Department of Signal Processing and Acoustics

Techniques for versatile spatial-audio reproduction in time-frequency domain

Mikko-Ville Laitinen

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 5/2014

Techniques for versatile spatial-audio reproduction in time-frequency domain

Mikko-Ville Laitinen

A doctoral dissertation completed for the degree of Doctor of Science in Technology to be defended, with the permission of the Aalto University School of Electrical Engineering, at a public examination held in lecture hall S1 of the school on February 28, 2014, at 12 noon.

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

Supervising professor
Professor Ville Pulkki

Thesis advisor
Professor Ville Pulkki

Preliminary examiners
Doctor Jeroen Breebaart, Dolby Laboratories, Australia
Doctor Juha Merimaa, Apple, USA

Opponent
Professor Athanasios Mouchtaris, Institute of Computer Science of the Foundation for Research and Technology - Hellas (FORTH-ICS), Greece and Department of Computer Science, University of Crete, Greece

Aalto University publication series
DOCTORAL DISSERTATIONS 5/2014

ISBN 978-952-60-5528-2
ISBN 978-952-60-5529-9 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5529-9

Unigrafia Oy
Helsinki 2014

Finland

441 697
Printed matter

Aalto University

Abstract

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

Author
Mikko-Ville Laitinen

Name of the doctoral dissertation
Techniques for versatile spatial-audio reproduction in time-frequency domain

Publisher School of Electrical Engineering

Unit Department of Signal Processing and Acoustics

Series Aalto University publication series DOCTORAL DISSERTATIONS 5/2014

Field of research Acoustics and audio signal processing

Manuscript submitted 31 July 2013 **Date of the defence** 28 February 2014

Permission to publish granted (date) 14 November 2013 **Language** English

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

Abstract

We can perceive many spatial aspects about the sounds around us. These include the direction, the distance, and the size of the sound source, as well as properties about the space inside which we are. Thus, reproduction of sound should take these spatial properties into account if natural perception of a sound scene is desired.

Directional audio coding (DirAC) is a recently proposed method for spatial sound reproduction. It operates in the time-frequency domain and aims to analyze the perceptually significant properties of the sound field. The analyzed parameters, namely the direction of arrival and the diffuseness, are used for manipulating recorded microphone signals in such a way that the perception of the reproduced sound field is equal to the original sound field. Subjective evaluations have shown that, compared to traditional methods, DirAC improves the perceived quality. However, DirAC was originally introduced for relatively limited use cases. This thesis presents methods to generalize the DirAC approach for more versatile use. The generalization is performed for three aspects: challenging spatial-sound scenarios, output systems, and input systems.

As DirAC is a parametric method, the resulting quality is signal dependent. Thus, challenging sound scenarios for DirAC processing were sought in order to improve the processing and to enable good quality with all kinds of signals. A few problematic cases were found, e.g., multiple simultaneous talkers in low-echoic conditions and applause-type signals. This thesis shows that the decorrelation processing used in DirAC increases the perceived spaciousness with certain signals. Alternative methods for these problematic cases are introduced showing improvement in the perceived quality based on subjective evaluation.

DirAC originally used loudspeakers for reproduction. As an addition to possible reproduction devices, a method for headphone reproduction is presented in this thesis. The method is based on binaural techniques and head tracking, and subjective evaluations show that natural spatial impression can be reproduced.

DirAC was originally developed to be used with B-format microphones, but in practice they are rarely used for recording. A method for more common spaced-microphone arrays, which is additionally shown to have some advantages compared to the B-format processing, is presented in this thesis. Furthermore, DirAC is extended to be used with legacy multi-channel signals, such as 5.1 surround, and even further to virtual-world spatial audio. Finally, a modular structure for DirAC processing is introduced. The structure allows several types of inputs to be used simultaneously without compromising the quality of reproduction.

Keywords spatial audio, multi-channel reproduction

ISBN (printed) 978-952-60-5528-2 **ISBN (pdf)** 978-952-60-5529-9

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

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

Pages 186 **urn** http://urn.fi/URN:ISBN:978-952-60-5529-9

Aalto-yliopisto

Tiivistelmä

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

Tekijä
Mikko-Ville Laitinen

Väitöskirjan nimi
Tekniikoita monipuoliseen tilaäänen toistamiseen aika-taajuusalueessa

Julkaisija Sähkötekniikan korkeakoulu

Yksikkö Signaalinkäsittelyn ja akustiikan laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 5/2014

Tutkimusala Akustiikka ja äänenkäsittelytekniikka

Käsikirjoituksen pvm 31.07.2013 **Väitöspäivä** 28.02.2014

Julkaisuluvan myöntämispäivä 14.11.2013 **Kieli** Englanti

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

Tiivistelmä

Havaitsemme ympärillämme olevista äänistä useita avaruudellisia ominaisuuksia, kuten äänilähteen suunnan, etäisyyden ja koon, ja lisäksi myös ominaisuuksia tilasta missä olemme. Näin ollen tilaäänen toistossa tulee ottaa nämä tilaa koskevat ominaisuudet huomioon, jos tavoitteena on luonnollinen havainto ääniympäristöstä.

Directional audio coding (DirAC) on äskettäin esitelty menetelmä tilaäänen toistamiseen. Se käsittelee ääntä aika-taajuusalueessa ja pyrkii analysoimaan äänikentästä havaintojen kannalta merkityksellisiä ominaisuuksia. Analysoituja parametreja, eli tulosuuntaa ja diffuusisuutta, käytetään äänitettyjen mikrofonisignaalien muokkaamiseen siten, että toistettu äänikenttä havaitaan samalla tavalla kuin alkuperäinen äänikenttä. Kuuntelukokeet ovat osoittaneet, että DirAC parantaa havaittua laatua verrattuna perinteisiin menetelmiin. DirAC esiteltiin kuitenkin alunperin verrattain suppeisiin käyttötarkoituksiin. Tämä väitöskirja esittää menetelmiä, joilla voidaan yleistää DirAC-tekniikan lähestymistapaa monipuolisempiin käyttötarkoituksiin. Yleistys tehdään kolmesta eri näkökulmasta: haastavat tilanteet tilaäänen kannalta, toistojärjestelmät ja sisäänmenojärjestelmät.

Koska DirAC on parametrinen menetelmä, toiston laatu riippuu signaalista. Tämän vuoksi tässä työssä etsittiin DirAC-toiston kannalta haastavia ääniskenaarioita prosessoinnin kehittämiseksi ja täten hyvän laadun mahdollistamiseksi kaikenlaisilla signaaleilla. Muutamia ongelmallisia tapauksia löydettiin, kuten monta samanaikaista puhujaa vähäkaikuisessa huoneessa ja taputuksia sisältävät signaalit. Tämä väitöskirja osoittaa, että DirAC-tekniikassa käytetty dekorrelointi lisää havaittua kaiuntaisuutta tietyillä signaaleilla. Vaihtoehtoisia menetelmiä esitetään näihin tilanteisiin, ja kuuntelukokeet osoittavat havaitun laadun paranevan.

Alunperin DirAC käytti toistoon kaiuttimia. Lisänä mahdollisiin toistojärjestelmiin tässä väitöskirjassa esitetään menetelmä kuuloketoistoon. Menetelmä perustuu binauraalisiin tekniikoihin ja päänseurantaan ja mahdollistaa luonnollisen tilantunnun toiston.

DirAC kehitettiin alunperin käytettäväksi B-formaattimikrofonien kanssa, mutta käytännössä niitä käytetään harvoin äänittämiseen. Tässä väitöskirjassa esitetään menetelmä yleisempien erillismikrofoniäänitysten käsittelyyn, josta lisäksi osoitetaan olevan etua verrattuna B-formaatin käsittelyyn. Lisäksi DirAC-tekniikkaa laajennetaan monikanavasignaalien, kuten 5.1-äänen, käsittelyyn ja jopa pidemmälle tilaäänentoistoon virtuaalimaailmoissa. Lopuksi esitetään modulaarinen rakenne DirAC-prosessointiin, joka mahdollistaa useanlaisten sisäänmenojen samanaikaisen käytön tinkimättä toiston laadusta.

Avainsanat tilaääni, monikanavainen toisto

ISBN (painettu) 978-952-60-5528-2 **ISBN (pdf)** 978-952-60-5529-9

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

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

Sivumäärä 186 **urn** http://urn.fi/URN:ISBN:978-952-60-5529-9

Preface

This work was carried out at the Department of Signal Processing and Acoustics, Aalto University, Finland, on a joint project with the Fraunhofer Institute for Integrated Circuits (IIS), Germany. During the period from August to October 2012, the work was carried out at Fraunhofer IIS.

I am grateful to my supervisor and instructor Professor Ville Pulkki. His guidance has been excellent, and I have really enjoyed working with him. He has given me endless amounts of ideas, and his professional expertise has been invaluable. I would also like to thank my former supervisor Professor Matti Karjalainen, who passed away in May 2010.

Working at the acoustics lab has been great. In addition to having loads of fun, having really talented and helpful colleagues has been a great aid while doing the research work. I would like to thank all of the current and the former personnel at the acoustics lab, especially, Tapani Pihlajamäki, Archontis Politis, Dr. Jukka Ahonen, Juha Vilkamo, Professor Cumhur Erkut, Stefan Lösler, Marko Takanen, Dr. Marko Hiipakka, Olli Santala, Dr. Ville Sivonen, Dr. Miikka Tikander, Ville Saari, Teemu Koski, Symeon Delikaris-Manias, Dr. Toni Hirvonen, Javier Goméz Bolaños, and Olli Rummukainen.

I would also like to thank my colleagues at Fraunhofer IIS, especially, Dr. Sascha Disch, Dr. Fabian Küch, Dr. Achim Kuntz, Dr. Markus Kallinger, and Oliver Thiergart. Working with you has been great, and I have really enjoyed the workshops and my research visit there.

Furthermore, I wish to thank Professor Tapio Lokki and the researchers at the Department of Media Technology, Aalto University for interesting discussions. I would also like to thank the pre-examiners of the thesis, Dr. Jeroen Breebaart and Dr. Juha Merimaa, for valuable suggestions to the manuscript.

The work was funded by Fraunhofer IIS and the Graduate School in Elec-

tronics, Telecommunications and Automation (GETA). The dissertation work was also supported by Tekniikan edistämissäätiö (TES). I wish to thank all financial supporters.

Finally, I would like to thank my parents, Mikko and Ritva, and my sister Maria for many kinds of support over the years. And above all, I want to express the deepest thanks to my lovely girlfriend Sonja for all the support.

Helsinki, December 30, 2013,

Mikko-Ville Laitinen

Contents

List of Publications

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

I Mikko-Ville Laitinen, Fabian Kuech, Sascha Disch, and Ville Pulkki. Reproducing applause-type signals with directional audio coding. *Journal of the Audio Engineering Society*, vol. 59, no. 1/2, pp. 29–43, January/February 2011.

II Mikko-Ville Laitinen and Ville Pulkki. Utilizing instantaneous direct-to-reverberant ratio in parametric spatial audio coding. In *Proceedings of the 133rd Convention of the Audio Engineering Society*, San Francisco, CA, USA, October 2012.

III Archontis Politis, Mikko-Ville Laitinen, Jukka Ahonen, and Ville Pulkki. Parametric spatial audio coding for spaced microphone array recordings. In *Proceedings of the 134th Convention of the Audio Engineering Society*, Rome, Italy, May 2013.

IV Mikko-Ville Laitinen and Ville Pulkki. Binaural reproduction for directional audio coding. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, pp. 337–340, October 2009.

V Mikko-Ville Laitinen, Tapani Pihlajamäki, Stefan Lösler, and Ville Pulkki. Influence of resolution of head tracking in synthesis of binaural audio. In *Proceedings of the 132nd Convention of the Audio Engineering*

Society, Budapest, Hungary, May 2012.

VI Mikko-Ville Laitinen and Ville Pulkki. Converting 5.1 audio recordings to B-format for directional audio coding reproduction. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, pp. 61–64, May 2011.

VII Mikko-Ville Laitinen, Tapani Pihlajamäki, Cumhur Erkut, and Ville Pulkki. Parametric time-frequency representation of spatial sound in virtual worlds. *ACM Transactions on Applied Perception*, vol. 9, no. 2, article 8, June 2012.

VIII Tapani Pihlajamäki, Mikko-Ville Laitinen, and Ville Pulkki. Modular architecture for virtual-world parametric spatial audio synthesis. In *Proceedings of the 49th International Conference of the Audio Engineering Society*, London, UK, February 2013.

Author's Contribution

Publication I: "Reproducing applause-type signals with directional audio coding"

The present author investigated causes for the problems with the applause-type signals. Based on the idea of the fourth author of the article, the present author developed and implemented the multi-resolution processing. In addition, the present author developed and implemented the transient-detection algorithm. The subjective evaluations were designed in collaboration with the fourth author and conducted by the present author. The results were analyzed by the present author, who was also mainly responsible for writing the article.

Publication II: "Utilizing instantaneous direct-to-reverberant ratio in parametric spatial audio coding"

The idea of using the estimation of the reverberant energy for the optimization of the decorrelation process was invented in collaboration with the co-author of the article. The present author developed and implemented the algorithms presented in this article. Furthermore, the present author designed and conducted the listening tests as well as analyzed the results and primarily wrote the article.

Publication III: "Parametric spatial audio coding for spaced microphone array recordings"

This study was based on the present author's idea of using spaced microphones with directional audio coding. The present author developed the

synthesis algorithm and the microphone arrangement. Furthermore, the present author designed the listening tests and analyzed the results. The present author also conducted the listening tests in collaboration with the first author of the article. Sections 2.2, 4.1, 4.2, 4.3, and 5 as well as parts of Sections 1, 2.1, and 6 were primarily written by the present author.

Publication IV: "Binaural reproduction for directional audio coding"

The algorithms presented in the article were developed in collaboration with the co-author. The present author implemented the algorithms. The listening tests were designed and the results were analyzed in collaboration with the co-author. The present author conducted the listening tests and primarily wrote the article.

Publication V: "Influence of resolution of head tracking in synthesis of binaural audio"

Planning of this study was done by the present author. The listening test was designed in collaboration with the third author of the article. The present author primarily wrote the article except Section 4.2.

Publication VI: "Converting 5.1 audio recordings to B-format for directional audio coding reproduction"

The present author developed the matrixing based methods suggested in the article and the crossfading between the outputs of them in collaboration with the co-author of the article. The present author developed the method to adjust the diffuseness properties of the resulting B-format signals in order to produce correct perception of coherence. All algorithms presented in the article were implemented by the present author, who was also mainly responsible for writing the article.

Publication VII: "Parametric time-frequency representation of spatial sound in virtual worlds"

This study is based on the idea of the third and the fourth author of the article. The present author developed the algorithms presented in

this article in collaboration with the fourth author. The algorithms were implemented by the present author, and he designed the listening tests in collaboration with the second author. Sections 2.2 and 3 were primarily written by the present author as well as parts of Sections 1, 2.1, 5, and 6.

Publication VIII: "Modular architecture for virtual-world parametric spatial audio synthesis"

The present author invented the method described in the article. The method was further developed in collaboration with the first author of the article. The present author took part in the writing of the article, especially in Sections 3 and 4.

