peripheral part of the model consists of components described earlier in Section 3.2. The transduction of sound from a source to the ears of a listener is realized by filtering the source signals either with HRTFs for simulations of anechoic listening conditions (Sections 4.3.1 and 4.3.2) or with measured BRIRs for studies in reverberant environments (Section 4.3.3). In multi-source scenarios, each source signal is first filtered with a pair of HRTFs or BRIRs corresponding to the simulated location of the source, and the resulting ear input signals are summed before the next processing stage. Since this study is limited to considering simulations at single critical bands, the frequency weighting effect of the subsequent middle ear stage has been discarded in the model.

Apart from the sinusoidal stimuli used in one case, all other simulations involve broadband stimuli, and the studied phenomena are not critically dependent on the absolute level. Consequently, the linear gammatone filterbank (see Section 3.2.3) has been deemed accurate enough for simulating the frequency analysis performed by the basilar membrane. Furthermore, for the sinusoidal stimuli, the shape of the filters does not affect the results for an auditory filter centered at the frequency of the sinusoidal. After passing the left and right ear signals through the gammatone filterbank, each resulting critical band signal is further processed using the neural transduction model of Bernstein *et al.* (1999) (see p. 52). The resulting nerve firing densities at the corresponding left and right ear critical bands are denoted by x_1 and x_2. These parts of the model are implemented using the freely available Matlab toolboxes from Slaney (1998) and Akeroyd (2001).

Internal noise is introduced into the model to account for the spontaneous activity of the auditory nerve and the resulting limited accuracy of the auditory system. For this purpose, independent Gaussian noise, filtered with the same gammatone filters[1] as the considered critical band signals, is added to each critical band signal before applying the model of neural transduction. The noise is statistically independent for each critical band, as well as for the left and right ears. For the critical band centered at 2 kHz, a sound pressure level (SPL) of 9.4 dB has been chosen according to Breebaart *et al.* (2001a), who fitted the level of the noise to describe detection performance near the threshold of hearing. For other critical bands, the level is scaled according to the hearing threshold curves[2] (ISO 389, 1975). For the 500 Hz band, an SPL of 14.2 dB is used.

4.2.2 Binaural processor

As discussed in Section 3.6, the exact physiological operation of binaural hearing is not known, and several different neural structures may be responsible for localization. The present study does not make any specific physiological assumptions about the binaural

[1]Note that using the same gammatone filter for the internal noise does not give firing densities proportional to the spontaneous rate, but is based on the assumption that the internal noise has a similar effect as an external masking noise which would go through the filters.

[2]This choice of levels of the internal noise is not necessarily proportional only to the *spontaneous* activity of the auditory nerve, but also may reflect activity due to noise caused by other physiological systems, such as the cardiac cycle (Soderquist and Lindsey, 1972).

processor. The only assumption made is that the processor extracts information that can be used by the upper stages of the auditory system for discriminating the ITD, ILD, and IC. Given this assumption, the proposed model computes the ITD, ILD, and IC directly. However, the ITD, ILD, and IC are defined with respect to critical band signals after applying the neural transduction, which often changes them relative to the physical values measured from the ear input signals.

The ITD and IC are estimated from the normalized cross-correlation function. Given x_1 and x_2 for a specific center frequency f_c, at the index of each sample n for a lag m (in samples), a running normalized cross-correlation (coherence) function is computed according to

$$\gamma(n,m) = \frac{a_{12}(n,m)}{\sqrt{a_{11}(n,m)a_{22}(n,m)}}\,, \tag{4.1}$$

where

$$\begin{aligned}
a_{12}(n,m) &= \alpha x_1(n-\max\{m,0\})x_2(n-\max\{-m,0\}) + (1-\alpha)a_{12}(n-1,m)\,,\\
a_{11}(n,m) &= \alpha x_1(n-\max\{m,0\})x_1(n-\max\{m,0\}) + (1-\alpha)a_{11}(n-1,m)\,,\\
a_{22}(n,m) &= \alpha x_2(n-\max\{-m,0\})x_2(n-\max\{-m,0\}) + (1-\alpha)a_{22}(n-1,m)\,.
\end{aligned}$$

The constant $\alpha \in [0,1]$ is the forgetting factor determining the time-constant of the exponentially decaying estimation window

$$T = \frac{1}{\alpha f_s}\,, \tag{4.2}$$

where f_s denotes the sampling frequency. $\gamma(n,m)$ is evaluated over time lags in the range of $m/f_s \in [-1,1]$ ms. The ITD (in samples) is estimated as the lag m of the maximum of the normalized cross-correlation function,

$$\tau(n) = \arg\max_m \gamma(n,m)\,. \tag{4.3}$$

Note that the time resolution of the computed ITD is limited by the sampling interval.

The normalization of the cross-correlation function is introduced to obtain an estimate of the IC, defined as the maximum value of the instantaneous normalized cross-correlation function,

$$c_{12}(n) = \max_m \gamma(n,m)\,. \tag{4.4}$$

This estimate describes the coherence of the left and right ear input signals. Due to x_1 and x_2 being half-wave rectified, it has in principle a range of $[0,1]$, where 1 occurs for perfectly coherent x_1 and x_2. However, due to the DC offset of the half-wave rectified signals, the values of c_{12} are typically higher than 0 even for independent (non-zero) x_1 and x_2. Thus, the effective range of the interaural coherence c_{12} is compressed from $[0,1]$ to $[k,1]$. The compression is more pronounced (larger k) at high frequencies, where the lowpass filtering of the half-wave rectified critical band signals yields signal envelopes with a higher DC offset than in the signal waveforms (Bernstein and Trahiotis, 1996; see also van de Par and Kohlrausch, 1995).

The ILD is computed as

$$\Delta L(n) = 10 \log_{10} \left(\frac{L_2(n, \tau(n))}{L_1(n, \tau(n))} \right) , \tag{4.5}$$

where

$$\begin{aligned} L_1(n,m) &= \alpha x_1^2(n - \max\{m, 0\}) + (1-\alpha) L_1(n-1, m) , \\ L_2(n,m) &= \alpha x_2^2(n - \max\{-m, 0\}) + (1-\alpha) L_2(n-1, m) . \end{aligned}$$

Note that due to the envelope compression in the peripheral model, the resulting ILD estimates will be smaller than the level differences between the ear input signals. For coherent ear input signals with a constant level difference, the estimated ILD (in dB) will be 0.46 times that of the physical signals.

The sum of the signal power of x_1 and x_2 that contributes to the estimated ITD, ILD, and IC cues at time index n is

$$p(n) = L_1(n, \tau(n)) + L_2(n, \tau(n)) . \tag{4.6}$$

As discussed in Section 3.4.2, choosing the time constant T is a difficult task since different experimental procedures may reflect different time integration. In this study, we have chosen to use a single-sided exponential time window with a time constant of 10 ms. This time constant is close to the smallest values found in the studies of the temporal resolution of the binaural hearing, and the time window is the same as that used in the temporal inhibition in the model of Lindemann (1986a).

4.2.3 Cue selection

The signals from the auditory periphery convey a vast amount of information to the higher stages of the auditory system. The focus of this study lies only in the analysis of the three inter-channel properties between left and right critical band signals defined in the previous section: ITD, ILD, and IC. It is assumed that at each time instant n, the information about the values of these three signal properties, $\{\Delta L(n), \tau(n), c_{12}(n)\}$, is available for further processing. Consider first the simple case of a single source in free-field. Whenever there is sufficient signal power, the source direction causally determines the nearly constant ITD and ILD, which appear between each left and right critical band signal with the same center frequency. In the following, the (average) ITDs and ILDs occurring in this scenario are called *free-field cues*. Furthermore, the free-field cues of a source with an azimuthal angle ϕ are denoted by τ_ϕ and ΔL_ϕ. It is assumed that this kind of one-source free-field scenario is the reference for the auditory system. In other words, in order for the auditory system to perceive auditory events at the directions of the sources, it must obtain ITD and/or ILD cues similar to the free-field cues corresponding to each source that is being discriminated. The most straightforward way to achieve this is to select the ITD and ILD cues at time instants when they are similar to the free-field cues. In the following, it is shown how this can be done with the help of the IC.

When several independent sources are concurrently active in free-field, the resulting cue triplets $\{\Delta L(n), \tau(n), c_{12}(n)\}$ can be classified into two groups: (1) Cues arising

at time instants when only one of the sources has non-negligible power in that critical band. These cues are similar to the free-field cues, i.e., the direction of the source is represented in $\{\Delta L(n), \tau(n)\}$, and $c_{12}(n) \approx 1$. (2) Cues arising when multiple sources have non-negligible power in a critical band. In such a case, the pair $\{\Delta L(n), \tau(n)\}$ does not represent the direction of any single source, unless the superposition of the source signals at the ears of the listener incidentally produces similar cues. Furthermore, when the two sources are assumed to be independent, the cues are fluctuating and $c_{12}(n) < 1$. These considerations motivate the following method for selecting the ITD and ILD cues. Given the set of all cue pairs, $\{\Delta L(n), \tau(n)\}$, we consider only the subset that occurs simultaneously with an IC larger than a certain threshold, $c_{12}(n) > c_0$. This subset is denoted by

$$\{\Delta L(n), \tau(n) | c_{12}(n) > c_0\}\,. \tag{4.7}$$

The same cue selection method is applicable for deriving the direction of a source while suppressing the directions of one or more reflections. When the "first wave front" arrives at the ears of a listener, the evoked ITD and ILD cues are similar to the free-field cues of the source, and $c_{12}(n) \approx 1$. Except for sinusoidal signals, as soon as the first reflection from a different direction arrives, the superposition of the source signal and the reflection results in cues that, most of the time, do not resemble the free-field cues of either the source or the reflection. At the same time, IC reduces to $c_{12}(n) < 1$, since the direct sound and the reflection superimpose as two signal pairs with different ITD and ILD. Thus, IC can be used as an indicator for whether ITD and ILD cues are similar to free-field cues of sources or not, while ignoring cues related to reflections.

For a given c_0, there are several factors determining how frequently $c_{12}(n) > c_0$. In addition to the number, strengths, and directions of the sound sources and room reflections, $c_{12}(n)$ depends on the specific source signals and on the critical band being analyzed. In many cases, the larger the c_0, the more similar the selected cues are to the free-field cues. However, there is a strong motivation to choose c_0 as small as possible while still getting accurate enough ITD and/or ILD cues, because this will lead to the cues being selected more often, and consequently to a larger proportion of the ear input signals contributing to the localization.

In this chapter, it is assumed that the auditory system adapts c_0 for each specific listening situation, i.e., for each scenario with a constant number of active sources at specific locations in a constant acoustical environment. Since the listening situations do not usually change very quickly, it is assumed that c_0 is adapted relatively slowly in time. In Section 4.3.2, it is also argued that such an adaptive process may be related to the buildup of the precedence effect. All simulations reported in this paper consider only one specific listening situation at a time. Therefore, for each simulation, a single constant c_0 is used.

4.2.4 Discussion

The physiological feasibility of the cue selection depends on the human sensitivity to changes in interaural correlation. The topic has been investigated by Pollack and Trittipoe (1959a,b), Gabriel and Colburn (1981), Grantham (1982), Koehnke *et al.*

(1986), Jain *et al.* (1991), Bernstein and Trahiotis (1997), Akeroyd and Summerfield (1999), Culling *et al.* (2001), and Boehnke *et al.* (2002). These investigations agree in that the sensitivity is highest for changes from full correlation, although the estimates of the corresponding just noticeable differences (JNDs) have a very large variance. For narrowband noise stimuli centered at 500 Hz, the reported JNDs of IC range from 0.0007 (Jain *et al.*, 1991, fringed condition) to 0.13 (Culling *et al.*, 2001) for different listeners and different stimulus conditions. The sensitivity has been generally found to be lower at higher frequencies. However, all of the cited studies have measured sensitivity to correlation of the ear input waveforms instead of correlation computed after applying a model of neural transduction. As discussed in Section 4.2.2, the model of Bernstein *et al.* (1999) reduces the range of IC, indicating overall lower JNDs of IC as defined in this chapter. Furthermore, the model has been specifically fitted to yield constant thresholds at different frequencies when applied to prediction of binaural detection based on changes in IC (see Section 3.6.1). With these considerations it can be concluded that at least the JNDs reported by Gabriel and Colburn (1981), Koehnke *et al.* (1986), and Jain *et al.* (1991) are within the range of precision needed for the simulations in Section 4.3.

The auditory system may not actually use a hard IC threshold for selecting or discarding binaural cues. Instead of pure selection, similar processing could be implemented as an IC-based weighting of ITD and ILD cues with a slightly smoother transition. It is interesting to note that such weighting of the ITDs would resemble the maximum likelihood estimator for the time difference used in physical microphone array localization (see Section 2.6.1)[3]. However, the simple selection criterion suffices to illustrate the potential of the proposed method. Note that the cue selection method also resembles a multiple looks approach (see Section 3.4.2), where the IC is used to determine which "looks" are considered in the localization judgment.

4.3 Simulation results

As mentioned earlier, it is assumed that in order to perceive an auditory event at a certain direction, the auditory system needs to obtain cues similar to the free-field cues corresponding to a source at that direction. In the following, the proposed cue selection is applied to several stimuli that have been used in previously published psychophysical studies. In all cases, both the selected cues as well as all cues prior to the selection are illustrated, and the implied directions are discussed in relation to the literature.

The effectiveness of the proposed cue selection is assessed using a number of statistical measures. The biases of the ITD and ILD cues with respect to the free-field cues τ_ϕ and ΔL_ϕ are defined as

$$\begin{aligned} b_\tau &= |\mathsf{E}\{\tau(n)\} - \tau_\phi| \\ b_{\Delta L} &= |\mathsf{E}\{\Delta L(n)\} - \Delta L_\phi| \,, \end{aligned} \tag{4.8}$$

[3]The cue selection method involving the possibility of IC-based weighting was originally developed without knowledge of this connection, which was pointed out by Professor Rainer Martin.

where $\mathsf{E}\{\cdot\}$ denotes expectation, and the corresponding standard deviations are given by

$$\begin{aligned} \sigma_\tau &= \sqrt{\mathsf{E}\{(\tau(n) - \mathsf{E}\{\tau(n)\})^2\}} \\ \sigma_{\Delta L} &= \sqrt{\mathsf{E}\{(\Delta L(n) - \mathsf{E}\{\Delta L(n)\})^2\}}\,. \end{aligned} \tag{4.9}$$

The biases and standard deviations are computed considering only the selected cues (Eq. 4.7). When there is more than one source to be discriminated, these measures are estimated separately for each source by grouping the selected cues at each time instant with the source known to have free-field cues closest to their current values.

For many cases, the larger the cue selection threshold c_0, the smaller the bias and standard deviation. The choice of c_0 is consequently a compromise between the similarity of the selected cues to the free-field cues and the proportion of the ear input signals contributing to the resulting localization. The proportion of the signals contributing to the localization is characterized with the fraction of power represented by the selected parts of the signals, given by

$$p_0 = \frac{\mathsf{E}\{p(n)w(n)\}}{\mathsf{E}\{p(n)\}}\,, \tag{4.10}$$

where $p(n)$ is defined in Eq. (4.6) and the weighting function $w(n)$ is

$$w(n) = \begin{cases} 1, & \text{if } c_{12}(n) > c_0 \\ 0, & \text{otherwise} \end{cases}\,. \tag{4.11}$$

As discussed in Section 3.3, the combination of conflicting ITD and ILD cues from the same frequency and over different frequencies is stimulus-dependent. In this chapter, no attempt will made to model the combination of the cues for the final localization judgment. Instead, the cue selection is considered independently at single critical bands displaying the extracted ITDs and ILDs separately. However, based on extensive simulations, the typical behavior of the cue selections appears to be fairly similar at different critical bands except for different values of c_0. For most simulations, the critical bands centered at 500 Hz and/or 2 kHz have been chosen as illustrative examples. At 500 Hz, the binaural processor operates on the input waveforms, whereas at 2 kHz the model of auditory periphery extracts the envelopes of the input signals and feeds them to the binaural processor (see Section 3.2.3). Where appropriate, results for other critical bands are also shown or briefly discussed. As mentioned earlier, the simulations are carried out with a single constant cue selection threshold c_0 for each case. It is assumed that the auditory system has already adapted c_0 to be effective for the specific listening situation. Unless otherwise noted, the specific c_0 was chosen such that a visual inspection of the simulation results implies an effective cue selection.

Two kinds of plots will be used to illustrate the cue selection. In some cases, the instantaneous ITD and ILD values are plotted as a function of time, marking the selected values. In other examples, the effect of the cue selection is visualized by plotting short-time estimates of *probability density functions* (PDFs) of the selected ITD and ILD cues. Unless otherwise noted, the PDFs are estimated by computing histograms

of ITD and ILD cues for a time span of 2 s[4]. The height of the maximum peak is normalized to one in all PDFs. In both types of plots, free-field cues resulting from simulations of the same source signals without concurrent sound sources or reflections, are also indicated[5].

Listening situations in free-field are simulated using HRTFs measured with the KEMAR dummy head with large pinnae, taken from the CIPIC HRTF Database (Algazi *et al.*, 2001). All simulated sound sources are located in the frontal horizontal plane, and, unless otherwise noted, all the stimuli are aligned to 60 dB SPL averaged over the whole stimulus length.

4.3.1 Independent sources in free-field

In this section, the cue selection method is applied to independent stimuli in an anechoic environment. As the first example, the operation of the selection procedure is illustrated in detail for the case of independent speech sources at different directions. Subsequently, simulation results of the effect of target-to-distracter ratio (TDR) on localization of the target stimulus are presented.

Concurrent speech

Localization of a speech target in the presence of one or more competing speech sources has been investigated psychophysically by Hawley *et al.* (1999) and Drullman and Bronkhorst (2000). Drullman and Bronkhorst (2000) utilized an anechoic virtual environment using both individualized and non-individualized HRTFs for binaural reproduction of the stimuli. They reported a slight but statistically significant degradation in localization performance when the number of competing talkers was increased beyond two. The experiment of Hawley *et al.* (1999), on the other hand, was conducted in a "sound-field room" (reverberation time of approximately 200 ms), as well as using headphone reproduction of the stimuli recorded binaurally in the same room. While not strictly anechoic, their results are also useful for evaluating the current simulation results. Hawley *et al.* (1999) found that apart from occasional confusions between the target and the distracters, increasing the number of competitors from 1 to 3 had no significant effect on localization accuracy. As discussed in Section 4.1, room reflections generally make the localization task more difficult, so a similar or a better result would be expected to occur in an anechoic situation. By comparison, the overall localization performance reported by Drullman and Bronkhorst (2000) was fairly poor, and the results may have been affected by a relatively complex task requiring listeners to recognize the target talker prior to judging its location.

Based on the previous discussion, the cue selection has to yield ITD and ILD cues similar to the free-field cues of each of the speech sources in order to correctly predict

[4]All simulations displaying PDFs have been rerun in order to correct a sign error on the ILD axis. Furthermore, while rerunning the simulations, it was discovered that the PDFs had been computed over 2 s intervals instead of 1.6 s, as stated earlier by Faller and Merimaa (2004). The new simulations display results practically identical to the earlier ones with the only difference in the simulations being different samples used for the internal noise.

[5]The Matlab code used for these simulations is available at `http://www.acoustics.hut.fi/software/cueselection/`.

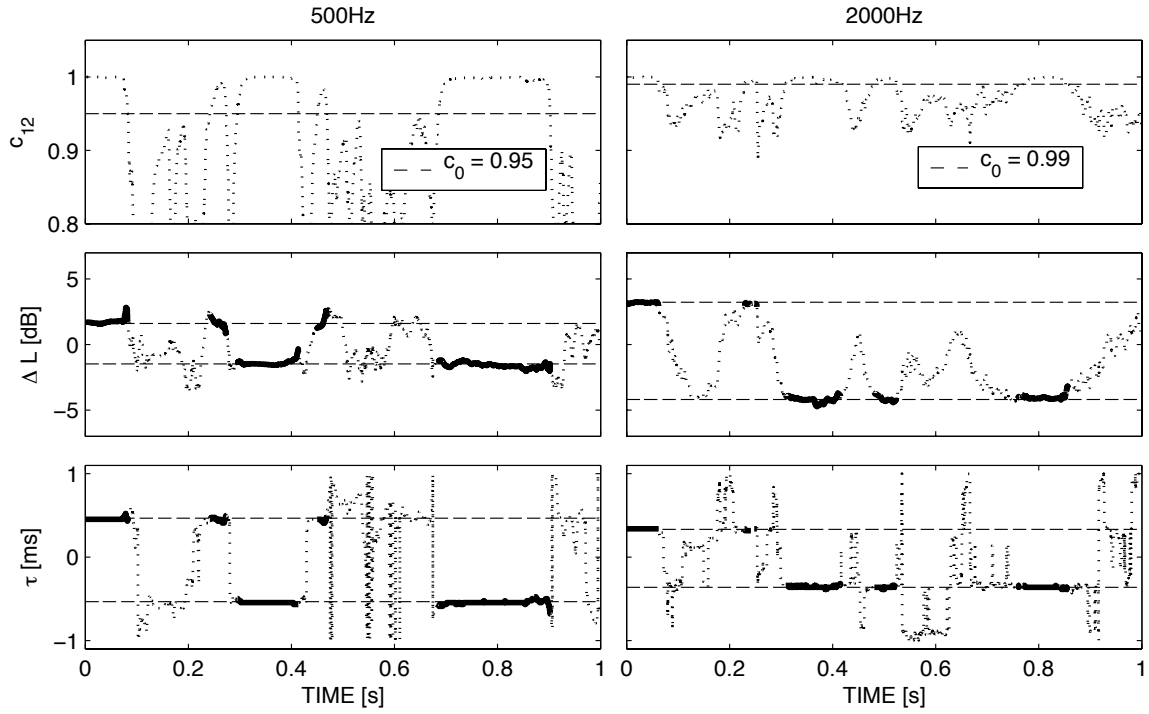

Figure 4.2: IC, ILD, and ITD as a function of time for two independent speech sources at ±40° azimuth. Left column: 500 Hz, and right column: 2 kHz critical band. The cue selection thresholds (top row) and the free-field cues of the sources (middle and bottom rows) are indicated with dashed lines. Selected cues are marked with bold solid lines.

the directions of the perceived auditory events. Three simulations were carried out with 2, 3, and 5 concurrent speech sources. The signal of each source consisted of a different phonetically balanced sentence from the Harvard IEEE list (IEEE, 1969) recorded by the same male speaker. As the first case, 2 speech sources were simulated at azimuthal angles of ±40°. Figure 4.2 shows the IC, ILD, and ITD as functions of time for the critical bands centered at 500 Hz and 2 kHz. The free-field cues that would occur with a separate simulation of the sources at the same angles are indicated with the dashed lines. The selected ITD and ILD cues (Eq. 4.7) are marked with bold solid lines. Thresholds of $c_0 = 0.95$ and $c_0 = 0.99$ were used for the 500 Hz and 2 kHz critical bands, respectively, resulting in 65 % and 54 % selected signal power (Eq. 4.10). The selected cues are always close to the free-field cues, implying perception of two auditory events located at the directions of the sources, as reported in the literature. As expected, due to the neural transduction, the IC has a smaller range at the 2 kHz critical band than at the 500 Hz critical band. Consequently, a larger c_0 is required.

The performance of the cue selection was assessed statistically as a function of c_0 for the same two speech sources and the critical bands with center frequencies of 250, 500, 1000, 2000, and 3000 Hz. Figure 4.3 shows the ITD and ILD biases (Eq. 4.8) and standard deviations (Eq. 4.9), as well as the fraction of signal power corresponding to the selected cues (Eq. 4.10) as a function of c_0. The biases and standard deviations

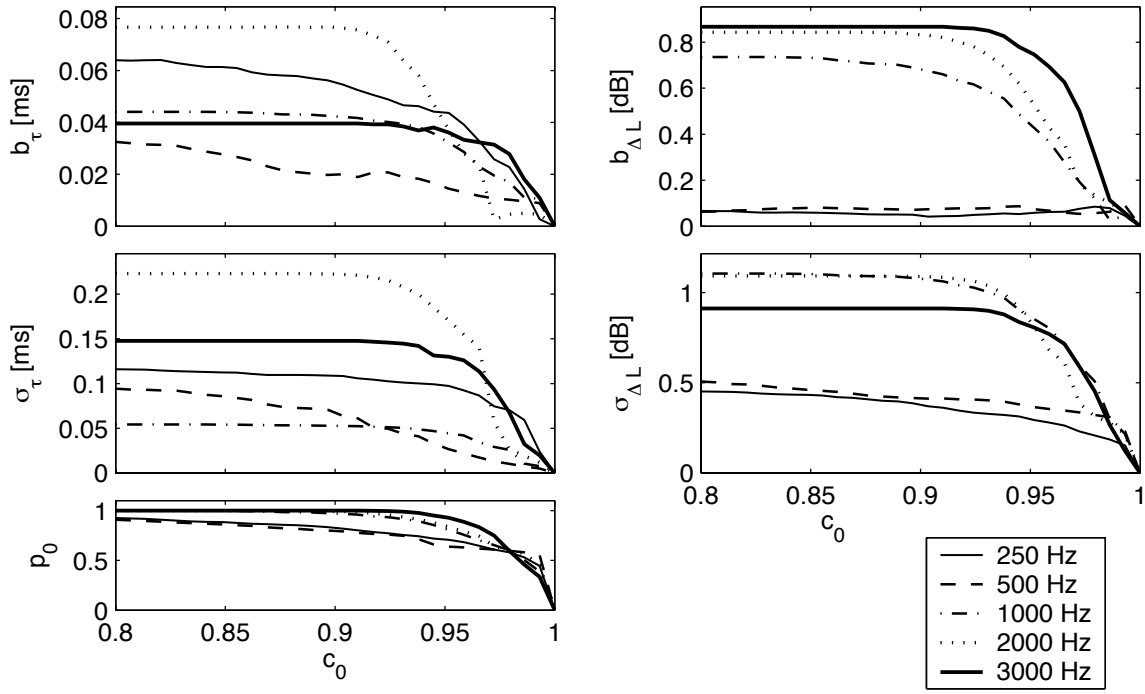

Figure 4.3: ITD and ILD bias (top panels), standard deviation (middle panels), and relative power (bottom left panel) of the selected signal portions as a function of the cue selection threshold c_0 for two independent speech sources. Data are shown for the 250, 500, 1000, 2000, and 3000 Hz critical bands.

were computed for both sources separately, as described earlier, and then averaged over 2 s of the signals. The graphs indicate that both the biases and the standard deviations decrease with increasing c_0. Thus, the larger the c_0, the closer the obtained cues are to the reference free-field values. Furthermore, the selected signal power decreases gradually until fairly high values of c_0. The general trend of having higher absolute ILD errors at high frequencies is related to the overall larger range of ILDs occurring at high frequencies due to more efficient head shadowing (see Section 3.2.1).

The simulation with 3 independent talkers was performed with speech sources at 0° and ±30° azimuth, and the simulation of 5 talkers with two additional sources at ±80° azimuth. In both cases, the results were fairly similar at different critical bands, so the data are only shown for the 500 Hz band. Panels (A) and (B) of Figure 4.4 show PDFs of ITD and ILD without the cue selection for the 3 and 5 speech sources, respectively, and panels (C) and (D) show similar PDFs of the selected cues. The selection threshold was set at $c_0 = 0.99$, corresponding to 54 % selected signal power for the 3 sources and 22 % for the 5 sources. In both cases, even the PDFs with all cues show ITD peaks at approximately correct locations, and the cue selection can be seen to enhance the peaks. With the cue selection, the widths of the peaks (i.e. the standard deviations of ITD and ILD) in the 3 source case are as narrow as in separate one-source free-field simulations, which implies robust localization of three auditory events corresponding to the psychophysical results of Hawley *et al.* (1999) and Drullman and Bronkhorst (2000). In the case of 5 sources, the peaks become

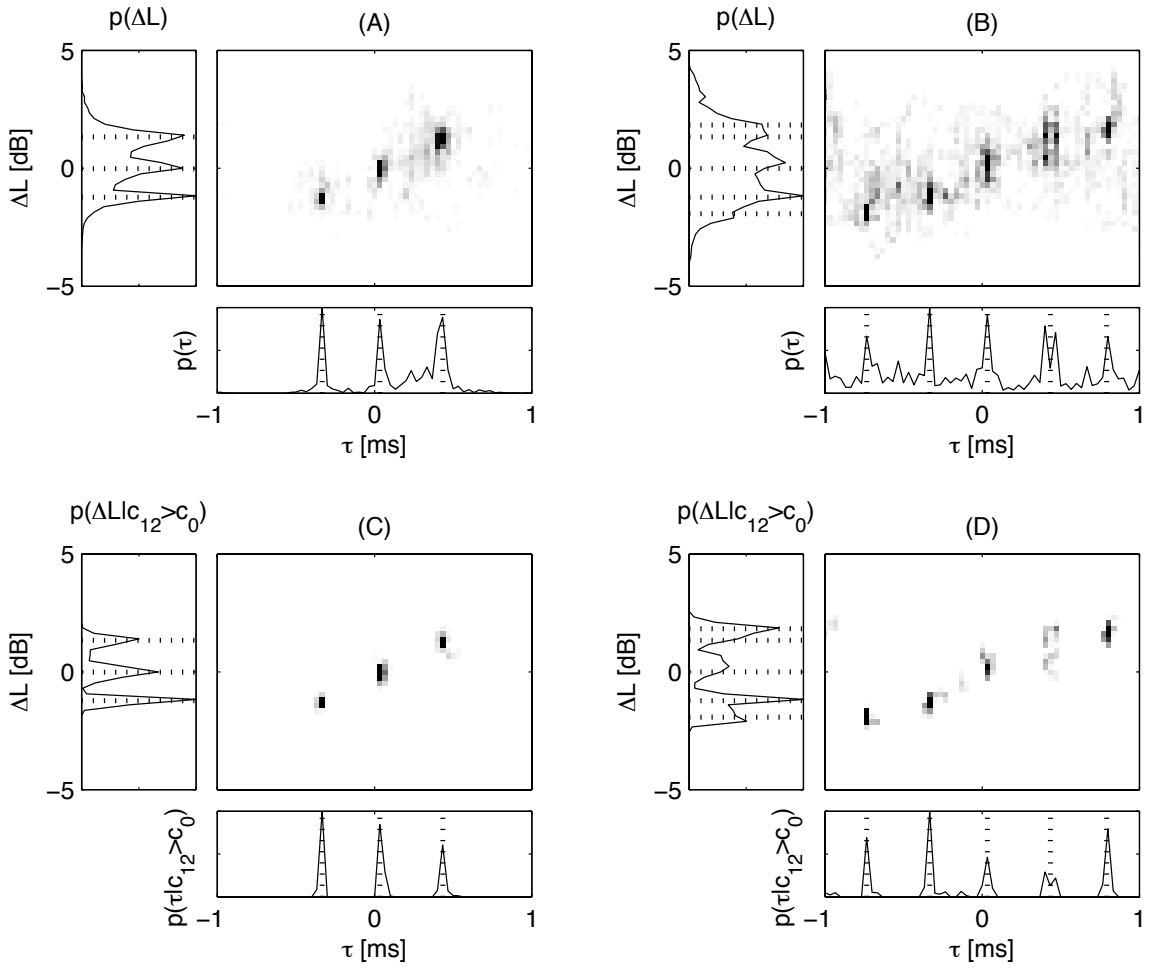

Figure 4.4: PDFs of ITD and ILD for 3 (A) and 5 (B) independent speech sources and corresponding PDFs when cue selection is applied (C and D). The values of the free-field cues for each source are indicated with dotted lines. Data are shown for the 500 Hz critical band.

slightly broader. The ITD peaks are still narrow and correctly located, but at the 500 Hz critical band, the range of ILD cues is insufficient for distinct peaks to appear along the ILD axis. This result is also in line with the classic duplex theory (Strutt [Lord Rayleigh], 1907; see also Section 3.3.1 in this thesis) of sound localization, which states that at low frequencies ITD cues are more salient than ILD cues.

Click-train and noise

Good and Gilkey (1996) and Good *et al.* (1997) studied the localization of a click-train target in the presence of a simultaneous noise distracter. Using loudspeaker reproduction in an anechoic chamber, localization performance was shown to degrade monotonously with a decreasing target-to-distracter ratio (TDR). The investigated TDRs were defined relative to the individual detection threshold of each listener for the case when the target sound was presented from the same direction as the dis-

tracter. With a target level just a few dB above the detection threshold, localization performance in the left–right direction was still found to be nearly as good as without the distracter. The degradation started earlier and was more severe for the up–down and front–back directions. The results for the left–right direction were later confirmed by Lorenzi *et al.* (1999), who conducted a similar experiment with sound sources in the frontal horizontal plane. However, the detection levels of Lorenzi *et al.* (1999) were slightly higher, possibly due to the utilization of a sound-treated chamber instead of a strictly anechoic environment. Furthermore, Lorenzi *et al.* (1999) found a degradation in performance when the stimuli were lowpass filtered at 1.6 kHz, unlike when the stimuli were highpass filtered at the same frequency.

A simulation was carried out with a white noise distracter directly in front of the listener and a click-train target with a rate of 100 Hz located at 30° azimuth. Assuming a detection level of −11 dB (the highest value in Good *et al.*, 1997), the chosen absolute TDRs of −3, −9, and −21 dB correspond to the relative TDRs of 8, 2, and −10 dB, respectively, as investigated by Good and Gilkey (1996). The PDFs for the critical band centered at 500 Hz did not yield a clear peak corresponding to the direction of the click train. Motivated by the fact that, in this case, higher frequencies are more important for directional discrimination (Lorenzi *et al.*, 1999), the 2 kHz critical band was investigated further. Panels (A)-(C) in Figure 4.5 show the resulting PDFs of ITD and ILD without the cue selection for the selected TDRs. The corresponding PDFs obtained by the cue selection (Eq. 4.7) are shown in panels (D)-(F). The selection thresholds for the panels (D)-(F) were $c_0 = 0.990$, $c_0 = 0.992$, and $c_0 = 0.992$, respectively, resulting in 3 %, 10 %, and 99 % of the signal power being represented by the selected cues.

The PDFs in Figure 4.5 imply that the target is localized as a separate auditory event for the TDRs of −3 dB and −9 dB. However, for the lowest TDR, the target click-train is no longer individually localizable, as also suggested by the results of Good and Gilkey (1996). In panels (A) and (B), ITD peaks can be seen to rise at regular intervals due to the periodicity of the cross-correlation function, while the cue selection suppresses the periodical peaks as shown in panels (D) and (E). Note that when the click-train is individually localizable, only the recovered ITD cues are close to the free-field cues of both sources, whereas a single broad ILD peak appears. This is in line with the findings of Braasch (2003) that in the presence of a distracter, ILDs are less reliable cues for localization, and that ITDs also gain more importance in the subjective localization judgment. The ITD peaks corresponding to the click-train are also shifted away from the distracter. Such a pushing effect caused by a distracter in front of the listener was observed for one listener in a similar experiment (Lorenzi *et al.*, 1999) and for most listeners when the target was an independent noise signal in the experiments of Braasch and Hartung (2002). On the contrary, pulling effects have been reported by Butler and Naunton (1964), Good and Gilkey (1996), and for two listeners by Lorenzi *et al.* (1999).

4.3.2 Precedence effect

This section illustrates the cue selection within the context of the precedence effect. Pairs of clicks are used to demonstrate the results for wideband signals. However,

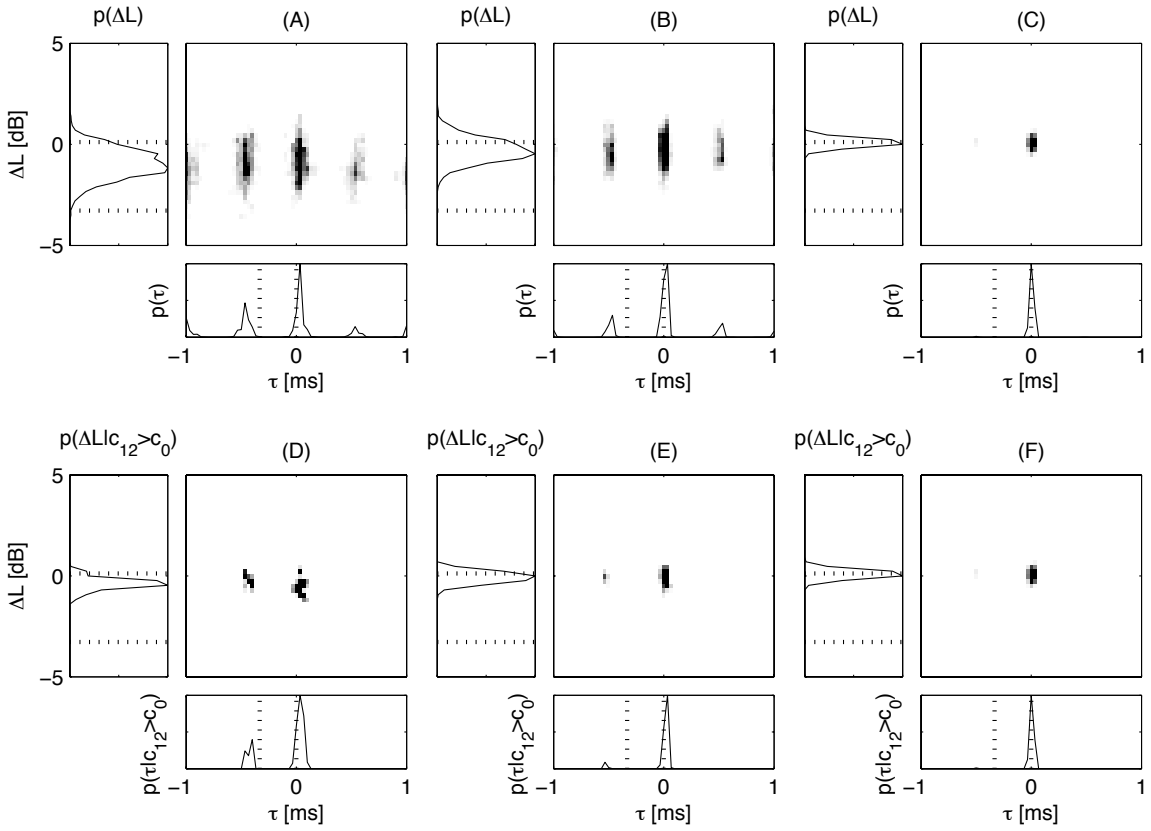

Figure 4.5: PDFs of ITD and ILD for a click-train and white Gaussian noise at different TDRs: −3, −9, −21 dB (A-C), and the corresponding PDFs when cue selection is applied (D-F). The values of the free-field cues are indicated with dotted lines. Data are shown for the 2 kHz critical band.

the simulations are still performed at individual critical bands, so in this context, any stimulus with a width of at least a critical band is effectively wideband. Furthermore, sinusoidal tones are simulated with different onset rates, and the cues obtained during the onset are shown to agree with results reported in the literature.