List of Abbreviations

AAC	advanced audio coding
BCC	binaural cue coding
DirAC	directional audio coding
DOA	direction of arrival
ERB	equivalent rectangular bandwidth
FFT	fast Fourier transform
HOA	higher-order Ambisonics
HRTF	head-related transfer function
IC	inter-aural coherence
ICC	inter-channel correlation
ICLD	inter-channel level difference
ICPD	inter-channel phase difference
ICTD	inter-channel time difference
ILD	inter-aural level difference
ITD	inter-aural time difference
OLA	overlap add
OTT	one-to-two
PS	parametric stereo
PTF	headphone transfer function
SASC	spatial audio scene coding
SIRR	spatial impulse response rendering
STFT	short-time Fourier transform
TTO	two-to-one
TTT	three-to-two, two-to-three
VBAP	vector-base amplitude panning
WFS	wave-field synthesis

List of Symbols

c	speed of sound
E	energy density
f	frequency
f_c	center frequency of a band
f_s	sampling frequency
g	gain
$\mathbf{I}$	intensity vector
K	number of the transmitted audio signals
k	frequency band index
L	length of a temporal frame
M	number of the input audio signals
N	number of the loudspeakers
n	temporal frame index
p	pressure
R	order of a harmonic component
r	distance, radius
S	frequency-domain audio signal
t	time
$\mathbf{u}$	particle-velocity vector
$\mathbf{V}$	vector $[X, Y, Z]/\sqrt{2}$
V	volume
W	omnidirectional microphone signal
X, Y, Z	dipole microphone signals
α	smoothing coefficient
θ	azimuth angle
ϕ	elevation angle
ψ	diffuseness
ω	angular frequency

1. Introduction

Spatial sound is a natural part of our everyday life. We perceive many spatial aspects about the sounds around us. These include the direction, the distance, and the size of the sound source, as well as properties about the space inside which we are. Thus, the reproduction of sound should take these spatial properties into account if natural perception of a sound scene is desired. The procedure of spatial-sound reproduction is visualized in Fig. 1.1, where there is a certain kind of microphone arrangement for capturing spatial sound inside a certain space. On the reproduction side, there is a certain kind of loudspeaker layout for reproducing spatial sound. The aim is that a human listener perceives the reproduced sound field as if he/she were present in the position of the microphones where the original recording was performed. The research question is how can we accomplish this.

Two-channel stereophonic reproduction techniques have already been used for decades to reproduce the spatial aspects, at least in some degree. They allow controlling the perceived direction of the auditory event by introducing amplitude and time differences between the reproduced loudspeaker signals. In amplitude panning, a mono signal is applied into two loudspeakers placed in front of the listener, and the individual gains of the loudspeakers are adjusted. Correspondingly, the sound scene can

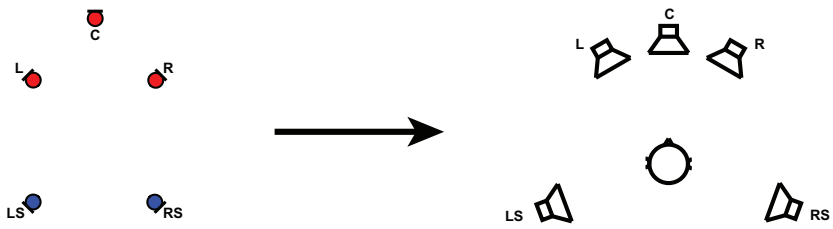

Figure 1.1. Procedure of spatial-sound reproduction. The sound scene is captured with microphones and reproduced with loudspeakers.

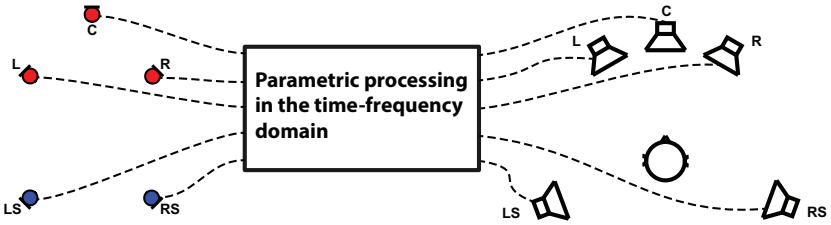

Figure 1.2. Perceptually significant properties of the sound field are analyzed in parametric methods. The microphone signals are manipulated based on them and reproduced using loudspeakers.

be captured by using coincident-pair microphone techniques that directly yield suitable amplitude differences. These techniques are still widely used. However, they have also some limitations: e.g., auditory events can be positioned only in between the loudspeakers, the perceived direction depends on the placement of the listener, and it is difficult to reproduce enveloping reverberation.

Multi-channel reproduction methods have been developed to address these issues. Compared to two-channel stereo, they allow more accurate and robust localization, the feel of space can be reproduced more naturally, and auditory events can be positioned all around the listener, at least in theory. Consequently, the multi-channel systems have gained interest, also in domestic use. However, the use of more loudspeakers introduces challenges for the recording techniques. Physical constraints of real microphones cause problems in localization, timbre, and the feel of space. To avoid these problems, the concept of parametric time-frequency processing has been introduced to spatial-sound processing.

The general idea in parametric methods is to analyze properties of the sound field that are significant to human perception of spatial sound and to use them to manipulate the captured microphone signals before being reproduced with loudspeakers (see Fig. 1.2). Furthermore, the parametric approach can be used for the compression of multi-channel signals by transmitting only a downmix of the signals alongside with information about the perceptually significant relationships between the original loudspeaker signals. Thus, parametric methods are based on understanding how the human auditory system perceives spatial sound.

This thesis deals with the reproduction of spatial sound. Particularly, parametric methods operating in the time-frequency domain are the topic of this thesis.

1.1 Aim of the thesis

When the research work for the thesis was started, directional audio coding (DirAC) [1] had recently been published. DirAC is a perceptually motivated parametric method for reproducing spatial sound, and it is based on a method for processing spatial impulse responses, known as spatial impulse response rendering (SIRR) [2]. At the time, DirAC was mainly implemented using B-format microphone signals as input and loudspeakers as output. Furthermore, the quality of reproduction had mainly been evaluated using sound scenes common in every-day life, but the most challenging scenarios had not been found. Nevertheless, great potential for more versatile use was seen in the method. Consequently, the aim of the research leading to this thesis was to enable excellent quality with many kinds of input and output methods, with real and virtual sound scenes, and also with challenging signals. The aim of the thesis itself may be seen as a generalization of the methods initially suggested in [1]. More specifically, the generalization is performed for three different aspects:

- A good reproduction system should be robust and should provide excellent perceptual quality with all kinds of spatial-sound scenarios. Hence, it is important to identify challenging cases for the reproduction method. Publications I, II, III, and VIII deal with identifying these cases and suggest solutions.

- A versatile reproduction system should enable reproduction with any output method. The original DirAC method used only loudspeakers as an output. Hence, DirAC processing is extended to headphone reproduction in Publications IV and V.

- A versatile reproduction system should also accept different kinds of inputs. The original DirAC method dealt only with capturing real sound scenes with a B-format microphone. Thus, extending DirAC processing to different input sources was seen important. Publication III presents a parametric method that uses spaced-microphone signals as an input. Publication VI extends DirAC processing from the microphone signals to legacy multi-channel signals, such as 5.1 surround. Publications VII and VIII extend the processing even further to virtual-world spatial audio.

1.2 Organization of the thesis

The thesis consists of an introductory part and eight articles published in peer-reviewed journals and conference proceedings. The introductory part begins by reviewing spatial sound as a physical phenomenon. Spatial sound as a perceptual phenomenon is discussed in Section 3. Section 4 gives an overview of traditional methods for spatial-sound reproduction. Section 5 discusses parametric techniques for spatial-sound reproduction and coding. Emphasis is given to DirAC, which was the starting point for the research in this thesis. Section 6 summarizes the main results of the publications presented in this thesis, and Section 7 concludes the thesis.

2. Physical background of spatial sound

Propagation of sound from a source to a listener in a room is typically divided into three parts: direct sound, early reflections, and late reverberation [3]. Fig. 2.1(a) shows an exemplary sound scene where the direct sound and a few reflections are illustrated. The corresponding impulse response is shown in Fig. 2.1(b). Direct sound travels in a straight line from the source to the listener. Thus, it reaches the listener first, ahead of early reflections from the walls, the floor, the ceiling, etc. These reflections travel longer paths than the direct sound. Therefore, they arrive later and have a lower sound pressure level. The density of the reflections as a function of time t (in s^{-1}) can be estimated by

$$N'_{\mathrm{refl}} = \frac{4\pi c^3 t^2}{V} \tag{2.1}$$

where c is the speed of sound and V is the volume of the room [4]. In addition, diffraction (bending of a wavefront around obstacles) can take place if the wavelength is large compared to the size of the obstacles, or the reflection can be diffusing, i.e., the sound scatters to all directions [3].

After a certain time instant, the density of the reflections is so dense in time that they are often modeled statistically and not as individual reflections. This time instant, referred to as the mixing time [5], is typically defined to be around 80 ms [6]. The reflections after the mixing time are called the late reverberation. The late reverberant field is typically considered to be diffuse [7]. An acoustic field is considered to be perfectly diffuse in a volume V if the energy density is the same in all points of this volume V [8]. In practice, a diffuse field can be assumed to consist of a very large number of plane waves arriving from all directions with equal probability and random phase relations. In addition, late reverberation can be assumed to be exponentially decaying in level [7] since the path of the arriving reflection is the longer the later it arrives to the measurement point.

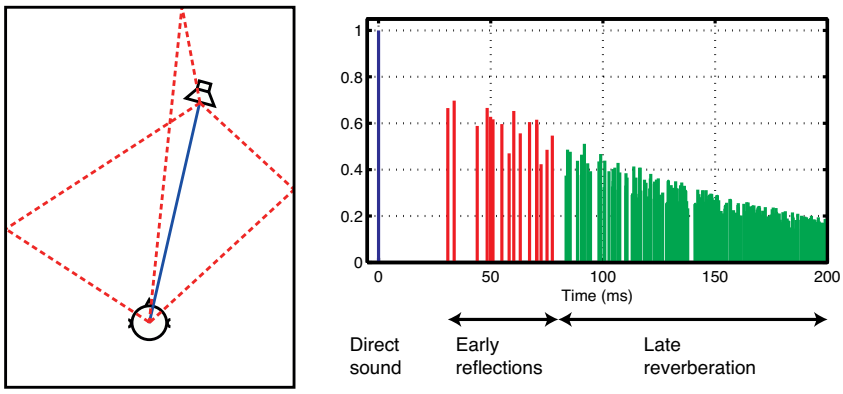

Figure 2.1. (a) Sound rays in a room: The direct path and a few first reflections are illustrated. (b) An exemplary impulse response, where the direct sound, the early reflections, and the late reverberation are separated. The absolute values are shown. The time instant when the direct sound reaches the receiver is set to 0 ms.

Due to the very large number of reflections (see Eq. 2.1 and Fig. 2.1), reverberation is too complex a phenomenon to be considered accurately reflection by reflection. Thus, simplified key figures for describing reverberation are needed. One of the most common descriptors is the reverberation time (T_{60}), which is defined as the period of time taken for the sound pressure in an enclosure to decay by 60 dB after a stationary test stimulus is switched off [9]. The relationship between the reverberation time (in seconds), the room size, and the absorption area A can be estimated using the well-known Sabine equation [9]

$$T_{60} = \frac{0.161 \cdot V}{A}. \tag{2.2}$$

T_{60} can be estimated by measuring the impulse response of the space, applying backwards integration to the response [10], fitting a line to the resulting curve, and computing the slope of the line. Relatively often, the decay does not have a constant slope for the 60 dB range due to measurement noise or non-ideal characteristics of the room. Thus, the reverberation time is typically computed from the late reverberation by finding a suitable portion where the decaying is taking place with a constant slope in the logarithmic scale and by interpolating the rest of the defined 60 dB decay. In addition, the reverberation time can be blindly estimated from any recorded signal, such as speech, by assuming certain properties of the excitation signal [11, 12, 13].

The reverberation time can be used, for example, in designing concert

halls and there are approximate suggestions for suitable values for different kinds of spaces [3]. Another useful objective descriptor is the early decay time (EDT), which is computed similarly to the reverberation time but based on the first 10 dB portion of the decay. It has been shown to be better related to subjective perception of the reverberation [3].

Physical properties of the sound field are discussed next. Sound can be presented using two quantities: pressure $p(t)$ and particle velocity $\mathbf{u}(t)$. Pressure is a scalar quantity, and particle velocity is a vector quantity. Using these two, the instantaneous intensity can be defined as [14]

$$\mathbf{I}(t) = p(t)\mathbf{u}(t), \tag{2.3}$$

and the instantaneous energy density as [15]

$$E(t) = \frac{1}{2}\rho\left[\frac{p^2(t)}{Z^2} + \mathbf{u}^2(t)\right], \tag{2.4}$$

where ρ is the density of the medium and Z is the acoustical impedance of the medium. Furthermore, the behavior of sound pressure can be presented using the general wave equation

$$\frac{\partial^2 p}{\partial t^2} = c^2 \frac{\partial^2 p}{\partial x^2}, \tag{2.5}$$

where x is position [16]. This equation gives sound pressure in one-dimensional case at any given location and time instant. In addition, if a diffuse field is assumed, it is possible to derive a closed-form solution for the coherence function of the pressure field between two points in the space. For three-dimensional sound fields the coherence is computed as [17, 8]

$$\gamma_{\mathrm{p}}(\omega, r) = \frac{sin(\omega r/c)}{\omega r/c}, \tag{2.6}$$

where ω is angular frequency, and r is the distance between the two points.

Apart from being able to compute some characteristic parameters about the sound field in a reverberant space, it would be useful to be able to model and to simulate the sound field. A well-known method for simulating reflections in a room is to use the image-source method [18]. The level, the delay, and the direction of the reflections are computed by mirroring the rooms at the boundaries and considering the reflections to be separate sources inside the mirrored rooms. Other methods for simulating reflections include, e.g., ray tracing, the finite-element method, and the finite-difference time-domain method [7]. Due to computational constraints, only the early reflections are typically computed using these methods. The late reverberation is often simulated as decaying random noise, where the decay envelope

for the considered acoustic environment is determined according to the prediction of energy decay curve [6]. In addition to simulations, the early reflections can also be localized by using microphone-array measurements of real spaces [19].

3. Human perception of spatial sound

The previous section discussed the physics of spatial sound. However, in this thesis, we are mostly interested in the reproduction of spatial sound for *humans*. Thus, we are only indirectly interested in the actual physical sound fields. Our main interest is in the *perception* of the spatial sound. This section discusses how our auditory system processes and perceives it.

Humans can perceive the direction and the distance of a sound source, as well as some spatial properties, such as the size and the reverberance of the room. Let us first consider the perception of the direction in the simplest case, i.e., a single sound source in anechoic conditions. After that, more complex auditory scenes are discussed and reverberation is also taken into account.

3.1 Single sound source in anechoic conditions

Let us assume that there is a listener in an anechoic room, and a sound source is located at 40° of azimuth in the horizontal plane (see Fig. 3.1). Two things can be seen in the figure: a) the sound arrives first to the right ear and after a short delay to the left ear, and b) the head is in between the left ear and the sound source, whereas there is a clear line between the right ear and the sound source. These properties are used by our auditory system to localize sound sources.

The time delay between the ears is typically known as the *inter-aural time difference* (ITD). The time difference translates into a frequency-dependent phase difference in our hearing, based on which our hearing can detect the direction of arrival in the left-right direction [20]. This cue is salient especially below 1.6 kHz [21]. Above that frequency the cycle time is shorter than the largest possible ITD value. Consequently, the phase differences between the ear signals do not provide unambiguous

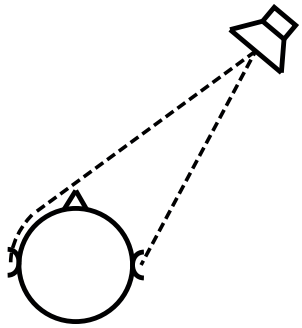

Figure 3.1. Sound source located 40 degrees to the right.

cues anymore. However, based on the temporal envelope of the pressure signal, the hearing is sensitive to ITD also above 1.6 kHz [22].

The shadowing of the head causes a difference in the levels of the ear-canal signals. This is known as the *inter-aural level difference* (ILD). Also ILD enables us to detect the direction of arrival in the left-right direction [22]. However, whereas ITD is a reliable cue at low frequencies, ILD provides a useful localization cue at high frequencies [20]. At low frequencies, the head does not give rise to shadowing due to the large wavelength, and thus the ILD cue is not available. At high frequencies, the wavelength is short enough, and the head causes significant level differences. Hence, according to the duplex theory [23], below 1.5 kHz the localization is based on ITD, and above it on ILD. More recent studies have shown that both cues actually have some effect below and above 1.5 kHz [20], but the duplex theory is a good starting point for describing the localization.

ITD and ILD provide information about the direction in the left-right direction, but there is no information, for example, whether the sound source is in front of or behind the listener. On the surface of so-called cone of confusion [22], the ITD and the ILD cues are similar, and based on these the listener cannot judge where the source is located on this cone. Thus, other cues, referred to as *monaural cues*, are needed. An impinging wave travels directly from the source to the ear, but there are also reflections from the pinnae, the head, and the torso of the listener. These reflections cause peaks and dips to the magnitude spectrum of the sound that reaches the eardrum, mostly at high frequencies. Based on the relative spectrum, humans are able to localize sound sources in the elevation direction [20]. It should be noted that in order to localize sound sources using the spectral cues, the cues should be distinguished from the spectral peaks and dips

inherent in the source. Thus, some kind of familiarity with the spectral shape of the sound has been found to help in localization [24]. Nevertheless, the familiarity might not be required in all cases, because a) the peaks and the dips produced by the pinnae are sharper than typically found in the spectrum of natural sounds at high frequencies, and b) the spectral cues are different for different ears [24].

Using ITD, ILD, and monaural cues, humans can localize sound sources both in the azimuth and the elevation direction. The accuracy of localization depends on the direction of the sound source. In addition, localization in the horizontal plane is more accurate than within cones of confusion. In the front, the minimum audible angle, i.e., the smallest perceivable change in the direction of the sound source, is about 1° in the azimuth direction and about 4° in the elevation direction [20]. The localization accuracy decreases on the sides [20].

3.2 Multiple sound sources and reverberant conditions

In a realistic scenario there are typically multiple simultaneous sound sources and also reverberation. Thus, considering only single directional cues is not enough. This section starts by considering the frequency and the time resolution of hearing. The perception of reverberation and multiple sources is discussed next, and the section ends with the discussion of the perception of distance.

3.2.1 Frequency and time resolution of hearing

Let us first consider how the sound pressure in the air is transformed to a perception of sound in the human auditory system. The pressure signal in the ear canal causes vibrations on the eardrum, and these vibrations are then transmitted through the ossicles in the middle ear to the cochlea [25]. The cochlea converts the mechanical vibrations to neural pulses with hair cells, which are organized tonotopically along the *basilar membrane* [25]. The base of the basilar membrane is relatively narrow and stiff, which causes it to respond best to high frequencies, whereas the apex of the basilar membrane is wider and less stiff and, thus, responds best to low frequencies [24]. Each point on the basilar membrane is tuned and responds with the greatest displacement to a certain frequency, which is called the characteristic frequency [24]. Using this property, the human

hearing can detect the frequency content of the stimulus. On the other hand, pitch perception has been suggested to be based on spectral pattern matching of neural activity [26] or on the temporal intervals of neural activity [27].

However, as the basilar membrane is a physical membrane, the displacement takes place at a wider area than only at the point of the characteristic frequency. Hence, the frequency resolution of human hearing is limited. It is often suggested that the peripheral auditory system behaves as if it contained a bank of overlapping band-pass filters [28], referred to as the *auditory filters*. The hearing can be assumed to handle broadband sound so that the partial sounds inside the auditory filter are analyzed as one entity. The width of these filters, often called the *critical bandwidth*, can be approximated to follow the equivalent rectangular bandwidth (ERB) [29]. ERB can be measured using a notched-noise method and has been found to follow the equation

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

where f_c is the center frequency of the band (in kHz) [30]. Correspondingly, an ERB scale can be created using

$$\xi_\mathrm{ERB} = 21.4 \log_{10}(4.37 f + 1), \tag{3.2}$$

where ξ_ERB is the number of ERBs and f is frequency (in kHz) [30]. ERB has been determined in monaural conditions. Nevertheless, several studies have shown that the critical bandwidth in binaural listening is equal or slightly larger than in monaural listening [31, 32, 33]. Thus, it is assumed that binaural frequency resolution can be described with sufficient accuracy using the ERB scale.

The temporal resolution of hearing is also of interest in complex listening scenarios. It is typically assumed to be of the order of 2 ms. The auditory system can be interpreted to contain a sliding temporal integrator with a corresponding window length for smoothing the neural signals [24]. Several studies support this assumption. For broadband noises, the threshold for detecting a temporal gap is about 2–3 ms [24], as it is also for distinguishing the order of two successive clicks differing in amplitude [24]. Furthermore, the auditory system can discriminate between two transient signals that have identical energy spectra but different phase spectra as long as the total duration of the signals exceeds about 2 ms [34]. Lastly, when single sinusoids are presented, the neural firing rate

of the hair cells inside the critical band shows a pulse for each period of the sinusoid at a temporal position corresponding to a certain value in the phase of the sinusoid. The temporal length of the pulse is about 0.5–1.0 ms [35]. Similar results were obtained also in Publication I, where the temporal resolution of 1.3 ms was found to be sufficient for analyzing and reproducing applause-type signals.

Correspondingly, it has been shown that the temporal resolution for binaural processing, such as detecting the directions of the sound sources, is significantly slower [33]. Values between about 15 and 100 ms have been suggested [36, 37]. Hence, due to the 'binaural sluggishness', the binaural cues have to be reproduced only with this temporal accuracy for faithful directional perception.

3.2.2 Precedence effect

As shown in Fig. 2.1, a direct sound within a room is followed by a large number of reflections. The directions of these reflections differ from that of the direct sound. However, humans are able to reliably localize sound sources also in reverberant conditions. The reason for this is that the first-arriving wavefront dominates our perception of direction. This effect is known as the *precedence effect* [38].

The effect of a reflection on the perception depends on the temporal lag [20]. Let us assume a single reflection with a level equal to that of the direct sound. In addition, let us assume that the directions of the direct sound and the reflection are symmetric in the left-right direction. If the reflection arrives at the same time as the direct sound, a single 'phantom' source is perceived in the middle of them. As the delay increases from 0 to 1 ms, the phantom source moves toward the direction of the direct sound. This phenomenon is typically referred to as summing localization [22]. If the delay is larger than 1 ms, the sound source is localized to the direction of the direct sound, with practically no contribution from the reflection. However, the reflection can affect other aspects of perception, such as loudness, spaciousness, and timbre. When the delay reaches a certain value, the echo threshold [22], the image breaks up into two distinct auditory images. The echo threshold depends on the type of the stimulus, and has been found to vary between 5 and 75 ms (about 35 ms for speech) [20]. In addition, it should be noted that the level of the reflections affects the precedence effect.

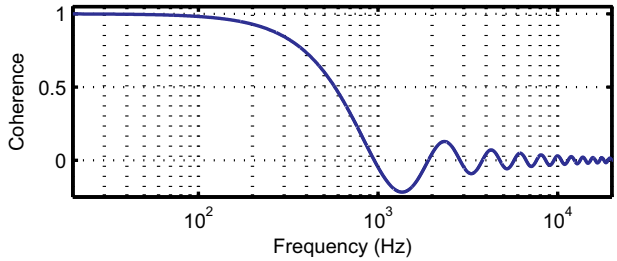

Figure 3.2. Approximated inter-aural coherence (IC) in a diffuse field as a function of frequency.

3.2.3 Inter-aural coherence

When a single source is presented in anechoic conditions, the ear canal signals are relatively coherent. Correspondingly, the coherence between two points in a diffuse field can be approximated using Eq. 2.6. The spacing between the ears is roughly 18 cm, so the *inter-aural coherence* (IC) is small at a wide frequency range (see Fig 3.2). Thus, IC can be seen as a potential cue for perceiving the reverberance and the diffuseness of the room. Experimental studies have shown that humans in fact can perceive the amount of coherence between the ear-canal signals [22]. If two signals with the coherence of 1 are reproduced using headphones, the perceived auditory event is a single, relatively small, image in the middle of the head. When the coherence is decreased, the size of the auditory image increases. When IC is about 0.4, the auditory event is perceived to fill the whole head. If the coherence is decreased even further, two spatially separated auditory events may appear, one at each ear.

Hence, if a realistic reproduction of spatial sound is desired, also the IC cues have to be reproduced correctly. How human auditory system actually detects IC is under discussion. A traditional view is based on mechanisms which can be conceived as a computation of inter-aural cross-correlation [39], whereas more recent studies are based on temporal fluctuation of localization cues [40]. Furthermore, fluctuation of the localization cues has been suggested to explain the perception of apparent source width and envelopment [41]. Nevertheless, no matter what the mechanism of human hearing for detecting differences in IC is, it is seen as important for the perception of spatial sound.

As discussed in Section 2, a diffuse field is defined to contain an infinite number of plane waves. However, it is not practical to have a very large

number of loudspeakers to reproduce all these waves. Hence, it is interesting, from a practical point of view, to know how many loudspeakers are needed to reproduce a sound field that is perceived equally as a diffuse field. It has been found that four loudspeakers, when placed optimally, are enough for obtaining a spatial impression resembling that of a diffuse field [42]. When distributed evenly, the required number has been found to be six [42].

3.2.4 Perception of distance

According to [20], at least the following cues have been identified to affect the perception of distance:

1. Sound pressure level (the greater the SPL, the shorter the judged distance).
2. The amount of reverberation (the greater the ratio of direct-to-reverberant energy in the received signal, the shorter the judged distance).
3. Spectral shape of the received signal (the greater the high-frequency content of the stimulus, the shorter the judged distance).
4. Binaural cues (for sources off midline and closer than about 1 m: the greater the ITD or ILD, the shorter the judged distance).

It should be noted that as cues 1 and 3 are relative, some familiarity about the target is required so that the SPL and the spectrum can be used for determining the distance. This was confirmed in [43], where sounds were reproduced at different distances in different kinds of rooms. The sound level was normalized to be equal in the listening position regardless of the distance of the sound source. In anechoic conditions, there was no correspondence between the physical and the perceived distance. Instead, the perceived distance was found to be correlated with the loudness of the sound, which was artificially controlled by the experimenters. On the contrary, listeners were able detect the distance of the sound source in normal listening rooms, where there was reverberation present, regardless of the loudness of the sound.

4. Reproduction of spatial sound

As discussed in Section 3, humans can perceive several properties of sound sources, including direction, size, and distance, as well as the reverberance of the room. The aim of spatial-sound reproduction methods is to create a sound field that causes a corresponding perception of these properties. These methods have many uses: controlling the perceived direction, creating reverberation, capturing and reproducing real sound scenes, and synthesizing realistic virtual acoustics. An overview of well-known methods for realizing these tasks is given in this section, and they form a foundation for the parametric methods discussed in Section 5.

4.1 Panning techniques

Panning techniques are used for controlling the perceived direction of single sound sources. Typically, the input is a mono recording of the source in low-echoic conditions, and the source is panned to a desired direction. In addition, Section 4.5 shows how these techniques can be used when rendering complex virtual scenarios with multiple sources and reverberation.

4.1.1 Amplitude panning

Amplitude panning [44] is the most common approach for controlling the perceived direction of a sound source. The basic idea is that a mono signal is applied into two loudspeakers, and the individual gains of the loudspeakers are adjusted (see Fig. 4.1). The gains (g_1, g_2) for the loudspeakers can be computed using the sine law

$$\frac{\sin\theta}{\sin\theta_0} = \frac{g_1 - g_2}{g_1 + g_2}, \tag{4.1}$$

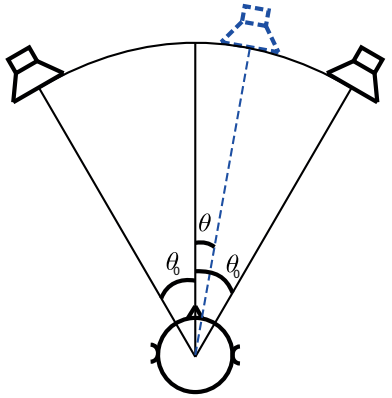

Figure 4.1. Amplitude panning. The perceived location of the virtual source can be controlled by adjusting the relative gains of the loudspeakers.

where $2\theta_0$ is the angle between the loudspeakers and θ is the desired angle of the virtual sound source [45]. The loudspeakers are placed symmetrically relative to the median plane. This equation is suitable if the listener is looking forward all the time and the effect of head shadowing is neglected. However, often this is not the case, but instead the listener turns his/her head towards the auditory event. For such cases, the tangent law has been suggested to be more correct [46]

$$\frac{\tan\theta}{\tan\theta_0} = \frac{g_1 - g_2}{g_1 + g_2}. \tag{4.2}$$

The tangent law is also more accurate when the effect of head shadowing is taken into account [47].

Even though only the level of the sound is controlled in amplitude panning, it has been found that amplitude panning actually translates into both ITD and ILD cues. The pressure signals from both loudspeakers reach both ears. These pressure signals are summed at the eardrum in each ear. When the summed signals are inspected, it can be seen that the time difference between the left and the right ear roughly matches the corresponding ITD of the desired virtual sound source at low frequencies and the level difference roughly matches the corresponding ILD at high frequencies [48]. Thus, humans perceive the direction of the auditory event as desired. The cues deviate slightly from each other depending on frequency, and this can result in spatial spreading of the virtual source [48]. Nevertheless, this spreading is typically not regarded as a major problem.

Conventional panning works only in the horizontal plane. However, it can be extended into three dimensions by using vector-base amplitude panning (VBAP) [49]. The input mono signal is applied into a triplet of loudspeakers

and the gains of these loudspeakers are controlled as in two-dimensional panning. The gains are computed using a vector-base formulation. The advantage of VBAP is that it can be used with arbitrary loudspeaker layouts, and the loudspeaker triplets are automatically formed by the algorithm. Furthermore, listening tests have shown that the perceived direction matches the desired direction relatively well in most cases [50].

4.1.2 Time-delay panning

As discussed already in Section 3.2.2, if there is a delay between the left and the right ear-canal signals, the perceived direction moves towards the preceding ear as the delay is increased from 0 to 1 ms. Introducing an inter-channel delay between the loudspeaker signals causes a similar effect [22]. This property can be used to control the perceived direction of the auditory image. However, as in the case of amplitude panning, the pressure signals from both loudspeakers reach both ears. Unfortunately, in the case of the time-delay panning, the localization cues produced by the summed ear-canal signals vary significantly depending on frequency [51, 52]. As a result, the auditory image is perceived to be blurred and unstable [51]. Hence, time-delay panning is rarely used as a panning tool.

4.2 Traditional microphone techniques

There is a large variety of recording techniques for spatial sound reproduction over multi-channel setups [53]. Typically, a number of microphones equal to the number of loudspeakers in the reproduction setup are used, and the microphone signals are directly routed to the corresponding loudspeakers. The aim is that the reproduced sound field would be perceived somewhat similarly to the original sound field. These techniques are based on the same psychoacoustical principles as the panning techniques.