Click pairs

In a classic precedence effect experiment (see Section 3.3.3), a lead/lag pair of clicks is presented to the listener. The leading click is first emitted from one direction, followed by another identical click from another direction after an *interclick interval* (ICI) of a few milliseconds. As discussed earlier, the directional perception changes depending on the ICI.

Figure 4.6 shows the IC, ILD, and ITD as a function of time for a click-train with a rate of 5 Hz analyzed at the critical bands centered at 500 Hz and 2 kHz. The lead source is simulated at 40° and the lag at −40° azimuth with an ICI of 5 ms. As expected based on earlier discussion, IC is close to one only when the lead sound is within the analysis time window. As soon as the lag reaches the ears of the listener, the superposition of the two clicks reduces the IC. The cues obtained by the selection

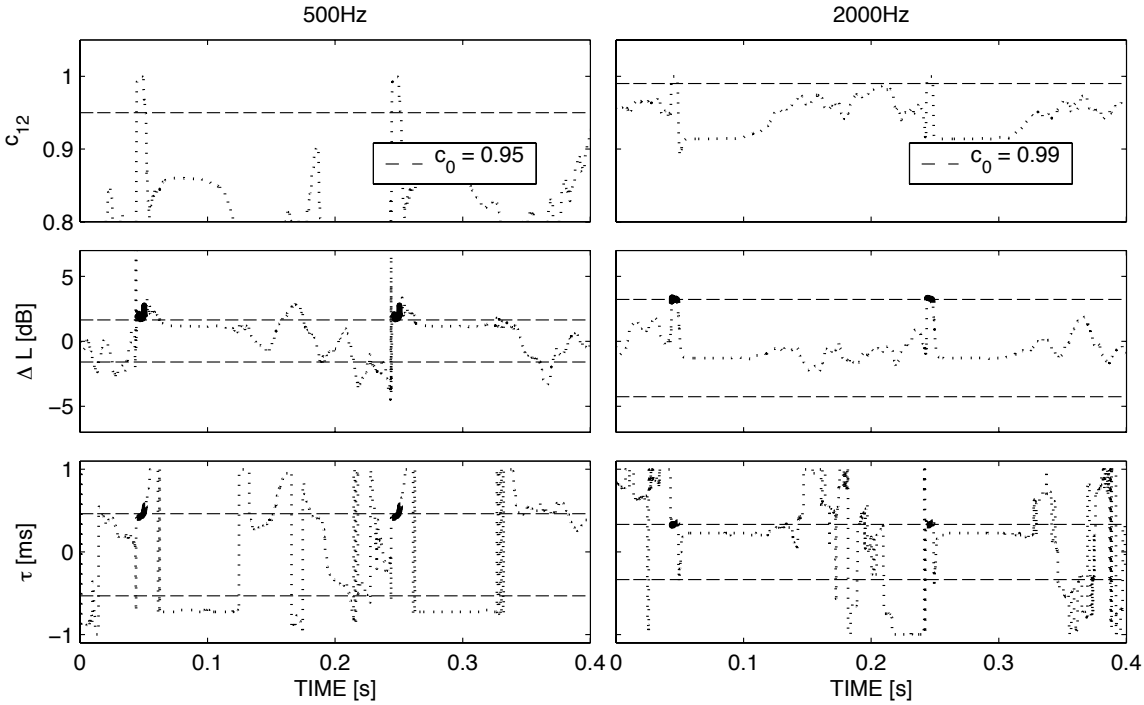

Figure 4.6: IC, ILD, and ITD as a function of time for a lead/lag click-train with a rate of 5 Hz and an ICI of 5 ms. Left column: 500 Hz, and right column: 2 kHz critical band. The cue selection thresholds (top row) and the free-field cues of the sources (middle and bottom rows) are indicated with dashed lines. Selected cues are marked with bold solid lines.

with $c_0 = 0.95$ for the 500 Hz and $c_0 = 0.985$ for the 2 kHz critical band are shown in the figure, and the free-field cues of both sources are indicated again with dashed lines. The selected cues are close to the free-field cues of the leading source, and the cues related to the lag are ignored, as is also known to happen based on the psychophysical studies reviewed earlier. The fluctuation in the cues before each new click pair is due to the internal noise of the model.

The performance of the cue selection was again assessed as a function of c_0 for the critical bands with center frequencies of 250, 500, 1000, 2000, and 3000 Hz. The statistical measures were calculated from a 2 s signal segment. Figure 4.7 shows ITD and ILD biases (Eq. 4.8) and standard deviations (Eq. 4.9), as well as the power of the selected cues (Eq. 4.10) as a function of c_0. Note that the biases and standard deviations were computed relative to only the free-field cues of the leading source, since localization of the lag should be suppressed if the selection works correctly. Both the biases and standard deviations decrease as c_0 increases. Thus, the larger the cue selection threshold c_0, the more similar the selected cues are to the free-field cues of the leading source.

At a single critical band, the energy of the clicks is spread over time due to the gammatone filtering and the model of neural transduction. Therefore, with an ICI of 5 ms, a large proportion of the critical band signals related to the clicks of a pair overlap, and only a small part of the energy of the lead click appears in the critical

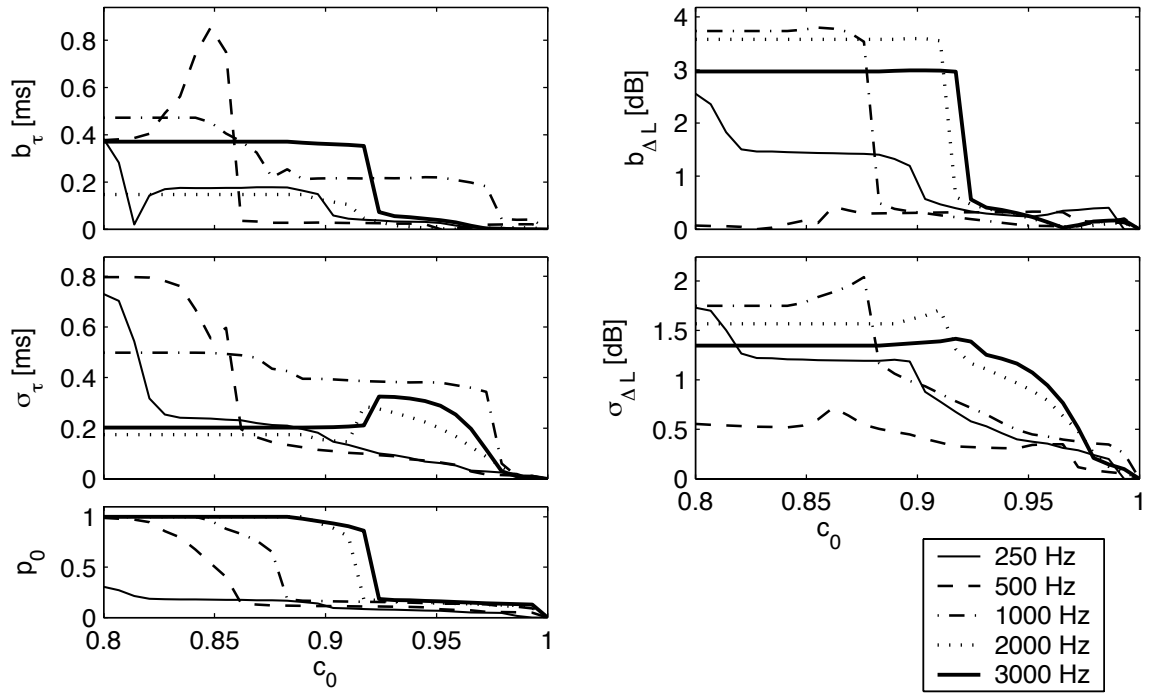

Figure 4.7: ITD and ILD bias, standard deviation, and relative power of the selected signal portions as a function of the cue selection threshold c_0 for a lead/lag click-train. Data are shown for the 250, 500, 1000, 2000, and 3000 Hz critical bands.

band signals before the lag. Consequently, the relative signal power corresponding to the selected cues is fairly low when requiring small bias and standard deviation, as can be seen in the left bottom panel of Figure 4.7.

The effect of ICI

The previous experiment was repeated for ICIs in the range of $0 - 20$ ms using the 500 Hz critical band. The chosen range of delays includes summing localization, localization suppression, and independent localization of both clicks without the precedence effect (see Section 3.3.3). For all previous simulations, a suitable c_0 was chosen as a compromise between similarity of the cues to free-field cues and how frequently cues are selected. Here, each ICI corresponds to a different listening situation, since the different delays of the lag imply different acoustical environments. It is thus expected that the most effective c_0 may also differ depending on ICI.

Several different criteria for determining the c_0 were assessed. Indeed, using the same c_0 for all ICIs did not yield the desired results. The criterion of adapting c_0 such that the relative power of the selected cues (Eq. 4.10) had the same value for each simulation did not provide good results either. Thus, a third criterion was adopted. The cue selection threshold c_0 was determined numerically for each simulation such that σ_τ (the narrowness of the peaks in the PDFs of ITD) was equal to 15 μs. This could be explained with a hypothetical auditory mechanism adapting c_0 in time with the aim of making the ITD and/or ILD standard deviation sufficiently small. Such

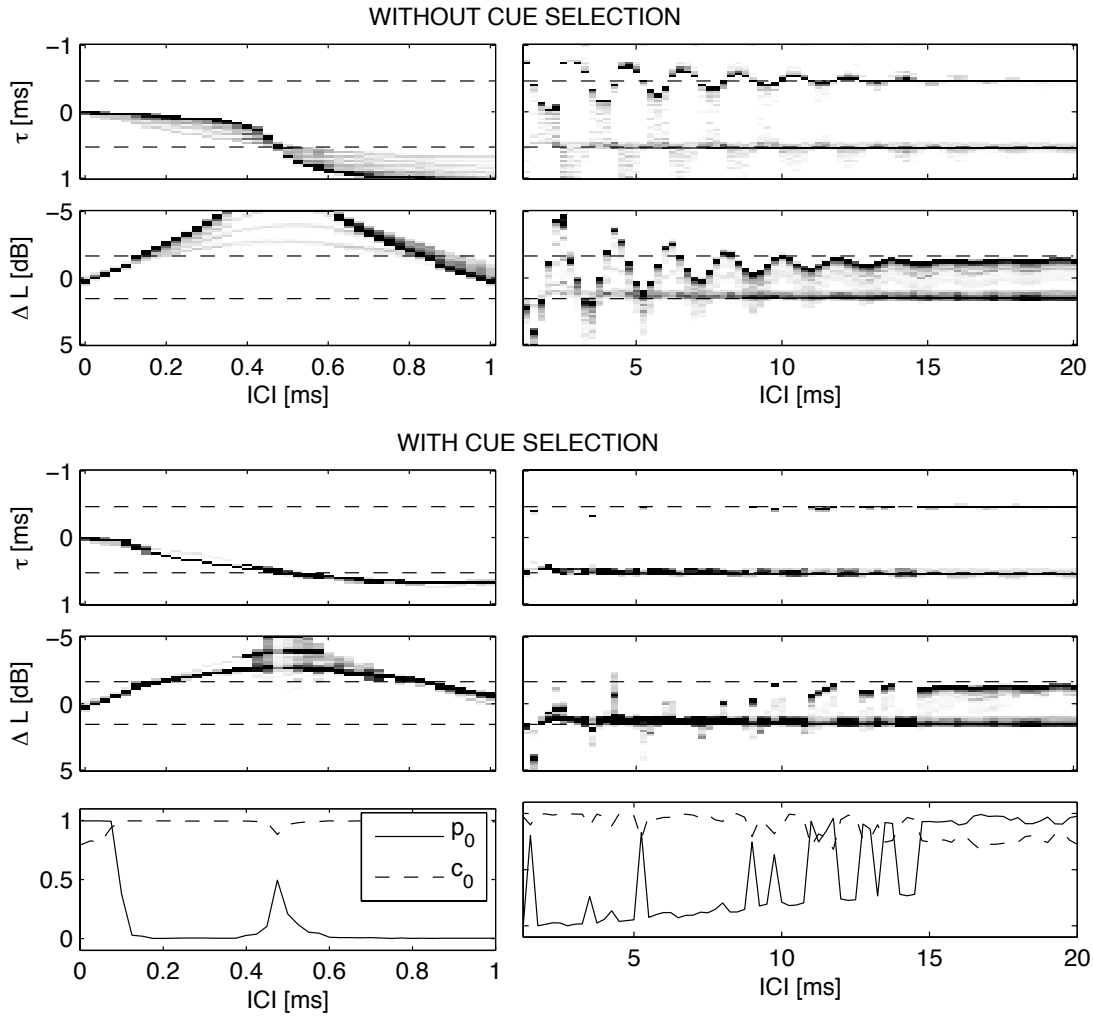

Figure 4.8: PDFs of ITD and ILD as a function of ICI for a click-train: without cue selection (rows 1 and 2) and with cue selection (rows 3 and 4). The cue selection threshold c_0 and relative power p_0 of the selected signal portion are shown in the bottom row.

small standard deviations would indicate small fluctuations of the selected cues in time and thus unambiguous localization of static auditory events. The resulting PDFs of ITD and ILD as a function of ICI with and without the cue selection are shown in Figure 4.8.

The PDFs without the cue selection (rows 1 and 2 in Figure 4.8) indicate two independently localized auditory events for most ICIs above 1 ms. Furthermore, the predicted directions depend strongly on the delay. On the contrary, apart from some secondary ILD peaks, the PDFs with the cue selection correctly predict all the three phases of the precedence effect (summing localization, localization suppression, and independent localization). At delays less than approximately 1 ms, the ITD peak moves to the side as the delay increases, as desired, but the ILD cues do not indicate the same direction as the ITD cues. However, this is also in line with existing psychophysical literature. Anomalies of the precedence effect have been observed in

listening tests with bandpass filtered clicks (Blauert and Cobben, 1978), suggesting a possible contribution of the extracted misleading ILDs to the localization judgment.

For delays within the range of approximately $1-10$ ms, there is only one significant peak in the PDFs, indicating localization in the direction of the lead. For larger delays, two peaks appear, suggesting two independently localized auditory events. The fusion of two clicks has been found to sometimes break down earlier, but 10 ms is within the range of reported critical thresholds for localization dominance (Litovsky *et al.*, 1999; Litovsky and Shinn-Cunningham, 2001). Note also that the selected ITD cues for the lag are actually stronger starting from an ICI of slightly above 10 ms. Indeed, Stellmack *et al.* (1999) found in a psychophysical lateralization experiment that starting from an ICI of 16 ms (the next tested value above 8 ms) up to 64 ms, the ITD of the lag was more easily discriminable than that of the lead, which is reflected in the current simulation results.

The bottom row of Figure 4.8 shows the selection threshold c_0 and the relative power p_0 of the signal corresponding to the selected cues as a function of the ICI. For most ICIs up to approximately 8 ms, the relative power of the selected signal portion almost vanishes. However, there are some characteristic peaks of p_0. The experiment was repeated for a number of critical bands in the range of 400 to 600 Hz with the observation that the peaks moved along the ICI axis as a function of the center frequency of the considered critical band. Otherwise, the general trends of the selected cues were very similar to those at the 500 Hz band in that they all strongly implied the three phases of the precedence effect. Thus, by considering a number of critical bands, the three phases of the precedence effect can indeed be explained by the cue selection such that at each ICI a signal portion with non-vanishing power is selected.

Discussion on the click pair simulations

For the previous simulation, it was hypothesized that the criterion for determining c_0 is the standard deviation of ITD and/or ILD. The computation of these quantities involves determining the number of peaks (i.e. the number of individually localized auditory events) adaptively in time, which might be related to the buildup of precedence. As explained in Section 3.3.3, a buildup occurs when a lead/lag stimulus with ICI close to the echo threshold is repeated several times. During the first few stimulus pairs, the precedence effect is not active and two auditory events are independently perceived. After the buildup, the clicks merge to a single auditory event in the direction of the lead. An adaptive process determining c_0 would require a certain amount of stimulus activity and time until an effective c_0 is determined and it could thus explain the time-varying operation of the precedence effect. Djelani and Blauert (2002) also showed that without stimulus activity the effect of the buildup decays slowly by itself, which supports the idea of an adaptive c_0. In order to model the direction-specific buildup, c_0 would also need to be defined as a function of direction. However, testing and developing the corresponding adaptation method is beyond the scope of this thesis and will be part of the future work.

Note also that the precedence effect simulations are a case where modeling the neural adaptation could have an effect on the results. As discussed earlier, Hartung

and Trahiotis (2001) were able to model the precedence effect for click pairs with short ICIs without any additional processing by using the neural transduction model of Meddis (1986, 1988); Meddis *et al.* (1990). However, their model did not produce correct results at single critical bands, but only averaged over a larger frequency range. Especially around 500 Hz, the results were considerably biased (Hartung and Trahiotis, 2001, Figure 5). Nevertheless, within the framework of the current model, the neural adaptation might turn out to be helpful because it would lower the level of the lagging click, and thus strengthen the cues produced by the lead.

Onset rate of a sinusoidal tone

Rakerd and Hartmann (1986) investigated the effect of the onset time of a 500 Hz sinusoidal tone on localization in the presence of a single reflection. In the case of a sinusoidal tone, the steady state ITD and ILD cues result from the coherent sum of the direct and reflected sound at the ears of a listener. Often these cues do not imply the direction of either the direct sound or the reflection. As discussed in Section 3.3.3, Rakerd and Hartmann (1986) found that the onset rate of the tone was a critical factor in determining how much the misleading steady state cues contributed to the localization judgment of human listeners. For fast onsets, localization was based on the correct onset cues, unlike when the level of the tone rose slowly. The cue selection mechanism cannot, as such, explain the discounting of the steady state cues, which always have an IC close to one. However, considering just the onsets, the following results reflect the psychophysical findings of Rakerd and Hartmann (1986).

Figure 4.9 shows the IC, ILD, and ITD as a function of time for a 500 Hz tone with onset times of 0, 5, and 50 ms. The simulated case corresponds approximately to the "WDB room" and "reflection source 6" condition reported by Rakerd and Hartmann (1986). The direct sound is simulated in front of the listener, and the reflection arrives with a delay of 1.4 ms from an azimuthal angle of 30°. A linear onset ramp is used, and the steady state level of the tone is set to 65 dB SPL. The ITD and ILD cues selected with a threshold of $c_0 = 0.93$ are marked with bold solid lines, and the free-field cues of the direct sound and the reflection are indicated with the dashed lines. Note that the direct sound reaches the ears of the listener at approximately 7 ms. For the onset times of 0 and 5 ms, the ITD and ILD cues are similar to the free-field cues at the time when the IC reaches the threshold. However, with the onset time of 50 ms, the ITD and ILD cues no longer correspond to the free-field cues, which is also suggested by the different localization in this condition (Rakerd and Hartmann, 1986).

In order to predict the final localization judgment, another selection mechanism would be needed to include only the localization cues at the time instant when the cue selection becomes effective. The dependence on the onset rate can be explained by considering the input signals of the binaural processor. During the onset, the level of the reflected sound follows that of the direct sound with a delay of 1.4 ms. Thus, the slower the onset, the smaller the difference. The critical moment is when the level of the direct sound rises high enough above the level of the internal noise to yield IC above the selection threshold. If the reflection has non-negligible power at that time, localization cues will already be biased to the steady state direction when the selection begins.

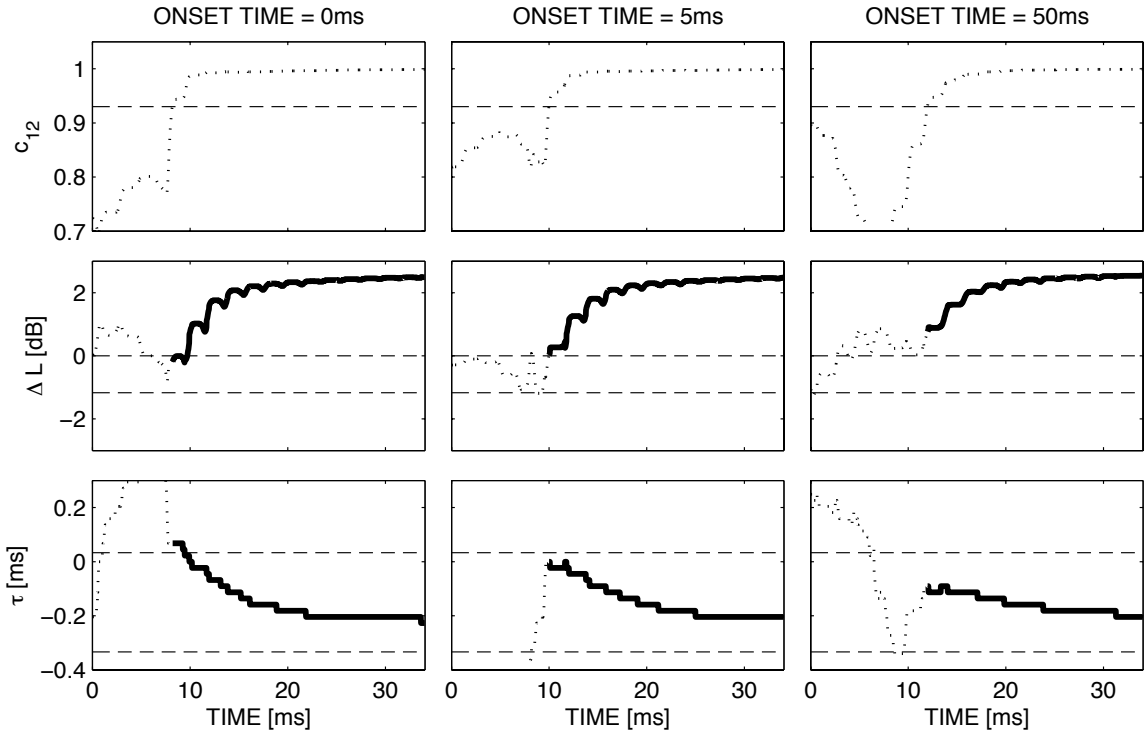

Figure 4.9: IC, ILD, and ITD as a function of time for a 500 Hz sinusoidal tone and one reflection. The columns from left to right show results for onset times of 0 ms, 5 ms, and 50 ms. The cue selection threshold of $c_0 = 0.95$ (top row) and the free-field cues of the source and the reflection (middle and bottom rows) are indicated with dashed lines. Selected cues are marked with bold solid lines. Data are shown for the 500 Hz critical band.

4.3.3 Independent sources in a reverberant environment

As a final test for the model, the localization of 1 and 2 speech sources was simulated in a reverberant environment. The utilized BRIRs were measured with a Neumann KU 80 dummy head in an empty lecture hall with reverberation times of 2.0 and 1.4 s at the octave bands centered at 500 and 2000 Hz, respectively. The same phonetically balanced speech samples as those used in Section 4.3.1 were convolved with the BRIRs simulating sources at 30° azimuth for the case of one source and ±30° for the two sources. The case of two talkers again included two different sentences uttered by the same male speaker. For computing the free-field cues, the BRIRs were truncated to 2.3 ms, such that the effect of the reflections was ignored.

The chosen hall is a difficult case for localization, due to an abundance of diffuse reflections from the tables and benches all around the simulated listening position. At the 500 Hz critical band, the ITD and ILD cues prior to the selection did not yield any meaningful data for localization. The cue selection resulted in high peaks close to the free-field cues, but it was not able to suppress all other peaks implying different directions. A subsequent investigation showed that these erroneous peaks appear at different locations at different critical bands. Thus, processing of localization information across critical bands should be able to further suppress them. At 2 kHz,

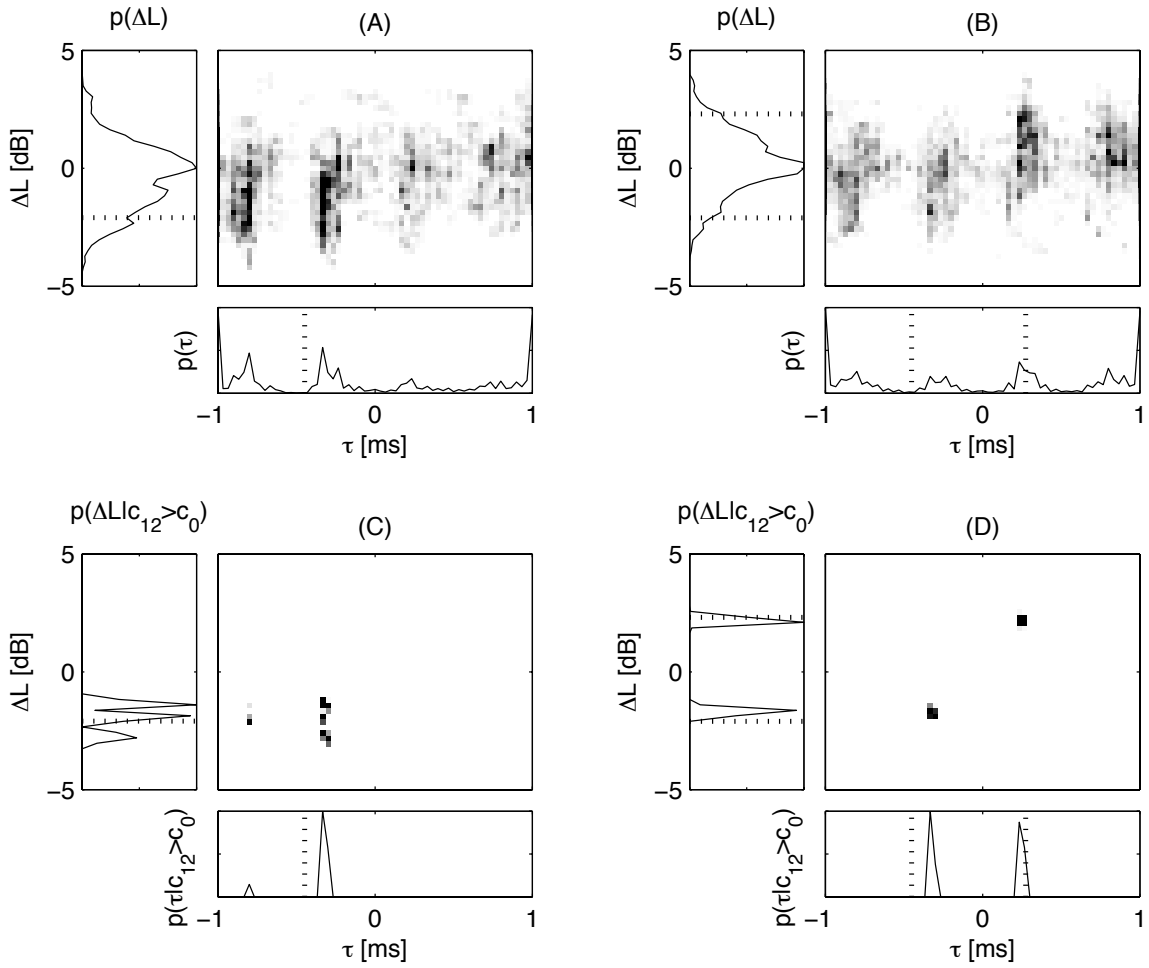

Figure 4.10: PDFs of ITD and ILD for 1 (A) and 2 (B) speech sources in a reverberant hall and the corresponding PDFs when cue selection is applied (C and D). The values of the free-field cues for each source are indicated with dotted lines. Data are shown for the 2 kHz critical band.

the results for a single critical band were clearer and will be illustrated here.

Panels (A) and (B) of Figure 4.10 show PDFs of ITD and ILD without the cue selection, and panels (C) and (D) show the corresponding PDFs of the selected cues. Since the cue selection in this case samples the ITD and ILD relatively infrequently, the PDFs were computed considering 3 s of the signals. However, similar results are obtained when the PDFs are computed from different time intervals. The cue selection criterion for both the 1 and 2 source scenarios was $c_0 = 0.99$, resulting in 1 % of the signal power corresponding to the selected cues. Without the cue selection, the PDFs do not yield much information for localization in either of the cases. Periodicity of the cross-correlation function is clearly visible and it is difficult to distinguish between the one and two source cases. However, with the cue selection, sharp peaks arise relatively close to the free-field cues. In the two-source case, the right source is practically correctly localized, whereas the ITD cues of the left source are slightly biased towards the center. Note that contrary to the results for independent speech sources in Section

4.3.1, the localization is in this case shifted towards the competing sound source. As discussed, this kind of a pulling effect has also been reported in psychoacoustical studies (e.g., Butler and Naunton, 1964; Good and Gilkey, 1996; Lorenzi *et al.*, 1999; Braasch and Hartung, 2002).

4.4 General discussion and future work

In the previous sections, the selection of ITD and ILD cues based on IC was introduced into a localization model and applied to simulations of a number of complex listening situations. In comparison to several existing localization models, a significant difference in the proposed method is the way that the signal power at each time instant affects the localization judgment. In models not designed for complex listening situations, the localization cues and subsequently the final localization judgment are often derived from a time window including the whole stimulus, or of a time integration of a binaural activity pattern computed with running non-normalized cross-correlation. In such cases, the contribution of each time instant to the final localization depends on the instantaneous power. In our approach, only the cues during the selected time instants contribute to localization. Thus, the model can, in many cases, neglect localization information corresponding to time instants with high power, if the power is high due to concurrent activity of several sound sources (or concurrent activity of sources and reflections). Nevertheless, the relative power of individual sources affects how often ITD and ILD cues corresponding to each source are selected.

After publishing the original paper on the current model (Faller and Merimaa, 2004), a kind of cue selection method was also found in physiological studies of the inferior colliculus[6] of the barn owl. The neurons measured by Keller and Takahashi (2005) responded strongly only when the instantaneous binaural cues created by two concurrent sources were similar to the free-field cues of the spatial receptive field of each neuron. In the model of Keller and Takahashi (2005), such time instants were identified based on the similarity of both ITD and ILD to the free-field cues. In this sense the approach resembles the model of Gaik (1993). Furthermore, cross-frequency processing was taken into account according to the sensitivity of each measured neuron to different critical bands. Nevertheless, as shown in this chapter, for a single critical band, the time instants when the instantaneous cues correspond to the free-field cues typically coincide with a high IC. Hence, apart from not considering cross-frequency interaction, the proposed model indeed extracts localization cues similar to those that were shown to be extracted physiologically by the barn owl in multi-source scenarios.

The proposed model also bears resemblance to earlier models of the precedence effect (see Section 3.6.3). The temporal inhibition in the model of Lindemann (1986a) tends to hold the highest peaks of the running cross-correlation function (calculated with the stationary inhibition that incorporates ILDs into the model). The higher a peak (i.e. the higher the IC at the corresponding time instant), the stronger the temporal inhibition will be. The cue selection achieves a somewhat similar effect without

[6]The inferior colliculus is located in the midbrain and it receives inputs, among other nuclei, from the superior olivary complex which is considered as the location where the binaural cues are extracted (see Section 3.6).

the need for an explicit temporal inhibition mechanism, since the localization suppression is directly related to the IC estimated with a similar time window. However, the effect can also be quite different in some scenarios. Whereas the model of Lindemann (1986a) only "remembers" the peaks corresponding to a high IC for a short time (time constant of 10 ms), the cue selection with a slowly varying c_0 can have a considerably longer memory. In the precedence effect conditions (Section 4.3.2), the cue selection also naturally derives most localization information from signal onsets, as is explicitly done in the model of Zurek (1987). However, the cue selection does not necessarily include all onsets.

The cross-correlation function has also been used in a somewhat similar manner in binaural front ends for automatic speech recognition (Usagawa *et al.*, 1996; Rateitschek, 1998; Palomäki *et al.*, 2004; Palomäki, 2005). These approaches use the magnitude of the cross-correlation function around a fixed lag position to identify time-frequency regions not corrupted by concurrent sound and/or reverberation. The methods thus effectively measure the similarity of the ITD to a known or estimated value corresponding to a chosen speech source. Motivated by the work presented in this chapter, Brown *et al.* (2006) also investigated direct utilization of an IC threshold (regardless of the position of the cross-correlation peak) for the selection of reliable speech features. When added to a statistical analysis of the instantaneous ITD and ILD cues, the thresholding led to an improved recognition rate in room conditions. However, the threshold had to be set to a lower value than those used in the current studies in order to include a sufficient amount of time-frequency regions for speech recognition purposes.

As discussed earlier, the cue selection threshold is a compromise between accuracy and frequency of occurance of the selected localization cues. The frequency of the time instants when the direct sound of only one source dominates within a critical band depends on the complexity of the listening situation. In the most complex simulated cases (e.g., multiple sources in a reverberant environment in Section 4.3.3), only a small fraction of the ear input signals contributed to localization, and new localization information was acquired relatively infrequently. We, nevertheless, propose that in the case of localization, it is the cues at these time instants that determine the perception. During the time when no cues are selected, the localization of the corresponding auditory events is assumed to be determined by the previously selected cues, which is in principle possible. Indeed, localization of sinusoidal tones based only on their onsets (Rakerd and Hartmann, 1985, 1986) and a related demonstration called the "Franssen effect" (Hartmann and Rakerd, 1989) show that a derived localization judgment can persist for several seconds after the related localization cues have occurred.

The cue selection mechanism could be seen to perform a function that Litovsky and Shinn-Cunningham (2001) have characterized as "a general process that enables robust localization not only in the presence of echoes, but whenever any competing information from a second source arrives before the direction of a previous source has been computed." For the purposes of this study, the ITD and IC cues were analyzed using a cross-correlation model, and the ILDs were computed directly from the signal levels. Nevertheless, a similar cue selection could also be implemented in other localization models, such as the EI model of Breebaart *et al.* (2001a) (see Section 3.6.2), where instead of specifying a lower bound of IC for the cue selection, an upper

bound of activity would need to be determined.

Throughout the chapter, the resulting ITD and ILD cues were also considered separately instead of deriving a combined localization judgment. As discussed earlier, the relative weights of ITDs and ILDs can change in complex listening situations, with ITDs often gaining more weight compared to free-field scenarios. It was also seen that in some cases the cue selection was able to extract more reliable ITD than ILD cues. In future work, it will be interesting to investigate whether quantitative measures of the reliability of the selected cues, such as the standard deviations used in this chapter, reflect the psychophysically determined relative importance of the ITDs and ILDs in the corresponding situations. Future work also includes listening experiments aiming at specifically testing the effect of IC on localization, and if that appears to be the case, to study whether an adaptive mechanism for adjusting the cue selection threshold exist. Furthermore, smoother weighting functions than the utilized hard threshold would need to be investigated.

As shown in this chapter, the cue selection model is able to simulate a number of psychophysical results by using a selection threshold adapted to each specific listening scenario. Although the chapter was limited to localization based on binaural cues, it should be mentioned that the precedence effect has also been observed in the median sagittal plane where the localization is based on spectral cues instead of interaural differences (Blauert, 1971; Litovsky *et al.*, 1997; Rakerd *et al.*, 2000; see also Section 3.3.3 in this thesis). Thus, the cue selection model does not fully describe the operation of the precedence effect. Furthermore, the model cannot as such explain the discounting of ITD and ILD cues occurring simultaneously with a high IC during the steady state of a sinusoidal tone presented in a room (Rakerd and Hartmann, 1985, 1986; Hartmann and Rakerd, 1989). The psychophysical results of Litovsky *et al.* (1997) show that the localization suppression is somewhat weaker in the median plane than in the horizontal plane, which could be interpreted as evidence for another suppression mechanism, possibly operating simultaneously with a binaural mechanism such as the proposed cue selection. Indeed, simulating all the results cited in this paragraph would appear to require some additional form of temporal inhibition.

4.5 Summary and conclusions

A cue selection mechanism for modeling binaural auditory localization in complex listening situations was proposed. The cue selection considers ITD and ILD cues only when the IC at the corresponding critical band is larger than a certain threshold. It was shown that at time instants when this occurs, ITD and ILD are likely to represent the direction of one of the sources. Thus, by looking at the different ITD and ILD values during the selected time instants, one can obtain information about the direction of each source. For the purposes of this study, the cue selection mechanism was implemented within the framework of a model with a fairly detailed simulation of the auditory periphery, whereas the remaining parts were analytically motivated. Nevertheless, it was pointed out that in principle the proposed cue selection method is physiologically feasible regarding the accuracy of the human auditory system in discriminating IC.