The microphone techniques can be roughly divided into two groups: coincident and spaced-microphone techniques. XY techniques are an example of two-channel coincident techniques. Two microphones with first-order directional patterns, such as a cardioid or a hypercardioid, are placed close to each other with angles varying from 60° to 180° [54] (see Fig. 4.2(a)). The differences between the microphone signals are mostly limited to level differences. Thus, the perception of sound reproduced with XY techniques is similar to amplitude panning (see Section 4.1.1), and good perceptual

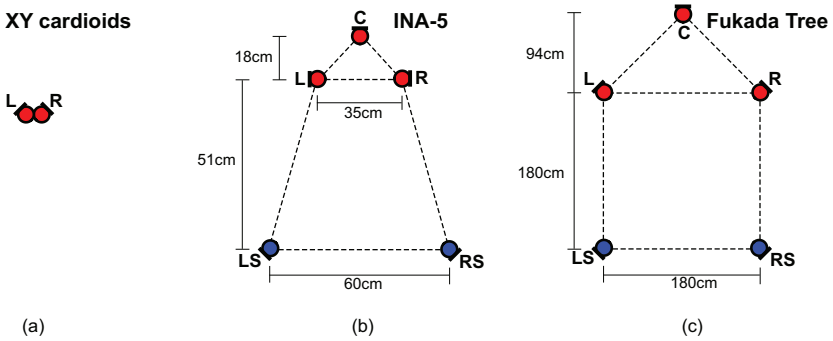

Figure 4.2. Example microphone arrays used for recording of stereo (a) and multi-channel audio (b), (c). The dimensions and the orientations of the microphones are based on [54, 55].

quality can be obtained [51]. However, using coincident techniques with multi-channel setups is more problematic. The broad directivity patterns result in high inter-channel coherence, which causes timbral and spatial artifacts (see Section 4.3.1 for more information).

The spaced-microphone techniques mostly do not suffer from excessive coherence. The spacing between the microphones ranges from about 30 cm to several meters [55] (see Figs. 4.2(b) and (c) for a few example arrays). Thus, the coherence between the microphones is low in a diffuse field (see Section 2), and the reverberation of the reproduced sound field is typically perceived as enveloping and spacious [55, 56]. However, the spacing between the microphones results in inter-channel time differences, and directional microphones cause inter-channel level differences. Thus, the spaced-microphone techniques can be seen as a combination of amplitude and time-delay panning techniques. As discussed in Section 4.1.2, the time differences cause ambiguous directional cues, and as a result, the perceived direction is vague and inaccurate [51, 54].

4.3 Sound-field reproduction techniques

Similarly to the panning techniques, the sound-field reproduction techniques can be used to position single sources. In addition, they can be used as microphone techniques for capturing real sound scenarios. Rendering complex virtual scenarios is discussed in Section 4.5.

4.3.1 Ambisonics

A sound field can be described using the wave equation, as discussed in Section 2. If spherical coordinates (r, θ, ϕ) are used, the solution to the wave equation can be written in terms of spherical Bessel functions and spherical harmonics

$$p(r, \theta, \phi, \omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} A_n^m(\omega) j_n(\omega r) Y_n^m(\theta, \phi), \tag{4.3}$$

where $A_n^m(\omega)$ contains the spherical harmonic coefficients, $j_n(\omega r)$ is the spherical Bessel function of the first kind, and $Y_n^m(\theta, \phi)$ is the spherical harmonic function [57]. The equation requires an infinite number of harmonics, so in practice the sound field has to be approximated by using a finite order of harmonics. The first-order Ambisonics [58, 59] is based on describing the sound field using only the zeroth and first order spherical harmonics, whereas in the higher-order Ambisonics (HOA) [58, 60, 61] harmonics up to Rth order are used.

The input to the first-order Ambisonics is the B-format signals. They consist of an omnidirectional signal W and X-, Y -, and Z-signals having the directional pattern of a dipole directed along the Cartesian axes [59]

$$\begin{aligned} W &= 1 \\ X &= \sqrt{2}\cos(\theta)\cos(\phi) \\ Y &= \sqrt{2}\sin(\theta)\cos(\phi) \\ Z &= \sqrt{2}\sin(\phi), \end{aligned} \tag{4.4}$$

where θ is the angle in the azimuth direction and ϕ in the elevation direction (see Fig. 4.3). The reproduction of the sound can be performed with any number of loudspeakers N, provided that $N \geq 4$. The B-format signals are fed to all loudspeakers after multiplying with a static mixing matrix, which is defined by the positioning of the loudspeakers. This is one of the advantages of Ambisonics: the B-format signals are a flexible way to transmit audio, since they are not bound to any specific loudspeaker setup. In HOA the reproduction is performed in a similar way, but the number of spherical harmonic functions and the minimum number of loudspeakers N_{min} is larger depending on the order R [58]

$$N_{\text{min}} = (R + 1)^2. \tag{4.5}$$

The reproduction is performed similarly using a mixing matrix.

Ambisonics can be used both as a panning technique and as a microphone technique. When using as a panning technique, the B-format signals can

Figure 4.3. Directional patterns of the B-format signals.

be computed using Eq. 4.4 for the first and the zeroth order and with corresponding equations for the higher orders. When using as a microphone technique, microphones with corresponding directivity patterns are required. Soundfield ST450 [62] is an example of a commercial first-order microphone, and Eigenmike [63] of a higher-order microphone.

An advantage of Ambisonics is that the reproduced sound field is physically correct within a certain area, for a certain frequency range. Thus, also the perception is correct inside this area at these frequencies. The downside is that the area is relatively small, depending on the order of the spherical harmonics available and the frequency. For single-position listening, the diameter of the required area is at least 20 cm to fit both ears inside the area. For first-order Ambisonics, the highest frequency for which accurate reproduction of the sound field can be obtained is about 500 Hz [57]. A significant amount of energy is typically found outside this frequency range. The required order for accurate broadband reproduction is 30, for which the minimum number of loudspeakers is 961 [57]. Thus, with practical loudspeaker layouts the sound field cannot be reproduced correctly for broadband signals at both ears. Moreover, if there are multiple listening positions, as there typically are, the suitable frequency range is smaller, or the required order is even higher.

As the reproduced sound field contains significant errors with practical systems, the question is how do humans perceive these errors. The loudspeaker signals are relatively coherent, which translates into higher IC than in the original sound field (see Section 3.2.3). Humans perceive this as timbral and spatial artifacts [1, 64], such as lack of envelopment. In addition, phasing effects are typically perceived when the head is moved.

4.3.2 Wave-field synthesis

Similarly to Ambisonics, the wave-field synthesis (WFS) [65, 66] aims at recreating the physical sound field. The basis of the wave-field synthesis is formed by Huygen's principle, stating that any point of a wave front can

be considered as a secondary source [65]. Thus, the wave-field of a virtual sound source can be synthesized with a number of secondary sources distributed on the surface of a bounded region. The problem with WFS is that the required number of loudspeakers is high for the reproduction of a broad frequency range. Thus, WFS is not very practical in typical use cases. However, if it is possible to have enough loudspeakers, it is possible to obtain the correct sound field for the selected frequency range.

4.4 Binaural techniques

As discussed in the previous sections, panning techniques aim at controlling the directional cues produced by the sound field, whereas Ambisonics and WFS aim at controlling the actual sound field. Binaural technologies can be seen as a third approach: controlling the pressure signal at the eardrum of a human listener. The basic idea is simple: the sound enters the auditory system through the eardrum, and if the sound pressure can be controlled at the eardrum, the perception of sound can also be fully controlled. (To be exact, it should be noted that sound as a physical phenomenon does not actually describe the *perception* of sound alone. Also other aspects, such as expectations and audio-visual interaction [67], have been found to influence the perception of sound. Nevertheless, these aspects are omitted in this section.)

Binaural techniques [68, 69, 22, 70] typically utilize so-called *head-related transfer functions* (HRTF). HRTF is a transfer function that, for a certain angle of incidence, describes the sound transmission from a free field to a point in the ear canal of a human subject [71]. It is measured for both ears. If anechoic recording of a sound source is convolved with a given HRTF pair and reproduced at the eardrums, the sound pressure, and thus also the perception, should be equal to the sound pressure caused by a sound source positioned in the direction where the HRTFs were measured. Hence, HRFTs enable the reproduction of auditory events positioned freely in 3D.

The question is how to measure and to reproduce the HRTFs. HRTFs are typically obtained from corresponding head-related impulse responses (HRIR), which can be measured using sinusoidal sweeps [72, 73, 74]. As we are interested in the sound pressure at the eardrum, it would be optimal to also measure it at the eardrum. However, measuring the response at the eardrum is difficult and also risky for human subjects [75]. Thus, HRTFs

are typically measured at the entrance of the ear canal, and the response of the ear canal is assumed to be direction independent [71]. The ear canal can be blocked or open [71]. Databases consisting of HRTFs measured from multiple subjects can be found, for example, in [76, 77]. An alternative to using human subjects is to use artificial heads [68] which are modeled based on 'average' human anatomy. The advantage is that the microphones can be freely positioned, and the subject stays still even for long-duration measurements.

Headphones are an obvious choice for reproducing binaural sound since the crosstalk between the channels is thus avoided. Special methods, such as 'stereo dipole' [78], are required when using loudspeakers. However, if sound convolved with HRTFs is directly reproduced with headphones, the signal at the eardrum contains also the response of the headphones, which has to be taken into account. Hence, *headphone transfer functions* (PTF) are typically measured [79]. PTFs are measured similarly to HRTFs, but the headphones instead of the loudspeaker are used as the source. The microphones have to be placed in the same position as when measuring the HRTFs since the sound pressure at the entrance of the ear canal varies significantly even with small movements [22]. In addition, the headphones have to be the same as the ones used for reproduction. Simplified, the binaural reproduction can be presented as

$$x_{\mathrm{binaural}}(f,i) = \frac{H_{\mathrm{HRTF}}(f,i,\theta,\phi)x_{\mathrm{in}}(f)}{H_{\mathrm{PTF}}(f,i)}, \tag{4.6}$$

where $x_{\mathrm{binaural}}(f,i)$ is the reproduced binaural signal, f is the frequency, i is the headphone channel, $H_{\mathrm{HRTF}}(f,i,\theta,\phi)$ is the measured HRTF, $x_{\mathrm{in}}(f)$ is the input signal to be spatialized, and $H_{\mathrm{PTF}}(f,i)$ is the measured PTF. Exact binaural synthesis, which is not discussed here, should take into account also other properties, such as the response of the loudspeaker and the pressure-division ratio (PDR) [71, 79]. Nevertheless, the aim is to obtain binaural signals that are similarly perceived as if there were a sound source at the direction (θ,ϕ). Interpolation techniques can be used in between the measured points [80]. Alternatively, amplitude panning can be used as presented in Publication IV.

One important aspect to take into account is the movement of the listener's head. If static HRTF rendering is used, the auditory image moves with the rotation of the head (see Fig. 4.4). This can be avoided by using *head tracking* [81, 82], in which the orientation of the listener's head is tracked and the HRTF in use is updated according to the orientation. This allows the position of the auditory event to be kept constant (see

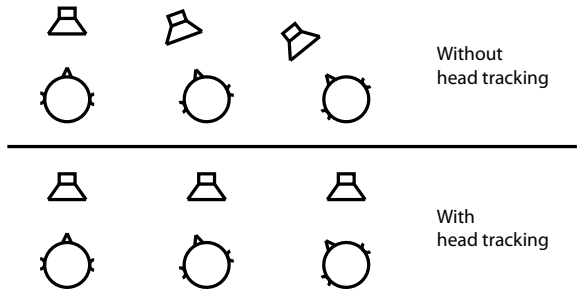

Figure 4.4. Effect of head tracking in binaural reproduction. Without head tracking the auditory event moves with the rotation of the head, whereas with head tracking the auditory event stays in its position.

Fig. 4.4), and also corresponds with natural listening conditions. Moreover, if head tracking is not used, binaural rendering often suffers from front-back confusion, i.e., the judgment of a sound stimulus as located on the opposite side of the inter-aural axis than the target position [82]. The use of head tracking has been reported to reduce this error significantly [82]. Furthermore, in Publication IV, head tracking was found to increase the naturalness of the spatial impression.

The influence of latency in head tracking was studied in [83]. It was found that the detection threshold for the latency is about 75 ms. Hence, the latency of the system should be smaller than this for optimal quality. Similar results were obtained in Publication V, where the update rate of 18 Hz was found sufficient for most listeners. In addition, it was found, in Publication V, that the tracking in azimuth direction is the most important and that tracking should follow all possible rotations without restrictions. Nevertheless, it was also found that even suboptimal tracking increases the perceived naturalness compared to no tracking at all.

4.5 Reproducing virtual acoustics

In virtual acoustics, the aim is to synthesize a realistic perception of a virtual space containing multiple sound sources and reverberation. The virtual space can be designed to resemble real-life spaces, but it can also be something else. The task of producing virtual acoustics can be divided into three stages: (a) definition, (b) modeling, and (c) reproduction [7]. The first task is to define the properties of the space, the sources, and the receivers in the virtual environment, e.g., the geometry of the room and the locations of the sources. The next task is to model the direct

sound and the reflections inside the room. The early reflections can be modeled using, for example, the image-source method [18], and the late reverberation as decaying random noise [6], as discussed in Section 2. The output of this stage consists of the gains, the delays, and the directions of each early reflection and the direct sound for each source-receiver pair and the statistical properties of the late reverberation. The final task is to reproduce these components. A straightforward method for that is to reproduce each reflection and the direct sound as separate single sources and the late reverberation as decaying diffuse noise [7]. The single sources can be created using, e.g., methods presented in the previous sections, such as amplitude panning, Ambisonics, WFS, or binaural techniques. Late reverberation can be reproduced using reverberators.

Hence, complex scenarios with multiple sources and reverberation can be reproduced as a sum of single anechoic sources. Interactive systems for reproduction of virtual acoustics have been presented, e.g., in [7, 84, 85]. These systems can be used, e.g., for the simulation of acoustics of real spaces. This aims to accurately reproduce the acoustics and can be used in designing the spaces. Another possible usage is to reproduce spatial sound in virtual environments, in games or movies for example. In these applications, it is not important to have accurate reproduction but, instead, plausible perception of the spatial sound. Furthermore, low computational complexity is often required. Thus, the computation of reflections is often simplified to lower the computation complexity.

One issue with virtual environments is the reproduction of spatially extended sources. One option is to use a large number of point sources [86], which is a viable option if computational complexity is not an issue. For real-time applications, some simplifications, such as clustering the point sources based on the perceptual importance [86], are needed. Creating spatially extended sources by applying different frequency bands into different directions is suggested in Publication VII. A similar approach was suggested in [87].

5. Parametric spatial-sound reproduction and coding in time-frequency domain

This section discusses the reproduction and the coding of spatial sound. In reproduction of spatial sound, the input to the system is often an actual physical sound field, and the aim is to recreate a realistic perception of this sound field by using loudspeakers or headphones. Correspondingly, in coding, the input to the system is a set of signals that has been created to be directly reproduced using loudspeakers or headphones, and the aim is to create a data stream that has a low bit rate but enables creating a set of signals which are perceived as the original signals when reproduced.

These two tasks are very different, but as later will be shown, similar approaches can be used for both of them. Furthermore, traditional methods for both tasks face challenges. As discussed in Section 4, good perceptual quality in spatial-sound reproduction can be obtained with traditional methods in many cases, but often there are problems in at least some aspects of reproduction: errors in timbre, localization, or feel of space, requirements for massive amounts of, or even non-existing, microphones and loudspeakers, etc. Correspondingly, there are well-established methods for the coding of single audio channels, such as MPEG-1 audio layer 3 (MP3) and advanced audio coding (AAC) [88]. However, the required bit rate for multi-channel signals is often too high, as the single-channel bit rate is directly multiplied by the number of the loudspeaker channels.

To avoid these problems, the concept of parametric time-frequency processing has been introduced to spatial-sound processing. Faller and Baumgarte introduced binaural cue coding (BCC) in 2001 [89], and they were shortly followed by Schuijers et al. with parametric stereo (PS) [90], Merimaa and Pulkki with what became directional audio coding (DirAC) [91], Herre et al. with MPEG surround [92], and Goodwin and Jot with spatial audio scene coding (SASC) [93]. The basic idea in all of these methods is the same: analyze certain properties of the sound field that are significant

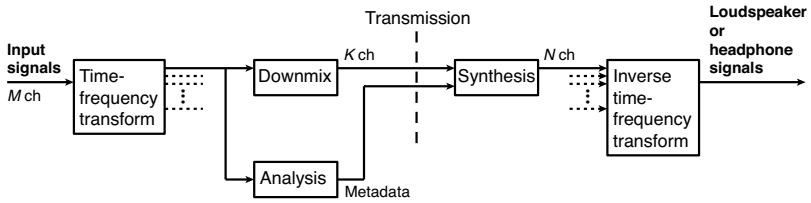

Figure 5.1. Generalized block diagram of parametric spatial-sound coding and reproduction methods. M is the number of input channels, K transmitted channels, and N output channels. Typically, $M = N > K$ in coding applications, whereas $M = K \leq N$ in reproduction applications. The processing is performed separately for each frequency band.

to human perception of spatial sound and use them to synthesize a sound field that is perceived equally as the original sound field.

A generalized block diagram of parametric spatial-sound coding and reproduction methods is presented in Fig. 5.1. It applies to all methods mentioned above. The processing begins with a time-frequency transform since the human hearing is analyzing sounds mostly in the frequency domain, as discussed in Section 3.2.1. The time and the frequency resolution of the transform should follow the properties of the human hearing. The next step is to analyze parameters related to spatial-sound perception. They are typically related to the directional and the coherence cues in our hearing (ITD, ILD, and IC, see Sections 3.1 and 3.2.3). The values of the analyzed parameters for each time-frequency tile form the metadata, which is sent along the audio signals. In the case of coding, the input audio signals are downmixed into one or two audio signals before the transmission to reduce the bit rate. Furthermore, the transmitted audio signals are typically core coded (e.g., using AAC) to further decrease the bit rate. These signals are upmixed in the synthesis phase using the metadata to create a sound field that is perceived equally as the original signals. Correspondingly, in the case of spatial-sound reproduction, downmixing is not typically required, and the metadata is used to enhance the microphone signals to obtain a sound field that is perceived as the sound field where the recording was performed.

The previously mentioned methods (BCC, PS, DirAC, MPEG surround, and SASC) are discussed in detail in the following sections, and also other similar methods are briefly mentioned. The work in this thesis has been conducted within the framework of DirAC, so most emphasis is given to it. However, most of the techniques suggested in this thesis are applicable to any parametric spatial-sound reproduction or coding method operating in

the time-frequency domain.

5.1 Directional audio coding

Directional audio coding (DirAC) [1] is a perceptually motivated method for reproducing spatial sound. The general idea is that there is no need for physically accurate sound-field reproduction for good audio quality. Instead, it is enough that perceptually significant properties of the original sound field are analyzed and that these properties are accurately reproduced. DirAC is based on the following assumptions about the interaction between the sound-field properties and the perceptual attributes they produce [2, 1]:

- The direction of the sound source transforms into ITD, ILD, and monaural cues (see Section 3.1).
- The diffuseness of the sound field transforms into IC cues (see Section 3.2.3).
- Timbre depends on the monaural spectrum together with ITD, ILD, and IC.
- The direction of arrival (DOA), the diffuseness, and the spectrum of sound measured in one position of interest with the temporal and the spectral resolution of human hearing (see Section 3.2.1) determines the auditory spatial image the listener perceives.

It is further assumed that at one time instant and at one critical band the auditory system is limited to decoding one cue for direction and another for inter-aural coherence [1]. This leads to the formulation of the DirAC processing:

- Capture the monaural spectrum. (Recording)
- Analyze the DOA and the diffuseness for each critical band with the temporal resolution of hearing. (DirAC analysis)
- Synthesize a sound field that has the correct DOA and diffuseness properties and the correct spectrum in one position with the corresponding resolutions. (DirAC synthesis)

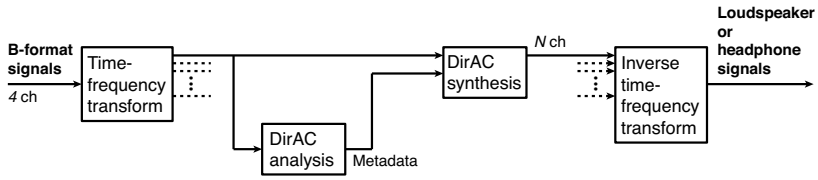

Figure 5.2. Block diagram of DirAC processing. The processing is performed separately for each frequency band. The block diagrams of the time-frequency transform, the DirAC analysis, and the DirAC synthesis are presented in Figs. 5.3, 5.4, and 5.5, respectively.

If the previously mentioned assumptions are correct and if these tasks can be fulfilled, the reproduced sound field should be perceived equally as the original sound field. The question is then how can we carry out these tasks. This is discussed in the following.

The block diagram of DirAC processing is presented in Fig. 5.2. The processing begins by recording the sound scene, and the microphone signals are time-frequency transformed in order to perform the processing separately for each time-frequency tile. The DirAC analysis is performed with these signals and the metadata is obtained as a result. The DirAC synthesis is performed using the recorded microphone signals and the analyzed metadata. The processing ends with the inverse time-frequency transform, and the resulting signals can be reproduced using loudspeakers or headphones. The different processing blocks are now discussed in detail in the following sections.

5.1.1 Recording

The sound scene is recorded using a B-format microphone (see Section 4.3.1). Recently, the use of other kinds of microphone inputs has also been suggested, e.g., XY cardioids [94], linear arrays [95], and A-format [96]. In addition, a method for using spaced-microphone recordings is presented in Publication III.

5.1.2 Time-frequency transform

Common ways for performing the transform include the short-time Fourier transform (STFT) and a filter bank [97]. In STFT, the sound is processed in overlapping time frames, which are transformed, using the fast Fourier transform (FFT), to frequency domain. The frequency-domain signal is manipulated with frequency-dependent gains which are computed in the

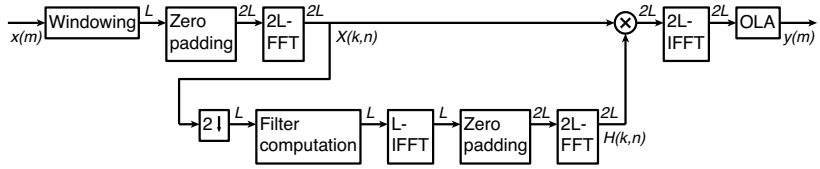

Figure 5.3. Block diagram of the time-frequency transform used in DirAC processing. m is discrete time, k is the frequency band index, and n is the temporal frame index. The input signal is processed in frames of the size L. The filter is computed in the DirAC analysis and synthesis stages. It should be noted that the process of zero padding, $2L$-size FFT, and decimation corresponds to computing a L-size FFT.

DirAC analysis and synthesis stages (i.e., the signal is filtered), and the result is transformed back to time domain via inverse FFT (IFFT). This method is commonly known as the overlap-add method (OLA) [98]. However, the filter is time variant in DirAC processing, which can cause crackling noise due to the circularity of FFT. Thus, a method for preventing this artifact was developed in this work (see Fig. 5.3). The suggested method is similar to the frequency-extension method presented in [99]. Effectively, the filter is computed in the frequency domain with the window size of L, and both the filter and the audio signals are zero-padded to $2L$ in the time domain before the multiplication in the frequency domain. This guarantees the absence of crackling artifacts due to the time-frequency processing. A summary of different STFT methods can be found in [98].

The parameters to be controlled in the transform are the window type, the window length, and the hop size of the window. Typically, a Hann window with the hop size of $L/2$ is used. The window size L is selected according to the required frequency and time resolutions. As discussed in Section 3.2.1, the frequency resolution of human hearing approximately follows the ERB bands, which are about 30 Hz at the lowest frequencies and become wider as the frequency is increased. Correspondingly, the temporal resolution is about 2 ms. Thus, the time-frequency transform should follow these resolutions. The temporal resolution of STFT is roughly the length of the window L, and the frequency resolution is $1/L$. Thus, it is not possible to obtain adequate frequency and time resolution at the same time. One option is to use a compromise, such as a 20-ms long window, as suggested in [97]. Alternatively, multi-resolution STFT processing can be used, as suggested in Publication I, in which the audio signal is divided into two or more frequency ranges and different window sizes are used for them. This enables high frequency resolution at low frequencies and

high temporal resolution at high frequencies. Publication I shows that using multi-resolution processing improves the perceived quality with critical signals. However, it should be noted that the temporal resolution of 20 ms would be enough for reproducing the binaural properties (see Section 3.2.1), but it is not enough for reproducing monaural properties with certain signals.

A filter bank can be optimized to have similar properties as STFT or the multi-resolution STFT processing. Similar quality of reproduction is expected in this case. For simplicity, only the STFT implementation is discussed in the following sections.

5.1.3 DirAC analysis

The aim of the DirAC analysis is to estimate the DOA and the diffuseness for each frequency band (see Fig. 5.4). The frequency bands are selected in way that they follow the ERB scale (see Eq. 3.2). The estimation is performed in the frequency domain using the B-format signals (see Section 4.3.1). The omnidirectional signal $W(k,n)$ is used for estimating the sound pressure, and the dipole signals $X(k,n)$, $Y(k,n)$, and $Z(k,n)$ form together a vector $\mathbf{V}(k,n) = [X,Y,Z]/\sqrt{2}$ that is used for estimating the relative particle velocity of sound. k is the frequency band index, and n is the temporal frame index. Using these variables, the active intensity $\mathbf{I}_\mathrm{a}$ and the instantaneous energy E_I can be estimated [2]

$$\mathbf{I}_\mathrm{a}(k,n) = -\mathrm{Re}\{W^*(k,n)\mathbf{V}(k,n)\}, \tag{5.1}$$

and

$$E_\mathrm{I}(k,n) = (|W(k,n)|^2 + ||\mathbf{V}(k,n)||^2)/2, \tag{5.2}$$

where $||\cdot||$ denotes the norm of the vector and $*$ complex conjugation (see Section 2 for the corresponding time-domain equations). The active intensity expresses the net flow of sound energy. The DOA vector is defined to point to the opposite direction of the active intensity vector

$$\mathbf{I}_\theta(k,n) = -\mathbf{I}_\mathrm{a}(k,n). \tag{5.3}$$

Correspondingly, the diffuseness ψ is estimated using the ratio between the active intensity and the energy [2, 100]

$$\psi(k,n) = 1 - \frac{||\mathrm{E}\{\mathrm{Re}\{W^*(k,n)\mathbf{V}(k,n)\}\}||}{\mathrm{E}\{|W(k,n)|^2 + ||\mathbf{V}(k,n)||^2\}/2}, \tag{5.4}$$

which is a real-valued number between zero and one, indicating whether the sound field is approximated to consist of direct sound only ($\psi = 0$), a

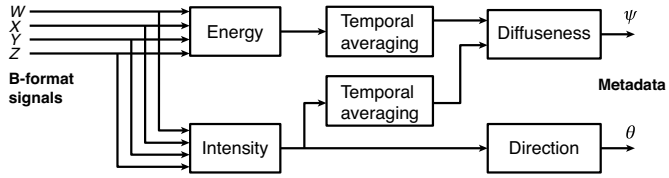

Figure 5.4. DirAC analysis.

diffuse field ($\psi = 1$), or partly direct and partly diffuse sound ($0 < \psi < 1$). $\mathrm{E}\{\}$ denotes the expectation operator, which is, in practice, typically realized with temporal averaging

$$\widehat{S}(k,n) = \alpha_{\mathrm{d}} \cdot S(k,n) + (1-\alpha_{\mathrm{d}})\widehat{S}(k,n-1), \tag{5.5}$$

where S is the variable to be smoothed, $\widehat{S}$ is the smoothed variable, and α_{d} is the smoothing coefficient for the diffuseness. Usually, α_{d} is chosen according to

$$\alpha_{\mathrm{d}} = \frac{L}{2\tau f_{\mathrm{s}}} \tag{5.6}$$

where L is the size of the STFT window, f_{s} is the sampling frequency, and τ is the time constant. A typical value of τ is about 50 ms. Anyhow, it is always selected in a way that the resulting α_{d} is smaller than 1. Other methods for estimating diffuseness can be found in [95, 100].

5.1.4 DirAC synthesis

A block diagram of the DirAC synthesis is presented in Fig. 5.5. The input to the processing is the B-format stream in the frequency domain and the metadata. First, a virtual microphone signal is computed for each loudspeaker direction as a weighted sum of the B-format signals

$$\begin{aligned} M_i(k,n) = 0.25 \cdot W(k,n) + 0.75 \cdot (\cos(\theta_i)\cos(\phi_i)X(k,n) + \\ \sin(\theta_i)\cos(\phi_i)Y(k,n) + \sin(\phi_i)Z(k,n)), \end{aligned} \tag{5.7}$$

where $M_i(k,n)$ is the virtual microphone signal of the loudspeaker channel i with the directional pattern of a hypercardiod [101], θ_i is the angle of the loudspeaker in azimuth direction, and ϕ_i in elevation direction. Also other directional patterns from a cardioid to a dipole have been used. The cardioid is optimal in a sense that all the sound pressure is captured with the same polarity. The dipole is optimal in a sense that the adjacent virtual microphones have the lowest mutual coherence. However, no significant differences in the produced quality have been observed, and the optimal directional pattern is, thus, still under debate. The virtual

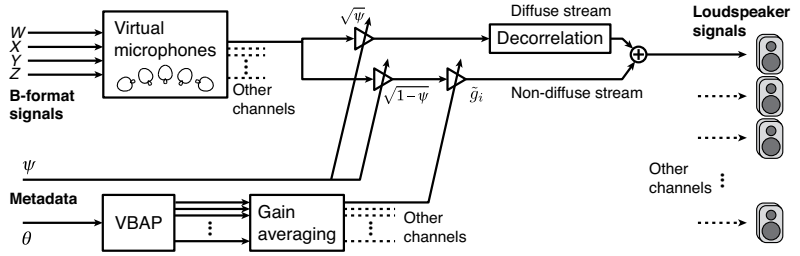

Figure 5.5. DirAC synthesis.

microphone signals are manipulated according to the metadata. DirAC can be interpreted to be an enhanced version of first-order microphone techniques, since the virtual microphones in DirAC are computed similarly as in the traditional systems, such as Ambisonics [59].

The aim of the synthesis stage is to recreate correct directional and IC cues. This is accomplished by dividing the virtual microphone signals into two streams: the nondiffuse stream and the diffuse stream. The nondiffuse stream includes mostly the part of the sound that has a certain direction. It is reproduced using amplitude-panning techniques, creating correct directional cues and high coherence between the ear-canal signals. Correspondingly, the diffuse stream includes mostly the reverberant and the ambient parts. It is reproduced using decorrelation techniques, which create loudspeaker signals that have low coherence between them. When these signals are reproduced, the coherence between the ear-canal signals is low (see Section 3.2.3). The DirAC synthesis is mixing between the nondiffuse and the diffuse streams according to the analyzed diffuseness, and, as a result, correct IC cues are produced. Moreover, the mixing, and also the reproduction of the different streams, is performed in a way that the energy is preserved, and, thus, a correct monaural spectrum is reproduced.

The reproduction of the different streams is now discussed in detail. The nondiffuse stream is reproduced as point sources by multiplying each virtual microphone signal with a loudspeaker-specific gain factor g_i. The gain factors are computed according to the analyzed direction transmitted in the metadata by using VBAP (see Section 4.1.1). This produces the same effect as panning, where only a single audio signal is used as an input, but it is less prone to nonlinear artifacts. The process is explained in more detail in [102]. Also other panning techniques, such as HOA or WFS, could be used, but only VBAP has been used so far.