The implemented binaural model with the proposed cue selection was verified with the results of a number of psychophysical studies from the literature. The simulations suggest relatively reliable localization of concurrent speech sources both in anechoic and reverberant environments. The simulated effect of the target-to-distracter ratio corresponds qualitatively to published results of localization of a click-train in the presence of a noise distracter. Furthermore, localization dominance is correctly reproduced for click pairs and for the onsets of sinusoidal tones. It was also hypothesized that the buildup of precedence may be related to the time the auditory system needs to find a cue selection threshold which is effective for the specific listening situation. As a final test, the model was applied for source localization in a reverberant hall with one and two speech sources. The results suggest that even in this most complex case, the model is able to obtain binaural cues corresponding to the directions of the sources.

Chapter 5

Spatial Impulse Response Rendering

5.1 Introduction

As outlined in Section 1.1.3, spatial sound reproduction may aim either for physical or perceptual authenticity, or plausibility. The attainable physical accuracy depends on the number of loudspeakers available for the reproduction and, correspondingly, on the number of microphones used in the recording. Depending on the reproduction technique, the accuracy may also be optimized either for a small sweet spot or for a larger listening area. This chapter is limited to reproduction with relatively small loudspeaker setups in the order of less than approximately twenty loudspeakers. This range covers current home theater systems and typical audio installations. Such loudspeaker setups cannot provide physically accurate spatial reproduction apart from a relatively small sweet spot. However, even a standard 5.1 loudspeaker setup (ITU-R BS.775-1, 1992–1994) is able to create a plausible surrounding sound field with fairly good perceptual accuracy especially in front of the listener.

The main problems of realistic multichannel loudspeaker reproduction are due to limitations in microphone technology, as will be argued in Section 5.2.2. Many recording techniques also require knowledge of the loudspeaker layout already in the recording phase, and conversion of the recorded signals for reproduction with a different loudspeaker system is, in general, not easy. The Spatial Impulse Response Rendering (SIRR) method (Merimaa and Pulkki, 2003, 2004, 2005; Pulkki *et al.*, 2004a,b; Pulkki and Merimaa, 2005, 2006) has been designed to overcome some of these problems by using a perceptually motivated analysis-synthesis approach. The goal of SIRR is as authentic perceptual reproduction as possible within the limits of the utilized recording and reproduction systems.

SIRR processing consists of an analysis of the direction of arrival and diffuseness of sound within frequency bands, followed by a synthesis yielding multichannel impulse responses that can be tailored for an arbitrary loudspeaker system. Although applicable also to general recording, SIRR is especially suitable for processing room responses for convolving reverberators (see Section 1.1.3) which is the context that SIRR will be placed into for the purposes of description in this chapter. Applications

to continuous sound will be briefly discussed in Section 5.7.

This chapter is largely based on the papers by Merimaa and Pulkki (2005) and Pulkki and Merimaa (2006) with some modified and extended discussion. The chapter is organized as follows: Section 5.2 briefly reviews common techniques for recording and reproduction of sound with small multichannel loudspeaker systems. The SIRR algorithm is described in Section 5.3, followed by a more detailed analysis of the alternative techniques for reproduction of diffuse sound in Section 5.4. Subjective evaluation of the SIRR algorithm in an anechoic environment and in a listening room is described in Sections 5.5 and 5.6, respectively. Finally, Section 5.7 includes some general discussion and conclusions.

5.2 Multichannel reproduction techniques

As mentioned earlier, in typical recording scenarios it is common to use *ad hoc* combinations of spot microphones, room microphones, and reverberators (see e.g., Theile, 2000). The resulting spatial reproduction often does not resemble performance in any existing venue but is rather constructed according to artistic considerations. However, systematic microphone techniques also exist and can be used either to directly record a performance or to capture room responses that can then be used to simulate recording in the same venue with the help of a convolving reverberator. This section is limited to a description of such systematic techniques.

Two important criteria will be used to characterize the reproduction techniques. In order to achieve perceptual authenticity, a reproduction method would need to be able to render *virtual sources* which are indistinguishable from corresponding real sources. Virtual sources can be positioned (usually) between loudspeakers and they are a result of the joint operation of several surrounding loudspeakers. In small multichannel setups, utilization of virtual sources is necessary to overcome the relatively coarse directional resolution offered directly by the loudspeaker locations. Another essential property is the ability of a reproduction method to create decorrelation of the loudspeaker signals, which will transfer to decorrelation of the ear input signals (e.g., Damaske and Ando, 1972; Kurozumi and Ohgushi, 1983; Tohyama and Suzuki, 1989) and affect the resulting spatial impression (see Section 3.5.2).

The discussion starts with a description of amplitude panning (Section 5.2.1), which is the most common technique for creating virtual sources from spot microphone signals. Amplitude panning will also be applied later in this chapter to reproduction of room responses. Subsequently, basic microphone techniques and some recent methods that can be seen as alternatives for the SIRR method are reviewed in Section 5.2.2.

5.2.1 Amplitude panning

According to the principles of summing localization (see Section 3.3.3), application of the same signal to two loudspeakers yields a virtual source between the loudspeakers. By imposing different gains or delays to the loudspeaker signals, the virtual source can be shifted towards the stronger or the leading signal. In basic amplitude panning, only the gains of two otherwise identical loudspeaker signals are varied. However, due

to a difference in the distance to the left and right ear, signals from two symmetrically placed loudspeakers have phase differences at the positions of the two ears. At low frequencies, the coherent summation of the loudspeaker signals results in an ITD depending on the relative levels of the loudspeaker signals.

In the absence of the head, the ITD at the positions of the ears can be shown to suggest a localization according to the traditional *sine law* (Blumlein, 1931; Bauer, 1961)

$$\frac{\sin\theta}{\sin\theta_0} = \frac{g_1 - g_2}{g_1 + g_2}\,, \tag{5.1}$$

where θ is the azimuth of the virtual source and g_1 and g_2 are the gains of the loudspeakers located at θ_0 and $-\theta_0$, respectively. However, if the listener turns his/her head following the virtual source, the *tangent law*

$$\frac{\tan\theta}{\tan\theta_0} = \frac{g_1 - g_2}{g_1 + g_2} \tag{5.2}$$

gives a better approximation for the direction of the virtual source (Bernfeld, 1973). Bennett *et al.* (1985) also showed that at low frequencies, when the diffraction of sound around the head is taken into account, the tangent law is more accurate than the sine law also for a fixed head. Pulkki (1997) has later reformulated and generalized the tangent law using computation of the gain factors with vector algebra. In the resulting *vector base amplitude panning* (VBAP) algorithm, the panning is performed between adjacent loudspeaker pairs for two-dimensional (2-D) panning or within non-overlapping loudspeaker triplets for three-dimensional (3-D) panning.

The localization of amplitude-panned virtual sources has also been studied psychophysically. Although the tangent law appears to give a reasonable approximation for the direction of a virtual source created by two loudspeakers symmetrically around the median plane, using loudspeakers in other directions may introduce bias into the localization of the virtual source (e.g., Theile and Plenge, 1977; Pulkki and Karjalainen, 2001; Pulkki, 2001b; see also Pulkki, 2002a). For a listener outside the sweet spot, the virtual sources are displaced towards the nearest loudspeaker, according to the precedence effect. Furthermore, at high frequencies, the ILD cues may suggest slightly different directions compared to the low-frequency ITDs (Pulkki and Karjalainen, 2001). Due to the inconsistency in the binaural cues, the perceived width of the amplitude-panned virtual sources also depends on the panning direction and the panned signal (see Pulkki *et al.*, 1999b; Pulkki, 1999). Furthermore, amplitude panning creates some timbral artifacts (Pulkki, 2001a; Ono *et al.*, 2001, 2002). In any position with different distances to the loudspeakers (which is always the case at least for one ear of the listener), summation of the two coherent signals with different delays results in comb filtering.

Based on the previous considerations, amplitude panning cannot be claimed to yield perceptually authentic virtual sources. However, the binaural decoloration (see Section 3.5.1) can be assumed to operate for amplitude panned sources, and reverberation of the listening room has been shown to reduce the resulting coloration (Pulkki, 2001a). Furthermore, the closer the loudspeakers are to each other, the smaller the potential reproduction errors are. The perceptual accuracy of amplitude panning thus increases with the number of loudspeakers in a reproduction setup.

5.2.2 Microphone techniques

In multichannel microphone recording, each microphone signal is typically reproduced with a single loudspeaker. Lipshitz (1986) divides analogous stereo microphone configurations into coincident, quasi-coincident and spaced setups. In coincident setups, a group of directional microphones are positioned as close to each other as possible, such that the sound from a single source is ideally captured in the same phase by all microphones. The level at which the sound from a single direction is recorded by each microphone depends thus on the directivities of the microphones and the subsequent reproduction resembles amplitude panning.

Ideally, the microphones in a coincident setup should have orientations and directivities corresponding to the loudspeaker configuration, so that non-diffuse sound from any direction would only be picked up by a few microphones and subsequently reproduced by a few loudspeakers close to the correct direction. Hence, using more loudspeakers would require narrower directivity patterns for the microphones. However, with conventional microphone technology, narrow enough broadband patterns cannot be achieved. The insufficient directional resolution results in sound from any direction always being reproduced by several loudspeakers. Consequently, the created virtual sources are spatially and timbrally less accurate than the loudspeaker system's reproduction capability using, for instance, amplitude panning (Pulkki, 2002b; Pulkki and Hirvonen, 2005). Furthermore, the wider the directivity patterns of the microphones are, the higher the correlation between the reproduced loudspeaker signals.

Ambisonics (Gerzon, 1973) was already mentioned in Section 2.3.2 as related to B-format recording. It can be seen as a a special form of coincident microphone techniques, where the spherical harmonic decomposition of a sound field is applied to reproduction. In theory, Ambisonics can accurately reproduce the physical sound field in a small sweet spot by the sympathetic operation of all loudspeakers. However, microphone technology limits the order of the B-format recording and, consequently, the attainable directional resolution and the size of the sweet spot[1]. The region that can be considered physically accurate in first-order reproduction is smaller than the size of a human head apart from frequencies well below 1 kHz (Poletti, 2005). Hence, in practice, the technique reduces to systematic use of a set of virtual coincident microphones that can be adjusted in the post-processing phase. The perceptual problems are also similar to those discussed above (see Pulkki, 2002b; Pulkki and Hirvonen, 2005).

In contrast to coincident techniques, spaced microphones are positioned at considerable distances from each other. The sound from a single source is thus captured in different phases by different microphones. The resulting reproduction is less sensitive to the location of the listener than that of coincident techniques. However, for most source directions, the spaced microphone techniques do not provide correct

[1] In spherical harmonic decomposition of the sound field, the directional resolution and the physical accuracy as a function of distance from the origin of the decomposition are, indeed, linked through the order at which the series is truncated. Approaching the origin, within a volume not including sound sources, components with increasingly high order (high directional resolution) have increasingly small magnitude as determined by the spherical Bessel functions of the first kind. The order needed for a physically accurate reconstruction of the sound field grows as a function of the product of frequency and the distance from the origin (Williams, 1999, Chapter 6; see also Rafaely, 2005; Poletti, 2005).

localization at any listening position (Lipshitz, 1986). Nevertheless, in a reverberant environment, the microphone signals will be decorrelated to a degree depending on the diffuseness of the sound field and the distance between the microphones (see Section 2.6.2). Based on earlier discussion, the decorrelation is known to contribute to the auditory source width (ASW) and listener envelopment (LEV). Indeed, spaced techniques have been found to yield, for instance, higher source and room width, source distance, envelopment, and room size (Berg and Rumsey, 2002)[2] than coincident techniques. The listeners of Berg and Rumsey (2002) also indicated preference for the spaced microphone techniques (see also Zacharov and Koivuniemi, 2001b).

The reported preference should not be interpreted as indisputable proof for a more authentic reproduction with spaced microphone techniques. In concert hall acoustics, halls yielding a high ASW and LEV have also been found to rank highly in preference (e.g., Hidaka *et al.*, 1995; Beranek, 1996). Hence, the high decorrelation of the loudspeaker channels could be a preferable artifact of the spaced microphone techniques. However, as mentioned earlier, insufficient directional resolution in coincident techniques produces too high correlation between the microphone channels, which suggests an insufficient reproduction of spatial impression. Quasi-coincident microphone techniques can be seen as one compromise between creating correct localization and spatial impression. Quasi-coincident microphones are placed close to each other with the inter-microphone distances typically comparable to the interaural distance. The characteristics of the reproduction also lie in between those of the coincident and spaced microphone techniques (see Pulkki, 2002b; Pulkki and Hirvonen, 2005, and for some related design principles, Williams and Le Dû, 1999, 2000; Williams, 2002, 2003).

Instead of reverting to spaced or quasi-coincident microphone techniques, which compromise the directional accuracy of the reproduction, recent developments in microphone array technology have also made it possible to somewhat increase the directional resolution in coincident multichannel recording. Higher-order B-format (Ambisonics) recording was already discussed in Section 2.3.2. Since the first implementation of SIRR (Merimaa and Pulkki, 2003), Laborie *et al.* (2004a,b) have also developed high directional resolution microphone systems tailored for 5.0 recording. Although promising, the high number of necessary microphones increases the costs of the microphone systems, and these optimized arrays are limited to 5.0 reproduction. Furthermore, the considerable amount of existing recordings and room response measurements with lower directional resolution makes it still worthwhile to investigate alternative methods, such as SIRR, for increasing the perceptual directional accuracy.

For the sake of completeness, wave field synthesis (Berkhout *et al.*, 1993) should also be mentioned. Wave field synthesis is based on physical principles aiming at a perfect reconstruction of a recorded sound field. However, to achieve this goal, a large microphone array and a high number of loudspeakers are needed. The limitation of this chapter to small multichannel loudspeaker systems rules out the possibility of using wave field synthesis. Nevertheless, with SIRR it would be possible to render

[2]Note that in the coincident setup of Berg and Rumsey (2002), only the microphones used for the three front channels of a 5.0 loudspeaker system were coincident, so the coincident results may partly reflect spaced techniques.

room responses measured with lower directional resolution for reproduction with wave field synthesis.

5.3 SIRR algorithm

Based on the previous discussion, it is desirable to increase the directional resolution of sound recording and reproduction, preferably such that the results can be auralized using any reproduction technique. It will be seen that, within certain limits, the SIRR method is able to do this using a room impulse response measurement with a lower directional resolution. SIRR can be formulated based directly on the psychoacoustical considerations in earlier parts of this thesis. The underlying assumptions are first presented in Section 5.3.1. The analysis and synthesis procedures operating according to these assumptions are described in Sections 5.3.2 and 5.3.3, respectively. The method is further illustrated with an application example in Section 5.3.4, followed by discussion in Section 5.3.5.

5.3.1 Assumptions

The first fundamental assumption for SIRR processing has already been stated several times: It is assumed that it is not necessary to perfectly reconstruct the physical sound field in order to be able to reproduce the auditory perception of an existing performance venue. Instead of sound field reconstruction, SIRR aims at as authentic recreation of localization of auditory events, spatial impression, and timbre as possible. As follows from Chapter 3, these features are known to depend on the following time- and frequency dependent properties of the ear input signals:

- ITD, ILD, and monaural localization cues
- IC
- Short-time spectrum

The most straightforward analysis method would, of course, be to measure these properties from the ears of a real listener or a dummy head in a desired sound field. However, the translation from the analyzed auditory cues to multichannel reproduction cannot be solved so easily. Furthermore, direct analysis and synthesis of binaural cues suffers from individual differences in HRTFs.

An easier way to approach the perceptual reproduction is to analyze and synthesize physical properties of the sound field that transform into the auditory cues. More specifically, it is assumed that:

1. Direction of arrival of sound transforms into ITD, ILD, and monaural localization cues.
2. Diffuseness of sound transforms into IC cues.
3. Together with the localization cues, the short-time spectrum in a measurement position determines the short-time spectra at both ears.

The validity of Assumption 1 was already verified in Section 3.2.1. Coherence related to diffuse or partly diffuse sound fields was also discussed in Sections 2.6.2 and 3.5.2 and Chapter 4. Moreover, Assumption 2 will be investigated with an auditory model simulation in Section 5.4.2. For a single static source (IC = 1), assumption 3 will also be correct. A lower IC, on the other hand, corresponds to different time-dependent fluctuations of the spectra at the two ears. If the analysis is performed within the temporal and spectral resolution of human hearing, it can be further assumed that the details of these fluctuations cannot be resolved but the perception is rather characterized by the overall short-time spectra at the ears. Thus the fourth assumption is:

4. When direction of arrival, diffuseness, and spectrum of sound are reproduced within the temporal and spectral resolution of human hearing, the reproduction is perceptually authentic.

Furthermore, the intended application of SIRR in convolving reverberators implies one more assumption:

5. If a room response is reproduced according to assumption 4, the perception of a sound signal convolved with the (multichannel) room response corresponds to the perception of the the signal presented in the room with the sound source used in the measurement of the response[3].

It is impossible to generally validate assumption 4 for all possible cases. Instead, assumptions 4 and 5 are tested together in two listening experiments presented in Sections 5.5 and 5.6. If the listeners cannot perceive the difference between an original sound field and SIRR reproduction, it can be presumed that assumptions 1–3 as well as 4 and 5 together are valid for the purposes of the current reverberator application.

5.3.2 SIRR analysis

SIRR analysis can be, in principle, performed with any multichannel microphone system and technique suitable for estimating the direction of arrival and diffuseness of sound within a sufficient frequency range (see Chapter 2). The analysis used in this chapter is based on energetic analysis of sound fields using B-format microphone measurements. This choice is motivated by the commercial availability of first-order B-format microphones. For time-frequency processing, a short-time Fourier transform (STFT) based scheme has been adopted, as is commonly done in audio coding. Similar processing could also be realized using an analysis-synthesis implementation of an auditory filter bank. However, Baumgarte and Faller (2003) found the computationally more efficient STFT implementation to perform equally well with an auditory filter bank in their experiments with the Binaural Cue Coding algorithm sharing some features with SIRR.

The analysis part of SIRR is illustrated in Figure 5.1. The input signals are first

[3]The directivity of the sound source may also affect the perception, so the measurements need to be performed using a sound source with desired directional characteristics.

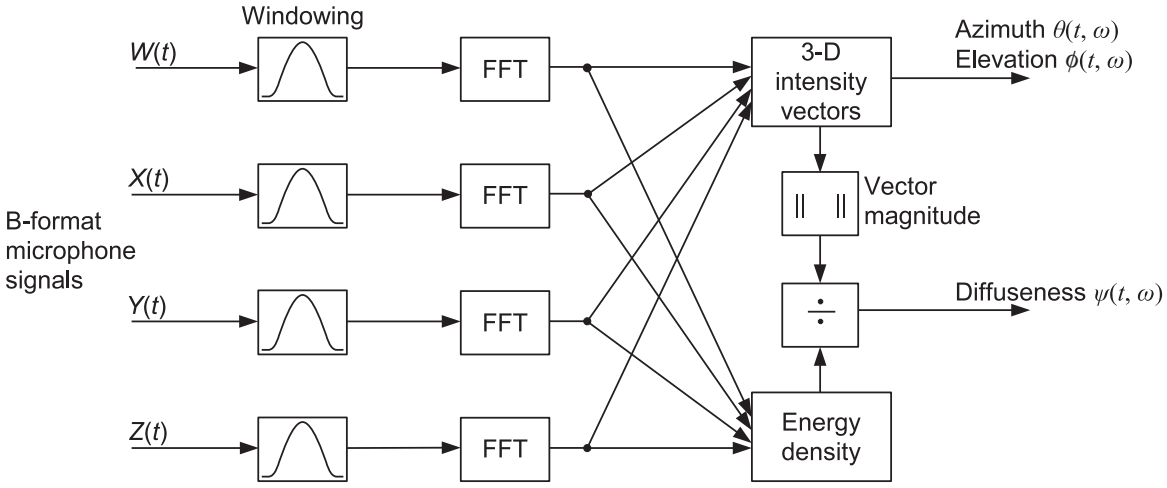

Figure 5.1: Directional energetic analysis of a B-format room response.

divided into short overlapping time windows, and the frequency distributions of the active intensity, energy density, and diffuseness are computed according to the equations derived in Section 2.4.8. The azimuth $\theta(t,\omega)$ and elevation $\phi(t,\omega)$ of the direction of arrival as a function of the time frame t and angular frequency ω are further computed as the opposite of the direction of the active intensity vector

$$\theta(t,\omega) = \begin{cases} \tan^{-1}\left[\dfrac{I_y(t,\omega)}{I_x(t,\omega)}\right], & \text{if } I_y(t,\omega) \geq 0 \\ \tan^{-1}\left[\dfrac{I_y(t,\omega)}{I_x(t,\omega)}\right] - 180°, & \text{if } I_y(t,\omega) < 0 \end{cases} \tag{5.3}$$

and

$$\phi(t,\omega) = \tan^{-1}\left[\frac{-I_z(t,\omega)}{\sqrt{I_x^2(t,\omega) + I_y^2(t,\omega)}}\right], \tag{5.4}$$

where $I_x(t,\omega)$, $I_y(t,\omega)$, and $I_z(t,\omega)$ are the components of the active intensity in the directions of the corresponding Cartesian coordinate axes.

As shown in Chapter 2, microphone systems often pose limits to the usable frequency range of the energetic analysis. At high frequencies, the direction of intensity vector may be systematically biased due to spatial aliasing. With microphone pair measurement techniques, the magnitude of the active intensity will also be estimated too low, yielding too high diffuseness estimates above the upper frequency limit. Furthermore, measurement errors at low frequencies may produce arbitrary values for both the direction of arrival and diffuseness. However, from a psychoacoustical point of view, erroneous analysis and subsequent reproduction of very low and very high frequencies is not a serious drawback. According to the discussion in Sections 3.3.2, 3.5.2, and 3.6.1 very low and high frequencies are less important for human localization and spatial impression. If desired, it is also possible to extrapolate the analysis data for low and/or high frequencies, which often yields approximately correct values for discrete room reflections.

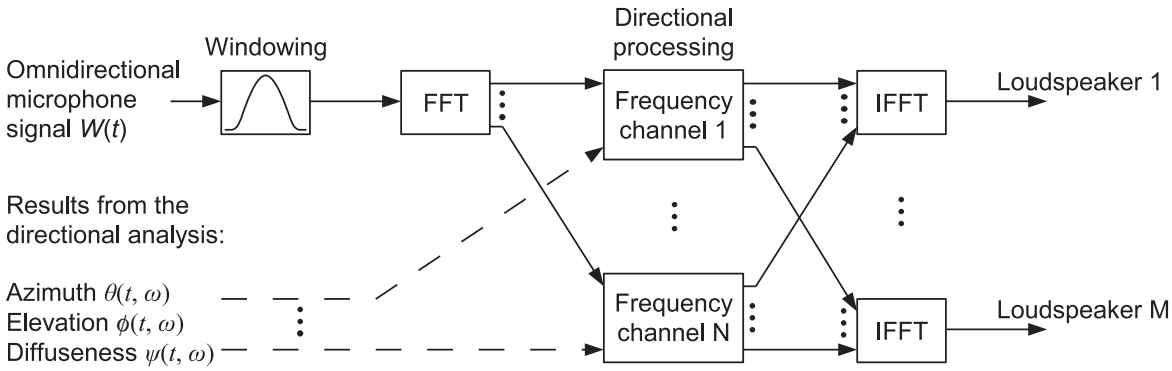

Figure 5.2: Directional synthesis based on an omnidirectional room response: time-frequency processing.

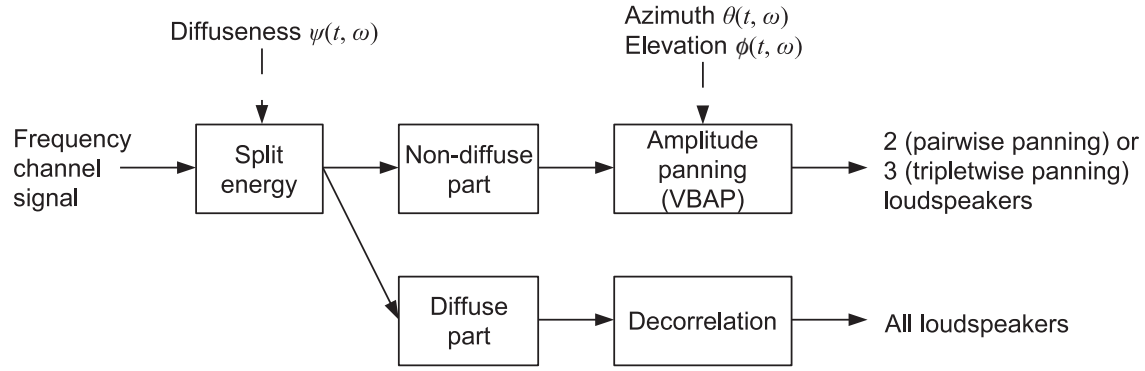

Figure 5.3: Directional synthesis based on an omnidirectional room response: processing of a single frequency channel.

5.3.3 SIRR synthesis with a multichannel loudspeaker system

The SIRR synthesis is based on manipulating an omnidirectional room response according to the analysis data. For this purpose, either the W signal from the B-format microphone measurement or another response measured with a desired microphone in the same position can be used. The omnidirectional response is processed with STFT using the same time-frequency resolution as in the directional analysis, and each time-frequency component is distributed to the loudspeaker channels. The time-frequency processing is illustrated in Figure 5.2, and the processing steps for a single frequency channel are shown in Figure 5.3. The energy of each incoming time-frequency component is first split into non-diffuse and diffuse parts according to the diffuseness estimate $\psi(t,\omega)$. The non-diffuse part of the omnidirectional signal $\sqrt{1-\psi(t,\omega)}W(t,\omega)$ is reproduced as accurately as possible from the correct direction, whereas the diffuse part is created by distributing the total diffuse energy $\psi(t,\omega)W^2(t,\omega)$ in a decorrelated form uniformly around the listener.

Non-diffuse synthesis

For the small multichannel loudspeaker setups considered in this chapter, amplitude panning using the VBAP algorithm (see Section 5.2.1) has been chosen as the method for the non-diffuse synthesis. Although some problems of amplitude panning were pointed out in Section 5.2.1, VBAP provides a good and robust solution for multichannel reproduction with a relatively low number of loudspeakers, and its precision increases with an increasing number of loudspeakers. SIRR is, however, not limited to amplitude panning. With a high number of loudspeakers, some other methods, e.g., higher-order Ambisonics[4] or wave field synthesis could provide better directional quality over a larger listening area. Furthermore, it would be possible to use HRTF filtering to position the time-frequency components in the correct direction in binaural headphone or loudspeaker listening.

The non-diffuse synthesis is performed separately for each time-frequency component. For a single time frame, this corresponds to deriving different linear (zero) phase filtered versions of the omnidirectional signal for each loudspeaker with the filters changing from one time frame to another. The actual implementation of the synthesis requires some care. To begin with, the linear phase filtering spreads the signal in time and may result in time domain aliasing. The aliasing can be prevented by zero-padding each time frame before performing the analysis and the overlap-add synthesis, as well as by smoothing the changes in the panning directions as a function of frequency, if necessary. Depending on the applied window function, the switching of consequent time frames at a single frequency from one loudspeaker to another may also produce audible clicks. If needed, the switching effects can be smoothed, for instance, by oversampling, i.e., by using time frames that are more overlapping than needed for perfect reconstruction. Several related implementational issues have been discussed in more detail by Faller and Baumgarte (2003).

Diffuse synthesis

The purpose of the diffuse synthesis is to recreate the reduced interaural coherence produced by the diffuse sound energy. With diffuse sound, the analyzed directions of arrival behave in a stochastic manner as a function of time and frequency. Applying the omnidirectional signal to such random directions (i.e., treating all sound as non-diffuse in the synthesis) results in some decorrelation if the analysis and synthesis are done at shorter time frames than the temporal resolution of the auditory system. In the context of diffuse synthesis, this decorrelation method will be denoted as *diffusion with only amplitude panning*. During the development of SIRR, reproductions of virtual reality impulse responses were often compared to the original samples, as explained in more detail in Section 5.5. According to informal testing, diffusion with only amplitude panning indeed produces a fairly authentic spatial impression. However, some temporal artifacts and instability were audible. Furthermore, at high frequencies the reproduced late reverberation was sometimes perceived shorter than the reference. According to the binaural cue selection model (see Chapter 4), the

[4]Note that although higher-order Ambisonics recording is difficult, it is possible to synthesize corresponding signals up to an arbitrarily high order.

instability of localization could be due to too high instantaneous coherence values[5]. Hence, more sophisticated diffusion methods seem to be required.

An alternative decorrelation technique involves designing specific decorrelation filters. The diffuse part of the omnidirectional response is then filtered with a different filter corresponding to each loudspeaker channel. Several filter structures have been proposed in the literature, including allpass filters (Gerzon, 1992; Breebaart *et al.*, 2004; see also Bouéri and Kyriakakis, 2004), rectangular noise bursts created in the frequency domain (Kendall, 1995), and filters modeling late reverberation (Faller, 2003). We have been experimenting with convolution of the omnidirectional response with short exponentially decaying noise bursts equalized to a flat magnitude spectrum with a minimum phase filter, as proposed by Hawksford and Harris (2002). This technique is denoted as *convolution diffusion*.

The convolution diffusion provides control over the decorrelation, as it is possible to determine the shape of the temporal spreading in the decorrelation. Furthermore, by adjusting the time constant of the exponentially decaying noise bursts, it is possible to modify the spatial impression. Short bursts appear to yield fairly transparent results, whereas using longer bursts makes the response sound more diffuse or "wet." However, although the bursts are equalized to a flat magnitude spectrum, the convolution diffusion may result in timbral artifacts. Since the filters are fixed, the phase relations of the decorrelated loudspeaker channels are always the same, and the summation of the loudspeaker signals at a listening position may produce spectral artifacts perceived as coloration. The problem is especially noticeable at low frequencies when using short noise bursts.

A third decorrelation method used in SIRR is denoted as *phase randomization*. The phase randomization is performed by creating continuous uncorrelated noise for each loudspeaker, and by setting the magnitude spectrum of each channel in each time window equal to the magnitude spectrum of the diffuse energy calculated for the corresponding loudspeaker. The method thus corresponds to replacing the diffuse parts of the room response with random noise samples with a similar time-frequency envelope. The phase randomization can create highly decorrelated signals and the time-variant nature of the algorithm avoids most timbral artifacts that might be caused by continuous coherent summation of the signals from different loudspeakers. However, the frequency domain equalization of the magnitudes in the STFT processing may again result in time domain aliasing, which manifests itself as nonlinear distortion.

The properties of the diffusion techniques will be investigated in more detail in Section 5.4, where a hybrid method is also derived.

5.3.4 Application example

In Figure 5.4, the SIRR analysis data are illustrated for a 100 ms sample of a concert hall response starting approximately 5 ms before the arrival of the direct sound. The response was measured in the Pori Promenadikeskus concert hall with an omnidirectional sound source on the stage and a Soundfield MKV microphone system at

[5]As mentioned earlier, the subjective experiences of unstable localization with early versions of SIRR and similar observations of Christof Faller related to the Binaural Cue Coding method actually led the two authors to develop the binaural cue selection model.

the raised rear part of the floor (response `s1_r4_sf` from the database of Merimaa *et al.*, 2005b). The topmost panel is for reference, presenting the envelope of the omnidirectional response $W(t)$. The two middle panels illustrate the time-frequency distribution of the active intensity, and the bottom panel shows the time-frequency distribution of the diffuseness estimate.

The format of the middle panels is identical to that described in Section 2.5. The vectors represent the direction and logarithmic magnitude of the propagation of sound energy, plotted on top of a sound pressure spectrogram calculated from the omnidirectional response W. The data are shown for a frequency range from 100 Hz to 5 kHz, which appears to give reasonable results with the utilized microphone system. The time-frequency representations have again been adjusted for illustrative purposes. The diffuseness and spectograms were calculated from 128-sample-long Hann windowed time frames at 48 kHz sampling frequency with zero-padded FFT and largely overlapping windows. The intensity vectors are plotted only for positions of local maxima within each frequency band in order to reduce the amount of shown data. Furthermore, the spectrograms and the intensity vectors have been thresholded such that only a range of 25 dB from the maxima is shown.

What can be seen from the analysis data is that the direct sound arrives at approximately 65 ms from a little below the horizontal plane. It is followed by two slightly bandlimited reflections close to each other: one from slightly above on the right side at approximately 80 ms, and one from the left at 85 ms. For these sound events, the diffuseness estimate shows low values across frequencies. Due to the highly diffuse design of the hall, the subsequent discrete reflections are already more constrained in frequency, and in the late part it is impossible to identify any single reflections.

The diffuseness estimate appears fairly random apart from the first few discrete events (see the white areas in the bottom panel corresponding to the dark areas in the middle panels). The plot includes a lot of statistical fluctuations during low energy parts and even during the measurement noise before the arrival of the direct sound, which makes visual inspection of the diffuseness difficult. However, this is the form in which the diffuseness data are used in the synthesis. Furthermore, plotting the data only for high-energy time-frequency components does not allow an improved visualization. On the contrary, such thresholding would make the visualization appear discontinuous because the diffuseness estimate would not have a constant value at the threshold.

A set of impulse responses computed with the SIRR algorithm from the previously illustrated analysis data is shown in Figure 5.5. The responses were computed for reproduction with a 5.0 loudspeaker system (ITU-R BS.775-1, 1992–1994). The topmost axis shows the same omnidirectional response W as depicted in the envelope and spectrogram plots in Figure 5.4. The five lower axes present the synthesized loudspeaker responses for the left surround (LS), left (L), center (C), right (R), and right surround (RS) speakers, respectively. The responses were processed at 48 kHz sampling rate using 128 sample (1.3 ms) Hann windowed time frames with altogether 128 samples of zero padding before and after the time window. For the diffuse synthesis, the phase randomization method was used. In order to realize 2-D reproduction, only the directional analysis data in the horizontal plane were utilized. However, the diffuseness estimate still included the vertical component. Furthermore, the direct sound was

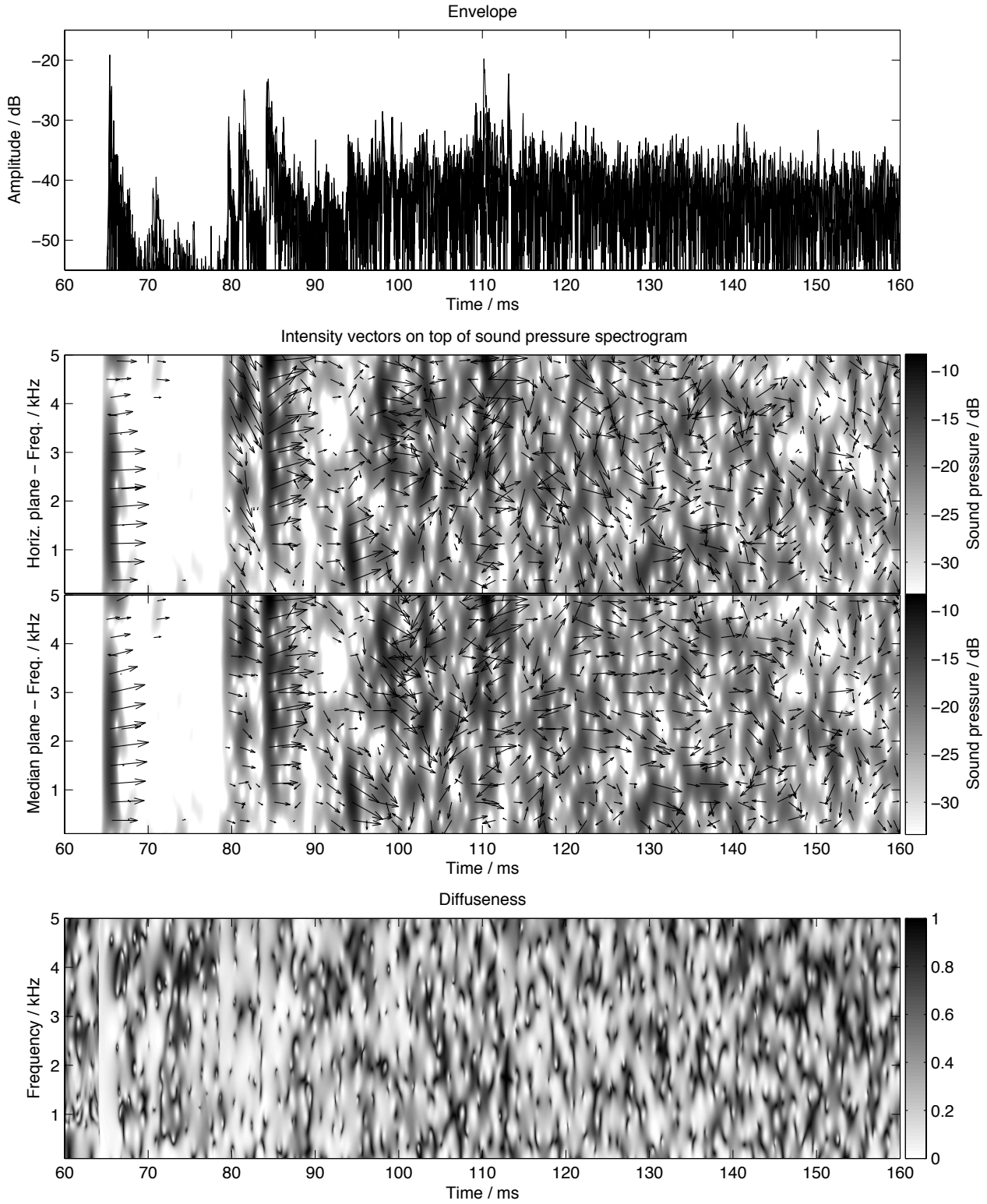

Figure 5.4: Analysis results of a concert hall response. Top: Hilbert envelope of the omnidirectional response. Middle panels: time-frequency distribution of active intensity vectors plotted on top of a sound pressure spectrogram. Bottom: time-frequency representation of the diffuseness estimate.