In many cases, the direction in metadata is subject to abrupt temporal changes [1]. To avoid artifacts, the gain factor $g_i(k,n)$ for the ith loudspeaker computed with VBAP is smoothed by temporal integration and weighted by the energy and the diffuseness

$$\begin{aligned}\widehat{g}_i(k,n) = \alpha_{\mathrm{g}}(k) \cdot g_i(k,n) \cdot E_{\mathrm{I}}(k,n) \cdot \sqrt{1-\psi(k,n)} \\ + (1-\alpha_{\mathrm{g}}(k)) \cdot \widehat{g}_i(k,n-1),\end{aligned} \tag{5.8}$$

where $\widehat{g}_i(k,n)$ is the smoothed gain factor, $g_i(k,n)$ is the non-smoothed gain factor, $E_{\mathrm{I}}(k,n)$ is the instant energy defined in Eq. 5.2, $\psi(k,n)$ is the diffuseness parameter defined in Eq. 5.4, and $\alpha_{\mathrm{g}}(k)$ is a smoothing coefficient

$$\alpha_{\mathrm{g}}(k) = \min\left[\frac{L}{2\tau(k)\cdot f_{\mathrm{s}}}, \alpha_{\max}\right], \tag{5.9}$$

where $\alpha_{\max}$ is the largest allowed value for α_{g}, and $\tau(k)$ is the frequency-dependent time constant

$$\tau(k) = \frac{C_{\mathrm{N}}}{f_{\mathrm{c}}(k)}, \tag{5.10}$$

where f_{c} is the center frequency of the band k and C_{N} is the number of cycle periods, which controls the value of the time coefficient. Typically used values include $C_{\mathrm{N}} = 50$ and $\alpha_{\max} = 0.7$. This effectively removes the artifacts and does not smooth the gain too much. The energy-weighted gain factors $\widehat{g}_i$ have to be normalized after the smoothing before being used for panning as actual gain factors $\tilde{g}_i$ so that the energy of the nondiffuse stream is preserved

$$\tilde{g}_i(k,n) = \frac{\widehat{g}_i(k,n)}{\sqrt{\sum_{i=1}^{N} \widehat{g}_i^2(k,n)}}. \tag{5.11}$$

The diffuse stream is reproduced by decorrelating the virtual microphone signals defined by Eq. 5.7 and reproducing them at the corresponding loudspeakers, as illustrated in Fig. 5.5. The diffuse stream is typically reproduced with equal gains for each loudspeaker. However, in the case of loudspeaker setups with uneven spacing, such as 5.1, it is possible to weight loudspeakers based on the angular density of the loudspeaker positioning (see Publication III). The decorrelation basically scrambles the phase spectrum making the coherence between the signals low. The virtual microphone signals are already incoherent to some degree, so they need to be decorrelated only mildly.

Decorrelation can be performed by using, e.g., noise bursts [1] or frequency-dependent delays which are static with time and different for different loudspeakers [102]. The frequency-dependent-delay method was used in

this thesis. The delays were selected randomly with the following restrictions. Below 1500 Hz, the maximum delay is 50 times the cycle time of the frequency band, and the upper limit is 100 ms. Above 1500 Hz, the maximum delay is always 50 ms. The minimum delay is 10 times the cycle time of the frequency band, with the lower limit of 5 ms. The delays are chosen so that the phase matches at the cross-over frequency where the delay is changed. The decorrelation is implemented for each loudspeaker channel by an FIR filter, which realizes these delays. The filtering is performed to time domain signals after IFFT. Decorrelation can also be implemented on sub-band signals, as depicted in Fig. 5.5. However, it was found that, with the STFT implementation, it is computationally more efficient to perform the decorrelation to time-domain signals.

Finally, the nondiffuse stream and the diffuse stream are summed together, and the resulting signal is reproduced using loudspeakers after the inverse time-frequency transform.

5.1.5 Applications of DirAC

There are many possible scenarios where DirAC can be applied. One of the obvious ones is the high-quality reproduction of recorded sound scenarios [1, 102]. The advantage of DirAC is that the processing is independent of the reproduction system. Arbitrary loudspeaker layouts can be used, and, in addition, headphone reproduction is possible, as shown in Publications IV and V.

Furthermore, DirAC can be used in audio coding. Instead of transmitting/storing all loudspeaker channels, it is possible to transmit/store a 1-N-channel downmix of the B-format signals and the metadata, or only the B-format signals [1]. The bit rate of the metadata required for sufficient quality is low in many applications [103], and the audio signals could be core coded using, for example, AAC [88], although this has not been implemented yet. A method for DirAC processing of legacy multi-channel signals, such as 5.1 surround, is presented in Publication VI. Thus, DirAC could be used as a generic audio format that would accept different kinds of input and output methods.

DirAC has also been suggested to be used for teleconferencing [104]. The directions of the participants can be rendered without a significant increase in the required bit rate compared to mono reproduction, which has been found to increase speech intelligibility [105].

In addition to processing continuous audio signals, also impulse re-

sponses can be processed with DirAC [2]. By convolving an anechoic recording with a DirAC-processed B-format impulse response, a perception of the auditory object as being in the space where the impulse response was measured is created. In this context, the technique is called spatial impulse response rendering (SIRR).

In this thesis, DirAC is extended to be used for spatial-sound synthesis in virtual worlds (see Publications VII and VIII). It is shown that DirAC can be used to position and to control the spatial extent of virtual sound sources with good audio quality. Furthermore, DirAC can be used to generate reverberation for N-channel horizontal listening with only two monophonic reverberators. Recently, also other features have been suggested to the virtual-world DirAC [106, 107].

In addition, the use of DirAC has also been suggested in the following applications: spatial filtering [108], source localization [109], spatial audio effects [110], binaural hearing aids [111], and sound-field speech audiometry [112].

5.1.6 Subjective evaluation

Subjective evaluation of DirAC reproduction has been performed using formal listening tests. Typically, multiple-stimulus tests (e.g., [113, 114]), where the task of the subject is to compare the suggested method to a reference, have been used. A reference can be created using room simulation such as the image-source method [18] (see Section 4.5). The directions of the early reflections in the simulated room are computed and discretized to the nearest loudspeakers of the reproduction system, and the late reverberation is simulated with Gaussian noise following the predicted energy decay curve. The audio signals for the reference scenarios are obtained by convolving monophonic audio signals with the produced impulse responses, and the B-format recordings can be simulated using these reference signals.

The results of the listening tests show that the perceived overall quality of the reproduction with typical sound scenarios is good for both loudspeaker [102] and headphone reproduction (see Publication IV). Furthermore, similar quality has been obtained in impulse-response rendering [115] and virtual-world reproduction (see Publication VII).

As DirAC is a parametric method, the resulting quality is signal dependent. Thus, in the research work leading to this thesis, signals that would be challenging for DirAC processing were sought in order to improve

the processing and to enable good quality with all kinds of signals. A few problematic cases were found, e.g., the case of multiple simultaneous talkers in low-echoic conditions (see Publications II, III, and VII) and that of applause-type signals (see Publication I). This thesis shows that the decorrelation processing used in DirAC increases the perceived spaciousness with certain signals. Alternative methods introduced for these problematic cases show improvement in the perceived quality based on subjective evaluation.

5.1.7 Challenges in DirAC processing

As discussed in Section 5.1.6, the resulting quality of DirAC processing is good with most of the signals, and even in problematic cases the quality can be improved with special processing. This section attempts to summarize the author's current views on challenges in DirAC processing.

Let us consider one time-frequency tile of captured sound. The content of this tile can be roughly divided to originate mainly from: (a) a single sound source, (b) multiple sound sources, or (c) a diffuse reverberant sound field. It should be noted that the multiple sources can consist of actual sources and strong reflections, and that the energy of a single source can be dominant in some time-frequency tiles even in the presence of multiple sources and reverberation. The requirements for perceptually correct rendering of spatial sound were discussed in the beginning of Section 5.1. How well these requirements are fulfilled in these cases is discussed next.

In case (a), amplitude panning creates correct ITD, ILD, and IC values, and the monaural spectrum of the sound is not affected. Thus, all the requirements are met, and the resulting quality is good. In case (c), decorrelation techniques produce low IC values and fluctuating ITD and ILD values, which corresponds well to a diffuse sound field. Furthermore, the decorrelation filter randomizes the phase spectrum. However, as the diffuse field consists of a very large number of reflections, also the original phase spectrum is relatively random. Hence, the phase randomization does not cause perceptual deterioration of quality.

Case (b), multiple sound sources, appears to cause the largest problems in DirAC processing. If the processing is performed with the frequency and the temporal resolution of binaural hearing, the produced ITD, ILD, and IC values should correspond to the original sound field. However, the problem is that reproducing the correct spectrum is difficult, especially the phase spectrum is easily distorted in DirAC processing. Sound originating

from multiple directions is analyzed as relatively diffuse in DirAC. As a result, a large portion of the sound is decorrelated. On the contrary to a diffuse sound field, the original phase spectrum is not necessarily random in the case of multiple sound sources. As a result, the phase randomization can cause perceivable differences.

There are a few possible solutions to this problem. One is to artificially lower the analyzed diffuseness. However, the analyzed direction is rapidly fluctuating, which can cause musical noise and crackling sounds when reproduced with panning. Another solution is to reproduce the diffuse stream without decorrelation. The original phase spectrum is preserved in this case, but unfortunately the produced IC values are too high, which can lead to a perception of coloration and reduced envelopment and spaciousness. Hence, a compromise is needed, and the optimal solution depends on the sound scenario. Publication II suggests a solution optimized for speech signals and Publication I for applause-type signals, which were found to be sensitive to phase modification. A step towards a more general solution is taken in [116], where an auditory model is suggested, which aims to predict when a phase modification causes perceivable differences. A solution from another point of view is suggested in [117], where higher-order microphones are used with DirAC. They allow obtaining more independent signal components, and analyzing and synthesizing multiple concurrent sources at different directions. Thus, the amount of decorrelation can be lower while still producing correct IC cues.

5.2 Binaural cue coding

Binaural cue coding (BCC) [118, 119] is a parametric method for multichannel audio coding. The general block diagram of spatial-sound processing presented in Fig. 5.1 applies to BCC. The input to the processing consists of 2-N-channel loudspeaker signals. As in DirAC, the audio is processed in the time-frequency domain following the frequency and the temporal resolution of human hearing.

The assumptions about the human perception of spatial sound are also similar to those concerning DirAC, i.e., the most important binaural cues are assumed to be ILD, ITD, and IC [118]. The variables translating into these cues are estimated from the loudspeaker signals in BCC, whereas in DirAC they are analyzed from the sound field. The following parameters are estimated between pairs of loudspeakers [118]: inter-channel level

difference (ICLD), inter-channel time difference (ICTD), and inter-channel correlation (ICC). These variables are assumed to translate into ILD, ITD, and IC cues when the loudspeaker signals are reproduced, respectively. The variables are computed for the sub-band signals $x_1(n)$ and $x_2(n)$ (representing frequency band k) of two audio channels with time index n, using the following equations [120]

- ICLD:

$$\Delta\Lambda_{12}(n) = 10\log_{10}\left(\frac{P_{x_2}(n)}{P_{x_1}(n)}\right), \tag{5.12}$$

- ICTD:

$$\Delta T_{12}(n) = \arg\max_d\{\Phi_{12}(d,n)\}, \tag{5.13}$$

- ICC:

$$\Gamma_{12}(n) = \max_d|\Phi_{12}(d,n)|, \tag{5.14}$$

where $P_{x_1}(n)$ and $P_{x_2}(n)$ are short-time estimates of the power of the signals $x_1(n)$ and $x_2(n)$, respectively, and $\Phi_{12}(d,n)$ is a short-time estimate of the normalized cross-correlation function

$$\begin{aligned}\Phi_{12}(d,n) &= \frac{P_{x_1x_2}(d,n)}{\sqrt{P_{x_1}(n-d_1)P_{x_2}(n-d_2)}}\\ d_1 &= \max\{-d,0\}\\ d_2 &= \max\{d,0\},\end{aligned} \tag{5.15}$$

where $P_{x_1x_2}(d,n)$ is a short-time estimate of the mean of $x_1(n-d_1)x_2(n-d_2)$.

The ICLD and the ICTD are computed between a reference channel (e.g., channel number 1) and the other channels, which yields $N-1$ ICLD and ICTD values for each time-frequency tile [119]. The relations between all channels can be synthesized using these values. On the contrary, ICC can have different values between all possible input pairs. If all coherence values are computed and transmitted, $N(N-1)/2$ values are required, which can result in too high bit rate [120]. Hence, ICC is computed only between the channels with most energy in each time-frequency tile [120]. This value is used to describe the overall coherence between all channels.

ICLD, ICTD, and ICC values for each time-frequency tile form the metadata, which can be compressed using quantization and coding techniques. The resulting bit rates are about 2 kbps for ICLDs and ICTDs for one channel pair and about 1.5 kbps for ICC [119]. The metadata is transmitted alongside with a mono downmix of the input channels. Thus, the required

bit rate can be significantly decreased compared to the transmission of N audio channels.

In the synthesis stage, the transmitted mono signal is duplicated into N signals, which are manipulated according to the analyzed ICLD, ICTD, and ICC values. ICLDs are synthesized by multiplying the signals with weighting factors w_i that have ratios equal to ICLDs [118]. ICTDs are synthesized by applying delays equal to ICTDs or by using all-pass filters [120]. ICCs are synthesized by adding randomization to the ICLD values [119] or by using decorrelation filters [120].

The resulting loudspeaker signals should have ICLD, ICTD, and ICC values equal to the original loudspeaker signals. Thus, also the perception should be equal to the original signals. Listening tests have shown that relatively good quality can be obtained, but transparent reproduction is not achieved [119, 120]. However, at low bit rates, the perceived quality is higher with BCC than with conventional single-channel coders [119].

5.3 Parametric stereo

Parametric stereo (PS) [121, 122] is the first employment of spatial audio coding technology in international standards and commercially available codecs [123] such as high-efficiency AAC (HE-AAC). PS is based on principles identical to those of BCC, and the differences between them are mostly found in certain implementation aspects and engineering choices [123]. Thus, PS is only briefly described here, and mostly its differences to BCC are highlighted. However, it should be noted that although the ICLD and the ICTD parameters were first introduced for BCC [89], the ICC parameter was first introduced for PS [90]. All three parameters are nowadays used in both of them.

The general block diagram of spatial-sound processing presented in Fig. 5.1 applies also to PS. The input contains always two channels in the case of PS. The analyzed parameters are otherwise the same as in BCC, but the inter-channel time difference is typically replaced by the inter-channel phase difference (ICPD) [121], which allows parameterizing out-of-phase signals. The input stereo signal is downmixed to a mono signal and transmitted alongside with the analyzed parameters. The synthesis is performed using methods similar to those in BCC. As in the case of BCC, the perceived quality obtained with PS is better than with single-channel coders, according to formal listening tests [121].

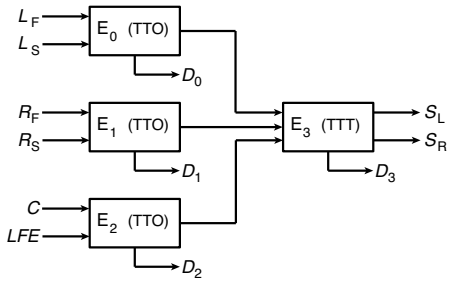

Figure 5.6. Tree configuration for stereo downmix of 5.1 signals using MPEG surround.

5.4 MPEG surround

MPEG surround [124, 125] is a backward-compatible method for parametric coding of multi-channel audio. It is based on the same psychoacoustical principles as BCC and PS (and also DirAC), but the processing is somewhat different. Thus, MPEG surround is discussed in a bit more detail.