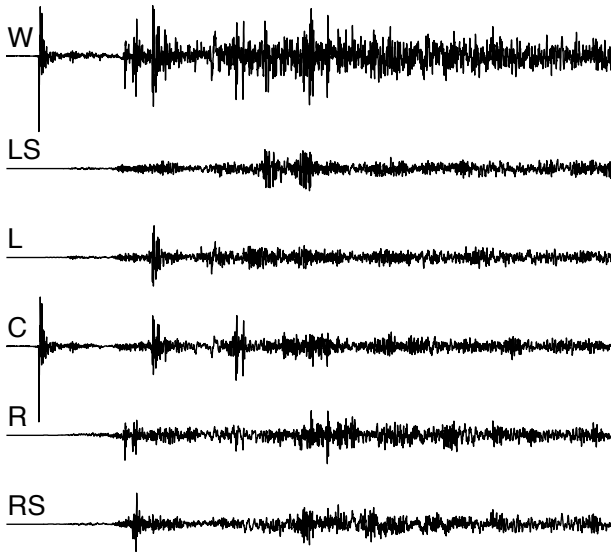

Figure 5.5: An omnidirectional impulse response W rendered to a 5.0 loudspeaker setup. The amplitudes of the responses are plotted as a function of time.

forced to the center speaker by modifying the analysis data before the synthesis. It can be seen that a major part of the first two reflections is conveyed to the rightmost loudspeakers, followed by a sound event from the front left, as expected based on the analysis data in Figure 5.4. The subsequent part is also distributed unevenly to the loudspeaker channels, whereas the late reverberation does not show large (statistical) deviations between the channels.

5.3.5 Discussion

Compared to established microphone techniques, the SIRR method can be characterized as follows. The reproduction of non-diffuse sound resembles coincident microphone techniques, where SIRR can be thought to adaptively narrow the microphone beams according to a chosen loudspeaker setup in order to achieve best possible directional accuracy. On the other hand, diffuse sound is reproduced as decorrelated, which is similar to spaced microphone techniques.

The time resolution used in the application example is actually considerably higher than that of the human hearing (see Section 3.4.2). Consequently, the analysis and synthesis bandwidth is larger than the critical bands at low frequencies. However, for room responses, these settings appear to yield good subjective results. The higher the time resolution, the more accurately the perceptually important early reflections can be individually processed. The importance of high time resolution will also be seen later in the assessment of SIRR's performance.

The motivation of SIRR starts from the psychoacoustical principles outlined in Section 5.3.1. Interestingly, after the publication of some first papers on SIRR, it was found that Farina and Ugolotti (1999) have independently proposed the concept of a very similar method based on theoretical considerations of energetic analysis and the principles of Ambisonics. However, their method was never developed to the stage

of a practical implementation including, for instance, the techniques for the diffuse synthesis. A significant difference is also that Farina and Ugolotti did not divide the B-format microphone signals into frequency bands, which is essential from the psychoacoustical point of view and has proved to be an important part of SIRR.

From earlier description, it is also clear that SIRR provides efficient means for the modification of room responses. In typical cases, it is possible to individually change the directions of arrival of some early reflections by simply modifying the corresponding analysis data prior to the synthesis. Similarly, the whole sound field can be rotated and the time-frequency envelope of the response can be easily weighted. Furthermore, the balance between diffuse and non-diffuse sound energy can be adjusted, and the parameters of the diffuse synthesis can be modified. Although such modifications of the analysis data generally reduce the perceptual authenticity of the reproduction, they may, nevertheless, be desirable for artistic purposes.

5.4 Diffuse synthesis

The three possible diffusion techniques introduced in the previous section are 1) diffusion with only amplitude panning, 2) phase randomization, and 3) convolution diffusion. As discussed, the results of using these three methods differ both in timbral and in spatial perception. In this section, the methods are investigated in more detail with simulations. Section 5.4.1 describes simulations related to the timbral effects and presents a hybrid diffusion method. Furthermore, the transformation of the diffuseness of the sound field into interaural coherence is studied with an auditory model in Section 5.4.2.

5.4.1 Timbral effects

As discussed in Section 3.5.1, the perceived timbre depends not only on the spectrum of sound but also on temporal properties and on the interaural differences. The effects of the temporal properties and interaural differences on timbre are not known well enough to be modeled. However, the overall spectral difference between an original and a reproduced sound field can be easily assessed. Thus, as a simplistic approximation of the timbral effects, the spectrum of the sum of the loudspeaker signals with different diffusion techniques was simulated and compared to the original omnidirectional spectrum. It is assumed that major timbral problems can be caught this way.

A measured B-format impulse response of a performance hall was used as the test case, and reproduction with a standard 5.0 loudspeaker setup (ITU-R BS.775-1, 1992–1994) in an anechoic environment was simulated at three listening positions. The distance between the center (best listening position) and the loudspeakers was set at 2.0 m, and the simulated listening positions were 0.2, 0.4, and 0.6 m right from the center. The best listening position was not used, since it would have yielded unrealistically good results for some of the systems. In reality, however, it would not be possible to perform the listening exactly in the best listening position, since the ears of a listener are not located on a single point. The resulting magnitude spectra in these

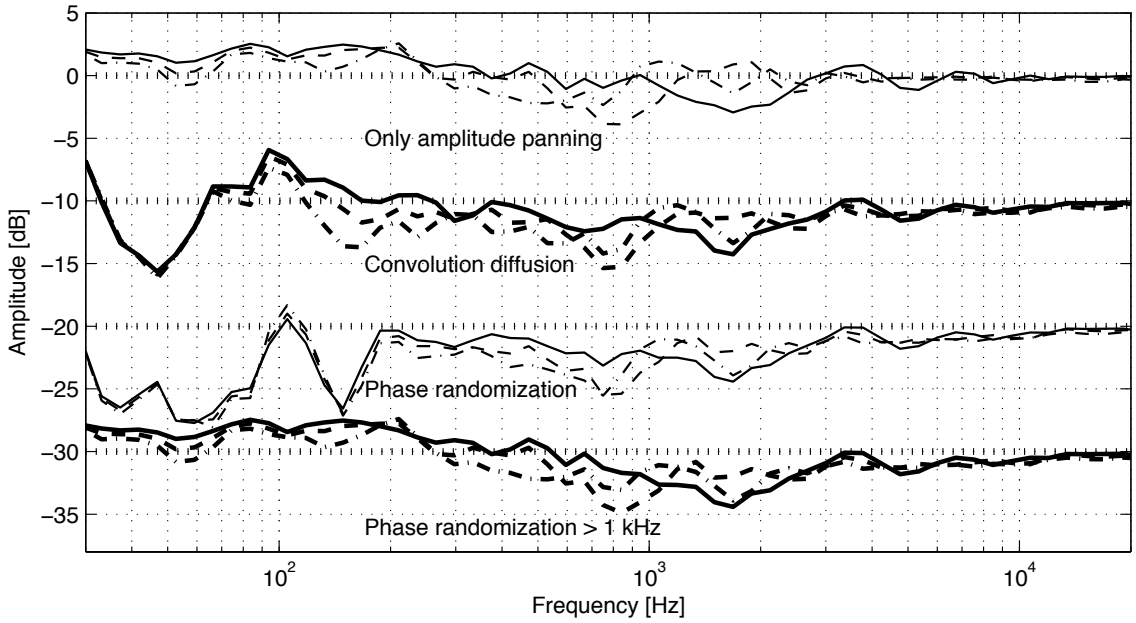

Figure 5.6: Deviations of the magnitude spectra of the sum of SIRR responses from a target response in 5.0 reproduction. The SIRR responses were computed with different diffusion methods for three listening positions located 0.2, 0.4, and 0.6 m right from the best listening position. The corresponding line types are solid for the 0.2 m, dashed for the 0.4 m, and dash-dotted for the 0.6 m case. The "phase randomization $>$ 1 kHz" curves are explained later on page 116.

positions were computed with FFT. The summed responses were smoothed with 1/3 octave band resolution (roughly approximating the frequency resolution of human hearing) and a similarly smoothed spectrum of the original omnidirectional target response was logarithmically subtracted from the summed responses. The results are shown in Figure 5.6, with the different SIRR implementations offset by 10 dB.

The topmost responses in Figure 5.6 correspond to SIRR with only amplitude panning. As described earlier, in these cases, the diffuse sound is actually not treated any differently from the non-diffuse sound. The simulated magnitude response varies between ± 3 dB from the reference, where the $+3$ dB amplification at low frequencies is due to coherent summation of the different loudspeaker channels in the power-normalized amplitude panning. Overall, the response is moderately smooth, suggesting a good reproduction of timbre. However, as described in Section 5.3.3, diffusion with only amplitude panning may create temporal and spatial artifacts and is, as such, not sufficient for the purposes of SIRR.

SIRR responses utilizing the convolution diffusion are plotted second in Figure 5.6, which illustrates the problem of the method: There are prominent deviations from the target response at frequencies below about 200 Hz. Due to limited length of the utilized noise samples, low frequency sound is not decorrelated enough to prevent coherent summation of the different loudspeaker channels. However, contrary to diffusion with only amplitude panning, there are fixed phase differences between the diffuse sound emitted from different loudspeakers, which can result in notches in the combined

response. Unfortunately, the length of the samples cannot be made arbitrarily long without producing audible differences in the temporal structure of the room response. In informal testing, exponentially decaying noise samples with time constants below about 50 ms were found not to prominently change the temporal character of the responses. A time constant of 50 ms was used in the simulation.

The phase randomization method is able to avoid the coherent summation at low frequencies. Nevertheless, in the phase randomization curves in Figure 5.6, there are also deviations from the reference at low frequencies. As mentioned in Section 5.3.3, the frequency-domain equalization of the windowed noise frames spreads the signal of each frame in the time domain. In order to prevent time-domain aliasing due to moderate temporal spreading (moderate equalization), zero padding of the time frames is used in the overlap-add processing. However, a computationally practical amount of zero padding is occasionally not sufficient, and aliasing occurs unless the equalization is limited.

Non-linear distortion

Possible time domain aliasing in the phase randomization results in non-linear distortion. The spectral effect of the distortion was examined further by processing a B-format room response with SIRR after convolution with a 500-ms-long, 100 Hz sinusoidal signal. The sinusoidal was turned on and off at zero-crossings and the same B-format impulse response as in the previous spectral study was used. In the subsequent SIRR processing, a sampling rate of 96 kHz and 256 sample (1.3 ms) time windows with additional 128 preceding and following zero samples were used. Reproduction with the same 5.0 loudspeaker setup in an anechoic environment was simulated at two listening positions, 0.2 and 0.4 m to the right from the best listening position, using diffusion with only amplitude panning and phase randomization. Finally, the spectra of both the original sample and the reproductions were calculated over the whole duration of the signal and smoothed with a 1/12-octave resolution.

The results are shown in Figure 5.7. It can be seen that even the omnidirectional reference includes sound energy outside 100 Hz due to the onset and offset of the sinusoidal signal. When the phase randomization is used, there is a prominent amount of additional off-frequency sound in the signal. When only amplitude panning is used, the level of distortion is reduced compared to phase randomization. However, some distortion remains due to the introduced fast changes in the panning directions, which correspond to random frequency-dependent amplitude modulation of the loudspeaker signals. In off-center listening positions, the summation of all loudspeaker signals does not fully cancel the modulations, which can be seen as slight spectral spreading.

The aliasing problem in the phase randomization could be solved by trading off the excessive time spreading to deviations from the ideal magnitude spectrum of each time frame, i.e., by effectively smoothing the frequency responses of the equalization filters. However, related to the current application, it is important to distinguish between distortion in a room response and in the sound convolved with the room response. Possible non-linear distortion in the SIRR-processed room response does not introduce non-linear distortion in an audio signal convolved with the response. Furthermore, in the case of wideband noise (such as the synthesized diffuse sound), the

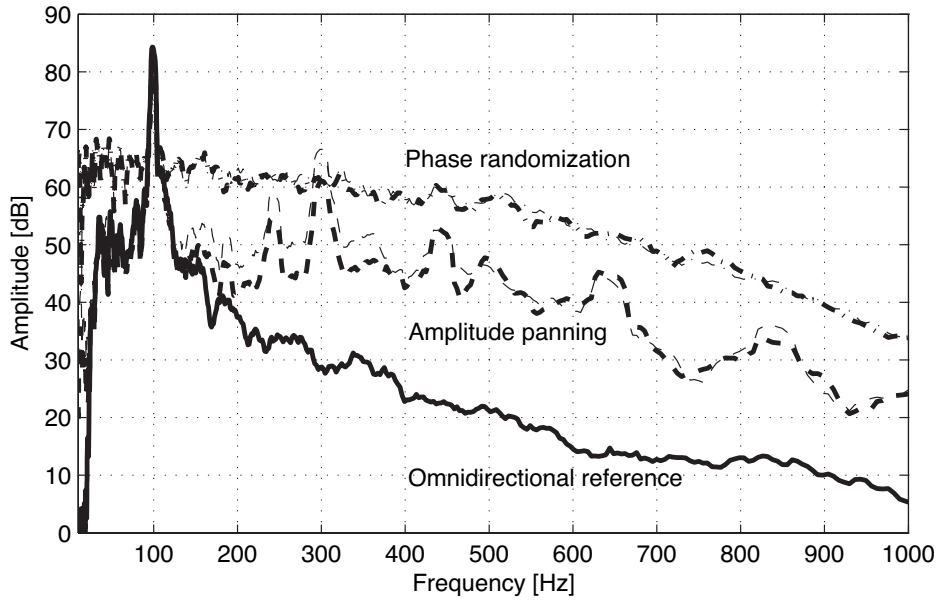

Figure 5.7: Distortion in SIRR processing. A 500-ms-long 100 Hz tone was convolved with a measured B-format impulse response before applying SIRR and the SIRR responses were computed using diffusion with only amplitude panning and phase randomization. The summed magnitude spectra are shown for the positions 0.2 m (thick lines) and 0.4 m (thin lines) right from the best listening position.

result of non-linear distortion is still noise but with a different spectral content. Thus, despite its non-linear nature, moderate distortion due to the phase randomization should only be viewed as deviations from the target magnitude response. In this sense, allowing the distortion can be considered as a design solution. Nevertheless, based on informal listening, the resulting spectral deviations seen earlier in Figure 5.6 are too large to be acceptable.

Hybrid diffusion method

Instead of limiting the equalization of the noise bursts synthesized in the phase randomization process, the spectral problems can be solved in a different way. As described earlier, diffusion with only amplitude panning does not have timbral problems, but it creates other artifacts, especially at high frequencies. A natural solution is thus to implement a hybrid method, which uses only amplitude panning at low frequencies and phase randomization or convolution diffusion at high frequencies. The frequency response of such a hybrid with phase randomization faded in between 800–1200 Hz is plotted undermost in Figure 5.6. The response follows the target well, although at low frequencies there is the positive bias of 1-3 dB caused by amplitude panning of both diffuse and nondiffuse sound. The hybrid method with phase randomization was chosen to be used in all simulations and listening tests described later in this chapter.

5.4.2 Diffuseness and interaural coherence

Assumption 2 in Section 5.3.1 stated that the diffuseness of a sound field will transform into interaural coherence cues. In this subsection, the recreation of IC with SIRR is studied using a binaural auditory model. This is done by comparing the ICs resulting from SIRR reproduction of a number of reference cases with different degrees of diffuseness to the ICs in the reference sound fields.

The investigated reference cases were simulated with a set of 36 equidistant sound sources positioned evenly around the listener in the horizontal plane, with each loudspeaker emitting independent 300 ms long white noise sequences. The level of the source at 30° azimuth was fixed and the levels of the other sources were adjusted to create different degrees of diffuseness. This roughly approximates listening to the noise emitted by the source at 30° azimuth in a reverberant environment with a sound-absorbing floor and ceiling. The root-mean-square (RMS) signal-to-noise ratio (SNR) between the nondiffuse sound (= signal of the sound source at 30°) and diffuse sound (= noise measured as sum of all other sound source signals) in the center point was varied from 30 dB to -24 dB at 6 dB steps. For the assessment of SIRR, a B-format microphone in the center of the reference loudspeaker setup was simulated. The resulting B-format responses were then processed for reproduction with a two-channel stereophonic (sources at ±30° azimuth), a standard 5.0 (ITU-R BS.775-1, 1992–1994), and a 12-channel setup using the hybrid diffusion method (see p. 116). The loudspeakers in the 12-channel system were positioned uniformly around the listener in the horizontal plane. The choice of a source at 30° azimuth was motivated by the fact that all the reproduction setups included a loudspeaker in that direction.

In the auditory simulations, the transduction of sound from the loudspeakers to the ears of the listener in the center of the reference and reproduction setups was simulated with HRTFs measured at the TKK Laboratory of Acoustics and Audio Signal Processing. The simulations were performed using the same model as in Section 4.2 (see Figure 4.1), with the exception that the IC was averaged over the full stimulus length of 300 ms. As earlier, the IC at each frequency band was computed as the maximum of the normalized cross-correlation function of the left and right ear signals. A mean value over the six sets of HRTFs and altogether 42 frequency bands in the ranges of 200–1000 Hz and 1–16 kHz was taken to get estimates of the average interaural coherence. These two frequency ranges were selected because they correspond to the frequency regions of different diffusion techniques in the hybrid method. Furthermore, the low frequency range approximately consists of frequencies where the ICs are computed from the stimulus waveforms and the high frequency range to frequencies where the coherence information is extracted from the envelopes of the left and right ear signals.

The resulting ICs as a function of SNR are shown in Figure 5.8. At low frequencies, the reference IC drops from 1.0 to 0.84 with decreasing SNR, and at high frequencies correspondingly from 0.997 to 0.976. With the stereo reproduction, the coherence does not reach the lowest values of the reference at either frequency range. At the low frequency range (diffusion with amplitude panning only), the ICs of both the 5.0 and 12-channel reproduction remain at slightly higher values than the reference, although the difference between the reference and the 12-channel reproduction is very small. At

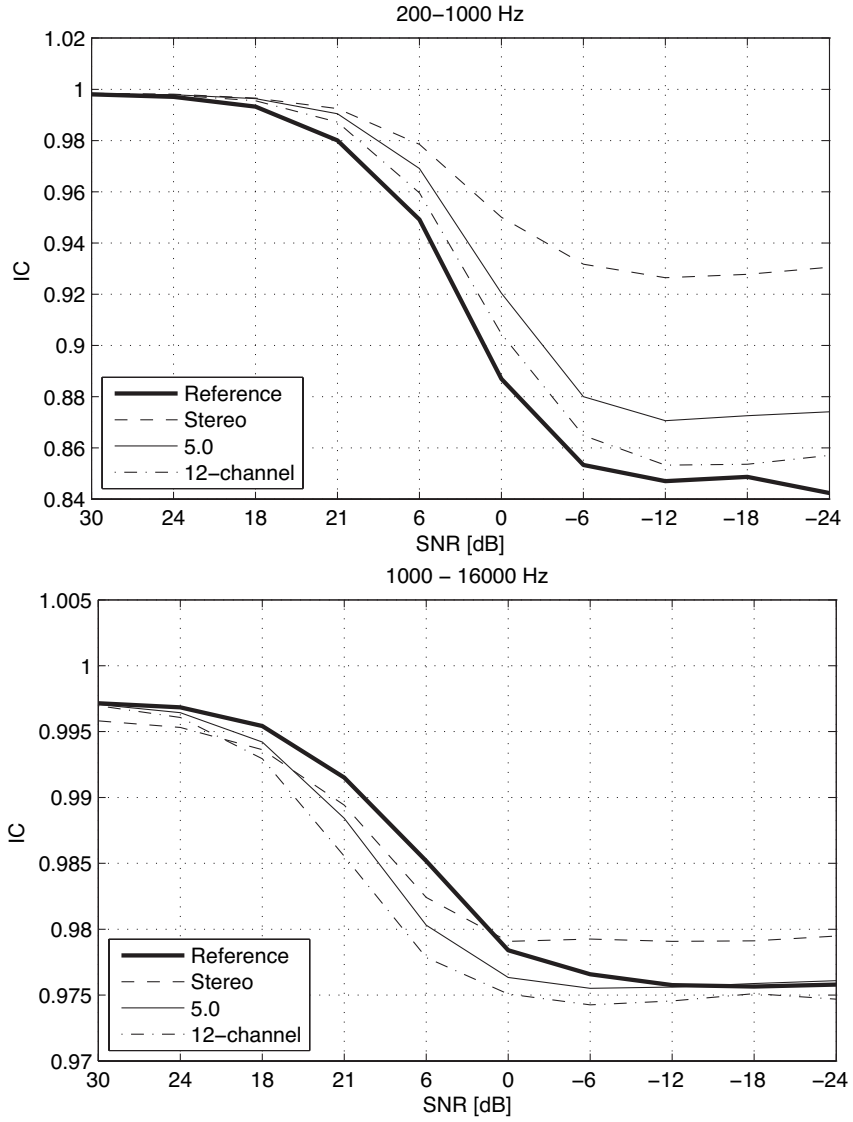

Figure 5.8: Dependence of the average IC on the ratio of nondiffuse and diffuse sound at two frequency ranges for reference cases and SIRR reproduction.

the high frequency range (phase randomization), all reproductions actually produce slightly lower IC than the reference at high SNRs. For the most diffuse cases with low SNRs, the 5.0 and 12-channel reproductions both match the reference IC.

The simulations show that the distribution of loudspeakers in the reproduction setup does indeed affect the resulting interaural coherence. This is in line with the findings of Damaske and Ando (1972) and Tohyama and Suzuki (1989), who measured interaural coherence with different loudspeaker setups, although without using a model of the inner ear. The deviations from the reference are also smaller when the target coherence is high, and the inaccuracy grows towards small target coherences. From a psychoacoustical point of view, such a trend is very desirable. It has been shown that human listeners are more sensitive to deviations from high coherence values than to deviations from low coherence (see Section 4.2.4).

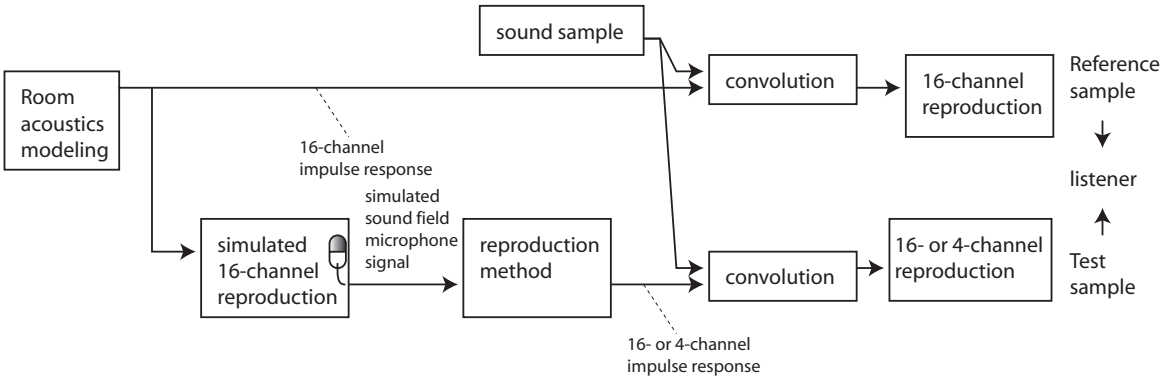

Figure 5.9: Method for investigating quality of spatial sound reproduction: A virtual reality sample is compared to its reproduction.

The simulations represent cases where the diffuse sound is distributed evenly around the listener. As discussed in Section 2.4.5, the diffuseness estimate also yields high values for sound fields such as standing waves where the sound energy is oscillating along only one dimension. In such cases, the IC in the original sound field depends on the orientation of the head of the listener, whereas the SIRR reproduction creates (within the limits of the loudspeaker setup) a direction-independent IC. However, there is no possibility to unambiguously measure such cases with a B-format microphone system. Furthermore, such cases are somewhat theoretical and, at least in large rooms, the room response will never include only a single standing wave within an analysis band (see the average frequency distribution of modes in Section 2.2.2). Thus, although in special cases the reproduction of IC may be incorrect, it can be concluded that in typical situations the hybrid diffusion method reproduces the interaural coherence fairly accurately.

5.5 Anechoic listening tests

The previous section studied parts of the SIRR algorithm with objective simulations. However, a method aiming for the perceptual reconstruction of a sound field can be fully evaluated only subjectively with listening tests. The purpose of the experiment reported in this section was to test the perceptual authenticity of the reproduction, i.e., to find out how close the SIRR reproduction can get to a reference, as well as to compare SIRR to the established Ambisonics reproduction.

In the evaluation of spatial sound reproduction, there is always the problem that the sound in a recording room cannot be directly compared to the reproduced sound in a listening room. In this study, the evaluation was done by first creating as natural-sounding virtual reality as possible and then by reproducing this virtual reality with SIRR and other techniques. This way, it was possible to present both the reference and the test samples consecutively in the same room. The procedure is illustrated in Figure 5.9. Although virtual reality responses are not fully authentic simulations of the modeled acoustical environments, they nevertheless resemble real room responses both physically and perceptually. The capability of the tested techniques to reproduce

the virtual responses should thus reflect their capability to reproduce real responses. Real measured responses will also be used in the second experiment reported in Section 5.6.

In this experiment, the listeners' task was to judge the differences between the chosen references and their reproductions. If the earlier assumptions are correct, SIRR reproduction should not be perceptually distinguishable from a reference. However, unideal properties of amplitude panning (see Section 5.2.1) or limitations of the diffuseness analysis in special cases (see Section 5.4.2) could still lead to slight degradation in the authenticity of the reproduction. It is also expected that Ambisonics will yield less authentic reproduction than SIRR.

5.5.1 Apparatus and stimuli

The test was conducted in an anechoic chamber. The stimuli were played back with a 3-D setup of 16 Genelec 1029A loudspeakers positioned as illustrated in Figure 5.10. Two listening positions were employed. The *reference listening position* was in the center of the loudspeaker setup, and it was thus expected to yield best results for the reproductions. The second position is denoted as *worst case position* chosen similarly as recommended in ITU-R BS.1116-1 (1997). The listening positions are shown in Figure 5.11.

Reference virtual reality

The reference acoustic virtual reality was created with the DIVA software[6] (Savioja *et al.*, 1999; Lokki, 2002), which models the direct sound and early reflections with the image-source method, and late reverberation statistically. The frequency-dependent air and surface absorption were modeled with digital filters. Direct sound and the modeled discrete reflections were applied to single loudspeakers, because using any multi-speaker spatialization method would have caused an abnormal response when recording the virtual environment with a microphone.

Three room geometries were chosen to serve as the reference responses: a large room with a reverberation time of 1.5 s, a medium-sized room with a reverberation time of 0.6 s, and a small room with a reverberation time of 0.3 s. With the image-source method, hundreds of image sources, resulting from reflections up to the 7th order, were computed in each case. Late reverberation was simulated using linearly rising and exponentially decaying uncorrelated noise samples reproduced from all loudspeakers. The level of the noise was fit to the early response. The linear rise began directly after the direct sound, and it was changed to exponential decay after 60 ms in the large room, 35 ms in the medium-sized room, and 15 ms in the small room. At frequencies above 4 kHz, the decay rate was faster in order to realize frequency-dependent reverberation times.

It is known that sound sources discplaced from the median plane constitute challenges to directional microphone techniques (Pulkki, 2002b). For this reason, the orientation of the listener in the chosen rooms was simulated such that the direct

[6]The modeling of the chosen acoustical environments was performed by Dr. Tapio Lokki.

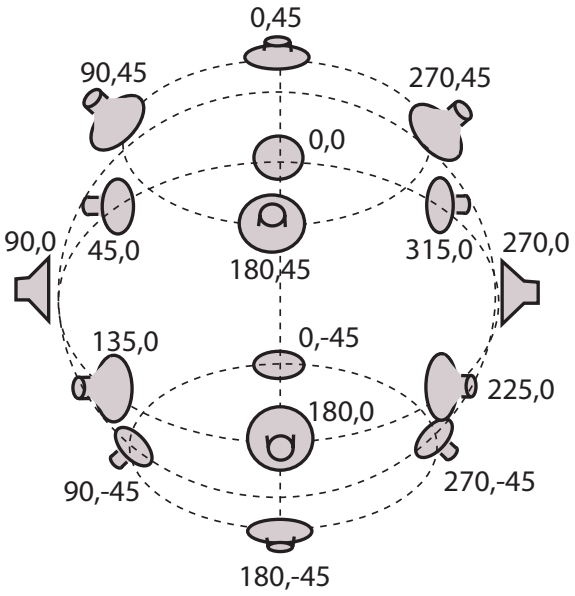

Figure 5.10: The 3-D loudspeaker system employed in the anechoic listening tests. The pairs of numbers describe the azimuth and elevation angles of each loudspeaker, respectively.

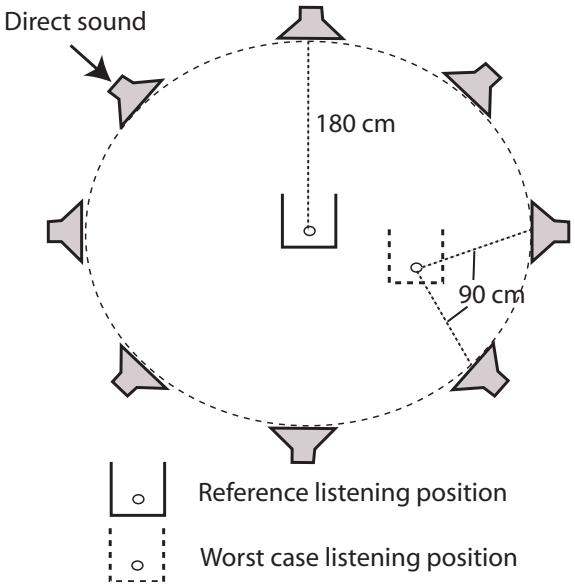

Figure 5.11: The listening positions employed in the anechoic listening tests as seen from above.

sound was emanating from an azimuth angle of 45° to better reveal the differences between the reproduction systems.

Test cases

In order to implement the test cases, the recording of the 16-channel virtual reality room responses described earlier was simulated with a B-format microphone in the center of the setup. A simulation of the recording, instead of physically measuring the impulse response of the virtual rooms produced with the 16 loudspeakers in the anechoic chamber, was preferred in order to avoid any differences in the reference and reproduction due to non-ideal properties of real B-format microphones and the loudspeakers. The virtually recorded responses were then processed with three reproduction methods:

- SIRR
- Phase randomization
- First-order Ambisonics

The SIRR responses were computed using the hybrid diffusion technique (see p. 116). The computations were performed at a sampling frequency of 48 kHz, using 1.3 ms time windows with an additional 1.3 ms of zero padding.

The phase randomization case corresponds to SIRR processing with the diffuseness set to maximum at all times. Thus, the phase randomization stimuli did not include any localization information at all, but all sound was reproduced in a decorrelated form from all loudspeakers. This case was added in order to test the importance of correct directional reproduction (or directional reproduction altogether) of the direct sound and discrete reflections.

First-order Ambisonics (see Sections 2.3.2 and 5.2.2) was selected for comparison, since it is an established technique using the same kind of microphone system as the current SIRR implementation. Contrary to SIRR and phase randomization, which used all the 16 loudspeakers in the reproduction, the Ambisonics samples were played back with only four loudspeakers in a standard quadraphonic setup. This choice was made since in previous experiments the quality of Ambisonics reproduction with 16 loudspeakers was found inferior to SIRR (Pulkki *et al.*, 2004a). Four is the minimum number of loudspeakers necessary for first-order Ambisonics reproduction. Using the minimum number was motivated by the need to reduce timbral artifacts due to the comb filtering resulting from the Ambisonics reproduction (see also related discussion later in Section 5.5.4). While it is possible that increasing the number of the loudspeakers from four to five or six could have still slightly improved the Ambisonics results (see Fredriksson and Zacharov, 2002), four loudspeakers allowed using a regular layout with the available reproduction setup. The subsequent test with four loudspeakers also gave considerably better results for Ambisonics than using all 16 loudspeakers (Pulkki *et al.*, 2004b). The Ambisonics method was formulated using hypercardioid directionality in the decoding stage, as proposed by Monro (2000).

It should be emphasized that the simulation of the B-format microphone may yield better results than can be achieved in reality, since practical microphones always suffer

from non-ideal frequency responses and directivity patterns, as well as from internal noise. However, using an ideal simulated microphone allows investigating the features of the actual reproduction methods disregarding other non-idealities in the recording process. For tests with real measured room responses, see Section 5.6.

Sound samples

The impulse responses of the virtual reality and the tested reproductions were convolved with two different anechoically recorded sounds: a drum sample with four consecutive snare drum shots, and a male talker pronouncing the words "in language." It was assumed that the drum shots would primarily reveal differences in spatial perception, and the speech sample would expose coloration due to the reproduction methods. The samples were played back at a comfortable listening level, and the loudness of all reproduction methods was subjectively equalized in the best listening position.

5.5.2 Procedure

The test procedure was A/B scale with hidden reference, as recommended in ITU-R BS.1284-1 (2003). The reference was always one of the virtual reality samples, and the test items included all reproductions of the same sample as well as the reference itself.

Before starting the test, the listeners received written instructions explaining the test procedure. They were asked to pay attention to coloration, localization of direct sound and reflections, envelopment by reverberation, as well as other artifacts, and to finally give a single overall difference rating based on deviations in these aspects between the reference and the test samples. The scale was selected according to the ITU impairment scale (ITU-R BS.1284-1, 2003): 5.0 = imperceptible, 4.0 = perceptible but not annoying, 3.0 = slightly annoying, 2.0 = annoying, and 1.0 = very annoying. The listeners were able to choose between 1 and 5 with increments of 0.2 using a small keyboard.

After reading the instructions, the listeners were allowed to freely listen to all samples for five minutes. After this familiarization, each subject conducted three test sessions. Each session consisted of two runs, one in both listening positions. One run took approximately 8 minutes and involved evaluation of each sample pair twice. A break of 5 minutes was held between successive sessions. The listening position for the first run was the same for each individual through all sessions. However, every second listener started the sessions in the reference listening position, and every other in the worst case listening position. The order of the sample pairs and the reference and test samples within each pair was randomized. Each grading was done after two consecutive presentations of the sample pair either as: reference - test [pause 1s] reference - test [evaluation], or as: test - reference [pause 1s] test - reference [evaluation].

Ten listeners participated in the experiment. The listeners were either faculty of the Laboratory of Acoustics and Audio Signal Processing or students of the Spatial Sound course at Helsinki University of Technology. None of the listeners reported any hearing deficiencies.

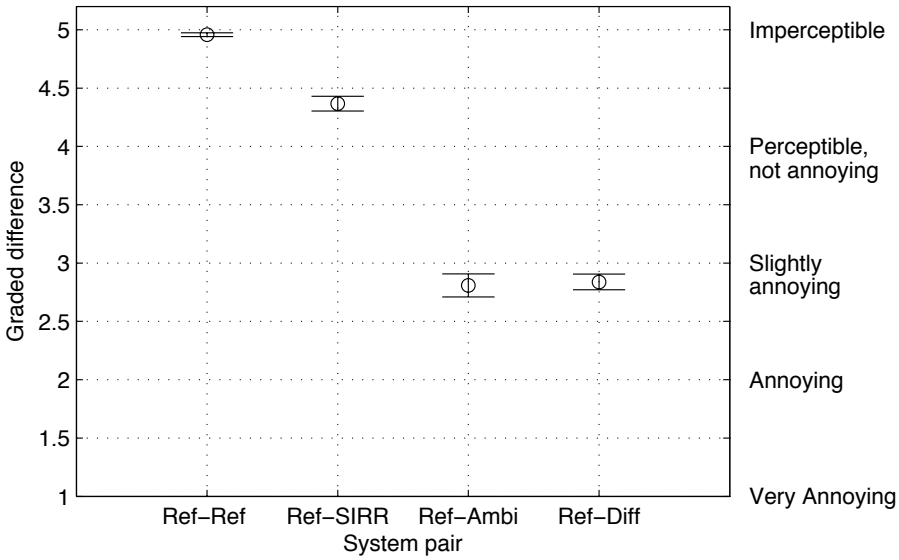

Figure 5.12: The average difference between the system pairs in the anechoic listening test.

5.5.3 Results

The results from the last two sessions were taken to data analysis, resulting thus in 4 evaluations of each sample pair per subject. The mean and variance of each subject within each session were normalized as recommended in ITU-R BS.1284-1 (2003). The normalized data were analyzed with analysis of variance (ANOVA). Checking for the assumptions of ANOVA showed that in this experiment the variances of the different cases were significantly different (Levene's test), and the distributions of the data did deviate statistically significantly from normal. Upon further inspection it was found that for the cases graded close to 5.0, the distributions were skewed. However, ANOVA is known to be robust for small violations of the previously mentioned assumptions. Furthermore, the most interesting effects were very strong, so the ANOVA results will, nevertheless, be used. The results for a model with all main effects and interactions up to the third order are shown in Table 5.1. In the following, the effects and interactions which were found to be significant and interesting in the context of the current experiment are discussed.

Main effects

The main effects include the compared system pair (SYSTPAIR), stimulus (SOUND), listening position (LISTP), virtual space (VSPACE) and repetition (REPE). The main effect of subject (PERS) was normalized out in the normalization of the means and variances. All other main effects except repetition were found significant. Upon further investigation, only the effect of system pair was found to be relevant in the context of the current investigation, and it is illustrated in Figure 5.12. The figure shows the mean values and 95% confidence intervals of mean, resulting in mean opinion scores (MOS) of the listeners for each system pair. It can be seen that the listeners have reliably graded the difference of reference and reference to be imperceptible.