The block diagram presented in Fig. 5.1 is applicable also in the case of MPEG surround, and the processing is performed in the time-frequency domain. The transform is typically performed using hybrid quadrature mirror filter banks (hybrid QMF) [92]. The analysis phase is somewhat different compared to BCC and PS. Instead of analyzing all parameters at once, the analysis in MPEG surround is performed using tree structures [124]. The elementary building blocks of the processing are the two-to-one (TTO) and the three-to-two (TTT) blocks that combine two or three channels into one or two channels and metadata. By combining these blocks it is possible to obtain any number of transmitted audio channels K from any number of input channels M.

For example, 5.1-surround signals can be downmixed into stereo by using the structure presented in Fig. 5.6 [124]. The six input signals, left front, right front, center, low-frequency enhancement, left surround, and right surround, are labeled L_f, R_f, C, ***LFE***, L_s, and R_s, respectively. Pairs of signals are first combined into single channels, and finally three channels are combined into two channels, S_L and S_R. In addition, each encoding stage produces metadata: D_0, D_1, D_2, and D_3.

The suggested solution has several advantages. First, the downmix signals are directly reproducible using two-channel systems, even if the decoder does not support MPEG surround. Second, the left and the right loudspeaker channels are separated also in the downmix. Typically, the surround signals contain uncorrelated reverberation aiming at creating

a perception of envelopment. By keeping these signals separated, the incoherence properties between the surround signals are maintained also in the synthesized multi-channel signals.

The content of the TTO and the TTT blocks is discussed next. The TTO encoder element is virtually identical to a PS coder [125]. However, the analyzed parameters are typically limited to ICLD and ICC. In addition, a residual signal can be transmitted. The residual signal represents the error associated with representing the two signals by their downmix and associated parameters and, in principle, enables the reconstruction of full multi-channel waveform in the decoder side [125]. The required bit rate determines whether the residual is transmitted. It can also be transmitted only for a certain frequency range.

TTT encoding can be performed using two alternative approaches: the prediction mode or the energy mode. In the prediction mode, two downmix signals, S_{L} and S_{R}, alongside with an auxiliary signal, S_{C}, are created using matrixing [124]. Using an inverse matrix, the original signals could be created with these signals, but in order to reduce the bit rate S_{C} is discarded and presented, instead, using two channel-prediction coefficients (CPC). The prediction error can be sent as an residual signal the way it is done in the TTO encoder, or it can be described using an ICC parameter in order to reduce the bit rate. Correspondingly, the energy mode describes the relations of the input signals simply by using two ICLD parameters.

The transmitted signal in MPEG surround contains K audio channels and metadata, which can contain the parameters that were referred to previously and/or residual signals. These signals are further compressed using the core coder. The synthesis phase is, conceptually, performed in a tree structure similarly to the encoding phase but in inverse order. In practice, the decoding is performed in a 'flattened' way in order to increase computational efficiency and to minimize decorrelation artifacts [125]. However, for simplicity, let us assume the decoding taking place in the one-to-two (OTT) and the two-to-three (TTT) blocks. The block diagram of the OTT decoding block is presented in Fig. 5.7. The stereo output is obtained by mixing the input signal with the decorrelated version of it based on the ICLD and the ICC parameters [125]. The residual signal is used, instead of the decorrelated signal, if it is available. The decoding in the TTT block is performed in a similar way.

Several listening tests have been arranged to evaluate the perceptual quality of MPEG surround (e.g., [124, 125]). The results of the listening

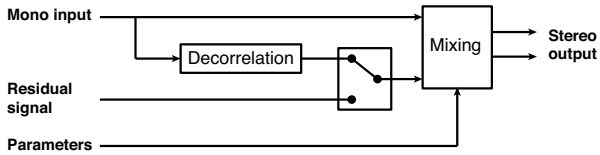

Figure 5.7. OTT decoding module in MPEG surround.

tests show that good audio quality can be obtained with bit rates as low as 64 kbps and that the quality is significantly better than the quality obtained with the coding of the audio channels separately [125]. At higher bit rates (e.g., 160 kbps), the perceived quality is excellent, although not transparent [125]. At these bit rates, the perceived quality is similar to channel-based coding, although slightly higher. At very high bit rates (e.g., 320 kbps), channel-based coding is assumed to enable the best quality [126]. Inspecting the results of each test sample individually, it can be seen that the perceived quality depends on the signal [124]. As in the case of DirAC, certain signals, such as the applause-type ones, appear to be more difficult for the coder. Reasons for this are suggested in Publication I.

In addition to coding, the use of MPEG surround has been suggested in other applications. Spatial audio object coding (SAOC) [127] can be used to interactively manipulate multi-channel signals, for example, by controlling the directions and the levels of individual audio objects, which can be useful in music remixing and teleconferencing. Furthermore, using microphone signals as an input to MPEG surround has been suggested in [128].

5.5 Spatial audio scene coding

Spatial audio scene coding (SASC) [129, 130] is a method for multi-channel audio coding and upmixing. In the previously mentioned methods (BCC, PS, and MPEG surround), the parameters are analyzed between channel pairs, whereas in SASC (and also in DirAC) universal cues about inter-channel relationships are analyzed. Thus, also upmixing can be performed flexibly. The general structure of the processing follows the block diagram of Fig. 5.1 also with SASC.

The analysis stage begins with a primary-ambient decomposition [130], which is based on principal component analysis (PCA). The multi-channel audio signals $S_i(k, n)$ are assumed to consist of a common primary compo-

nent $Q(k,n)$ and independent ambient components $A_i(k,n)$

$$S_i(k,n) = w_i(k,n)Q(k,n) + A_i(k,n), \tag{5.16}$$

where $w_i(k,n)$ is the weight of the primary component for the loudspeaker channel i. The analyzed primary and ambient components are processed differently. The next phase is to compute directions for both components. The directional estimation is based on Gerzon vectors [131]. The direction of the primary component is obtained using a Gerzon localization vector

$$\mathbf{g}_{\mathrm{P}} = \frac{\sum_{i=1}^{M} |w_i(k,n)|\mathbf{q}_i}{\sum_{i=1}^{M} |w_i(k,n)|}, \tag{5.17}$$

and the direction of the ambient component using a Gerzon energy vector

$$\mathbf{g}_{\mathrm{A}} = \frac{\sum_{i=1}^{M} |A_i(k,n)|^2\mathbf{q}_i}{\sum_{i=1}^{M} |A_i(k,n)|^2}, \tag{5.18}$$

where $\mathbf{q}_i$ is a unit vector pointing to loudspeaker i. The directions and the primary-ambient weights are transmitted alongside with a downmix of the input signals. In the synthesis phase, the primary component is reproduced using, for example, amplitude panning [129]. The ambient component is reproduced as a combined result of decorrelation and panning techniques, depending on the length of the ambient localization vector.

5.6 Comparison of the methods

As discussed in the previous sections, BCC, PS, DirAC, MPEG surround, and SASC have many similarities. They all operate in the time-frequency domain and aim at recreating correct ITD, ILD, and IC cues. Furthermore, they all divide the sound into two streams: the 'directional/primary' and the 'diffuse/incoherent/ambient' stream. Thus, it is assumed that most of the methods suggested for DirAC processing in this thesis could also be applied in other similar parametric techniques. BCC, PS, MPEG surround, and SASC primarily operate with loudspeaker signals, whereas DirAC primarily operates with recorded sound fields, but these roles have also been mixed by introducing loudspeaker-signal processing for DirAC (see Publication VI) and microphone processing for MPEG surround [128]. Moreover, even a method for combining DirAC and MPEG surround has been suggested [132].

One might ask that if these methods are similar, which one is the best. There is no straightforward answer to this. DirAC and SASC are better in

a way that the universal cues are not bound to any reproduction system and that the amount of the required parameter values is smaller, especially for systems with a large number of loudspeakers. On the other hand, BCC, PS, and MPEG surround can be assumed to enable better quality in some cases since there is more information available. A significant factor to the resulting quality is also the exact implementation and the use of other tools, such as the guided envelope shaping (GES) tool in MPEG surround [125]. Furthermore, many of these techniques appear to have problems with the same kinds of signals, for example with applause-type signals (see [133] and Publication I).

5.7 Other related parametric techniques

The previously mentioned methods are probably the most widely known parametric methods for spatial-audio processing. However, there are also many other interesting approaches. A few of them are now briefly presented.

In [134], ICLD, ICTD, and ICC parameters are estimated as in BCC, PS, and MPEG surround, but the analysis and the synthesis are implemented differently. Spatial squeezing surround audio coding (S^3AC) [135] is based on transmitting directional information about multi-channel audio as level differences in a stereo signal. A method for modifying the directional responses of a coincident pair of microphones by parametric processing is suggested in [136]. In high angular resolution planewave expansion (HARPEX) [137], the sound field is presented using two plane waves. The method presented in [138] analyzes the DOA using a tetrahedral microphone array. MPEG unified speech and audio coding (USAC) [139] combines MPEG surround with efficient single-channel coding of audio and speech. The method presented in [140] estimates the DOAs of active sound sources and, using beamforming, performs source separation.

5.8 Binaural versions

The description of the methods presented in the previous sections assumed loudspeaker reproduction. However, also binaural versions of most of them are available [141, 142, 143]. Binaural reproduction for DirAC is suggested in Publication IV.

6. Summary of publications

This section summarizes the publications in this thesis.

Publication I: "Reproducing applause-type signals with directional audio coding"

Applause-type signals are known to be challenging for parametric multi-channel coding and spatial-audio reproduction. Publication I investigates this phenomenon in the context of directional audio coding (DirAC). It is suggested that the main reason for these artifacts is that the temporal resolution within the processing is too coarse. More specifically, applause signals typically contain either multiple transients from random directions within one analysis window or transients accompanied with other sounds or reverberation. In these cases, the diffuseness value computed in the DirAC analysis is too high, which causes temporal artifacts in the reproduction, as the transients are smeared temporally by the decorrelation processing applied to the diffuse stream in the DirAC synthesis. By using carefully tuned temporal resolution at all frequencies, the individual transients are analyzed as nondiffuse components and the artifacts are significantly less audible.

This article proposes a multi-resolution STFT implementation of DirAC for optimizing temporal resolution: input signals are first divided into several frequency regions, and each region is processed with STFT using different window lengths. This modification results in a significant audio-quality improvement compared to a DirAC implementation using a frequency-independent window length. However, it is found that even though most of the transients are analyzed as nondiffuse with multi-resolution STFT some of them remain in the diffuse stream, decreasing the perceived sharpness of claps. This can be prevented by processing the

transients in the diffuse stream separately. This modification provides an additional quality improvement in the reproduction. Formal listening tests confirm the improvement, due to the suggested methods, in the perceived quality.

Publication II: "Utilizing instantaneous direct-to-reverberant ratio in parametric spatial audio coding"

Scenarios with multiple simultaneous sources in an acoustically dry room may be challenging for parametric spatial sound reproduction techniques such as DirAC. It is found in Publication II that, especially in the case of speech signals, the processing causes a perception of added reverberation. It is suggested that decorrelation used in DirAC processing causes the added-room effect. The effect can be mitigated by turning off the decorrelator, and it is found that good overall quality is perceived with dry signals with this approach. However, in the case of reverberant signals, lack of decorrelation causes other perceivable artifacts, such as differences in the perception of room and timbre. Based on these results, a new model for DirAC reproduction, reverberation-extraction DirAC (RE-DirAC), is suggested.

In conventional DirAC processing, the sound is divided into nondiffuse and diffuse streams. In RE-DirAC, the diffuse stream is further divided into reverberant and non-reverberant parts. Decorrelation is applied for the reverberant part, whereas the non-reverberant part is reproduced without decorrelation. The division into reverberant and non-reverberant parts is performed using instantaneous direct-to-reverberant ratio, which can be estimated with the help of blind dereverberation techniques. The results of formal listening tests show that perceptually good audio quality can be obtained using this approach for both dry and reverberant scenarios (see Fig. 6.1). In addition, the quality is better than or as good as that of the traditional DirAC method, either with (DcOn) or without (DcOff) decorrelation in all cases.

Publication III: "Parametric spatial audio coding for spaced microphone array recordings"

Spaced-microphone arrays are often used for multi-channel recording of music performances. The coherence between the microphone channels is

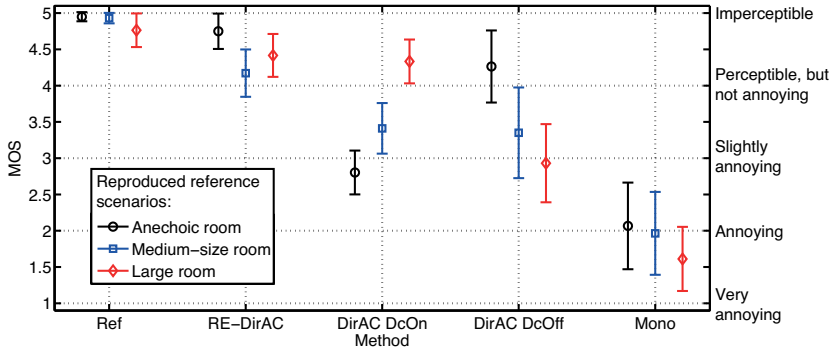

Figure 6.1. Perceived impairment in the listening test in Publication II. Means and 95% confidence intervals are shown.

low in a reverberant field due to microphone spacing, which translates into a perception of a pleasant 'enveloping' sound when reproduced with a multichannel system, at the expense of accurate localization of sound sources. Publication III presents a parametric method to process spaced-microphone recordings. The method, which is based on the principles of DirAC, uses the knowledge of the array configuration and the frequency-dependent microphone patterns. Directional analysis combined with panning techniques is used to improve the perceived localization without harming the pleasant enveloping qualities of spaced recordings. The results of formal listening tests show that, compared to traditional methods, the suggested method improves the perceived overall quality.

In addition, it is shown that the quality of conventional DirAC processing can be improved by using spaced microphones as an input. Publications I and II showed that the decorrelation processing decreases the perceived quality with certain signals. Due to low inter-channel coherence, the amount of required decorrelation is lower in the case of spaced microphones. The results of formal listening tests confirm that the perceived quality is improved.

Furthermore, the suggested method extends DirAC processing from coincident to spaced recordings, making DirAC more versatile as a scheme for spatial-sound processing.

Publication IV: "Binaural reproduction for directional audio coding"

The original implementations of DirAC used only loudspeakers as an output. Publication IV introduces binaural reproduction for DirAC utilizing

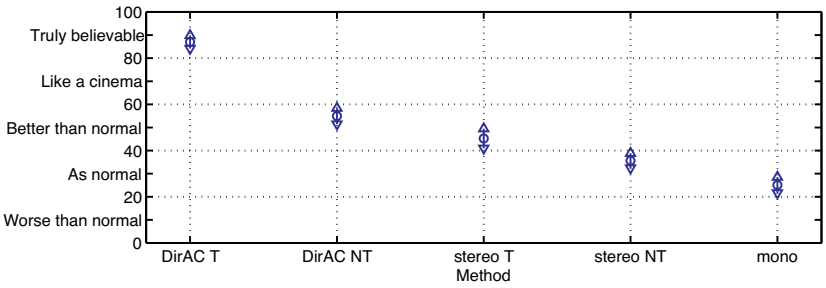

Figure 6.2. Spatial impression of reproduction in the listening test in Publication IV. Means and 95% confidence intervals are shown. T = tracking on, NT = tracking off.

head-related transfer functions (HRTF) and head tracking. In practice, real loudspeakers of the conventional DirAC method are replaced by virtual loudspeakers implemented with HRTFs. In addition, the metadata and the virtual microphones are modified dynamically according to the head-tracking data.

The results of formal listening tests show that 'Excellent' overall quality and 'Truly believable' spatial impression can be obtained with the suggested method (see Fig. 6.2). Many listeners commented informally after the tests that the externalization worked so well that they did not know whether the sound was coming from the loudspeakers present in the room or from the headphones. The improvement compared to traditional stereo reproduction is clear. In addition, the use of head tracking with DirAC significantly increases the perceived quality.