Tests of Between-Subjects Effects

Dependent Variable: DIFFER

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2569.51[a]	557	4.613	26.577	.000
Intercept	26898.1	1	26898.1	154964.4	.000
SYSTPAIR	1707.60	3	569.202	3279.264	.000
REPE	.414	3	.138	.796	.496
VSPACE	24.916	2	12.458	71.772	.000
SOUND	147.864	1	147.864	851.867	.000
LISTP	122.289	1	122.289	704.525	.000
PERS	.000	9	.000	.000	1.000
SYSTPAIR * REPE	1.266	9	.141	.811	.606
SYSTPAIR * VSPACE	36.295	6	6.049	34.851	.000
SYSTPAIR * SOUND	44.466	3	14.822	85.393	.000
SYSTPAIR * LISTP	150.324	3	50.108	288.681	.000
SYSTPAIR * PERS	67.547	27	2.502	14.413	.000
REPE * VSPACE	.601	6	.100	.577	.749
REPE * SOUND	1.311	3	.437	2.517	.057
REPE * LISTP	.799	3	.266	1.534	.204
REPE * PERS	5.050	27	.187	1.078	.358
VSPACE * SOUND	14.695	2	7.348	42.331	.000
VSPACE * LISTP	2.646	2	1.323	7.622	.001
VSPACE * PERS	7.259	18	.403	2.323	.001
SOUND * LISTP	.278	1	.278	1.602	.206
SOUND * PERS	40.428	9	4.492	25.879	.000
LISTP * PERS	21.002	9	2.334	13.444	.000
SYSTPAIR * REPE * VSPACE	3.353	18	.186	1.073	.374
SYSTPAIR * REPE * SOUND	3.790	9	.421	2.426	.010
SYSTPAIR * REPE * LISTP	3.334	9	.370	2.134	.024
SYSTPAIR * REPE * PERS	13.756	81	.170	.978	.534
SYSTPAIR * VSPACE * SOUND	7.779	6	1.297	7.470	.000
SYSTPAIR * VSPACE * LISTP	2.638	6	.440	2.533	.019
SYSTPAIR * VSPACE * PERS	17.825	54	.330	1.902	.000
SYSTPAIR * SOUND * LISTP	10.361	3	3.454	19.897	.000
SYSTPAIR * SOUND * PERS	25.115	27	.930	5.359	.000
SYSTPAIR * LISTP * PERS	48.819	27	1.808	10.417	.000
REPE * VSPACE * SOUND	.394	6	7.E-02	.378	.893
REPE * VSPACE * LISTP	1.295	6	.216	1.244	.281
REPE * VSPACE * PERS	8.546	54	.158	.912	.657
REPE * SOUND * LISTP	.586	3	.195	1.124	.338
REPE * SOUND * PERS	5.115	27	.189	1.091	.341
REPE * LISTP * PERS	5.079	27	.188	1.084	.350
VSPACE * SOUND * LISTP	4.E-02	2	2.E-02	.127	.881
VSPACE * SOUND * PERS	7.446	18	.414	2.383	.001
VSPACE * LISTP * PERS	4.461	18	.248	1.428	.109
SOUND * LISTP * PERS	2.723	9	.303	1.743	.075
Error	236.411	1362	.174		
Total	29704.0	1920			
Corrected Total	2805.92	1919			

a. R Squared = .916 (Adjusted R Squared = .881)

Table 5.1: ANOVA results for the anechoic listening test.

The difference between the SIRR reproduction and the virtual reality has been, on the average, graded to a value of 4.4 (perceptible but not annoying), i.e., almost imperceptible. Both Ambisonics and the phase randomization produce a value of 2.8, which is characterized as slightly annoying on the ITU scale.

Interactions between variables

In the ANOVA results, many interactions were found to be statistically significant. Most of them were not of great interest and will not be discussed here. However, two interactions were considered relevant. The dependence of the results of a system pair on the listening position (SYSTPAIR*LISTPOS, F(3,1362)=288, $p < 0.001$) is plotted in the upper panel in Figure 5.13. It can be seen that with Ambisonics there is a very large increase in perceived difference when moving from the reference listening position to the worst case listening position; the MOS value drops from 3.5 to 2.1. According to listeners' comments and the authors' informal listening, the difference between the reference stimuli and their Ambisonics reproductions was perceived mainly as sound coloration in the reference position. In the worst case position, the coloration is still present but the localization of the direct sound and the reflections collapses to the nearest loudspeaker. This happens because in Ambisonics the direct sound and early reflections are present in practically all loudspeaker channels due to the limited directional resolution, and they arrive first to the listening position from the nearest loudspeakers. The precedence effect (see Section 3.3.3) then causes the sound to be localized in the direction of the nearest loudspeaker.

With SIRR and phase randomization, the effect of the listening position is smaller than with Ambisonics. With SIRR, the perceived directions of the sound sources did not change prominently at different listening positions. Nevertheless, the MOS value changes from 4.5 to 4.2. According to informal listening of the SIRR reproductions, a slight change in timbre and small spatial artifacts were perceived in the worst case listening position. With the phase randomization technique, on the other hand, the MOS value changes from 3.0 to 2.6 when moving to the worst case position, which is also due to an increase in the spatial artifacts similar to what occurred with SIRR.

The dependence of the MOS of each system pair on the reproduced virtual space (SYSTPAIR*VSPACE, F(6,1362)=34.8, $p < 0.001$) is shown in the lower panel of Figure 5.13. The most interesting interaction happens with SIRR: The MOS value drops from 4.7 to 4.1 when changing from the largest virtual room to the smallest one. The average values for SIRR in the large virtual space and in the reference listening position are 4.74 and 4.89 for the drum and speech sounds, respectively. These are very high values, and it can be said that in the large virtual room and the reference listening position, the SIRR reproduction could not be distinguished from the original. In the smaller rooms, slight timbral and spatial artifacts emerged, especially for the snare drum sound.

5.5.4 Discussion

The results show that in the idealized conditions, SIRR reproduces the perception of a virtual room in a very faithful fashion. With the large room, the perceived difference

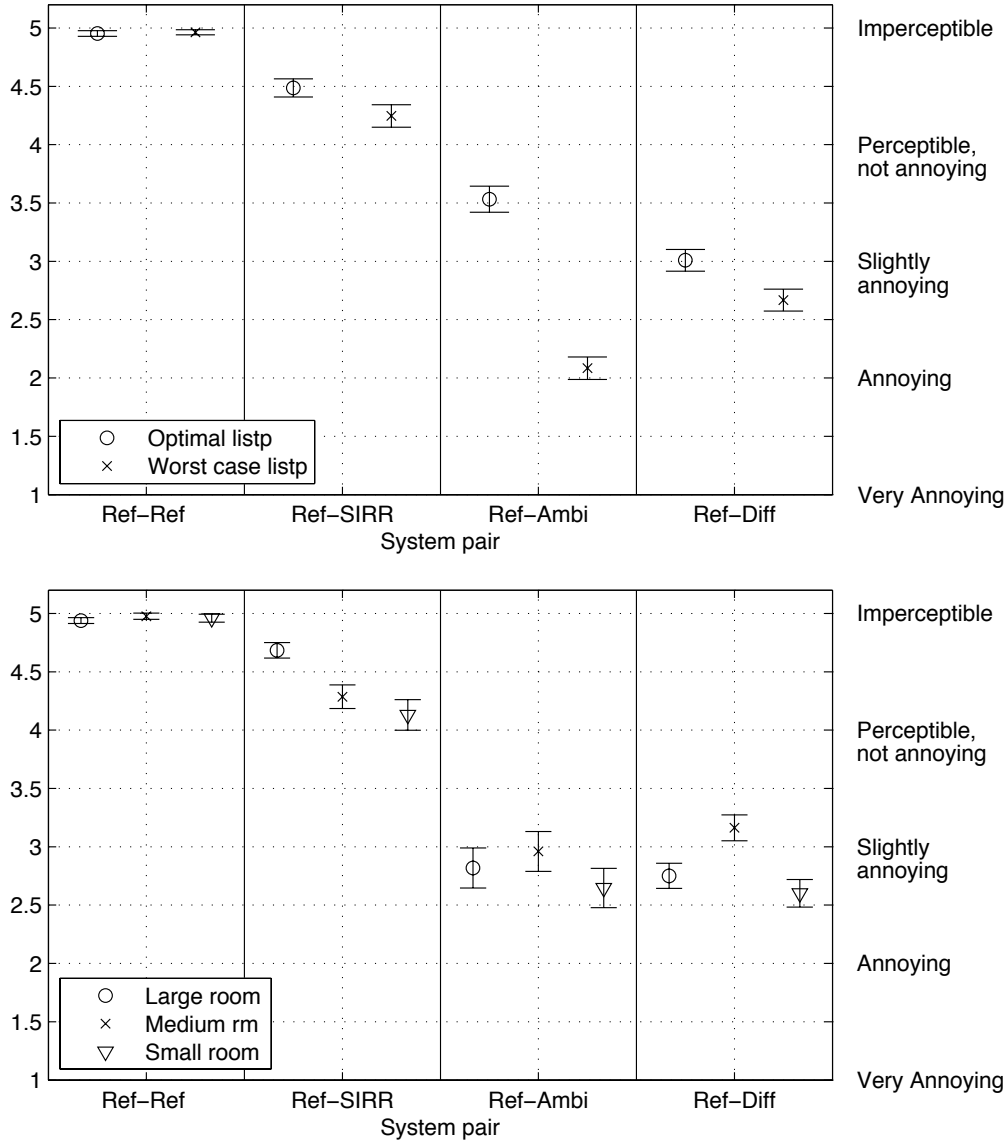

Figure 5.13: The dependence of the judged difference between the system pairs on the listening position (upper panel) and on the virtual space (lower panel) in the anechoic listening test.

is minor or imperceptible, indicating at best an authentic reproduction. However, based on the slightly degraded results for the two smaller rooms, the validity of the assumptions in Section 5.3.1 must be questioned. Nevertheless, it could be that the room-dependent performance of SIRR is not due to inadequate or wrong assumptions, but due to the chosen implementation.

In the smaller rooms there were more early reflections arriving within the duration of the time window used in the SIRR processing (see the dependence of the average density of reflections on volume of the room in Section 2.2.2). In such cases, SIRR is not able to physically reconstruct the discrete sound events and the diffuse synthesis might not be accurate enough to always recreate an authentic perception. In the

current implementation, STFT processing was also used, although human hearing does not have uniform time-frequency resolution. Especially the broad frequency resolution at low frequencies might create some audible artifacts. The choice of the computationally efficient STFT method was motivated by the equal performance of STFT and auditory filterbank analysis and synthesis in the Binaural Cue Coding method (Baumgarte and Faller, 2003). However, with more recent advances in analysis and synthesis of diffuse sound in SIRR (and in Binaural Cue Coding, see Faller, 2003), a new comparison would be needed. This is left for future research.

In contrast to SIRR, with phase randomization, the directional information of the direct sound and discrete reflections is totally lost. Moreover, there might be some coloration at low frequencies due to the phase randomization artifacts. However, in informal listening, the difference in timbre was less salient than the differences in spatial perception. Hence, it can be argued, not surprisingly, that the reproduction of non-diffuse sound as point-like virtual sources is an important part of SIRR processing.

As mentioned earlier, the differences in the best listening position between first-order Ambisonics reproduction and the reference were most prominent as changes in timbre, and in the worst case position as changes both in timbre and spatial aspects. The major difference between SIRR and first-order Ambisonics is that in Ambisonics the same sound is reproduced at different levels with practically all loudspeakers. As discussed earlier, the same sound signal traveling via multiple paths of different lengths to each ear produces comb-filter effects. These effects are more pronounced if there are more loudspeakers around the listener, as the results of a similar listening tests with 16 loudspeakers showed (Pulkki *et al.*, 2004a). In SIRR, the coherence between the loudspeakers is low, and, consequently, the comb-filter effects are minimal. It is assumed that this is the reason why SIRR was found superior to Ambisonics in this test.

Despite the small impairments in the reproduction of the smaller rooms, SIRR thus appears as a plausible reproduction method. However, it is questionable whether the current results are valid in other listening conditions. In reverberant environments, the comb-filter effects produced by the coherence between the loudspeaker channels are less prominent than in anechoic conditions, which might decrease the difference between SIRR and Ambisonics. Furthermore, the room responses and the recording of the responses were simulated, and the number of loudspeakers used for the reproduction was larger than what is typically utilized in, e.g., home theater setups. Hence, in the next section, SIRR is tested in a more practical application.

5.6 Listening tests in listening room

The second experiment to investigate the differences between SIRR and competing systems was conducted in a standard listening room with more conventional loudspeaker setups than in the first experiment and with using real measured room responses. However, as argued in Section 5.5, such an experiment does not allow presenting a reference stimulus. Consequently, a different test method had to be chosen.

As opposed to authenticity, the listening test reported in this section assesses the plausibility of the reproduction based on expectations of the listeners. Since the goal

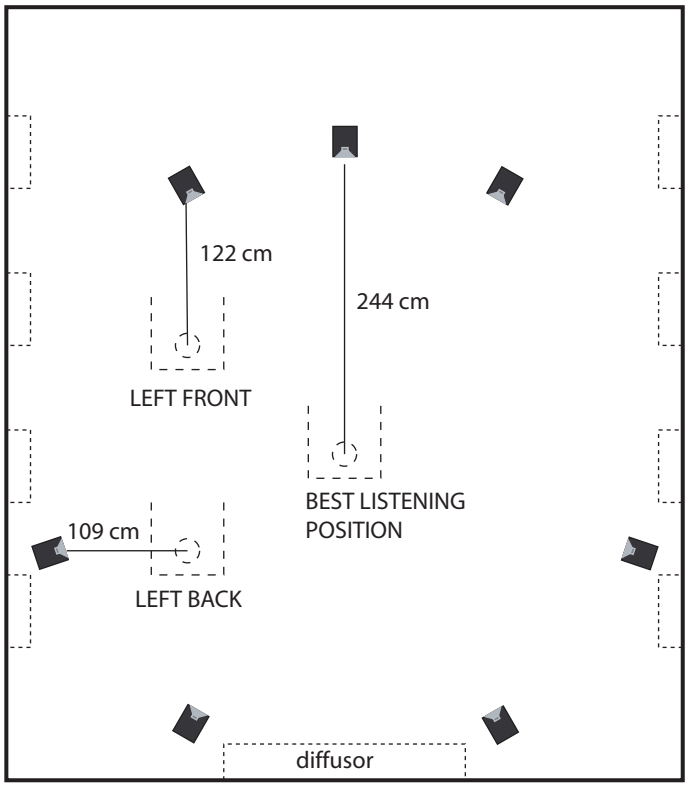

Figure 5.14: The loudspeaker configuration and listening positions in the listening room.

of SIRR is to realize as authentic reproduction and it was applied to reproduction of natural acoustical environments, naturalness was considered as a suitable attribute to be used in the assessment (see also Fredriksson and Zacharov, 2002). The hypothesis of the experiment is that SIRR yields a more natural reproduction than the established Ambisonics method, which uses the same microphone system. Based on the previous experiment, it is also expected that the difference between SIRR and Ambisonics is somewhat larger in off-center listening positions than in the sweet spot.

5.6.1 Apparatus and stimuli

The listening test was designed to correspond as closely as possible to methods typically used by recording engineers. The test was conducted in a listening room fulfilling the ITU-R BS.1116-1 (1997) recommendation for multichannel listening. The loudspeaker system consisted of a standard 5.0 setup (ITU-R BS.775-1, 1992–1994) with Genelec 1030A loudspeakers in the directions of 0°, ±30° and ±110° azimuth. Two additional speakers were added at ±150° to form an extended setup which is denoted as 7.0 in this section. The drivers of the loudspeakers were at a distance of 244 cm from the center of the setup (denoted as the *best listening position*). Furthermore, the distance between the drivers and the nearest wall varied from 55 cm to 140 cm. The loudspeaker setup is illustrated in Figure 5.14. For studying the quality of the reproduction outside the best listening position, two worst case listening positions,

as recommended in ITU-R BS.1116-1 (1997), were also chosen. These positions are denoted as *left back* and *left front*, and they are also shown in Figure 5.14.

Room responses

Two different acoustical environments (denoted as rooms from here on) were chosen for reproduction: a concert hall with a reverberation time of 2.25 s, and an entertainment center with a reverberation time of 7.0 s. The responses of these rooms are part of the library of the Waves IR-360 multichannel convolving reverberator, and they have been measured using a B-format (Soundfield) microphone.

The measurement data for both rooms include two responses measured at a single receiver location with sound sources in two different directions. In order to reverberate a stereophonic signal, both measured responses were processed for reproduction with the chosen loudspeaker setups. Each of the two resulting sets of multichannel responses thus contained one response per each loudspeaker. For reverbarating a monophonic (single-channel) signal, only one of the response sets was used. For stereophonic source material, the left channel was convolved with the responses corresponding to the measurements with the leftmost source, and the right channel with the responses corresponding to the measurements with the rightmost source. The resulting reproduction of the stereophonic channels thus corresponds to listening to the stereophonic sound material in the measurement room with loudspeakers in the measurement source positions. The stereophonic application does not exactly correspond to recording the sound events contained in the stereophonic mix in the space where the room responses were measured. However, it is how the responses are commonly used in a recording studio and, consequently, considered as a suitable method for evaluating the plausibility of the reproduction.

Test cases

When using a reverberator, the whole audio signal may be fed through the device. However, the dry source signal is often directly positioned to a desired direction, and the reverberator is used only to add the effect of a room instead of also affecting the direct sound. Both methods were tested in this experiment. Altogether, five different reproduction methods were employed to process the measured responses:

- SIRR 5.0
- SIRR 7.0
- First-order Ambisonics
- First-order Ambisonics with the direct sound applied to single loudspeaker(s) (Ambitail)
- Multichannel mono with the direct sound applied to single loudspeaker(s) (Omnitail)

All other responses were rendered for 5.0 reproduction except for SIRR 7.0. Furthermore, all responses were computed at 96 kHz sampling rate, and presented downsampled to 48 kHz.

The SIRR processing was realized with 1.3 ms time windows plus an additional 1.3 ms of zero padding and using the hybrid diffusion method (see p. 116). Since the direct sound of all responses was practically non-diffuse, there would not have been any difference between SIRR and SIRR with direct sound applied separately to the same direction. Hence, only one condition with direct sound of the left and right responses forced to the left and right front speakers, respectively, was tested with both loudspeaker configurations.

As already mentioned, first-order Ambisonics was again selected for comparison as a method using the same microphone system as the current implementation of SIRR. Unfortunately, there does not seem to be a consensus as to which kind of Ambisonics decoding should be used for a 5.0 setup. Several hypercardioid and cardioid decodings with different weightings were tested informally (see discussion later in Section 5.6.4). Finally, the implementation used in the Waves IR-360 reverberator (Waves Inc., 2005) was selected, as it was assumed that the parameters had been tuned to get as good sound quality as possible. It was also considered of interest to compare the two different treatments of the direct sound with Ambisonics. As mentioned earlier, the loudspeaker channels resulting from first-order Ambisonics are highly correlated due to the limited directional resolution. Hence, Ambisonics reproduction results in coloration and spatial artifacts in the direct sound unless it is treated separately. In the Ambitail system, the first 4 ms of the responses starting from the direct sound were applied to the left or right front loudspeakers. In order to prevent prominent timbral differences from occurring between Ambitail and other systems, the measured direct sound was used instead of replacing the direct sound with an impulse.

In the Omnitail test case, only the omnidirectional response W from the B-format microphone was used. The direct sound of the left and right responses were applied to the left and right front loudspeakers, as with Ambitail, and the rest of the response was applied coherently to all loudspeakers. This method is computationally very inexpensive, since only one convolution is needed. It was also assumed that Omnitail would produce the worst results out of all the test cases.

Informal listening suggested that the test cases produced somewhat different direct-to-reverberant ratios and loudnesses. This happened because of the summation of parts of the responses in or out of phase, depending on the reproduction method. Consequently, the energy ratio between the direct sound and the rest of the multichannel response in Ambisonics, Ambitail, and Omnitail was adjusted to be the same as with SIRR. Furthermore, the total energy of the responses of the different reproduction methods was equalized. In careful informal listening, it was found that these adjustments led to the perceived loudnesses and direct-to-reverberant ratios being very similar in the best listening position, although differences in other positions remained. It would have also been possible to make individual alignments for each listening position. However, the position-dependent differences were considered as inherent properties of the reproduction systems that were to be compared.

Sound samples

Three sound samples were reverberated with the responses of all test cases. The first sample is denoted as *speech+click*, and it consisted of a male speaker making a count

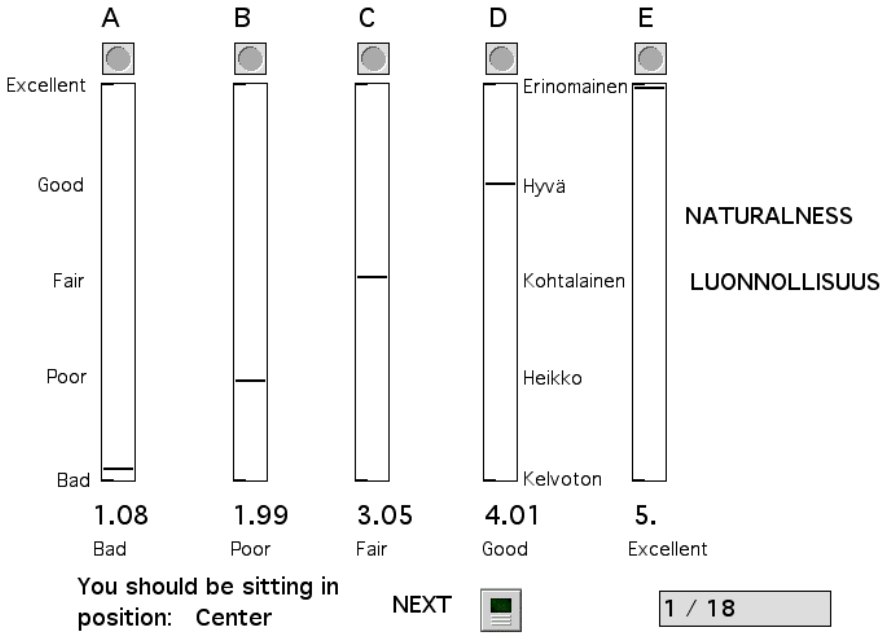

Figure 5.15: GUI used in the listening room tests.

in verbally and by clicking drum sticks as in: *one-click-two-click-one-two(click)-three-four(click)*. This monophonic sample was recorded with a close-up microphone in moderately dry acoustics and panned fully to the left stereophonic channel. Thus the sample was only convolved with one set of the room responses.

The second sound sample was an eight-second-long stereo mix of a Latin band, denoted as *latin*. The sample consisted of seven percussion instruments, piano, and double bass. All instruments were acoustically dry and taken from the loops included in Apple's GarageBand software (Apple, 2005).

The third sample was an eight-second long excerpt from a two-channel anechoic recording of an orchestra playing the introduction of the Marriage of Figaro by Mozart. This sample is denoted as *orchestra* and it was recorded as a part of the Sound Preference Audition Room project (Yoshimasa Electronic, 2003).

5.6.2 Procedure

The test was arranged as a multiple comparison task using a graphical user interface (GUI) as illustrated in Figure 5.15. The method resembles the MUSHRA procedure (ITU-R BS.1534-1, 2001-2003) with the modification that there was no reference. A similar task has also been used by Fredriksson and Zacharov (2002). In the GUI, the listeners were asked to grade the naturalness of each sample. The GUI was projected to the front wall of the listening room via a window in the back wall of the room. The listeners used the GUI with a mouse only.

Each multiple comparison included the five systems with the same single sound sample and room response, evaluated in a single listening position. Hence, 18 separate comparisons were needed to complete the experiment. The listeners were able to switch at will between the stimuli during the playback. The switching was imple-

mented by successive 20 ms fade-out of the previous and fade-in of the new stimulus. The gradings were given on a continuous ITU quality scale (ITU-R BS.1284-1, 2003; ITU-R BS.1534-1, 2001-2003) shown on the GUI in Finnish and in English. Before each comparison, the test program asked the listener to move to a specific listening position. The order of the comparisons and the order of the stimuli in each comparison were randomized.

As mentioned earlier, there was no possibility to present a natural reference sample and thus ask for a simple difference rating. The test assumed that the listeners had an expectation of what the reproduced acoustical environments should sound like and that they would be able to judge the naturalness of the reproductions accordingly. Such an approach is, of course, prone to bias and noise since the internal references of the listeners necessarily vary. Before the actual testing, the listeners were familiarized with the task and the stimuli. They first received written instructions and they were shown photographs of the two acoustical environments. In a subsequent guided demonstration, different stimuli were played back to the listeners who were asked to pay attention to a selection of attributes and to describe them verbally.

The attributes to be considered both in the demonstration and in the actual test were:

- Timbre of sound sources
- Localization
- Ensemble width
- Perceived distance
- Reverberation:
  - Does reverberation surround the listener from all directions?
  - Timbre of the reverberant tail
- Reflections

These are attributes that were expected to change between the different reproductions (see Zacharov and Koivuniemi, 2001b; Berg and Rumsey, 2002). Similar attributes have also been found to be related to different aspects of naturalness (Berg and Rumsey, 2000a). Although these attributes' relation to the judged overall naturalness is not known, and may differ from one listener to another, listing the attributes was considered important in order to somewhat unify the evaluation. Otherwise, each listener could have concentrated on a different subset of the attributes, thus neglecting unnaturalness in some other respect. In the demonstration, none of the samples was identified. The conductor of the experiment was well aware of the importance of not imposing his own opinions on the listeners through commenting on the samples. Despite potential biasing effects, the short demonstration was also considered important in order to familiarize the subjects with the listed attributes. With these preparations, the subjects were assumed to know better where to direct their attention.

After the familiarization, the listeners were allowed to freely listen to all the stimuli in all listening positions for about 10 minutes. The free listening was followed by a training session consisting of three multiple comparisons. Finally, the full test with the 18 multiple comparisons was conducted. The whole procedure including the familiarization session was fairly long, taking two hours per listener on average, and thus each sample was evaluated only once by each listener. A break of ten minutes was held in the middle of the final test.

A total of 21 listeners completed the test. All subjects had at least some interest in and experience with sound, as they were either staff of the Laboratory of Acoustics and Audio Signal Processing or students on the Communication Acoustics course. None of the listeners reported any hearing deficiencies.

5.6.3 Results

Preliminary analysis showed that most listeners had graded the different reproduction systems to significantly different naturalness values. However, despite giving different ratings to individual stimuli, one listener had rated all the reproduction methods very close to 3, on average. It was concluded that she was either not able to detect the differences between the test cases or that she was guessing, so her data were left out of the analysis. The means and variances of the data from the remaining 20 listeners were first equalized and then subjected to ANOVA. Checking for the assumptions for ANOVA showed that in this experiment the variances of different cases were not significantly different (Levene's test) but the distributions did deviate significantly from normal distribution. However, based on visual inspection of the data, the normality assumption was not seriously violated, so it was considered safe to use ANOVA. The results are shown in Table 5.2, and the most important findings are discussed in the following.

Results for different reproduction systems

As expected, the reproduction system (SYSTEM) had the largest effect on the results ($F(4,1004)=340$, $p < 0.001$). The means and the 95% confidence intervals of the means averaged over all data for each reproduction method are plotted in the upper panel of Figure 5.16. As anticipated, the Omnitail system produced the lowest naturalness. Ambisonics and Ambitail were graded to values of 2.7 and 2.9, respectively. However, their difference was not found statistically significant in a Tukey's A posthoc test. SIRR 5.0 was graded on the average to 3.2, and SIRR 7.0 to 3.5. In posthoc tests, the means of both SIRR systems were found to differ from all the other systems and from each other. Thus, the hypothesis that SIRR yields more natural reproduction than Ambisonics can be confirmed. The effect of the number of loudspeakers will be further discussed in Section 5.6.4.

Effect of listening position

The most interesting effect of the listening position (LISTPOS) is its interaction with the reproduction methods. Nevertheless, the listening position was also found to

Tests of Between-Subjects Effects

Dependent Variable: VALUE

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1196.07[a]	795	1.504	5.098	.000
Intercept	15033.2	1	15033	50936.1	.000
PERS	.000	19	.000	.000	1.000
SOUND	19.593	2	9.797	33.194	.000
LISTPOS	137.717	2	68.859	233.310	.000
ACOUST	3.509	1	3.509	11.889	.001
SYSTEM	402.098	4	100.52	340.602	.000
PERS * SOUND	40.330	38	1.061	3.596	.000
PERS * LISTPOS	82.583	38	2.173	7.363	.000
PERS * ACOUST	17.478	19	.920	3.117	.000
PERS * SYSTEM	56.191	76	.739	2.505	.000
SOUND * LISTPOS	2.511	4	.628	2.127	.075
SOUND * ACOUST	.798	2	.399	1.351	.259
SOUND * SYSTEM	6.799	8	.850	2.880	.004
LISTPOS * ACOUST	10.241	2	5.120	17.349	.000
LISTPOS * SYSTEM	154.486	8	19.311	65.430	.000
ACOUST * SYSTEM	4.731	4	1.183	4.008	.003
PERS * SOUND * LISTPOS	47.182	76	.621	2.103	.000
PERS * SOUND * ACOUST	17.502	38	.461	1.561	.017
PERS * SOUND * SYSTEM	54.458	152	.358	1.214	.050
PERS * LISTPOS * ACOUST	24.223	38	.637	2.160	.000
PERS * LISTPOS * SYSTEM	71.746	152	.472	1.599	.000
PERS * ACOUST * SYSTEM	25.320	76	.333	1.129	.218
SOUND * LISTPOS * ACOUST	2.307	4	.577	1.954	.099
SOUND * LISTPOS * SYSTEM	7.152	16	.447	1.515	.087
SOUND * ACOUST * SYSTEM	4.519	8	.565	1.914	.055
LISTPOS * ACOUST * SYSTEM	2.592	8	.324	1.098	.362
Error	296.318	1004	.295		
Total	16525.5	1800			
Corrected Total	1492.38	1799			

a. R Squared = .801 (Adjusted R Squared = .644)

Table 5.2: ANOVA results for the listening room test.

have a significant main effect ($F(38,1004)=68.9, p<0.001$). The mean grades for the different listening positions are plotted in the lower panel of Figure 5.16. The best overall naturalness, 3.2, is obtained in the best listening position, as expected. In the left front position, the average naturalness decreases only slightly to a value of 3.0, whereas in the left back position the naturalness has a MOS value of only 2.5.

The interaction between the listening position and the reproduction system (LISTPOS * SYSTEM, $F(6,1004)=65.4, p<0.001$) was also found highly significant. This effect is illustrated in Figure 5.17. The significances of the differences between each data group were also tested by running one-way ANOVA separately over the data from each listening positions with setting SYSTEM as the only variable, and then computing the Tukey's A posthoc test. The results were equal to those obtained by visually comparing the 95% confidence intervals of means.

SIRR 7.0 appears to provide the most uniform naturalness in all listening positions. In the left back listening position, it is a full grade above the other methods, being the only system that yields a reasonably natural reproduction. However, in the best and left front positions, the difference from SIRR 5.0 is not statistically significant.

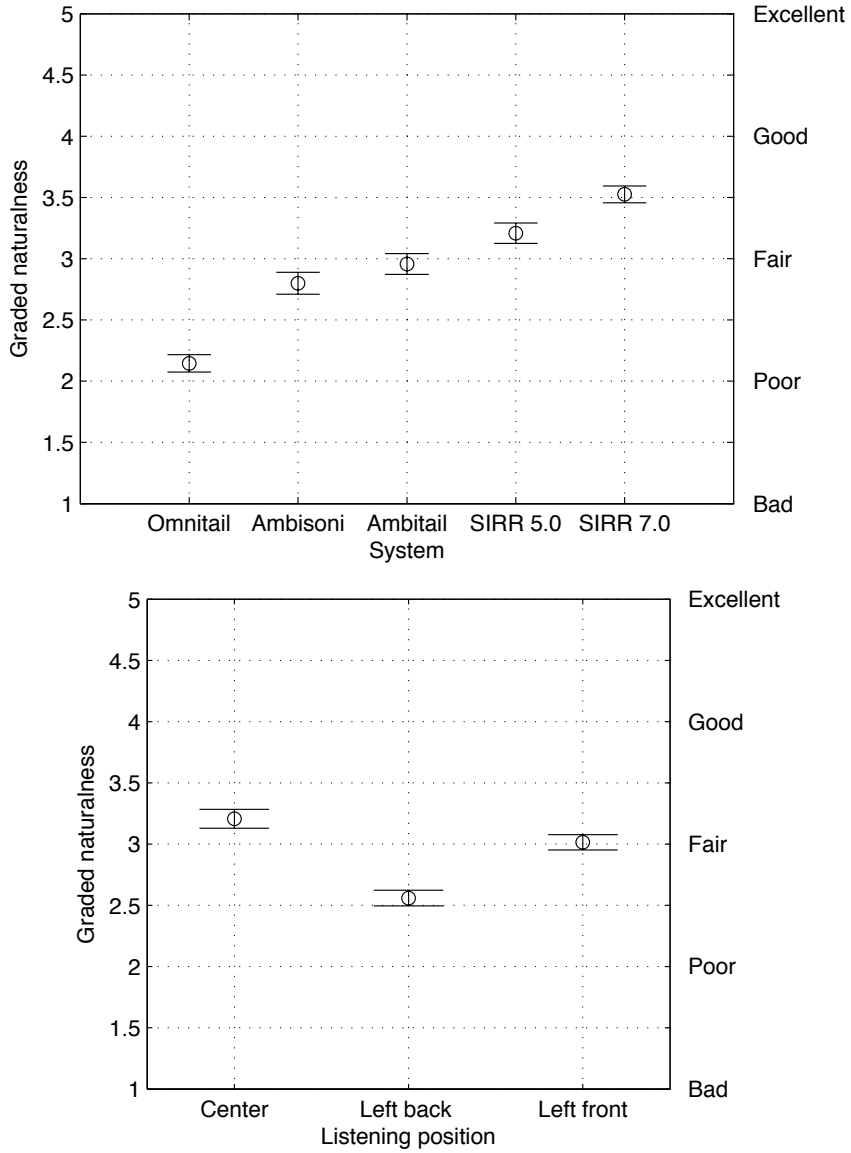

Figure 5.16: The dependence of the graded naturalness on the reproduction method and on the listening position in the listening room test.

The order of the Ambisonics, Ambitail, and SIRR 5.0 systems is the same in all listening positions, with Ambisonics sounding least natural and SIRR 5.0 most natural. In the left back listening position, the gradings for the Ambisonics and Ambitail systems are below even the Omnitail system. The difference between the Ambisonics and Ambitail systems is statistically significant only in the best listening position. SIRR 5.0, on the other hand, has been graded significantly more natural than Ambisonics in all listening positions, and in the left back and left front positions, also more natural than the Ambitail system. The advantages of SIRR compared to Ambisonics are thus larger in off-center listening positions, as expected, although for the 5.0 systems the difference is not very pronounced.

In all cases, the Omnitail system has been graded to values below 2.3. However,

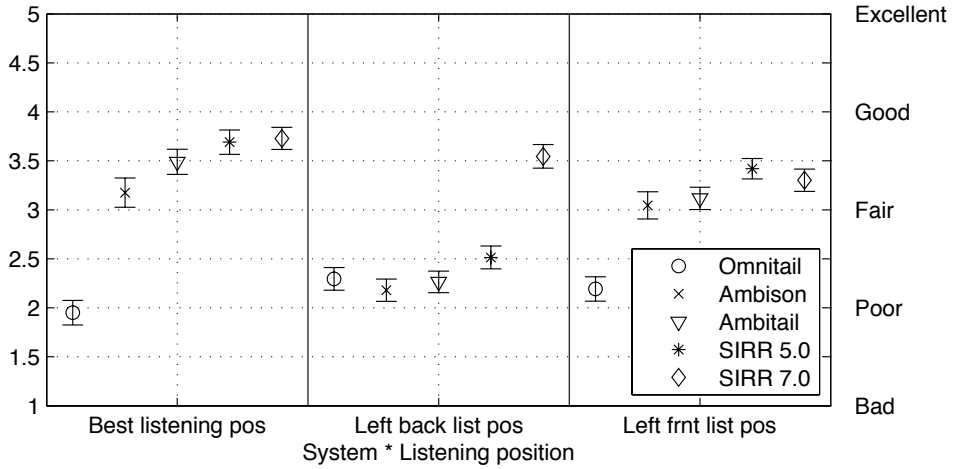

Figure 5.17: The dependence of the graded naturalness on the reproduction system at different listening positions in the listening room test.

contrary to the other reproduction methods, the naturalness of the Omnitail system is higher in the left back and left front positions than in the best listening position.

5.6.4 Discussion

In the anechoic listening tests (see Section 5.5), the listeners graded the difference between the virtual reality and SIRR reproduction at best to imperceptible (4.8) on the ITU impairment scale. Although the test presented in this section is not directly comparable and the scales are different (the current test used the ITU quality scale), it is interesting to note that in the listening room, the best obtained values were on average only 3.7 (between fair and good but not excellent naturalness). There are several possible reasons for the different gradings in the two tests. Since in the current experiment the listeners did not know what the chosen real acoustical environments sound like, they may have expected something different or they may have been cautious about rating anything as excellent without knowing what the actual target was. However, the lower values compared to the anechoic test may also be due to non-idealities in the technology used to measure the real responses, or due to the smaller number of loudspeakers used in the SIRR reproduction.

Some more insight into the effect of the number of loudspeakers can be gained by considering the gradings of the SIRR 5.0 and SIRR 7.0 systems in the current test. In the left back listening position, all reproduction systems except SIRR 7.0 were judged poor. It appears that in this position, the left back loudspeaker was perceived annoyingly loud with all the 5-channel reproduction methods. With SIRR 7.0, the sound energy was spread more evenly behind the listener, which produced considerably better naturalness. Somewhat similar experiences were gained in preliminary testing with a 7.0 Ambitail system. Unfortunately, this Ambitail system had to be left out of the formal experiment for two reasons: The amount of samples had to be kept to minimum since each listener could only be occupied for a maximum of 2.5 hours, and 7.0 decoding was not supported by the software finally chosen to realize the Ambitail

system. Nevertheless, it can be concluded that the 5.0 loudspeaker setup is too sparse behind the listener to produce optimal sound quality on a listening area extending far backwards or to the sides from the best listening position.

Several other results can be also discussed further based on comments from the listeners and on informal listening to the stimuli by the authors. The worst-performing Omnitail reproduction system was perceived overall as colored and lacking directional information. In the best listening position, the reverberation was for several subjects localized inside the head, or elevated. The colorations were also strongest in the best listening position, which explains the exceptional preference for the off-center listening positions with the Omnitail system.

Ambisonics was found to produce overall blurred localization of direct sound, less defined spatial impression, and to have a slightly colored reverberant tail. However, in informal listening, the colorations seemed less annoying than in the anechoic chamber, as hypothesized earlier in Section 5.5.4. In the best listening position, the diffuse reverberation was found to surround the listener evenly. Nevertheless, as already mentioned, the left surround speaker was perceived to be too dominant in the left back listening position. Part of the dominance of the left surround speaker might have been due to the selected Ambisonics decoding coefficients. Since the utilized implementation is known to apply more sound energy to the surround than to the frontal loudspeakers, basic hypercardioid decoding was also tried in the preliminary testing. The hypercardioid coefficients distribute the reverberant sound with equal weight to all loudspeakers, and indeed it was found that with these coefficients, the naturalness was increased in the left back position. However, in the best listening position, the reverberation was perceived narrower and concentrated too close the median plane, and in the left front position there was more coloration than with the Waves implementation.

As expected, the Ambitail system produced better directional accuracy for the sound sources than Ambisonics, although the difference in naturalness was statistically significant only in the best listening position. In this position, the Ambitail and SIRR 5.0 systems were actually almost indistinguishable even to the authors. The most notable distinction was a difference in the timbre of the reverberant tail. In the other listening positions, the differences were, nevertheless, clearly audible, with SIRR typically performing better.

Overall, both SIRR systems were perceived to have the best directional accuracy of all the test cases. Except for the left back position for SIRR 5.0, the reverberation also surrounded the listeners evenly. Interestingly, in SIRR reproduction the reverberant tail was found to sound brighter than with the Ambitail or Ambisonics systems. Two listeners commented afterwards that they had preferred the darker sound of some of the stimuli, and the results of these listeners showed that they had graded the Ambisonics systems more natural than SIRR.

5.7 Discussion and conclusions

Spatial Impulse Response Rendering (SIRR) is a method for reproduction of measured room responses with an arbitrary multichannel loudspeaker system. Compared

to conventional microphone techniques, SIRR is able to improve the quality of the reproduction by using a perceptually motivated analysis-synthesis procedure. However, instead of operating directly on human binaural cues, SIRR considers physical properties of the sound field that transform into the binaural cues. The time-dependent direction of arrival and diffuseness of a room response are analyzed within frequency bands. Based on this analysis data and an omnidirectional measurement, a multichannel response for a chosen loudspeaker setup is synthesized using two different methods: Non-diffuse parts of the response are positioned as sharply as possible in the correct direction, whereas diffuse parts are applied to all loudspeakers in a decorrelated form.

The validity of the underlying perceptual assumptions was tested both objectively and subjectively. It was shown that the applied analysis and synthesis methods recreate interaural coherence reasonably well in typical cases. The authenticity of SIRR reproduction was assessed in an anechoic listening test. Ideal measurement of virtual reality environments was simulated and the resulting room responses were reproduced with SIRR and other techniques. The task of the listeners was to grade the difference between the reference virtual reality and the reproductions. SIRR performed best out of the tested methods, yielding for the large virtual room at best an imperceptible difference compared to the reference. With the smaller rooms, some artifacts were created by the SIRR processing, but in all cases the listeners rated the difference to the reference as not annoying.

In the listening room test, real measured B-format responses of two existing spaces were reproduced with SIRR and other systems using a standard 5.0 loudspeaker setup and an extended setup with seven loudspeakers. The test addressed the plausibility of the reproduction, and the task of the listeners was to grade the naturalness of the samples based on their expectations. The results showed that SIRR yields also in this case better results than the other tested methods. The difference compared to the other methods was most notable in off-center listening positions. It was also found that in the left back listening position, the naturalness of the reproduction was improved by introducing the two additional rear loudspeakers to the reproduction system.

The SIRR method was described only in the context of processing room responses. However, the method is not limited to such reverberator applications. The room responses are an easy application because the important early reflections can be reproduced as discrete sound events. In processing continuous reverberant sound, the reflections appear convolved with the source signal and cannot be as easily separated. Consequently, similar artifacts appear as in processing the responses of small rooms. Furthermore, the phase randomization method cannot be used as such because in a continuous sound application it would result in audible non-linear distortion. Nevertheless, first promising steps towards processing continuous sound were taken by Pulkki and Merimaa (2005). At its current state, SIRR applied to continuous sound cannot in all cases produce better results than existing reproduction methods. However, it would already be an attractive choice for audio coding, since multichannel sound for arbitrary loudspeaker setups could be transmitted as a single channel and side information. Further development in processing continuous sound is subject to future work.

Chapter 6

Summary and Conclusions

After a brief introduction to the general fields of research related to the current work, the thesis started with two chapters consisting mainly of background information. Chapter 2 provided an overview of physical analysis of spatial sound. The discussion on microphone techniques showed that it is difficult to create high directivity over a large bandwidth. However, using directional microphone systems is not the only possible method for directional analysis. The energetic properties of sound fields were derived theoretically, and measurement methods based on B-format responses were developed. The energetic analysis was also applied to visualization of directional room responses.

Chapter 3 concentrated on the operation of human hearing and the perception of spatial sound. It was established that the auditory periphery realizes a frequency analysis which can be modeled with existing knowledge. Furthermore, the temporal resolution of the spatial hearing is limited. In complex listening situations (in the presence of multiple concurrent sound sources and/or room reflections), the auditory system is, however, usually able to individually localize the sound sources while suppressing the localization of room reflections. Nevertheless, existing binaural models have difficulties in predicting the resulting localization.

Building upon existing auditory models, a novel binaural modeling mechanism for explaining auditory localization in complex listening situations was proposed in Chapter 4. The proposed cue selection mechanism considers the interaural time difference (ITD) and interaural level difference (ILD) cues at time instants when the interaural coherence (IC) is high. It was shown that at such time instants, the ITD and ILD are likely to correspond to the direction of one of the sound sources. The correspondence of the extracted localization cues to the results of a number of psychophysical studies from the literature was verified by a number of simulations.

The perceptual knowledge and analysis tools from the earlier chapters were further applied to development of the Spatial Impulse Response Rendering (SIRR) method in Chapter 5. SIRR makes it possible to overcome some of the limitations related to insufficient directional resolution of microphones by using a perceptually motivated analysis-synthesis procedure. The analysis consists of determining the direction of arrival and diffuseness of sound as a function of time and frequency. These data and an omnidirectional signal are then used to recreate the perceptually important spatial features of the analyzed sound. Moreover, the resulting reproduction can be tailored

to any chosen multichannel loudspeaker system. The performance of the SIRR method was evaluated in two listening tests within the context of a convolving reverberator application. In idealized conditions, it was shown that the SIRR reproduction can be at best indistinguishable from an original sound field. Furthermore, in the second experiment, the SIRR reproduction of real measured room responses was found to create the most natural spatial impression out of the tested methods.

Bibliography

Abel, J. S. and Begault, D. R. (**1998**). "Methods for room acoustic analysis using a monopole-dipole microphone array," in *Proc. Int. Conf. on Noise Control Engineering (InterNoise 98)*, Christchurch, New Zealand. Paper 123.

Abhayapala, T. D. and Ward, D. B. (**2002**). "Theory and design of high order sound field microphones using spherical microphone array," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, Vol. 2, pp. 1949–1952.

Akeroyd, M. A. (**2001**). "A binaural cross-correlogram toolbox for MATLAB," `http://www.biols.susx.ac.uk/home/Michael_Akeroyd/download2.html`.

Akeroyd, M. A. and Bernstein, L. R. (**2001**). "The variation across time of sensitivity to interaural disparities: Behavioral measurements and quantitative analyses," *J. Acoust. Soc. Am.* **110**, 2516–2526.

Akeroyd, M. A. and Summerfield, A. Q. (**1999**). "A binaural analog of gap detection," *J. Acoust. Soc. Am.* **105**, 2807–2820.

Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. (**2001**). "The CIPIC HRTF Database," in *Proc. IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust.*, New Paltz, NY, USA, pp. 99–102. Available at `http://interface.cipic.ucdavis.edu/CIL_html/CIL_HRTF_database.htm`.

Allen, J. B. and Berkley, D. A. (**1979**). "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.* **65**, 943–950.

Ando, Y. and Kurihara, Y. (**1986**). "Nonlinear response in evaluating the subjective diffuseness of sound fields," *J. Acoust. Soc. Am.* **80**, 833–836.

Aoshima, N. (**1981**). "Computer-generated pulse signal applied for sound measurement," *J. Acoust. Soc. Am.* **69**, 1484–1488.

Apple (**2005**). "iLife – GarageBand," `http://www.apple.com/ilife/garageband/`.

Asano, F., Suzuki, Y., and Sone, T. (**1990**). "Role of spectral cues in median plane localization," *J. Acoust. Soc. Am.* **88**, 159–168.

Avendano, C., Algazi, V. R., and Duda, R. O. (**1999**). "A head-and-torso model for low-frequency binaural elevation effects," in *Proc. IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust.*, New Paltz, NY, USA, pp. 179–182.

Barron, M. (**1971**). "The subjective effects of first reflections in concert halls — The need for lateral reflections," *J. Sound Vib.* **15**, 475–494.

Barron, M. (**1995a**). "Bass sound in concert auditoria," *J. Acoust. Soc. Am.* **97**, 1088–1098.

Barron, M. (**1995b**). "Interpretation of early decay times in concert auditoria," *Acus-*

tica **81**, 320–331.

Barron, M. (**2001**). "Late lateral energy fractions and the envelopment question in concert halls," *Appl. Acoust.* **62**, 185–202.

Barron, M. and Lee, L.-J. (**1988**). "Energy relations in concert auditoriums. I," *J. Acoust. Soc. Am.* **84**, 618–628.

Barron, M. and Marshall, A. H. (**1981**). "Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure," *J. Sound Vib.* **77**, 211–232.

Bauer, B. B. (**1961**). "Phasor analysis of some stereophonic phenomena," *J. Acoust. Soc. Am.* **33**, 1536–1539.

Baumgarte, F. and Faller, C. (**2003**). "Binaural Cue Coding. Part I: Psychoacoustic fundamentals and design principles," *IEEE Trans. Speech Audio Proc.* **11**, 509–519.

Bech, S. (**1995**). "Timbral aspects of reproduced sound in small rooms. I," *J. Acoust. Soc. Am.* **97**, 1717–1726.

Bech, S. (**1996**). "Timbral aspects of reproduced sound in small rooms. II," *J. Acoust. Soc. Am.* **99**, 3539–3549.

Bech, S. (**1998**). "Spatial aspects of reproduced sound in small rooms," *J. Acoust. Soc. Am.* **103**, 434–445.

Becker, J. and Sapp, M. (**2001**). "Synthetic soundfields for the rating of spatial perceptions," *Appl. Acoust.* **62**, 217–228.

Begault, D. R. (**1992**). "Perceptual effects of synthetic reverberation on three-dimensional audio systems," *J. Audio Eng. Soc.* **40**, 895–904.

Begault, D. R. and Abel, J. S. (**1998**). "Studying room acoustics using a monopole-dipole microphone array," in *Proc. 16th Int. Congr. Acoust./135th Meeting of Acoust. Soc. Am.*, Seattle, WA, USA, pp. 369–370.

Begault, D. R., Wenzel, E. M., and Anderson, M. R. (**2003**). "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *J. Audio Eng. Soc.* **49**, 904–916.

Bell, A. and Fletcher, N. H. (**2004**). "The cochlear amplifier is a surface acoustic wave: "Squirting" waves between rows of outer hair cells?," *J. Acoust. Soc. Am.* **116**, 1016–1024.

Ben-Hador, R. and Neoran, I. (**2004**). "Capturing manipulation and reproduction of sampled acoustic impulse responses," in *AES 117th Convention*, San Francisco, CA, USA. Preprint 6232.

Benesty, J. (**2000**). "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Am.* **107**, 384–391.

Bennett, J. C., Barker, K., and Edeko, F. O. (**1985**). "A new approach to the assessment of stereophonic sound system performance," *J. Audio Eng. Soc.* **33**, 314–321.

Beranek, L. L. (**1996**). Concert and Opera Halls — How They Sound, Acoustical Society of America, Woodbury, NY, USA.

Berdugo, B., Doron, M. A., Rosenhouse, J., and Azhari, H. (**1999**). "On direction finding of an emitting source from time delays," *J. Acoust. Soc. Am.* **105**, 3355–

3363.

Berg, J. and Rumsey, F. (**1999a**). "Identification of perceived spatial attributes of recordings by repertory grid technique and other methods," in *AES 106th Convention*, Munich, Germany. Preprint 4924.

Berg, J. and Rumsey, F. (**1999b**). "Spatial attribute identification and scaling by repertory grid technique and other methods," in *Proc. AES 16th Int. Conf.*, Rovaniemi, Finland, pp. 51–66.

Berg, J. and Rumsey, F. (**2000a**). "Correlation between emotive, descriptive and naturalness attributes in subjective data relating to spatial sound reproduction," in *AES 109th Convention*, Los Angeles, CA, USA. Preprint 5206.

Berg, J. and Rumsey, F. (**2000b**). "In search of the spatial dimensions of reproduced sound: Verbal protocol analysis and cluster analysis of scaled verbal descriptors," in *AES 108th Convention*, Paris, France. Preprint 5139.

Berg, J. and Rumsey, F. (**2001**). "Verification and correlation of attributes used for describing the spatial quality of reproduced sound," in *Proc. AES 19th Int. Conf.*, Bavaria, Germany.

Berg, J. and Rumsey, F. (**2002**). "Validity of selected spatial attributes in the evaluation of 5-channel microphone techniques," in *AES 112th Convention*, Munich, Germany. Preprint 5593.

Berkhout, A. J., de Vries, D., and Boone, M. M. (**1980**). "A new method to acquire impulse responses in concert halls," *J. Acoust. Soc. Am.* **68**, 179–183.

Berkhout, A. J., de Vries, D., and Sonke, J. J. (**1997**). "Array technology for acoustic wave field analysis in enclosures," *J. Acoust. Soc. Am.* **102**, 2757–2770.

Berkhout, A. J., de Vries, D., and Vogel, P. (**1993**). "Acoustic control by wave field synthesis," *J. Acoust. Soc. Am.* **93**, 2764–2778.

Bernfeld, B. (**1973**). "Attempts for better understanding of the directional stereophonic listening mechanism," in *AES 44th Convention*, Rotterdam, The Netherlands. Paper C-4.

Bernstein, L. R. and Trahiotis, C. (**1996**). "The normalized correlation: Accounting for binaural detection across center frequency," *J. Acoust. Soc. Am.* **100**, 3774–3784.

Bernstein, L. R. and Trahiotis, C. (**1997**). "The effects of randomizing values of interaural disparities on binaural detection and on discrimination of interaural correlation," *J. Acoust. Soc. Am.* **102**, 1113–1120.

Bernstein, L. R. and Trahiotis, C. (**2002**). "Enhancing sensitivity to interaural delays at high frequencies by using "transposed stimuli"," *J. Acoust. Soc. Am.* **112**, 1026–1036.

Bernstein, L. R. and Trahiotis, C. (**2003**). "Enhancing interaural-delay-based extents of laterality at high frequencies by using "transposed stimuli"," *J. Acoust. Soc. Am.* **113**, 3335–3347.

Bernstein, L. R. and Trahiotis, C. (**2004**). "The apparent immunity of high-frequency "transposed" stimuli to low-frequency binaural interference," *J. Acoust. Soc. Am.* **116**, 3062–3069.

Bernstein, L. R. and Trahiotis, C. (**2005**). "Measures of extents of laterality for high-

frequency "transposed" stimuli under conditions of binaural interference," *J. Acoust. Soc. Am.* **118**, 1626–1635.

Bernstein, L. R., Trahiotis, C., Akeroyd, M. A., and Hartung, K. (**2001**). "Sensitivity to brief changes of interaural time and interaural intensity," *J. Acoust. Soc. Am.* **109**, 1604–1615.

Bernstein, L. R., van de Par, S., and Trahiotis, C. (**1999**). "The normalized interaural correlation: Accounting for NoSπ thresholds obtained with Gaussian and "low-noise" masking noise," *J. Acoust. Soc. Am.* **106**, 870–876.

Blauert, J. (**1969/70**). "Sound localization in the median plane," *Acustica* **22**, 205–213.

Blauert, J. (**1971**). "Localization and the law of the first wavefront in the median plane," *J. Acoust. Soc. Am.* **50**, 466–470.

Blauert, J. (**1972**). "On the lag of lateralization caused by interaural time and intensity differences," *Audiology* **11**, 265–270.

Blauert, J. (**1997**). Spatial Hearing: The Psychophysics of Human Sound Localization, revised edn, The MIT Press, Cambridge, MA, USA.

Blauert, J. and Cobben, W. (**1978**). "Some consideration of binaural cross correlation analysis," *Acustica* **39**, 96–104.

Blauert, J. and Divenyi, P. L. (**1988**). "Spectral selectivity in binaural contralateral inhibition," *Acustica* **66**, 267–274.

Blauert, J. and Lindemann, W. (**1986a**). "Auditory spaciousness: Some further psychoacoustic analyses," *J. Acoust. Soc. Am.* **80**, 533–542.

Blauert, J. and Lindemann, W. (**1986b**). "Spatial mapping of intracranial auditory events for various degrees of interaural coherence," *J. Acoust. Soc. Am.* **79**, 806–813.

Blauert, J., Lehnert, H., Sahrhage, J., and Strauss, H. (**2000**). "An interactive virtual-environment generator for psychoacoustic research I: Architecture and implementation," *ACUSTICA – Acta Acustica* **86**, 94–102.

Blauert, J., Möbius, U., and Lindemann, W. (**1986**). "Supplementary psychoacoustical results on auditory spaciousness," *Acustica* **59**, 292–293.

Blesser, B. (**2001**). "An interdisciplinary synthesis of reverberation viewpoints," *J. Audio Eng. Soc.* **49**, 867–903.

Blumlein, A. D. (**1931**) U.K. patent 394,325. Reprinted in Stereophonic Techniques (Audio Engineering Society, New York, 1986).

Bodden, M. (**1998**). "Auditory models for spatial impression, envelopment, and localization," in *Proc. AES 15th Int. Conf.*, Copenhagen, Denmark, pp. 150–156.

Bodlund, K. (**1976**). "A new quantity for comparative measurements concerning the diffusion of stationary sound fields," *J. Sound Vib.* **44**, 191–207.

Boehnke, S. E., Hall, S. E., and Marquadt, T. (**2002**). "Detection of static and dynamic changes in interaural correlation," *J. Acoust. Soc. Am.* **112**, 1617–1626.

Bouéri, M. and Kyriakakis, C. (**2004**). "Audio signal decorrelation based on a critical band approach," in *AES 117th Convention*, San Francisco, CA, USA. Preprint 6291.

Bourennane, S. and Bendjama, A. (**2002**). "Locating wide band acoustic sources using higher order stastics," *Appl. Acoust.* **63**, 235–251.

Braasch, J. (**2002**). "Localization in the presence of a distracter and reverberation in the frontal horizontal plane. II. Model algorithms," *ACUSTICA – Acta Acustica* **88**, 956–969.

Braasch, J. (**2003**). "Localization in the presence of a distracter and reverberation in the frontal horizontal plane. III. The role of interaural level differences," *ACUSTICA – Acta Acustica* **89**, 674–692.

Braasch, J. (**2005**). "Modelling of binaural hearing," in J. Blauert, ed., *Communication Acoustics*, Springer-Verlag, Berlin Heidelberg New York, pp. 75–108.

Braasch, J. and Blauert, J. (**2003**). "The precedence effect for noise bursts of different bandwidths. II. Comparison of model algorithms," *Acoust. Sci. & Tech.* **24**, 293–303.

Braasch, J. and Hartung, K. (**2002**). "Localization in the presence of a distracter and reverberation in the frontal horizontal plane. I. Psychoacoustical data," *ACUSTICA – Acta Acustica* **88**, 942–955.

Braasch, J., Blauert, J., and Djelani, T. (**2003**). "The precedence effect for noise bursts of different bandwidths. I. Psychoacoustical data," *Acoust. Sci. & Tech.* **24**, 233–241.

Bradley, J. S. (**1994**). "Comparison of concert hall measurements of spatial impression," *J. Acoust. Soc. Am.* **96**, 3525–3535.

Bradley, J. S. and Soulodre, G. A. (**1995a**). "The influence of late arriving energy on spatial impression," *J. Acoust. Soc. Am.* **97**, 2263–2271.

Bradley, J. S. and Soulodre, G. A. (**1995b**). "Objective measures of listener envelopment," *J. Acoust. Soc. Am.* **98**, 2590–2597.

Brand, A., Behrend, O., Marquardt, T., McAlpine, D., and Grothe, B. (**2002**). "Precise inhibition is essential for microsecond interaural time difference coding," *Nature* **417**, 543–547.

Brandenburg, K. and Bosi, M. (**1997**). "Overview of MPEG audio: Current and future standards for low-bit-rate audio coding," *J. Audio Eng. Soc.* **45**, 4–21.

Brandenburg, K. and Stoll, G. (**1994**). "ISO-MPEG-1 audio: A generic standard for coding of high-quality digital audio," *J. Audio Eng. Soc.* **42**, 780–792.

Brandstein, M. S., Adcock, J. E., and Silverman, H. F. (**1997**). "A closed-form location estimator for use with room environment microphone arrays," *IEEE Trans. Speech Audio Proc.* **5**, 45–50.

Brandstein, M. S. and Silverman, H. F. (**1997**). "A practical methodology for speech source localization with microphone arrays," *Computer Speech and Language* **11**, 91–126.

Breebaart, J., van de Par, S., and Kohlrausch, A. (**1999**). "The contribution of static and dynamically varying ITDs and IIDs to binaural detection," *J. Acoust. Soc. Am.* **106**, 979–992.

Breebaart, J., van de Par, S., and Kohlrausch, A. (**2001a**). "Binaural processing model based on contralateral inhibition. I. Model structure," *J. Acoust. Soc. Am.* **110**, 1074–1088.

Breebaart, J., van de Par, S., and Kohlrausch, A. (**2001b**). "Binaural processing model based on contralateral inhibition. II. Dependence on spectral parameters," *J. Acoust. Soc. Am.* **110**, 1089–1104.

Breebaart, J., van de Par, S., and Kohlrausch, A. (**2001c**). "Binaural processing model based on contralateral inhibition. III. Dependence on temporal parameters," *J. Acoust. Soc. Am.* **110**, 1105–1117.

Breebaart, J., van de Par, S., and Kohlrausch, A. (**2002**). "A time-domain binaural signal detection model and its predictions for temporal resolution data," *ACUSTICA – Acta Acustica* **88**, 110–112.

Breebaart, J., van de Par, S., Kohlrausch, A., and Schuijers, E. (**2004**). "High-quality parametric spatial audio coding at low bitrates," in *AES 116th Convention*, Berlin, Germany. Preprint 6072.

Broadhurst, A. D. (**1982**). "Sparse volume array for architectural acoustic measurements," *Acustica* **50**, 33–38.

Bronkhorst, A. W. (**1995**). "Localization of real and virtual sound sources," *J. Acoust. Soc. Am.* **98**, 2542–2553.

Bronkhorst, A. W. and Houtgast, T. (**1999**). "Auditory distance perception in rooms," *Nature* **397**, 517–520.

Brown, G. J., Harding, S., and Barker, J. P. (**2006**). "Speech separation based on the statistics of binaural auditory features," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*.

Brüggen, M. (**2001**). "Coloration and binaural decoloration in natural environments," *ACUSTICA – Acta Acustica* **87**, 400–406.

Buchholz, J. M. and Mourjopoulos, J. (**2004a**). "A computational auditory masking model based on signal-dependent compression. I. Model description and performance analysis," *ACUSTICA – Acta Acustica* **90**, 873–886.

Buchholz, J. M. and Mourjopoulos, J. (**2004b**). "A computational auditory masking model based on signal-dependent compression. II. Model simulations and analytical approximations," *ACUSTICA – Acta Acustica* **90**, 87–900.

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

Buell, T. N. and Hafter, E. R. (**1991**). "Combination of binaural information across frequency bands," *J. Acoust. Soc. Am.* **90**, 1894–1900.

Buell, T. N. and Trahiotis, C. (**1994**). "Detection of interaural delay in bands of noise: Effects of spectral interference combined with spectral uncertainty," *J. Acoust. Soc. Am.* **95**, 3568–3573.

Buell, T. N. and Trahiotis, C. (**1997**). "Recent experiments concerning the relative potency and interaction of interaural cues," in R. H. Gilkey and T. R. Anderson, eds, *Binaural and Spatial Hearing in Real and Virtual Environments*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, pp. 139–149.

Buell, T. N., Trahiotis, C., and Bernstein, L. R. (**1994**). "Lateralization of bands of noise as a function of combinations of interaural intensitive differences, interaural temporal differences, and bandwidth," *J. Acoust. Soc. Am.* **95**, 1482–1489.

Butler, R. A. and Naunton, R. F. (**1964**). "Role of stimulus frequency and duration in the phenomenon of localization shifts," *J. Acoust. Soc. Am.* **36**, 917–922.

Buus, S. (**1999**). "Temporal integration and multiple looks, revisited: Weights as a function of time," *J. Acoust. Soc. Am.* **105**, 2466–2475.

Carter, G. C. (**1987**). "Coherence and time delay estimation," *Proc. IEEE* **75**, 236–255.

Chan, Y. T. and Ho, K. C. (**1994**). "A simple and efficient estimator for hyperbolilc location," *IEEE Trans. Signal Processing* **42**, 1905–1915.

Chandler, D. W. and Grantham, D. W. (**1992**). "Minimum audible movement angle in the horizontal plane as a function of stimulus frequency and bandwidth, source azimuth, and velocity," *J. Acoust. Soc. Am.* **91**, 1624–1636.

Cherry, E. C. and Sayers, B. M. A. (**1956**). " "Human 'cross-correlator' "—A technique for measuring certain parameters of speech perception," *J. Acoust. Soc. Am.* **28**, 889–895.

Chiang, Y.-C. and Freyman, R. L. (**1998**). "The influence of broadband noise on the precedence effect," *J. Acoust. Soc. Am.* **104**, 3039–3047.

Chu, W. T. (**1981**). "Comments on the coherent and incoherent nature of a reverberant sound field," *J. Acoust. Soc. Am.* **69**, 1710–1715.

Chung, J. Y. (**1978**). "Cross-spectral method of measuring acoustic intensity without error caused by instrument phase mismatch," *J. Acoust. Soc. Am.* **64**, 1613–1616.

Clifton, R. K. (**1987**). "Breakdown of echo suppression in the precedence effect," *J. Acoust. Soc. Am.* **82**, 1834–1835.

Clifton, R. K. and Freyman, R. L. (**1989**). "Effect of click rate and delay on breakdown of the precedence effect," *Percept. Psychophys.* **46**, 139–145.

Clifton, R. K. and Freyman, R. L. (**1997**). "The precedence effect: Beyond echo suppression," in R. H. Gilkey and T. R. Anderson, eds, *Binaural and Spatial Hearing in Real and Virtual Environments*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, pp. 233–255.

Clifton, R. K., Freyman, R. L., Litovsky, R. Y., and McCall, D. (**1994**). "Listeners' expectations about echoes can raise or lower echo threshold," *J. Acoust. Soc. Am.* **95**, 1525–1533.

Cokely, J. A. and Hall, J. W. (**1991**). "Frequency resolution for diotic and dichotic listening conditions compared using the bandlimiting measure and a modified bandlimiting measure," *J. Acoust. Soc. Am.* pp. 1331–1339.

Colburn, H. S. (**1977**). "Theory of binaural interaction based on auditory-nerve data. II. Detection of tones in noise," *J. Acoust. Soc. Am.* **61**, 525–533.

Colburn, H. S. (**1996**). "Computational models of binaural processing," in H. Hawkins and T. McMullen, eds, *Auditory Computation*, volume in A. Popper and R. Fay, eds., *Springer Handbook on Auditory Research*, pp. 332–400.

Colburn, H. S. and Durlach, N. I. (**1978**). "Models of binaural interaction," in E. C. Carterette and M. P. Friedman, eds, *Handbook of Perception*, Vol. IV, Academic Press, Inc., New York, NY, USA, pp. 467–518.

Coleman, P. D. (**1962**). "Failure to localize the source distance of an unfamiliar sound," *J. Acoust. Soc. Am.* **34**, 345–346.

Coleman, P. D. (**1968**). "Dual role of frequency spectrum in determination of auditory

distance," *J. Acoust. Soc. Am.* **44**, 631–632.

Cook, R. K., Waterhouse, R. V., Berendt, R. D., Edelman, S., and Thompson, Jr., M. C. (**1955**). "Measurement of correlation coefficients in reverberant sound fields," *J. Acoust. Soc. Am.* **27**, 1072–1077.

Cotterell, P. S. (**2002**). *On the Theory of the Second-Order Soundfield Microphone*, PhD thesis, Instrumentation and Signal Processing Research Group, Department of Cybernetics, The University of Reading. Available at `http://www.ambisonic.net/links.html`.

Cox, T. J., Davies, W. J., and Lam, Y. W. (**1993**). "The sensitivity of listeners to early sound field changes in auditoria," *Acustica* **79**, 27–41.

Craven, P. G. and Gerzon, M. A. (**1977**). "Coincident microphone simulation covering three dimensional space and yielding various directional outputs," U.S. Patent Number 4,042,779.

Cremer, L. and Müller, H. A. (**1982a**). Principles and Applications of Room Acoustics, Vol. 1, Applied Science Publishers Ltd, Essex, England.

Cremer, L. and Müller, H. A. (**1982b**). Principles and Applications of Room Acoustics, Vol. 2, Applied Science Publishers Ltd, Essex, England.

Culling, J. F. and Colburn, H. S. (**2000**). "Binaural sluggishness in the perception of tone sequences and speech in noise," *J. Acoust. Soc. Am.* **107**, 517–527.

Culling, J. F. and Summerfield, A. Q. (**1998**). "Measurement of the binaural temporal window using a detection task," *J. Acoust. Soc. Am.* **103**, 3540–3553.

Culling, J. F., Colburn, H. S., and Spurchise, M. (**2001**). "Interaural correlation sensitivity," *J. Acoust. Soc. Am.* **110**, 1020–1029.

Dai, H. and Wright, B. A. (**1995**). "Detecting signals of unexpected or uncertain durations," *J. Acoust. Soc. Am.* **98**, 798–806.

Dai, H. and Wright, B. A. (**1999**). "Predicting the detectability of tones with unexpected durations," *J. Acoust. Soc. Am.* **105**, 2043–2046.

Dalenbäck, B.-I., Kleiner, M., and Svensson, P. (**1994**). "A macroscopic view of diffuse reflection," *J. Audio Eng. Soc.* **42**, 793–807.

Damaske, P. and Ando, Y. (**1972**). "Interaural crosscorrelation for multichannel loudspeaker reproduction," *Acustica* **27**, 232–238.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997a**). "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," *J. Acoust. Soc. Am.* **102**, 2892–2905.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997b**). "Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration," *J. Acoust. Soc. Am.* **102**, 2906–2919.

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

Davies, W. J. and Cox, T. J. (**2000**). "Reducing seat dip attenuation," *J. Acoust. Soc. Am.* **108**, 2211–2218.

de Bree, H.-E. (**2003**). "An overview of Microflown technologies," *ACUSTICA – Acta*

Acustica **89**, 163–172.

de Vries, D., Hulsebos, E. M., and Baan, J. (**2001**). "Spatial fluctuations in measures for spaciousness," *J. Acoust. Soc. Am.* **110**, 947–954.

Divenyi, P. L. (**1992**). "Binaural suppression of nonechoes," *J. Acoust. Soc. Am.* **91**, 1078–1084.

Dizon, R. M. and Litovsky, R. Y. (**2004**). "Localization dominance in the median-sagittal plane: Effect of stimulus duration," *J. Acoust. Soc. Am.* **115**, 3142–3155.

Djelani, T. and Blauert, J. (**2001**). "Investigations into the build-up and breakdown of the precedence effect," *ACUSTICA – Acta Acustica* **87**, 253–261.

Djelani, T. and Blauert, J. (**2002**). "Modelling the direction-specific build-up of the precedence effect," in *Proc. Forum Acusticum*, Sevilla, Spain.

Drullman, R. and Bronkhorst, A. W. (**2000**). "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation," *J. Acoust. Soc. Am.* **107**, 2224–2235.

Duda, R. O. (**1997**). "Elevation dependence of the interaural transfer function," in R. H. Gilkey and T. R. Anderson, eds, *Binaural and Spatial Hearing in Real and Virtual Environments*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, pp. 49–75.

Duda, R. O. and Martens, W. L. (**1998**). "Range dependence of the response of a spherical head model," *J. Acoust. Soc. Am.* **104**, 3048–3058.

Durlach, N. I. (**1963**). "Equalization and cancellation theory of binaural masking-level differences," *J. Acoust. Soc. Am.* **35**, 1206–1218.

Durlach, N. I. (**1966**). "On the application of the EC model to interaural jnd's," *J. Acoust. Soc. Am.* **40**, 1392–1397.

Dye, Jr., R. H. (**1997**). "The relative contributions of targets and distractors in judgments of laterality based on interaural differences of level," in R. H. Gilkey and T. R. Anderson, eds, *Binaural and Spatial Hearing in Real and Virtual Environments*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, pp. 151–168.

Dye, Jr., R. H., Niemiec, A. J., and Stellmack, M. A. (**1994a**). "Discrimination of interaural envelope delays: The effect of randomizing component starting phase," *J. Acoust. Soc. Am.* **95**, 463–470.

Dye, Jr., R. H., Stellmack, M. A., Grange, A. N., and Yost, W. A. (**1996**). "The effect of distractor frequency on judgments of target laterality based on interaural delays," *J. Acoust. Soc. Am.* **99**, 1096–1107.

Dye, Jr., R. H., Yost, W. A., Stellmack, M. A., and Sheft, S. (**1994b**). "Stimulus classification procedure for assessing the extent to which binaural processing is spectrally analytic or synthetic," *J. Acoust. Soc. Am.* **96**, 2720–2730.

Eargle, J. (**2001**). The Microphone Book, Focal Press, Woburn, MA, USA.

Ebata, M., Sone, T., and Nimura, T. (**1968**). "On the perception of direction of echo," *J. Acoust. Soc. Am.* **44**, 542–547.

Elko, G. W. (**2000**). "Steerable and variable first-order differential microphone array," U.S. Patent Number 6,041,127.

Elko, G. W., West, J. E., Kubli, R. A., and McAteer, J. P. (**1994**). "Image-derived

second-order differential microphones," *J. Acoust. Soc. Am.* **95**, 1991–1997.

Evjen, P., Bradley, J. S., and Norcross, S. G. (**2001**). "The effect of late reflections from above and behind on listener envelopment," *Appl. Acoust.* **62**, 137–153.

Fahy, F. J. (**1977**). "Measurement of acoustic intensity using the cross-spectral density of two microphone signals," *J. Acoust. Soc. Am.* **62**, 1057–1059.

Fahy, F. J. (**1989**). Sound Intensity, Elsevier Science Publishers Ltd., Essex, England.

Faller, C. (**2003**). "Parametric multi-channel audio coding: Synthesis of coherence cues," *IEEE Trans. Speech Audio Proc.* Submitted Dec. 2003, accepted for publication.

Faller, C. and Baumgarte, F. (**2001**). "Efficient representation of spatial audio using perceptual parametrization," in *Proc. IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust.*, pp. 199–202.

Faller, C. and Baumgarte, F. (**2003**). "Binaural Cue Coding. Part II: Schemes and applications," *IEEE Trans. Speech Audio Proc.* **11**, 520–531.

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

Fang, B. T. (**1990**). "Simple solutions for hyperbolic and related position fixes," *IEEE Trans. Aerosp. Electron. Syst.* **26**, 748–753.

Farina, A. (**2000**). "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *AES 108th Convention*, Paris, France. Preprint 5093.

Farina, A. (**2001a**). "Anechoic measurement of the polar plot of B-format microphones," `http://pcfarina.eng.unipr.it/Public/Soundfield/St-250_vs_-BruelKjaer.PDF`.

Farina, A. (**2001b**). "Anechoic measurement of the polar plots of a Soundfield MKV B-format microphone," `http://pcfarina.eng.unipr.it/Public/Soundfield/MKV-PolarPatterns.PDF`.

Farina, A. and Ugolotti, E. (**1999**). "Subjective comparison between stereo dipole and 3D Ambisonic surround systems for automotive applications," in *Proc. AES 16th Int. Conf.*, Rovaniemi, Finland, pp. 532–543.

Farrar, K. (**1979a**). "Soundfield microphone: Design and development of microphone and control unit," *Wireless World* **85**, 48–50.

Farrar, K. (**1979b**). "Soundfield microphone: Detailed functioning of control unit," *Wireless World* **85**, 99–103.

Flanagan, J. L. (**1985**). "Use of acoustic filtering to control the beamwidth of steered microphone arrays," *J. Acoust. Soc. Am.* **78**, 423–428.

Flanagan, J. L., Johnston, J. D., Zahn, R., and Elko, G. W. (**1985**). "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.* **78**, 1508–1518.

Fredriksson, M. and Zacharov, N. (**2002**). "Natural reproduction of music and environmental sounds," in *AES 112th Convention*, Munich, Germany. Preprint 5581.

Freyman, R. L., Clifton, R. K., and Litovsky, R. Y. (**1991**). "Dynamic processes in

the precedence effect," *J. Acoust. Soc. Am.* **90**, 874–884.

Furuya, H., Fujimoto, K., Ji, C. Y., and Higa, N. (**2001**). "Arrival direction of late sound and listener envelopment," *Appl. Acoust.* **62**, 125–136.

Gabriel, K. J. and Colburn, H. S. (**1981**). "Interaural correlation discrimination: I. Bandwidth and level dependence," *J. Acoust. Soc. Am.* **69**, 1394–1401.

Gade, A. C. (**1989**). "Investigations of musicians' room acoustic conditions in concert halls. Part II: Field experiments and synthesis of results," *Acustica* **69**, 249–262.

Gaik, W. (**1993**). "Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling," *J. Acoust. Soc. Am.* **94**, 98–110.

Gardner, B. and Martin, K. (**1994**). "HRTF measurements of a KEMAR dummy-head microphone," Technical Report 280, MIT Media Lab Perceptual Computing. Available at `http://sound.media.mit.edu/KEMAR.html`.

Gardner, M. B. (**1973a**). "Some monaural and binaural facets of median plane localization," *J. Acoust. Soc. Am.* **54**, 1489–1495.

Gardner, M. B. (**1973b**). "Some single- and multiple-source localization effects," *J. Audio Eng. Soc.* **21**, 430–437.

Gardner, W. G. (**1998**). "Reverberation algorithms," in M. Kahrs and K. Brandenburg, eds, *Applications of Digital Signal Processing to Audio and Acoustics*, Kluwer Academic Publishers, Norwell, MA, USA, pp. 85–131.

Gardner, W. G. and Martin, K. D. (**1995**). "HRTF measurements of a KEMAR," *J. Acoust. Soc. Am.* **97**, 3907–3908.

Geisler, C. D. and Cai, Y. (**1996**). "Relationships between frequency-tuning and spatial-tuning curves in the mammalian cochlea," *J. Acoust. Soc. Am.* **99**, 1550–1555.

Gerzon, M. A. (**1973**). "Periphony: With-height sound reproduction," *J. Audio Eng. Soc.* **21**, 2–10.

Gerzon, M. A. (**1975**). "The design of precisely coincident microphone arrays for stereo and surround sound," in *AES 50th Convention*, London, UK. Paper L-20.

Gerzon, M. A. (**1992**). "Signal processing for simulating realistic stereo images," in *AES 93rd Convention*, San Francisco, CA, USA. Preprint 3423.

Giguère, C. and Abel, S. M. (**1993**). "Sound localization: Effects of reverberation time, speaker array, stimulus frequency, and stimulus rise/decay," *J. Acoust. Soc. Am.* **94**, 769–776.

Giguère, C. and Woodland, P. C. (**1994a**). "A computational model of the auditory periphery for speech and hearing research. I. Ascending path," *J. Acoust. Soc. Am.* **95**, 331–342.

Giguère, C. and Woodland, P. C. (**1994b**). "A computational model of the auditory periphery for speech and hearing research. II. Descending path," *J. Acoust. Soc. Am.* **95**, 343–349.

Glasberg, B. R. and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.* **47**, 103–138.

Glasberg, B. R. and Moore, B. C. J. (**2000**). "Frequency selectivity as a function of level and frequency measured with uniformly exciting notched noise," *J. Acoust.*

Soc. Am. **108**, 2318–2328.

Glasberg, B. R., Moore, B. C. J., and Nimmo-Smith, I. (**1984**). "Comparison of auditory filter shapes derived with three different maskers," *J. Acoust. Soc. Am.* **75**, 536–544.

Good, M. D. and Gilkey, R. H. (**1996**). "Sound localization in noise: The effect of signal to noise ratio," *J. Acoust. Soc. Am.* **99**, 1108–1117.

Good, M. D., Gilkey, R. H., and Ball, J. M. (**1997**). "The relation between detection in noise and localization in noise in the free field," in R. H. Gilkey and T. R. Anderson, eds, *Binaural and Spatial Hearing in Real and Virtual Environments*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, pp. 349–376.

Gover, B. N. (**2002**). "Microphone array measurement system for analysis of directional and spatial variations of sound fields," *J. Acoust. Soc. Am.* **112**, 1980–1991.

Grantham, D. W. (**1982**). "Detectability of time-varying interaural correlation in narrow-band noise stimuli," *J. Acoust. Soc. Am.* **72**, 1178–1184.

Grantham, D. W. (**1984a**). "Discrimination of dynamic interaural intensity differences," *J. Acoust. Soc. Am.* **76**, 71–76.

Grantham, D. W. (**1984b**). "Interaural intensity discrimination: Insensitivity at 1000 Hz," *J. Acoust. Soc. Am.* **75**, 1191–1194.

Grantham, D. W. (**1986**). "Detection and discrimination of simulated motion of auditory targets in the horizontal plane," *J. Acoust. Soc. Am.* **79**, 1939–1949.

Grantham, D. W. (**1996**). "Left–right asymmetry in the buildup of echo suppression in normal-hearing adults," *J. Acoust. Soc. Am.* **99**, 1118–1123.

Grantham, D. W. and Wightman, F. L. (**1978**). "Detectability of varying interaural temporal differences," *J. Acoust. Soc. Am.* **63**, 511–523.

Grantham, D. W. and Wightman, F. L. (**1979**). "Detectability of a pulsed tone in the presence of a masker with time-varying interaural correlation," *J. Acoust. Soc. Am.* **65**, 1509–1517.

Green, D. M. (**1966**). "Interaural phase effects in the masking of signals of different duration," *J. Acoust. Soc. Am.* **39**, 720–724.

Greenwood, D. D. (**1961**). "Critical bandwidth and the frequency coordinates of the basilar membrane," *J. Acoust. Soc. Am.* **33**, 1344–1356.

Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species—29 years later," *J. Acoust. Soc. Am.* **87**, 2592–2605.

Griesinger, D. (**1992**). "IALF — Binaural measures of spatial impression and running reverberation," in *AES 92nd Convention*, Vienna, Austria. Preprint 3292.

Griesinger, D. (**1997**). "The psychoacoustics of apparent source width, spaciousness and envelopment in performance spaces," *ACUSTICA – Acta Acustica* **83**, 721–731.

Griesinger, D. (**1999**). "Objective measures of spaciousness and envelopment," in *Proc. AES 16th Int. Conf.*, Rovaniemi, Finland, pp. 27–41.

Griffiths, T. D. and Warren, J. D. (**2004**). "What is an auditory object?," *Nature Rev. Neurosci.* **5**, 887–892.

Grothe, B. (**2003**). "New roles for synaptic inhibition in sound localization," *Nature Rev. Neurosci.* **4**, 1–11.

Haas, H. (**1949**). *Über den Einfluss eines Einfachechos auf die Hörsamkeit von Sprache*, PhD thesis, University of Göttingen, Göttingen, Germany. "The Influence of a Single Echo on the Audibility of Speech (transl.)," *J. Audio Eng. Soc.* **20**, 146–159 (**1972**).

Hafter, E. R. and Carrier, S. C. (**1972**). "Binaural interaction in low-frequency stimuli: The inability to trade time and intensity completely," *J. Acoust. Soc. Am.* **51**, 1852–1862.

Hahn, W. R. and Tretter, S. A. (**1973**). "Optimum processing for delay-vector estimation in passive signal arrays," *IEEE Trans. Inf. Theory* **IT-19**, 608–614.

Hall, J. W., Tyler, R. S., and Fernandes, M. A. (**1983**). "Monaural and binaural auditory frequency resolution measured using bandlimited noise and notched-noise masking," *J. Acoust. Soc. Am.* **73**, 894–898.

Hammershøi, D. and Møller, H. (**1996**). "Sound transmission to and within the human ear canal," *J. Acoust. Soc. Am.* **100**, 408–427.

Hammershøi, D. and Møller, H. (**2005**). "Binaural technique — Basic methods for recording, synthesis, and reproduction," in J. Blauert, ed., *Communication Acoustics*, Springer-Verlag, Berlin Heidelberg New York, pp. 223–254.

Hanyu, T. and Kimura, S. (**2001**). "A new objective measure for evaluation of listener envelopment focusing on the spatial balance of reflections," *Appl. Acoust.* **62**, 155–184.

Hartmann, W. M. (**1983**). "Localization of sound in rooms," *J. Acoust. Soc. Am.* **74**, 1380–1391.

Hartmann, W. M. (**1997**). "Listening in a room and the precedence effect," in R. H. Gilkey and T. R. Anderson, eds, *Binaural and Spatial Hearing in Real and Virtual Environments*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, pp. 349–376.

Hartmann, W. M. and Rakerd, B. (**1989**). "Localization of sound in rooms, IV: The Franssen effect," *J. Acoust. Soc. Am.* **86**, 1366–1373.

Hartmann, W. M. and Wittenberg, A. (**1996**). "On the externalization of sound images," *J. Acoust. Soc. Am.* **99**, 3678–3688.

Hartung, K. and Trahiotis, C. (**2001**). "Peripheral auditory processing and investigations of the "precedence effect" which utilize successive transient stimuli," *J. Acoust. Soc. Am.* **110**, 1505–1513.

Hawksford, M. O. J. and Harris, N. (**2002**). "Diffuse signal processing and acoustic source characterization for applications in synthetic loudspeaker arrays," in *AES 112th Convention*, Munich, Germany. Preprint 5612.

Hawley, M. L., Litovsky, R. Y., and Colburn, H. S. (**1999**). "Speech intelligibility and localization in a multi-source environment," *J. Acoust. Soc. Am.* **105**, 3436–3448.

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

Heinz, M. G., Colburn, H. S., and Carney, L. H. (**2002**). "Quantifying the implications of nonlinear cochlear tuning for auditory-filter estimates," *J. Acoust. Soc. Am.* **111**, 996–1011.

Heinz, M. G., Zhang, X., Bruce, I. C., and Carney, L. H. (**2001**). "Auditory nerve

model for predicting performance limits of normal and impaired listeners," *Acoust. Res. Letters Online* **2**, 91–96.

Henning, G. B. (**1974**). "Detectability of interaural delay in high-frequency complex waveforms," *J. Acoust. Soc. Am.* **55**, 84–90.

Henning, G. B. (**1980**). "Some observations on the lateralization of complex waveforms," *J. Acoust. Soc. Am.* **68**, 446–454.

Hess, W. (**2006**). *Time-Variant Binaural-Activity Characteristics as Indicator of Auditory Spatial Attributes*, PhD thesis, Institut für Kommunikationsakustik, Ruhr-Universität Bochum, Germany.

Hewitt, M. J. and Meddis, R. (**1991**). "An evaluation of eight computer models of mammalian inner hair-cell function," *J. Acoust. Soc. Am.* **90**, 904–917.

Heyser, R. C. (**1967**). "Acoustical measurements by time delay spectrometry," *J. Audio Eng. Soc.* **15**, 370–382.

Hicks, M. L. and Bacon, S. P. (**1999**). "Psychophysical measures of auditory nonlinearities as a function of frequency in individuals with normal hearing," *J. Acoust. Soc. Am.* **105**, 326–338.

Hidaka, T., Beranek, L. L., and Okano, T. (**1995**). "Interaural cross-correlation, lateral fraction, and low- and high-frequency sound level measures of acoustical quality in concert halls," *J. Acoust. Soc. Am.* **98**, 988–1007.

Hill, N. I. and Darwin, C. J. (**1996**). "Lateralization of a perturbed harmonic: Effects of onset asynchrony and mistuning," *J. Acoust. Soc. Am.* **100**, 2352–2364.

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

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

Holm, S., Austeng, A., Iranpour, K., and Hopperstad, J.-F. (**2001**). "Sparse sampling in array processing," in F. Marvasti, ed., *Nonuniform Sampling: Theory and Practice*, Kluwer Academic/Plenum Publishers, New York, NY, USA, pp. 787–833.

Holube, I., Kinkel, M., and Kollmeier, B. (**1998**). "Binaural and monaural auditory filter bandwidths and time constants in probe tone detection experiments," *J. Acoust. Soc. Am.* **104**, 2412–2425.

Hudde, H. (**2005**). "A functional view on the peripheral human hearing organ," in J. Blauert, ed., *Communication Acoustics*, Springer-Verlag, Berlin Heidelberg New York, pp. 47–74.

Hudde, H. and Engel, A. (**1998a**). "Measuring and modeling basic properties of the human middle ear and ear canal. Part I: Model structure and measuring techniques," *ACUSTICA – Acta Acustica* **84**, 720–738.

Hudde, H. and Engel, A. (**1998b**). "Measuring and modeling basic properties of the human middle ear and ear canal. Part II: Ear canal, middle ear cavities, eardrum and ossicles," *ACUSTICA – Acta Acustica* **84**, 894–913.

Hudde, H. and Engel, A. (**1998c**). "Measuring and modeling basic properties of the human middle ear and ear canal. Part III: Eardrum impedances, transfer functions and model calculations," *ACUSTICA – Acta Acustica* **84**, 1091–1109.

Humanski, R. A. and Butler, R. A. (**1988**). "The contribution of the near and far ear toward localization of sound in the sagittal plane," *J. Acoust. Soc. Am.* **83**, 2300–2310.

IEEE (**1969**). "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.* **17**, 137–148.

Irino, T. and Patterson, R. D. (**2001**). "A compressive gammachirp auditory filter for both physiological and psychophysical data," *J. Acoust. Soc. Am.* **109**, 2008–2022.

ISO 3382 (**1997**). "Acoustics — Measurement of the reverberation time of rooms with reference to other acoustical parameters," International Standards Organization.

ISO 389 (**1975**). "Acoustics — Standard reference zero for the calibration of pure-tone audiometers," International Standards Organization.

ISO 9613-1 (**1993**). "Acoustics — Attenuation of sound during propagation outdoors — Part 1: Calculation of the absorption of sound by the atmosphere," International Standards Organization.

ITU-R BS.1116-1 (**1997**). "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," International Telecommunication Union Radiocommunication Assembly.

ITU-R BS.1284-1 (**2003**). "General methods for the subjective assessment of sound quality," International Telecommunication Union Radiocommunication Assembly.

ITU-R BS.1534-1 (**2001-2003**). "Method for the subjective assessment of intermediate quality level of coding systems," International Telecommunication Union Radiocommunication Assembly.

ITU-R BS.775-1 (**1992–1994**). "Multichannel stereophonic sound system with and without accompanying picture," International Telecommunication Union Radiocommunication Assembly.

Jacobsen, F. (**2002**). "A note on finite difference estimation of acoustic particle velocity," *J. Sound Vib.* **256**, 849–859.

Jacobsen, F., Cutanda, V., and Juhl, P. M. (**1998**). "A numerical and experimental investigation of the performance of sound intensity probes at high frequencies," *J. Acoust. Soc. Am.* **103**, 953–961.

Jacovitti, G. and Scarano, G. (**1993**). "Discrete time techniques for time delay estimation," *IEEE Trans. Signal Processing* **41**, 525–533.

Jagger, D. S. (**1984**). "Recent developments and improvements in soundfield microphone technology," in *AES 75th Convention*, Paris, France. Preprint 2064.

Jain, M., Gallagher, D. T., Koehnke, J., and Colburn, H. S. (**1991**). "Fringed correlation discrimination and binaural detection," *J. Acoust. Soc. Am.* **90**, 1918–1926.

Jeffress, L. A. (**1948**). "A place theory of sound localization," *J. Comp. Physiol. Psych.* **41**, 35–39.

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

Johnson, D. H. (**1980**). "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones," *J. Acoust. Soc. Am.* **68**, 1115–1122.

Joris, P. X. and Yin, T. C. T. (**1995**). "Envelope coding in the lateral superior olive. I. Sensitivity to interaural time differences," *J. Neurophys.* **73**, 1043–1062.

Jot, J.-M., Cerveau, L., and Warusfel, O. (**1997**). "Analysis and synthesis of room reverberation based on a statistical time-frequency model," in *AES 103rd Convention*, New York, NY, USA. Preprint 4629.

Juhl, P. and Jacobsen, F. (**2004**). "A note on measurement of sound pressure with intensity probes," *J. Acoust. Soc. Am.* **116**, 1614–1620.

Karjalainen, M. (**1985**). "A new auditory model for the evaluation of sound quality of audio systems," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 608–611.

Karjalainen, M. (**1996**). "Binaural auditory model for sound quality measurements and spatial hearing studies," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 985–988.

Kates, J. M. (**1985**). "A central spectrum model for the perception of coloration in filtered Gaussian noise," *J. Acoust. Soc. Am.* **77**, 1529–1534.

Keller, C. H. and Takahashi, T. T. (**2005**). "Localization and identification of concurrent sounds in the owl's auditory space map," *J. Neurosci.* **25**, 10446–10461.

Kendall, G. S. (**1995**). "The decorrelation of audio signals and its impact on spatial imagery," *Comput. Music J.* **19**, 71–87.

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

Kleiner, M., Dalenbäck, B.-I., and Svensson, P. (**1993**). "Auralization—An overview," *J. Audio Eng. Soc.* **41**, 861–875.

Knapp, C. H. and Carter, G. C. (**1976**). "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Processing* **ASSP-24**, 320–327.

Koehnke, J., Colburn, H. S., and Durlach, N. I. (**1986**). "Performance in several binaural-interaction experiments," *J. Acoust. Soc. Am.* **79**, 1558–1562.

Kohlrausch, A. (**1988**). "Auditory filter shape derived from binaural masking experiments," *J. Acoust. Soc. Am.* **84**, 573–583.

Kohlrausch, A. (**1990**). "Binaural masking experiments using noise maskers with frequency-dependent interaural phase differences. I: Influence of signal and masker duration," *J. Acoust. Soc. Am.* **88**, 1737–1748.

Kohlrausch, A. and Fassel, R. (**1997**). "Binaural masking level differences in nonsimultaneous masking," in R. H. Gilkey and T. R. Anderson, eds, *Binaural and Spatial Hearing in Real and Virtual Environments*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, pp. 169–190.

Kohlrausch, A. and van de Par, S. (**2005**). "Audio–visual interaction in the context of multi-media applications," in J. Blauert, ed., *Communication Acoustics*, Springer-Verlag, Berlin Heidelberg New York, pp. 109–138.

Koivuniemi, K. and Zacharov, N. (**2001**). "Unravelling the perception of spatial sound reproduction: Language development, verbal protocol analysis and listener train-

ing," in *AES 111th Convention*, New York, NY, USA. Preprint 5424.

Kollmeier, B. and Gilkey, R. H. (**1990**). "Binaural forward and backward masking: Evidence for sluggishness in binaural detection," *J. Acoust. Soc. Am.* **87**, 1709–1719.

Kollmeier, B. and Holube, I. (**1992**). "Auditory filter bandwidths in binaural and monaural listening conditions," *J. Acoust. Soc. Am.* **92**, 1889–1901.

Korenbaum, V. I. (**1992**). "Comments on "Unidirectional, second-order gradient microphone"," *J. Acoust. Soc. Am.* **92**, 583–584.

Krumbholz, K. and Nobbe, A. (**2002**). "Buildup and breakdown of echo suppression for stimuli presented over headphones—The effects of interaural time and level differences," *J. Acoust. Soc. Am.* **112**, 654–663.

Kulkarni, A. and Colburn, H. S. (**1998**). "Role of spectral detail in sound-source localization," *Nature* **396**, 747–749.

Kummer, W. H. (**1992**). "Basic array theory," *Proc. IEEE* **80**, 127–140.

Kurozumi, K. and Ohgushi, K. (**1983**). "The relationship between the cross-correlation coefficient for two-channel acoustic signals and sound image quality," *J. Acoust. Soc. Am.* **74**, 1726–1733.

Kuttruff, H. (**2000**). Room Acoustics, 4th edn, Spon Press, London, England.

Kuttruff, H. and Schmitz, A. (**1994**). "Measurement of sound intensity by means of multi-microphone probes," *Acustica* **80**, 388–396.

Kuwada, S., Batra, R., and Fitzpatrick, D. C. (**1997**). "Neural processing of binaural temporal cues," in R. H. Gilkey and T. R. Anderson, eds, *Binaural and Spatial Hearing in Real and Virtual Environments*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, pp. 399–425.

Laborie, A., Bruno, R., and Montoya, S. (**2003**). "A new comprehensive approach of surround sound recording," in *AES 114th Convention*, Amsterdam, The Netherlands. Preprint 5717.

Laborie, A., Bruno, R., and Montoya, S. (**2004a**). "Designing high spatial resolution microphones," in *AES 117th Convention*, San Francisco, CA, USA. Preprint 6231.

Laborie, A., Bruno, R., and Montoya, S. (**2004b**). "High spatial resolution multichannel recording," in *AES 116th Convention*, Berlin, Germany. Preprint 6116.

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

Langendijk, E. H. A., Kistler, D. J., and Wightman, F. L. (**2001**). "Sound localization in the presence of one or two distracters," *J. Acoust. Soc. Am.* **109**, 2123–2134.

Langhans, A. and Kohlrausch, A. (**1992**). "Spectral integration of broadband signals in diotic and dichotic masking experiments," *J. Acoust. Soc. Am.* **91**, 317–326.

Lehnert, H. and Blauert, J. (**1992**). "Principles of binaural room simulation," *Appl. Acoust.* **36**, 259–291.

Liberman, M. C. (**1978**). "Auditory-nerve responses from cats raised in a low-noise chamber," *J. Acoust. Soc. Am.* **63**, 442–455.

Lindemann, W. (**1986a**). "Extension of a binaural cross-correlation model by means of contralateral inhibition. I. Simulation of lateralization for stationary signals," *J. Acoust. Soc. Am.* **80**, 1608–1622.

Lindemann, W. (**1986b**). "Extension of a binaural cross-correlation model by means of contralateral inhibition. II. The law of the first wave front," *J. Acoust. Soc. Am.* **80**, 1623–1630.

Lindevald, I. M. and Benade, A. H. (**1986**). "Two-ear correlation in the statistical sound fields of rooms," *J. Acoust. Soc. Am.* **80**, 661–664.

Lipshitz, S. P. (**1986**). "Stereo microphone techniques... Are the purists wrong?," *J. Audio Eng. Soc.* **34**, 716–744.

Litovsky, R. Y. and Macmillan, N. A. (**1994**). "Sound localization precision under conditions of the precedence effect: Effects of azimuth and standard stimuli," *J. Acoust. Soc. Am.* **96**, 752–758.

Litovsky, R. Y. and Shinn-Cunningham, B. G. (**2001**). "Investigation of the relationship among three common measures of precedence: Fusion, localization dominance, and discrimination suppression," *J. Acoust. Soc. Am.* **109**, 346–358.

Litovsky, R. Y., Colburn, H. S., Yost, W. A., and Guzman, S. J. (**1999**). "The precedence effect," *J. Acoust. Soc. Am.* **106**, 1633–1654.

Litovsky, R. Y., Rakerd, B., Yin, T. C. T., and Hartmann, W. M. (**1997**). "Psychophysical and physiological evidence for a precedence effect in the median sagittal plane," *J. Neurophys.* **77**, 2223–2226.

Lokki, T. (**2002**). *Physically-Based Auralization—Design, Implementation, and Evaluation*, PhD thesis, Helsinki University of Technology, Telecommunications Software and Multimedia Laboratory.

Lokki, T. (**2005**). "Subjective comparison of four concert halls based on binaural impulse responses," *Acoust. Sci. & Tech.* **26**, 200–203.

Lokki, T., Merimaa, J., and Pulkki, V. (**2003–2006**). "A method for reproducing natural or modified spatial impression in multichannel listening," Patent applications FI 20030294, JP 2006502072, and US 10/547,151.

Lopez-Poveda, E. A. and Meddis, R. (**2001**). "A human nonlinear cochlear filterbank," *J. Acoust. Soc. Am.* **110**, 3107–3118.

Lopez-Poveda, E. A., Plack, C. J., and Meddis, R. (**2003**). "Cochlear nonlinearity between 500 and 8000 Hz in listeners with normal hearing," *J. Acoust. Soc. Am.* **113**, 951–960.

Lorenzi, C., Gatehouse, S., and Lever, C. (**1999**). "Sound localization in noise in normal-hearing listeners," *J. Acoust. Soc. Am.* **105**, 1810–1820.

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

Makous, J. C. and Middlebrooks, J. C. (**1990**). "Two-dimensional sound localization by human listeners," *J. Acoust. Soc. Am.* **87**, 2188–2200.

Mann, III, J. A. and Tichy, J. (**1991**). "Acoustic intensity analysis: Distinguishing energy propagation and wave-front propagation," *J. Acoust. Soc. Am.* **90**, 20–25.

Mann, III, J. A., Tichy, J., and Romano, A. J. (**1987**). "Instantaneous and time-averaged energy transfer in acoustic fields," *J. Acoust. Soc. Am.* **82**, 17–30.

Marshall, A. H. (**1967**). "A note on the importance of room cross-section in concert

halls," *J. Sound Vib.* **5**, 100–112.

Marshall, A. H. and Barron, M. (**2001**). "Spatial responsiveness in concert halls and the origins of spatial impression," *Appl. Acoust.* **62**, 91–108.

Martin, K. D. (**1997**). "Echo suppression in a computational model of the precedence effect," in *Proc. IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust.*, New Paltz, NY, USA.

Mason, R. and Rumsey, F. (**2002**). "A comparison of objective measurements for predicting selected subjective spatial attributes," in *AES 112th Convention*, Munich, Germany. Preprint 5591.

Mason, R., Brookes, T., and Rumsey, F. (**2005**). "Frequency dependency of the relationship between perceived auditory source width and the interaural cross-correlation coefficient for time-invariant stimuli," *J. Acoust. Soc. Am.* **117**, 1337–1350.

Mason, R., Rumsey, F., and de Bruyn, B. (**2001a**). "An investigation of interaural time difference fluctuations, part 1: The subjective spatial effect of fluctuations delivered over headphones," in *AES 110th Convention*, Amsterdam, The Netherlands. Preprint 5383.

Mason, R., Rumsey, F., and de Bruyn, B. (**2001b**). "An investigation of interaural time difference fluctuations, part 2: Dependence of the subjective effect on audio frequency," in *AES 110th Convention*, Amsterdam, The Netherlands. Preprint 5389.

Mason, R., Rumsey, F., and de Bruyn, B. (**2001c**). "An investigation of interaural time difference fluctuations, part 3: The subjective effect of fluctuations in continuous stimuli delivered over loudspeakers," in *AES 111th Convention*, New York, NY, USA. Preprint 5457.

Mason, R., Rumsey, F., and de Bruyn, B. (**2001d**). "An investigation of interaural time difference fluctuations, part 4: The subjective effect of fluctuations in decaying stimuli delivered over loudspeakers," in *AES 111th Convention*, New York, NY, USA. Preprint 5458.

McAlpine, D., Jiang, D., and Palmer, A. R. (**2001**). "A neural code for low-frequency sound localization in mammals," *Nature Neurosci.* **4**, 396–401.

McCall, D. D., Freyman, R. L., and Clifton, R. K. (**1998**). "Sudden changes in spectrum of an echo cause a breakdown of the precedence effect," *Percept. Psychophys.* **60**, 593–601.

McFadden, D. and Pasanen, E. G. (**1976**). "Lateralization at high frequencies based on interaural time differences," *J. Acoust. Soc. Am.* **59**, 634–639.

Meddis, R. (**1986**). "Simulation of mechanical to neural transduction in the auditory receptor," *J. Acoust. Soc. Am.* **79**, 702–711.

Meddis, R. (**1988**). "Simulation of auditory-neural transduction: Further studies," *J. Acoust. Soc. Am.* **83**, 1056–1063.

Meddis, R., Hewitt, M. J., and Shackleton, T. M. (**1990**). "Implementation details of a computation model of the inner hair-cell/auditory-nerve synapse," *J. Acoust. Soc. Am.* **87**, 1813–1816.

Meddis, R., O'Mard, L. P., and Lopez-Poveda, E. A. (**2001**). "A computational algorithm for computing nonlinear auditory frequency selectivity," *J. Acoust. Soc. Am.*

109, 2852–2861.

Mehrgardt, S. and Mellert, V. (**1977**). "Transformation characteristics of the external human ear," *J. Acoust. Soc. Am.* **61**, 1567–1576.

Merimaa, J. (**2002**). "Applications of a 3-D microphone array," in *AES 112th Convention*, Munich, Germany. Preprint 5501.

Merimaa, J. and Hess, W. (**2004**). "Training of listeners for evaluation of spatial attributes of sound," in *AES 117th Convention*, San Francisco, CA, USA. Preprint 6237.

Merimaa, J. and Pulkki, V. (**2003**). "Perceptually-based processing of directional room responses for multichannel loudspeaker reproduction," in *Proc. IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust.*, New Paltz, NY, USA, pp. 51–54.

Merimaa, J. and Pulkki, V. (**2004**). "Spatial Impulse Response Rendering," in *Proc. 7th Int. Conf. on Digital Audio Effects (DAFx-04)*, Naples, Italy, pp. 139–144.

Merimaa, J. and Pulkki, V. (**2005**). "Spatial Impulse Response Rendering I: Analysis and synthesis," *J. Audio Eng. Soc.* **53**, 1115–1127.

Merimaa, J., Hess, W., and Blauert, J. (**2005a**). "Der auditive Raumeindruck in zwei Konzertsälen stark unterschiedlicher Größe," in *Fortschritte der Akustik DAGA'05*, Munich, Germany, pp. 669–670. In German.

Merimaa, J., Lokki, T., Peltonen, T., and Karjalainen, M. (**2001**). "Measurement, analysis, and visualization of directional room responses," in *AES 111th Convention*, New York, NY, USA. Preprint 5449.

Merimaa, J., Peltonen, T., and Lokki, T. (**2005b**). "Concert hall impulse responses — Pori, Finland," `http://www.acoustics.hut.fi/projects/poririrs/`.

Mershon, D. H. (**1997**). "Phenomenal geometry and the measurement of perceived auditory distance," in R. H. Gilkey and T. R. Anderson, eds, *Binaural and Spatial Hearing in Real and Virtual Environments*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, pp. 257–274.

Meyer, J. (**2001**). "Beamforming for a circular microphone array mounted on spherically shaped objects," *J. Acoust. Soc. Am.* **109**, 185–193.

Meyer, J. and Elko, G. (**2002**). "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, Vol. 2, pp. 1781–1784.

Middlebrooks, J. C. (**1992**). "Narrow-band sound localization related to external ear acoustics," *J. Acoust. Soc. Am.* **92**, 2607–2624.

Middlebrooks, J. C. (**1997**). "Spectral shape cues for sound localization," in R. H. Gilkey and T. R. Anderson, eds, *Binaural and Spatial Hearing in Real and Virtual Environments*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, pp. 77–97.

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

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

Middlebrooks, J. C. and Green, D. M. (**1990**). "Directional dependence of interaural

envelope delays," *J. Acoust. Soc. Am.* **87**, 2149–2162.

Middlebrooks, J. C., Makous, J. C., and Green, D. M. (**1989**). "Directional sensitivity of sound-pressure levels in the human ear canal," *J. Acoust. Soc. Am.* **86**, 89–108.

Møller, H. (**1992**). "Fundamentals of binaural technology," *Appl. Acoust.* **36**, 171–218.

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

Møller, H., Sørensen, M. F., Jensen, C. B., and Hammershøi, D. (**1996**). "Binaural technique: Do we need individual recordings," *J. Audio Eng. Soc.* **44**, 451–469.

Monro, G. (**2000**). "In-phase corrections for Ambisonics," in *Proc. Int. Computer Music Conf.*, Berlin, Germany, pp. 292–295.

Moore, B. C. J. (**1987**). "Distribution of auditory-filter bandwidths at 2 kHz in young normal listeners," *J. Acoust. Soc. Am.* **81**, 1633–1635.

Moore, B. C. J. and Glasberg, B. R. (**1981**). "Auditory filter shapes derived in simultaneous and forward masking," *J. Acoust. Soc. Am.* **70**, 1003–1014.

Moore, B. C. J. and Glasberg, B. R. (**1983**). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Am.* **74**, 750–753.

Moore, B. C. J., Glasberg, B. R., Plack, C. J., and Biswas, A. K. (**1988**). "The shape of the ear's temporal window," *J. Acoust. Soc. Am.* **83**, 1102–1116.

Moore, B. C. J., Glasberg, B. R., van der Hejden, M., Houtsma, A. J. M., and Kohlrausch, A. (**1995**). "Comparison of auditory filter shapes obtained with notched-noise and noise-tone maskers," *J. Acoust. Soc. Am.* **97**, 1175–1182.

Moore, B. C. J., Peters, R. W., and Glasberg, B. R. (**1990**). "Auditory filter shapes at low center frequencies," *J. Acoust. Soc. Am.* **88**, 132–140.

Moore, B. C. J., Vickers, D. A., Plack, C. J., and Oxenham, A. J. (**1999**). "Interrelationship between different psychoacoustic measures assumed to be related to the cochlear active mechanism," *J. Acoust. Soc. Am.* **106**, 2761–2778.

Morimoto, M. (**2001**). "The contribution of two ears to the perception of vertical angle in sagittal planes," *J. Acoust. Soc. Am.* **109**, 1569–1603.

Morimoto, M. and Maekawa, Z. (**1988**). "Effects of low frequency components on auditory spaciousness," *Acustica* **66**, 190–196.

Morimoto, M. and Maekawa, Z. (**1989**). "Auditory spaciousness and envelopment," in *Proc. 13th Int. Congr. Acoust.*, Vol. 2, Belgrade, Yugoslavia, pp. 215–218.

Morimoto, M., Iida, K., and Furue, Y. (**1993**). "Relation between auditory source width in various sound fields and degree of interaural cross-correlation," *Appl. Acoust.* **38**, 291–301.

Morimoto, M., Iida, K., and Sakagami, K. (**2001**). "The role of reflections from behind the listener in spatial impression," *Appl. Acoust.* **62**, 109–124.

Morimoto, M., Sugiura, S., and Iida, K. (**1994**). "Relation between auditory source width in various sound fields and degree of interaural cross-correlation: Confirmation by constant method," *Appl. Acoust.* **42**, 233–238.

Morimoto, M., Ueda, K., and Kiyama, M. (**1995**). "Effects of frequency characteristics of the degree of interaural cross-correlation and sound pressure level on the auditory

source width," *Acustica* **81**, 20–25.

Müller, S. and Massarani, P. (**2001**). "Transfer-function measurement with sweeps," *J. Audio Eng. Soc.* **49**, 443–471.

Müller, S. and Massarani, P. (**2005**). "Transfer-function measurement with sweeps: Director's cut including previously unreleased material and some corrections," `http://www.anselmgoertz.de/Page10383/Monkey_Forest_dt/Manual_dt/aes-swp-english.PDF`.

Nélisse, H. and Nicolas, J. (**1997**). "Characterization of a diffuse field in a reverberant room," *J. Acoust. Soc. Am.* **101**, 3517–3524.

Nelson, D. A. and Schroder, A. C. (**2004**). "Peripheral compression as a function of stimulus level and frequency region in normal-hearing listeners," *J. Acoust. Soc. Am.* **115**, 2221–2233.

Nelson, D. A., Schroder, A. C., and Wojtczak, M. (**2001**). "A new procedure for measuring peripheral compression in normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.* **110**, 2045–2064.

Nielsen, S. H. (**1993**). "Auditory distance perception in different rooms," *J. Audio Eng. Soc.* **41**, 755–770.

Novo, P. (**2005**). "Auditory virtual environments," in J. Blauert, ed., *Communication Acoustics*, Springer-Verlag, Berlin Heidelberg New York, pp. 277–297.

Nuetzel, J. M. and Hafter, E. R. (**1976**). "Lateralization of complex waveforms: Effects of fine structure, amplitude, and duration," *J. Acoust. Soc. Am.* **60**, 1339–1346.

Nuttall, A. L. and Dolan, D. F. (**1993**). "Two-tone suppression of inner hair cell and basilar membrane responses in the guinea pig," *J. Acoust. Soc. Am.* **93**, 390–400.

Okano, T. (**2002**). "Judgments of noticeable differences in sound fields of concert halls caused by intensity variations in early reflections," *J. Acoust. Soc. Am.* **111**, 217–229.

Okano, T., Beranek, L. L., and Hidaka, T. (**1998**). "Relations among interaural cross-correlation coefficient (IACC_E), lateral fraction (LF_E), and apparent source width (ASW) in concert halls," *J. Acoust. Soc. Am.* **104**, 255–265.

Okubo, H., Otani, M., Ikezawa, R., Komiyama, S., and Nakabayashi, K. (**2000**). "A system for measuring the directional room acoustical parameters," *Appl. Acoust.* **62**, 203–215.

Olive, S. E. and Toole, F. E. (**1989**). "The detection of reflections in typical rooms," *J. Audio Eng. Soc.* **37**, 539–553.

Olson, H. F. (**1973**). "Gradient loudspeakers," *J. Audio Eng. Soc.* **21**, 86–93.

Olson, H. F. (**1991**). Acoustical Engineering, Professional Audio Journals, Inc., Philadelphia, PA, USA. Republication of earlier work published by D. Van Nostrand Company, Inc., 1957.

Ono, K., Pulkki, V., and Karjalainen, M. (**2001**). "Binaural modeling of multiple sound source perception: Methodology and coloration experiments," in *AES 111th Convention*, New York, NY, USA. Preprint 5446.

Ono, K., Pulkki, V., and Karjalainen, M. (**2002**). "Binaural modeling of multiple sound source perception: Coloration of wideband sound," in *AES 112th Convention*,

Munich, Germany. Preprint 5550.

Osman, E. (**1971**). "A correlation model of binaural masking level differences," *J. Acoust. Soc. Am.* **50**, 1494–1511. Erratum: *J. Acoust. Soc. Am.* **51**, 669 (**1972**).

Osman, E. (**1973**). "Correlation model of binaural detection: Interaural amplitude ratio and phase variation for signal," *J. Acoust. Soc. Am.* **54**, 386–389.

Oxenham, A. J. and Plack, C. J. (**1997**). "A behavioral measure of basilar-membrane nonlinearity in listeners with normal and impaired hearing," *J. Acoust. Soc. Am.* **101**, 3666–3675.

Palomäki, K. (**2005**). *Studies on Auditory Processing of Spatial Sound and Speech by Neuromagnetic Measurements and Computational Modeling*, PhD thesis, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing.

Palomäki, K. J., Brown, G. J., and Wang, D. (**2004**). "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Communication* **43**, 361–378.

Patterson, R. D. (**1974**). "Auditory filter shape," *J. Acoust. Soc. Am.* **55**, 802–809.

Patterson, R. D. (**1976**). "Auditory filter shapes derived with noise stimuli," *J. Acoust. Soc. Am.* **59**, 640–654.

Patterson, R. D. and Nimmo-Smith, I. (**1980**). "Off-frequency listening and auditory-filter asymmetry," *J. Acoust. Soc. Am.* **67**, 229–245.

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

Pellegrini, R. (**2001a**). *A Virtual Reference Listening Room as an Application of Auditory Virtual Environments*, PhD thesis, Institut für Kommunikationsakustik, Ruhr-Universität Bochum, Germany.

Pellegrini, R. S. (**2001b**). "Quality assessment of auditory virtual environments," in *Proc. Int. Conf. on Auditory Displays (ICAD)*, Espoo, Finland, pp. 161–168.

Pelorson, X., Vian, J.-P., and Polack, J.-D. (**1992**). "On the variability of room acoustical parameters: Reproducibility and statistical validity," *Appl. Acoust.* **37**, 175–198.

Peltonen, T., Lokki, T., Goutarbès, B., Merimaa, J., and Karjalainen, M. (**2001**). "A system for multi-channel and binaural room response measurements," in *AES 110th Convention*, Amsterdam, The Netherlands. Preprint 5289.

Perrett, S. and Noble, W. (**1997**). "The effect of head rotations on vertical plane sound localization," *J. Acoust. Soc. Am.* **102**, 2325–2332.

Perrott, D. R. and Musicant, A. D. (**1977**). "Minimum auditory movement angle: Binaural localization of moving sound sources," *J. Acoust. Soc. Am.* **62**, 1463–1466.

Perrott, D. R. and Pacheco, S. (**1989**). "Minimum audible angle thresholds for broadband noise as a function of delay between the onset of the lead and lag signals," *J. Acoust. Soc. Am.* **85**, 2669–2672.

Perrott, D. R. and Saberi, K. (**1990**). "Minimum audible angle thresholds for sources varying in both elevation and azimuth," *J. Acoust. Soc. Am.* **87**, 1728–1731.

Perrott, D. R., Strybel, T. Z., and Manligas, C. L. (**1987**). "Conditions under which

the Haas precedence effect may or may not occur," *J. Audit. Res.* **27**, 59–72.

Pickles, J. O. (**1988**). An Introduction to the Pysiology of Hearing, 2nd edn, Academic Press, London, England.

Pierce, A. D. (**1989**). Acoustics: An Introduction to Its Physical Principles and Applications, Acoustical Society of America, Woodbury, NY, USA.

Plack, C. J. and Drga, V. (**2003**). "Psychophysical evidence for auditory compression at low characteristic frequencies," *J. Acoust. Soc. Am.* **113**, 1574–1586.

Plack, C. J. and Moore, B. C. J. (**1990**). "Temporal window shape as a function of frequency and level," *J. Acoust. Soc. Am.* **87**, 2178–2187.

Plack, C. J. and Oxenham, A. J. (**1998**). "Basilar-membrane nonlinearity and the growth of forward masking," *J. Acoust. Soc. Am.* **103**, 1598–1608.

Plack, C. J. and Oxenham, A. J. (**2000**). "Basilar-membrane nonlinearity estimated by pulsation threshold," *J. Acoust. Soc. Am.* **107**, 501–507.

Plomp, P. and Levelt, W. J. M. (**1965**). "Tonal consonance and critical bandwidth," *J. Acoust. Soc. Am.* **38**, 548–560.

Plomp, P. and Steeneken, H. J. M. (**1968**). "Interference between two simple tones," *J. Acoust. Soc. Am.* **43**, 883–884.

Polack, J. D. (**1993**). "Playing billiards in the concert hall: The mathematical foundations of geometrical room acoustics," *Appl. Acoust.* **38**, 235–244.

Poletti, M. A. (**2000**). "A unified theory of horizontal holographic sound systems," *J. Audio Eng. Soc.* **48**, 1155–1182.

Poletti, M. A. (**2005**). "Three-dimensional surround sound systems based on spherical harmonics," *J. Audio Eng. Soc.* **53**, 1004–1025.

Pollack, I. (**1978**). "Temporal switching between binaural information sources," *J. Acoust. Soc. Am.* **63**, 550–558.

Pollack, I. and Trittipoe, W. (**1959a**). "Binaural listening and interaural noise cross correlation," *J. Acoust. Soc. Am.* **31**, 1250–1252.

Pollack, I. and Trittipoe, W. (**1959b**). "Interaural noise correlation: Examination of variables," *J. Acoust. Soc. Am.* **31**, 1616–1618.

Potter, J. M., Bilsen, F. A., and Raatgever, J. (**1995**). "Frequency dependence of spaciousness," *Acta Acustica* **3**, 417–427.

Pulkki, V. (**1997**). "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc.* **45**, 456–466.

Pulkki, V. (**1999**). "Uniform spreading of amplitude panned virtual sources," in *Proc. IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust.*, New Paltz, NY, USA.

Pulkki, V. (**2001a**). "Coloration of amplitude-panned virtual sources," in *AES 110th Convention*, Amsterdam, The Netherlands. Preprint 5402.

Pulkki, V. (**2001b**). "Localization of amplitude-panned virtual sources II: Two- and three-dimensional panning," *J. Audio Eng. Soc.* **49**, 753–767.

Pulkki, V. (**2002a**). "Compensating displacement of amplitude-panned virtual sources," in *Proc. AES 22nd Int. Conf.*, Espoo, Finland, pp. 186–195.

Pulkki, V. (**2002b**). "Microphone techniques and directional quality of sound reproduction," in *AES 112th Convention*, Munich, Germany. Preprint 5500.

Pulkki, V. and Hirvonen, T. (**2005**). "Localization of virtual sources in multichannel audio reproduction," *IEEE Trans. Speech Audio Proc.* **13**, 105–118.

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

Pulkki, V. and Lokki, T. (**2003**). "Visualization of edge diffraction," *Acoust. Res. Letters Online* **4**, 118–123.

Pulkki, V. and Merimaa, J. (**2005**). "Spatial impulse response rendering: Listening tests and applications to continuous sound," in *AES 118th Convention*, Barcelona, Spain. Preprint 6371.

Pulkki, V. and Merimaa, J. (**2006**). "Spatial Impulse Response Rendering II: Reproduction of diffuse sound and listening tests," *J. Audio Eng. Soc.* **54**, 3–20.

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

Pulkki, V., Karjalainen, M., and Välimäki, V. (**1999b**). "Localization, coloration, and enhancement of amplitude-panned virtual sources," in *Proc. AES 16th Int. Conf.*, Rovaniemi, Finland, pp. 257–278.

Pulkki, V., Merimaa, J., and Lokki, T. (**2004a**). "Multi-channel reproduction of measured room responses," in *Proc. 18th Int. Congr. Acoust.*, Vol. II, Kyoto, Japan, pp. 1273–1276.

Pulkki, V., Merimaa, J., and Lokki, T. (**2004b**). "Reproduction of reverberation with spatial impulse response rendering," in *AES 116th Convention*, Berlin, Germany. Preprint 6057.

Pumphrey, H. C. (**1993**). "Design of sparse arrays in one, two, and three dimensions," *J. Acoust. Soc. Am.* **93**, 1620–1628.

Puria, S., Peake, W. T., and Rosowski, J. J. (**1997**). "Sound-pressure measurements in the cochlear vestibule of human-cadaver ears," *J. Acoust. Soc. Am.* **101**, 2754–2770.

Rafaely, B. (**2005**). "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech Audio Proc.* **13**, 135–143.

Rakerd, B. and Hartmann, W. M. (**1985**). "Localization of sound in rooms, II: The effects of a single reflecting surface," *J. Acoust. Soc. Am.* **78**, 524–533.

Rakerd, B. and Hartmann, W. M. (**1986**). "Localization of sound in rooms, III: Onset and duration effects," *J. Acoust. Soc. Am.* **80**, 1695–1706.

Rakerd, B., Hartmann, W. M., and Hsu, J. (**2000**). "Echo suppression in the horizontal and median sagittal planes," *J. Acoust. Soc. Am.* **107**, 1061–1064.

Rateitschek, K. (**1998**). *Ein binauraler Signalverarbeitungsansatz zur robusten maschinellen Spracherkennung in lärmerfüllter Umgebung*, PhD thesis, Institut für Kommunikationsakustik, Ruhr-Universität Bochum, Germany. In German.

Recio, A. and Rhode, W. S. (**2000**). "Basilar membrane responses to broadband stimuli," *J. Acoust. Soc. Am.* **108**, 2281–2298.

Recio, A., Rich, N. C., Narayan, S. S., and Ruggero, M. A. (**1998**). "Basilar-membrane responses to clicks at the base of the chinchilla cochlea," *J. Acoust. Soc. Am.* **103**, 1972–1989.

Reed, M. C. and Blum, J. J. (**1990**). "A model for the computation and encoding of

azimuthal information by the lateral superior olive," *J. Acoust. Soc. Am.* **88**, 1442–1453.

Rhode, W. S. and Recio, A. (**2000**). "Study of mechanical motions in the basal region of the chinchilla cochlea," *J. Acoust. Soc. Am.* **107**, 3317–3332.

Rhode, W. S. and Recio, A. (**2001a**). "Basilar-membrane response to multicomponent stimuli in chinchilla," *J. Acoust. Soc. Am.* **110**, 981–994.

Rhode, W. S. and Recio, A. (**2001b**). "Multicomponent stimulus interactions observed in basilar-membrane vibration in the basal region of the chinchilla cochlea," *J. Acoust. Soc. Am.* **110**, 3140–3154.

Riederer, K. A. J. (**2005**). *HRTF Analysis: Objective and Subjective Evaluation of Measured Head-Related Transfer Functions*, PhD thesis, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing.

Risset, J.-C. and Wessel, D. L. (**1999**). "Exploration of timbre by analysis and synthesis," in D. Deutsch, ed., *Psychology of Music*, 2nd edn, Academic Press, London, England, pp. 113–170.

Robert, A. and Eriksson, J. L. (**1999**). "A composite model of the auditory periphery for simulating responses to complex sounds," *J. Acoust. Soc. Am.* **106**, 1852–1864.

Roffler, S. K. and Butler, R. A. (**1967**). "Factors that influence the localization of sound in the vertical plane," *J. Acoust. Soc. Am.* **43**, 1255–1259.

Rosen, S., Baker, R. J., and Darling, A. (**1998**). "Auditory filter nonlinearity at 2 kHz in normal hearing listeners," *J. Acoust. Soc. Am.* **103**, 2539–2550.

Ross, S. (**1996**). "A functional model of the hair cell–primary fiber complex," *J. Acoust. Soc. Am.* **99**, 2221–2238.

Ruggero, M. A., Rich, N. C., and Robles, L. (**1997**). "Basilar-membrane responses to tones at the base of the chinchilla cochlea," *J. Acoust. Soc. Am.* **101**, 2151–2163.

Rumsey, F. (**2002**). "Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm," *J. Audio Eng. Soc.* **50**, 651–666.

Saberi, K. and Hafter, E. R. (**1997**). "Experiments on auditory motion discrimination," in R. H. Gilkey and T. R. Anderson, eds, *Binaural and Spatial Hearing in Real and Virtual Environments*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, pp. 315–327.

Saberi, K. and Perrott, D. R. (**1990**). "Minimum audible movement angles as a function of sound source trajectory," *J. Acoust. Soc. Am.* **88**, 2639–2644.

Saberi, K., Tirtabudi, P., Petrosyan, A., Perrot, D. R., and Strybel, T. Z. (**2003**). "Detection of dynamic changes in interaural delay," *ACUSTICA – Acta Acustica* **89**, 333–338.

Sabine, W. C. (**1900**). "Reverberation," *The American Architect and The Engineering Record.* Reprinted in W. C. Sabine, *Collected Papers on Acoustics*, Peninsula Publishing, Los Altos, CA, USA (**1993**).

Savioja, L., Huopaniemi, J., Lokki, T., and Väänänen, R. (**1999**). "Creating interactive virtual acoustic environments," *J. Audio Eng. Soc.* **47**, 675–705.

Sayers, B. M. and Cherry, E. C. (**1957**). "Mechanism of binaural fusion in the hearing of speech," *J. Acoust. Soc. Am.* **29**, 973–987.

Schiffrer, G. and Stanzial, D. (**1994**). "Energetic properties of acoustic fields," *J. Acoust. Soc. Am.* **96**, 3645–3653.

Schmidt, R. O. (**1986**). "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.* **AP-34**, 276–280.

Schroeder, M. R. (**1954a**). "Die statistischen Parameter der Frequenzkurven von grossen Räumen," *Acustica* **4**, 594–600. "Statistical Parameters of the Frequency Response Curves of Large Rooms (transl.)," *J. Audio Eng. Soc.* **35**, 299–306 (**1987**).

Schroeder, M. R. (**1954b**). "Eigenfrequenzstatistik und Anregungsstatistik in Räumen," *Acustica* **4**, 456–468. "Normal Frequency and Excitation Statistics in Rooms (transl.)," *J. Audio Eng. Soc.* **35**, 307–316 (**1987**).

Schroeder, M. R. (**1962**). "Frequency-correlation functions of frequency responses in rooms," *J. Acoust. Soc. Am.* **12**, 1819–1823.

Schroeder, M. R. (**1965**). "New method of measuring reverberation time," *J. Acoust. Soc. Am.* **37**, 409–412.

Schroeder, M. R. (**1966**). "Complementarity of sound buildup and decay," *J. Acoust. Soc. Am.* **40**, 549–551.

Schroeder, M. R. (**1979**). "Integrated-impulse method measuring sound decay without using impulses," *J. Acoust. Soc. Am.* **66**, 497–500.

Schroeder, M. R. (**1996**). "The 'Schroeder frequency' revisited," *J. Acoust. Soc. Am.* **99**, 3240–3241.

Searle, C. L., Braida, L. D., Cuddy, D. R., and Davis, M. F. (**1975**). "Binaural pinna disparity: Another auditory localization cue," *J. Acoust. Soc. Am.* **57**, 448–455.

Seraphim, H.-P. (**1961**). "Über die Wahrnehmbarkeit mehrerer Rückwürfe von Sprachschall," *Acustica* **11**, 80–91. In German.

Sessler, G. M. and West, J. E. (**1992**). "Reply to "Comments on 'Unidirectional, second-order gradient microphone' "," *J. Acoust. Soc. Am.* **92**, 584.

Sessler, G. M., West, J. E., and Kubli, R. A. (**1989**). "Unidirectional, second-order gradient microphone," *J. Acoust. Soc. Am.* **86**, 2063–2066.

Sever, Jr., J. C. and Small, Jr., A. M. (**1979**). "Binaural critical masking bands," *J. Acoust. Soc. Am.* **66**, 1343–1350.

Shackleton, T. M., Meddis, R., and Hewitt, M. J. (**1992**). "Across frequency integration in a model of lateralization," *J. Acoust. Soc. Am.* **91**, 2276–2279.

Shailer, M. J., Moore, B. C. J., Glasberg, B. R., Watson, N., and Harris, S. (**1990**). "Auditory filter shapes at 8 and 10 kHz," *J. Acoust. Soc. Am.* **88**, 141–148.

Shamma, S. A. and Morrish, K. A. (**1987**). "Synchrony suppression in complex stimulus responses of a biophysical model of the cochlea," *J. Acoust. Soc. Am.* **81**, 1486–1498.

Shamma, S. A., Chadwick, R. S., Wilbur, W. J., Morrish, K. A., and Rinzel, J. (**1986**). "A biophysical model of cochlear processing: Intensity dependence of pure tone responses," *J. Acoust. Soc. Am.* **80**, 133–145.

Shamma, S. A., Shen, N., and Gopalaswamy, P. (**1989**). "Sterausis: Binaural processing without neural delays," *J. Acoust. Soc. Am.* **86**, 989–1006.

Shaw, E. A. G. (**1974**). "Transformation of sound pressure level from the free field to

the eardrum in the horizontal plane," *J. Acoust. Soc. Am.* **56**, 1848–1861.

Shaw, E. A. G. (**1997**). "Acoustical features of the human external ear," in R. H. Gilkey and T. R. Anderson, eds, *Binaural and Spatial Hearing in Real and Virtual Environments*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, pp. 25–47.

Shinn-Cunningham, B. (**2000**). "Learning reverberation: Considerations for spatial auditory displays," in *Proc. Int. Conf. on Auditory Displays (ICAD)*, Atlanta, Georgia, USA, pp. 126–134.

Shinn-Cunningham, B. G., Santarelli, S., and Kopco, N. (**2000**). "Tori of confusion: Binaural localization cues for sources within reach of a listener," *J. Acoust. Soc. Am.* **107**, 1627–1636.

Shinn-Cunningham, B. G., Zurek, B. G., and Durlach, N. I. (**1993**). "Adjustment and discrimination measurements of the precedence effect," *J. Acoust. Soc. Am.* **93**, 2923–2932.

Shinn-Cunningham, B. G., Zurek, P. M., Durlach, N. I., and Clifton, R. K. (**1995**). "Cross-frequency interactions in the precedence effect," *J. Acoust. Soc. Am.* **98**, 164–171.

Singh, P. K., Ando, Y., and Kurihara, Y. (**1994**). "Individual subjective diffuseness responses of filtered noise sound fields," *Acustica* **80**, 471–477.

Slaney, M. (**1993**). "An efficient implementation of the Patterson-Holdsworth auditory filter bank," Technical Report 35, Apple Computer. Available at `http://rvl4.ecn.purdue.edu/~malcolm/apple/tr35/`.

Slaney, M. (**1998**). "Auditory toolbox: Version 2," Technical Report 1998-010, Interval Research Corporation. Available at `http://rvl4.ecn.purdue.edu/~malcolm/interval/1998-010/`.

Soderquist, D. R. and Lindsey, J. W. (**1972**). "Physiological noise as a masker of low frequencies: The cardiac cycle," *J. Acoust. Soc. Am.* **52**, 1261–1219.

Sommers, M. S. and Gehr, S. E. (**1998**). "Auditory suppression and frequency selectivity in older and younger adults," *J. Acoust. Soc. Am.* **103**, 1067–1073.

Soulodre, G. A., Lavoie, M. C., and Norcross, S. G. (**2003**). "Objective measures of listener envelopment in multichannel surround systems," *J. Audio Eng. Soc.* **51**, 826–840.

Stan, G.-B., Embrechts, J.-J., and Archambeau, D. (**2002**). "Comparison of different impulse response measurement techniques," *J. Audio Eng. Soc.* **50**, 249–262.

Stanzial, D. and Prodi, N. (**1997**). "Measurements of newly defined intensimetric quantities and their physical interpretation," *J. Acoust. Soc. Am.* **102**, 2033–2039.

Stanzial, D., Bonsi, D., and Prodi, N. (**2000**). "Measurement of new energetic parameters for the objective characterization of an opera house," *J. Sound Vib.* **232**, 193–211.

Stanzial, D., Prodi, N., and Schiffrer, G. (**1996**). "Reactive acoustic intensity for general fields and energy polarization," *J. Acoust. Soc. Am.* **99**, 1868–1876.

Stellmack, M. A. and Lufti, R. A. (**1996**). "Observer weighting of concurrent binaural information," *J. Acoust. Soc. Am.* **99**, 579–587.

Stellmack, M. A., Dye, Jr., R. H., and Guzman, S. J. (**1999**). "Observer weighting of

binaural information in source and echo clicks," *J. Acoust. Soc. Am.* **105**, 377–387.

Stern, Jr., R. M. and Colburn, H. S. (**1978**). "Theory of binaural interaction based on auditory-nerve data. IV. A model for subjective lateral position," *J. Acoust. Soc. Am.* **64**, 127–140.

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

Stern, R. M. and Trahiotis, C. (**1997**). "Models of binaural perception," in R. H. Gilkey and T. R. Anderson, eds, *Binaural and Spatial Hearing in Real and Virtual Environments*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, pp. 499–531.

Stern, R. M., Zeiberg, A. S., and Trahiotis, C. (**1988**). "Lateralization of complex binaural stimuli: A weighted-image model," *J. Acoust. Soc. Am.* **84**, 156–165. Erratum: *J. Acoust. Soc. Am.* **90**, 2202 (**1991**).

Strutt [Lord Rayleigh], J. W. (**1907**). "On our perception of sound direction," *Phil. Mag.* **13**, 214–232.

Strybel, T. Z. and Perrott, D. R. (**1984**). "Discrimination of relative distance in the auditory modality: The success and failure of the loudness discrimination hypothesis," *J. Acoust. Soc. Am.* **76**, 318–320.

Stuller, J. A. and Hubing, N. (**1997**). "New perspectives for maximum likelihood time-delay estimation," *IEEE Trans. Signal Processing* **45**, 513–525.

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

Sumner, C. J., Lopez-Poveda, E. A., O'Mard, L. P., and Meddis, R. (**2003a**). "Adaptation in a revised inner-hair cell model," *J. Acoust. Soc. Am.* **113**, 893–901.

Sumner, C. J., O'Mard, L. P., Lopez-Poveda, E. A., and Meddis, R. (**2003b**). "A nonlinear filter-bank model of the guinea-pig cochlear nerve: Rate responses," *J. Acoust. Soc. Am.* **113**, 3264–3274.

Suzuki, Y., Asano, F., Kim, H.-Y., and Sone, T. (**1995**). "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," *J. Acoust. Soc. Am.* **97**, 1119–1123.

Svensson, U. P., Fred, R. I., and Vanderkooy, J. (**1999**). "An analytic secondary source model of edge diffraction impulse responses," *J. Acoust. Soc. Am.* **106**, 2331–2344.

Takahashi, D. (**1997**). "Seat dip effect: The phenomena and the mechanism," *J. Acoust. Soc. Am.* **102**, 1326–1334.

Theile, G. (**2000**). "Multichannel natural recording based on psychoacoustic principles," in *AES 108th Convention*, Paris, France. Preprint 5156.

Theile, G. and Plenge, G. (**1977**). "Localization of lateral phantom sources," *J. Audio Eng. Soc.* **25**, 196–200.

Tobias, J. V. and Zerlin, S. (**1959**). "Lateralization threshold as a function of stimulus duration," *J. Acoust. Soc. Am.* **31**, 1591–1594.

Tohyama, M. and Suzuki, A. (**1989**). "Interaural cross-correlation coefficients in stereo-reproduced sound fields," *J. Acoust. Soc. Am.* **85**, 780–786.

Tollin, D. J. and Henning, G. B. (**1998**). "Some aspects of the lateralization of echoed sound in man. I. The classical interaural-delay based precedence effect," *J. Acoust. Soc. Am.* **104**, 3030–3038.

Tollin, D. J. and Henning, G. B. (**1999**). "Some aspects of the lateralization of echoed sound in man. II. The role of the stimulus spectrum," *J. Acoust. Soc. Am.* **105**, 838–849.

Torres, R. R., Svensson, U. P., and Kleiner, M. (**2000**). "Computation of edge diffraction for more accurate room acoustics auralization," *J. Acoust. Soc. Am.* **109**, 600–610.

Trahiotis, C. and Stern, R. M. (**1989**). "Lateralization of bands of noise: Effects of bandwidth and differences of interaural time and phase," *J. Acoust. Soc. Am.* **86**, 1285–1293.

Trahiotis, C., Bernstein, L. R., and Akeroyd, M. A. (**2001**). "Manipulating the "straightness" and "curvature" of patterns of interaural cross correlation affects listeners' sensitivity to changes in interaural delay," *J. Acoust. Soc. Am.* **109**, 321–330.

Ueda, K. and Morimoto, M. (**1995**). "Estimation of auditory source width (ASW): I. ASW for two adjacent 1/3 octave band noises with equal band level," *J. Acoust. Soc. Jpn.* **16**, 77–83.

Usagawa, T., Bodden, M., and Rateitschek, K. (**1996**). "A binaural model as a front-end for isolated word recognition," in *Proc. 4th Int. Conf. Spoken Language*, Vol. 4, pp. 2352–2355.

Väänänen, R. (**2003**). *Parametrization, Auralization, and Authoring of Room Acoustics for Virtual Reality Applications*, PhD thesis, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing.

van de Par, S. and Kohlrausch, A. (**1995**). "Analytical expressions for the envelope correlation of certain narrow-band stimuli," *J. Acoust. Soc. Am.* **98**, 3157–3169.

van de Par, S. and Kohlrausch, A. (**1997**). "A new approach to comparing binaural masking level differences at low and high frequencies," *J. Acoust. Soc. Am.* **101**, 1671–1680.

van de Par, S., Trahiotis, C., and Bernstein, L. R. (**2001**). "A consideration of the normalization that is typically included in correlation-based models of binaural detection," *J. Acoust. Soc. Am.* **109**, 830–833.

van der Hejden, M. and Trahiotis, C. (**1998**). "Binaural detection as a function of interaural correlation and bandwidth of masking noise: Implications for estimates of spectral resolution," *J. Acoust. Soc. Am.* **103**, 1609–1614.

Vassilantonopoulos, S., Hatziantoniou, P., Worley, J., Mourjopoulos, J., and Merimaa, J. (**2005**). "The acoustics of Roman Odeion of Patras: Comparing simulations and acoustic measurements," in *Proc. Forum Acusticum*, Budapest, Hungary, pp. 2197–2202.

Viemeister, N. F. and Wakefield, G. H. (**1991**). "Temporal integration and multiple looks," *J. Acoust. Soc. Am.* **90**, 858–865.

Wakuda, A., Furuya, H., Fujimoto, K., Isogai, K., and Anai, K. (**2003**). "Effects of arrival direction of late sound on listener envelopment," *Acoust. Sci. & Tech* **24**, 179–185.

Wallach, H. (**1939**). "On sound localization," *J. Acoust. Soc. Am.* **10**, 270–274.

Wallach, H. (**1940**). "The role of head movements and vestibular and visual cues in sound localization," *J. Exp. Psychol* **27**, 339–368.

Wallach, H., Newman, E. B., and Rosenzweig, M. R. (**1949**). "The precedence effect in sound localization," *The American Journal of Psychology* **LXII**, 315–336. Reprinted in *J. Audio Eng. Soc.* **21**, 817–826 (**1973**).

Wang, H. and Kaveh, M. (**1985**). "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. Acoust., Speech, Signal Processing* **ASSP-33**, 823–831.

Waves Inc. (**2005**). "Preserving acoustics for posterity," `http://www.acoustics.net/`.

Wax, M. and Kailath, T. (**1983**). "Optimum localization of multiple sources by passive arrays," *IEEE Trans. Acoust., Speech, Signal Processing* **ASSP-31**, 1210–1218.

Weber, D. L. (**1977**). "Growth of masking and the auditory filter," *J. Acoust. Soc. Am.* **62**, 424–429.

Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. (**1993**). "Localization using nonindividualized head-related transfer functions," *J. Acoust. Soc. Am.* **94**, 111–123.

Wightman, F. L. and Kistler, D. J. (**1989a**). "Headphone simulation of free-field listening. I: Stimulus synthesis," *J. Acoust. Soc. Am.* **85**, 858–867.

Wightman, F. L. and Kistler, D. J. (**1989b**). "Headphone simulation of free-field listening. II: Psychophysical validation," *J. Acoust. Soc. Am.* **85**, 868–878.

Wightman, F. L. and Kistler, D. J. (**1992**). "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoust. Soc. Am.* **91**, 1648–1661.

Wightman, F. L. and Kistler, D. J. (**1997a**). "Factors affecting the relative salience of sound localization cues," in R. H. Gilkey and T. R. Anderson, eds, *Binaural and Spatial Hearing in Real and Virtual Environments*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, pp. 1–23.

Wightman, F. L. and Kistler, D. J. (**1997b**). "Monaural sound localization revisited," *J. Acoust. Soc. Am.* **101**, 1050–1063.

Wightman, F. L. and Kistler, D. J. (**1999**). "Resolution of front–back ambiguity in spatial hearing by listener and source movement," *J. Acoust. Soc. Am.* **105**, 2841–2853.

Williams, E. G. (**1999**). Fourier Acoustics — Sound Radiation and Nearfield Acoustical Holography, Academic Press, London, England.

Williams, M. (**2002**). "Multichannel microphone array design: Segment coverage analysis above and below the horizontal reference plane," in *AES 112th Convention*, Munich, Germany. Preprint 5567.

Williams, M. (**2003**). "Multichannel sound recording practice using microphone arrays," in *Proc. AES 24th Int. Conf.*, Banff, Canada, pp. 1–16.

Williams, M. and Le Dû, G. (**1999**). "Microphone array analysis for multichannel sound recording," in *AES 107th Convention*, New York, NY, USA. Preprint 4997.

Williams, M. and Le Dû, G. (**2000**). "Multichannel microphone array design," in *AES 108th Convention*, Paris, France. Preprint 5157.

Woszczyk, W. R. (**1984**). "A microphone technique applying the principle of second-order-gradient unidirectionality," *J. Audio Eng. Soc.* **32**, 507–530.

Yama, M. F. and Small, Jr., A. M. (**1983**). "Tonal masking and frequency selectivity for the monaural and binaural hearing systems," *J. Acoust. Soc. Am.* **73**, 285–290.

Yang, X. and Grantham, D. W. (**1997a**). "Cross-spectral and temporal factors in the precedence effect: Discrimination suppression of the lag sound in free-field," *J. Acoust. Soc. Am.* **102**, 2973–2983.

Yang, X. and Grantham, D. W. (**1997b**). "Echo suppression and discrimination suppression aspects of the precedence effect," *Percept. Psychophys.* **59**, 1108–1117.

Yasin, I. and Plack, C. J. (**2003**). "The effects of a high-frequency suppressor on tuning curves and derived basilar membrane response functions," *J. Acoust. Soc. Am.* **114**, 322–332.

Yasin, I. and Plack, C. J. (**2005**). "Psychophysical tuning curves at very high frequencies," *J. Acoust. Soc. Am.* **118**, 2498–2506.

Yin, T. C. T. and Chan, J. C. K. (**1990**). "Interaural time sensitivity in medial superior olive of cat," *J. Neurophys.* **64**, 465–488.

Yin, T. C. T., Joris, P. X., Smith, P. H., and Chan, J. C. K. (**1997**). "Neuronal processing for coding interaural time disparities," in R. H. Gilkey and T. R. Anderson, eds, *Binaural and Spatial Hearing in Real and Virtual Environments*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, pp. 427–445.

Yoshimasa Electronic (**2003**). "Sound preference audition room," `http://www.ymec.com/hp/pref2/`.

Yost, W. A. (**1985**). "Prior stimulation and the masking-level difference," *J. Acoust. Soc. Am.* **78**, 901–907.

Yost, W. A. and Dye, R. (**1988**). "Discrimination of interaural differences of level as a function of frequency," *J. Acoust. Soc. Am.* **83**, 1846–1851.

Zacharov, N. and Koivuniemi, K. (**2001a**). "Audio descriptive analysis & mapping of spatial sound displays," in *Proc. Int. Conf. on Auditory Displays (ICAD)*, Espoo, Finland, pp. 95–104.

Zacharov, N. and Koivuniemi, K. (**2001b**). "Unravelling the perception of spatial sound reproduction: Analysis & external preference mapping," in *AES 11th Convention*, New York, NY, USA. Preprint 5423.

Zacharov, N. and Koivuniemi, K. (**2001c**). "Unravelling the perception of spatial sound reproduction: Techniques and experimental design," in *Proc. AES 19th Int. Conf.*, Schloss Elmau, Germany.

Zahorik, P. (**2002a**). "Assessing auditory distance perception using virtual acoustics," *J. Acoust. Soc. Am.* **111**, 1832–1846.

Zahorik, P. (**2002b**). "Direct-to-reverberant energy ratio sensitivity," *J. Acoust. Soc. Am.* **112**, 2110–2117.

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

Zhang, M. and Er, M. H. (**1996**). "An alternative algorithm for estimating and tracking talker location by microphone arrays," *J. Audio Eng. Soc.* **44**, 729–736.

Zhang, X., Heinz, M. G., Bruce, I. C., and Carney, L. H. (**2001**). "A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression," *J. Acoust. Soc. Am.* **109**, 648–670.

Zhou, B. (**1995**). "Auditory filter shapes at high frequencies," *J. Acoust. Soc. Am.* **98**, 1935–1942.

Zurek, P. M. (**1979**). "Measurement of binaural echo suppression," *J. Acoust. Soc. Am.* **66**, 1750–1757.

Zurek, P. M. (**1987**). "The precedence effect," in W. A. Yost and G. Gourevitch, eds, *Directional Hearing*, Springer-Verlag, New York, NY, USA, pp. 85–105.

Zwicker, E. (**1961**). "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)," *J. Acoust. Soc. Am.* **33**, 248.

Zwicker, E. and Fastl, H. (**1990**). Psychoacoustics: Facts and Models, Springer-Verlag, Berlin Heidelberg.

Zwicker, E. and Terhardt, E. (**1980**). "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Am.* **68**, 1523–1525.

Zwicker, E., Flottorp, G., and Stevens, S. S. (**1957**). "Critical band width in loudness summation," *J. Acoust. Soc. Am.* **29**, 548–557.

Zwiers, M. P., van Opstal, A. J., and Paige, G. D. (**2003**). "Plasticity in human sound localization induced by compressed spatial vision," *Nature Neurosci.* **6**, 175–181.

Zwislocki, J. (**1962**). "Analysis of the middle-ear function. Part I: Input impedance," *J. Acoust. Soc. Am.* **34**, 1514–1523.

HELSINKI UNIVERSITY OF TECHNOLOGY

LABORATORY OF ACOUSTICS AND AUDIO SIGNAL PROCESSING

35 T. I. Laakso, V. Välimäki, M. Karjalainen, U. K. Laine: Crushing the Delay—Tools for Fractional Delay Filter Design. 1994

36 J. Backman, J. Huopaniemi, M. Rahkila (toim.): Tilakuuleminen ja auralisaatio. Akustiikan seminaari 1995.

37 V. Välimäki: Discrete-Time Modeling of Acoustic Tubes Using Fractional Delay Filters. 1995

38 T. Lahti: Akustinen mittaustekniikka. 2. korjattu painos. 1997

39 M. Karjalainen, V. Välimäki (toim.): Akustisten järjestelmien diskreettiaikaiset mallit ja soittimien mallipohjainen äänisynteesi. Äänenkäsittelyn seminaari 1995

40 M. Karjalainen (toim.): Aktiivinen äänenhallinta. Akustiikan seminaari 1996

41 M. Karjalainen (toim.): Digitaaliaudion signaalinkäsittelymenetelmiä. Äänenkäsittelyn seminaari 1996

42 M. Huotilainen, J. Sinkkonen, H. Tiitinen, R. J. Ilmoniemi, E. Pekkonen, L. Parkkonen, R. Näätänen: Intensity Representation in the Human Auditory Cortex. 1997

43 M. Huotilainen: Magnetoencephalography in the Study of Cortical Auditory Processing. 1997

44 M. Karjalainen, J. Backman, L. Savioja (toim.): Akustiikan laskennallinen mallintaminen. Akustiikan seminaari 1997

45 V. Välimäki, M. Karjalainen (toim.): Aktiivisen melunvaimennuksen signaalinkäsittelyalgoritmit. Äänenkäsittelyn seminaari 1997

46 T. Tolonen: Model-Based Analysis and Resynthesis of Acoustic Guitar Tones. 1998

47 H. Järveläinen, M. Karjalainen, P. Maijala, K. Saarinen, J. Tanttari: Työkoneiden ohjaamomelun häiritsevyys ja sen vähentäminen. 1998

48 T. Tolonen, V. Välimäki, M. Karjalainen: Evaluation of Modern Sound Synthesis Methods. 1998

49 M. Karjalainen, V. Välimäki (toim.): Äänenlaatu. Akustiikan seminaari 1998

50 V. Välimäki , M. Karjalainen (toim.): Signaalinkäsittely audiotekniikassa, akustiikassa musiikissa. Äänenkäsittelyn seminaari 1998

51 M. Karjalainen: Kommunikaatioakustiikka. 1998

52 M. Karjalainen (toim.): Kuulon mallit ja niiden sovellutukset. Akustiikan seminaari 1999

53 J. Huopaniemi: Virtual Acoustics And 3-D Sound In Multimedia Signal Processing. 1999

HELSINKI UNIVERSITY OF TECHNOLOGY
LABORATORY OF ACOUSTICS AND AUDIO SIGNAL PROCESSING

ISBN 951-22-8290-9
ISSN 1456-6303