Publication V: "Influence of resolution of head tracking in synthesis of binaural audio"

In Publication IV, the use of head tracking in binaural synthesis of spatial sound was found to increase the quality of reproduction. The required quality of a head-tracking system for this purpose is studied in Publication V. A listening test was performed to evaluate the effect of four common sources of error in head-tracking systems, namely, degrees of freedom in listener orientation, angle restriction, tracking stability, and decreasing the update rate. Using the binaural version of DirAC, B-format recordings of natural sound events were reproduced and played back over headphones with head tracking. The listeners rated the naturalness of binaural reproduction in different head-tracking conditions.

Based on the results, it is clear that azimuth tracking is the most important degree of freedom to implement. The other directions have only little importance. Furthermore, restriction of tracking angle seemed to cause perceptible degradation with all tested parameter values. Thus, it is suggested that head tracking should not be restricted into a small area if perceptual naturalness is desired. The results of the additive-random-bias test showed that some of the listeners were relatively insensitive to the tracking instability. However, since the confidence intervals of the means were large in the results of this test, further tests are required to assess the accepted amount of bias. In addition, it was found that lower update rate of the tracking system affects the quality.

Publication VI: "Converting 5.1 audio recordings to B-format for directional audio coding reproduction"

Publication VI extends DirAC processing from the microphone signals to legacy multi-channel signals such as 5.1 surround. A method to transform 5-channel surround sound signals to B-format is proposed, which provides unaltered spatial qualities when reproduced with DirAC. The proposed method simulates anechoic B-format recordings of the signals with two different virtual loudspeaker configurations, which are combined based on time-frequency analysis of the diffuseness of the virtual sound field. The resulting B-format signals are further modified based on the diffuseness of these signals.

The suggested solution provides both accurate localization and the perception of correct spaciousness and envelopment. In addition, the resulting B-format signals can be mixed with other B-format signals that can be real recordings, virtual signals, or other converted signals. Based on informal listening tests, the B-format conversion was not found to degrade the audio quality compared to the original signals when the resulting B-format signals were reproduced with DirAC.

Publication VII: "Parametric time-frequency representation of spatial sound in virtual worlds"

Publication VII extends DirAC processing to virtual-world spatial audio. In virtual worlds, tasks for a spatial-audio renderer include synthesizing the direction and the spatial extent of sound sources, as well as creating

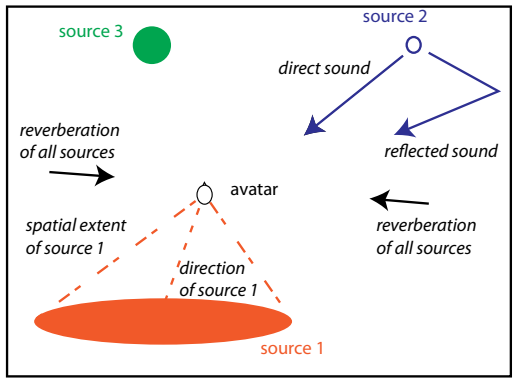

Figure 6.3. A virtual world with three sound sources. The user should auditorily perceive the sources in the directions as they are relative to the avatar, the spatial extents of the sources, the reflections from nearby surfaces, and also the surrounding reverberation generated by all sources.

a realistic perception of reflections and reverberation (see Fig. 6.3). In the proposed method, a DirAC-monosynth block is introduced for carrying out these tasks. The input to the block is the desired direction and the extent of the source, and the output is a mono DirAC stream containing one audio signal alongside with metadata. Each source is processed with a separate DirAC-monosynth block. Different mono DirAC streams can be efficiently merged by creating corresponding B-format streams and summing them. The resulting signals can be reproduced with DirAC using arbitrary loudspeaker setups or headphones.

The directions of the sources are synthesized by controlling the DOA values in the metadata. The spatial extent is synthesized without increased computational complexity by distributing different frequency bands to different directions and controlling the diffuseness parameter. Furthermore, reverberation can be efficiently created for arbitrary horizontal loudspeaker layouts by using only two single-channel reverberators and applying the result to the dipole signals of the intermediate B-format signals. The perceptual quality of the suggested methods was studied with formal listening tests, and it was found that the quality produced was good in direct comparison with reference cases.

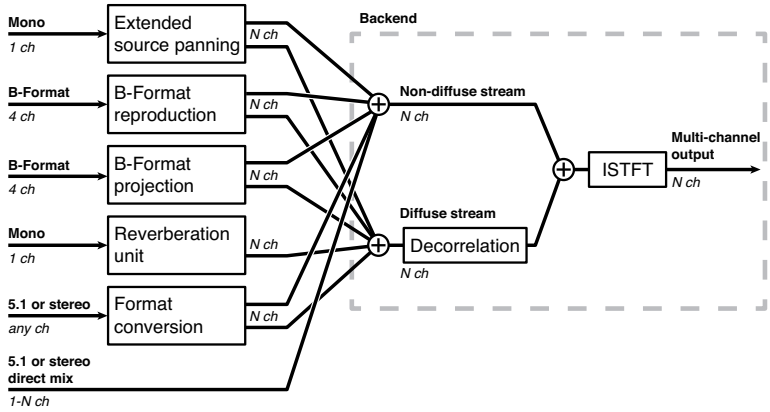

Figure 6.4. Modular architecture proposed in Publication VIII. Different frontends and the unified backend are shown. The lines signify different multi-channel signals. The number of channels is marked next to each signal path.

Publication VIII: "Modular architecture for virtual-world parametric spatial audio synthesis"

Publication VII suggested a method for DirAC processing in virtual-world applications. Although the resulting quality is very good in most cases, it was noticed that some auditory scenes, for example the ones containing multiple sources in acoustically dry conditions, are not produced optimally. Publication VIII presents a restructured version of the virtual-world DirAC to avoid these problems.

It is shown in the article that these kinds of scenarios are analyzed as being relatively diffuse and that the decorrelation processing applied to the signals causes a perception of added spaciousness. It is suggested that this can be avoided by introducing a new modular structure which keeps sources separated in the processing and thus avoids the problems in the earlier algorithm. However, a full DirAC processing of the different sources separately would increase the computational complexity significantly. To maintain relatively low computational complexity, the most complex tasks should be done as few times as possible. In the suggested solution, the different sources are combined just before decorrelation and inverse time-frequency transform. This reduces computational complexity while maintaining the desired property of multiple separate directions for each time-frequency tile.

The block diagram of the suggested modular design is presented in Fig. 6.4. The processing is divided into separate frontend modules and a

unified backend. Different sound sources can be spatialized using separate extended-source-panning blocks. In addition, by using corresponding frontends, for example B-format signals and legacy audio formats can be added as separate components. Moreover, new modules can be easily added and the existing modules can be modified using, for example, the methods suggested in this thesis, as long as they are consistent with the used time-frequency transform. Thus, the suggested new architecture enables very versatile processing of spatial sound.

7. Conclusions

Directional audio coding (DirAC) is a method for spatial-sound reproduction. It operates in the time-frequency domain and aims to analyze the perceptually significant properties of the sound field. The analyzed parameters, namely the direction of arrival and the diffuseness, are used for manipulating the microphone signals in a way that the perception of the reproduced sound field is equal to the original sound field. Subjective evaluations have shown that DirAC improves the perceived quality compared to traditional methods. However, DirAC was originally introduced for relatively limited use cases. This thesis has presented methods to generalize the DirAC approach for more versatile use. The generalization was performed for three aspects: challenging spatial-sound scenarios, output systems, and input systems.

As DirAC is a parametric method, the resulting quality is signal dependent. Thus, challenging signals for DirAC processing were sought in order to improve the processing and to enable good quality with all kinds of signals. A few problematic cases were found, e.g., multiple simultaneous talkers in low-echoic conditions and applause-type signals. It was shown in this thesis that decorrelation processing, which basically scrambles the phase spectrum, increases the perceived spaciousness and smears transients with certain signals. However, it was also shown that without decorrelation, many signals, especially the reverberant ones, are perceived to be lacking spaciousness and envelopment, and also timbral errors are perceived, due to too high inter-aural coherence.

Thus, it is suggested in this thesis that the amount of decorrelation should be minimized but in a way that the problems associated with too high coherence are avoided. A few possible approaches were presented in this thesis. In Publication I, temporal resolution was increased by introducing multi-resolution STFT processing to DirAC. The suggested

solution decreases the analyzed diffuseness, and thus, the amount of decorrelation. Especially the amount of transients and onsets present in the diffuse stream is decreased. Moreover, a transient-detection algorithm was introduced in order to further reduce the decorrelation of the transients. The lack of decorrelation for transient-like components was not found to cause a perception of too high coherence. In Publication II, the diffuse stream was divided into reverberant and non-reverberant parts based on reverberant-energy estimation. Only the reverberant part was decorrelated. The suggested solution was found to provide excellent quality in both acoustically dry and reverberant scenarios. In Publication III, spaced microphones were used as an input to the processing. The coherence between the microphones is relatively low in a diffuse field. Hence, the decorrelation is not required for the most of the frequency range. In Publication VIII, different sound sources were processed separately, thus avoiding the increase in the analyzed diffuseness due to multiple simultaneous sources. With real recordings this is not possible due to the limited directionality of the microphones, but in virtual worlds the different sources can be processed separately. In addition, the suggested solution is computationally efficient.

DirAC originally used loudspeakers as an output. As an addition to possible reproduction devices, a method for headphone reproduction was presented in Publication IV. The method is based on binaural techniques and head tracking, and subjective evaluations showed that 'Excellent' overall quality and 'Truly believable' spatial impression can be obtained with the suggested method. Furthermore, it was shown that the effect of head tracking is very significant to both the overall quality and the spatial impression. As the head tracking was found important, the required quality of the head-tracking system for this purpose was studied in Publication V. It was shown that the main requirement for the head-tracking system is to obtain an unrestricted azimuth angle with high enough update rate. In addition, the accuracy of the tracking was found to affect the quality.

DirAC was originally developed for capturing real sound scenes with a B-format microphone. As that is a relatively narrow scope, this thesis extended DirAC processing to different input sources. Using spaced-microphone recordings with DirAC was suggested in Publication III. Compared to traditional spaced-microphone techniques, the presented method was shown to improve the perceived quality by offering improved and stable localization cues. This was achieved by applying panning based on the time-frequency analysis of the sound field. As discussed earlier, the method

also improves the quality compared to conventional DirAC processing due to the reduced amount of decorrelation. Publication VI extended DirAC processing from the microphone signals to legacy multi-channel signals such as 5.1 surround. The proposed method analyses the virtual sound field generated by 5.1 audio content in the time-frequency domain, and B-format signals are created based on the analysis. The method provides unaltered spatial qualities when the resulting B-format signals are reproduced with DirAC. In addition, the resulting B-format signals can be mixed with other B-format signals originating from various sources. Publications VII and VIII extended DirAC processing to virtual-world spatial audio. Formal listening tests were used to show that DirAC can be used to position and to control the spatial extent of virtual sound sources with good audio quality. It was also shown that DirAC can be used to generate reverberation for N-channel horizontal listening with only two monophonic reverberators without prominent loss in quality when compared to the quality obtained with N-channel reverberators.

The outcome of this thesis is a versatile spatial-sound processing scheme that enables excellent quality with many kinds of signals originating from various inputs and being reproduced with various outputs. In addition, it was illustrated in this thesis that DirAC processing is based on principles of human perception resembling those in many other spatial-sound reproduction and coding techniques. Thus, even though this work was performed in the context of DirAC, it is suggested that the methods developed in this thesis could be applied in a variety of parametric spatial-sound processing techniques.

Bibliography

[1] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, vol. 55, pp. 503–516, June 2007.

[2] J. Merimaa and V. Pulkki, "Spatial impulse response rendering I: Analysis and synthesis," *J. Audio Eng. Soc.*, vol. 53, pp. 1115–1127, December 2005.

[3] M. Barron, *Auditorium Acoustics and Architectural Design*. E & FN Spon, London, 1993.

[4] M. Karjalainen and H. Järveläinen, "More about this reverberation science: Perceptually good late reverberation," in *AES 111th Convention*, (New York, NY, USA), September 2001.

[5] J.-M. Jot, L. Cerveau, and O. Warusfel, "Analysis and synthesis of room reverberation based on a statistical time-frequency model," in *AES 103rd Convention*, (New York, NY, USA), September 1997.

[6] E. A. Lehmann and A. M. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 1429–1439, August 2010.

[7] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating interactive virtual acoustic environments," *J. Audio Eng. Soc.*, vol. 47, pp. 675–705, September 1999.

[8] H. Nélisse and J. Nicolas, "Characterization of a diffuse field in a reverberant room," *The Journal of the Acoustical Society of America*, vol. 101, pp. 3517–3524, June 1997.

[9] W. C. Sabine, *Collected Papers on Acoustics*. Cambridge: Harvard U.P., 1922.

[10] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Am.*, vol. 47, pp. 409–412, 1965.

[11] K. Lebart, J.-M. Boucher, and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica United with Acustica*, vol. 87, pp. 359–366, 2001.

[12] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien, Jr., C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *J. Acoust. Soc. Am.*, vol. 114, pp. 2877–2892, November 2003.

[13] P. Kendrick, F. F. Li, T. J. Cox, Y. Zhang, and J. A. Chambers, "Blind estimation of reverberation parameters for non-diffuse rooms," *Acta Acustica United with Acustica*, vol. 93, pp. 760–770, 2007.

[14] F. J. Fahy, *Sound Intensity*. Elsevier Science Publishers Ltd, 1989.

[15] G. Schiffrer and D. Stanzial, "Energetic properties of acoustic fields," *J. Acoust. Soc. Am.*, vol. 96, pp. 3645–3653, December 1994.

[16] L. E. Kinsler and A. R. Frey, *Fundamentals of Acoustics*. John Wiley and Sons, Inc., 1950.

[17] F. Jacobsen, "The diffuse sound field," Tech. Rep. 27, Technical University of Denmark, 1979.

[18] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustic," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, April 1979.

[19] S. Tervo, *Localization and tracing of early acoustic reflections*. PhD thesis, Aalto University, Espoo, Finland, 2012.

[20] B. C. J. Moore, ed., *Hearing*. Academic Press, 1995.

[21] J. L. Schiano, C. Trahiotis, and L. R. Bernstein, "Laterazation of low-frequency tones and narrow bands of noise," *J. Acoust. Soc. Am.*, vol. 79, pp. 1563–1570, May 1986.

[22] J. Blauert, *Spatial Hearing*. The MIT Press, 1997.

[23] Lord Rayleigh, "On our perception of sound direction," *Philosophical Magazine*, vol. 13, pp. 214–232, 1907.

[24] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. Academic Press, 1982.

[25] J. Blauert, ed., *Communications Acoustics*. Springer, 2005.

[26] F. L. Wightman, "The pattern-transformation model of pitch," *J. Acoust. Soc. Am.*, vol. 54, no. 2, pp. 407–416, 1973.

[27] J. C. R. Licklider, "A duplex theory of pitch perception," *Experientia*, vol. 7, no. 4, pp. 128–134, 1951.

[28] H. Fletcher, "Auditory patterns," *Rev. Mod. Phys.*, vol. 12, pp. 47–65, January 1940.

[29] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Am.*, vol. 74, pp. 750–753, September 1983.

[30] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, pp. 103–138, August 1990.

[31] A. Kohlrausch, "Auditory filter shape derived from binaural masking experiments," *J. Acoust. Soc. Am.*, vol. 84, pp. 573–583, August 1988.

[32] M. van der Heijden and C. Trahiotis, "Binaural detection as a function of interaural correlation and bandwidth of masking noise: Implications for estimates of spectral resolution," *J. Acoust. Soc. Am.*, vol. 103, pp. 1609–1614, March 1998.

[33] I. Holube, M. Kinkel, and B. Kollmeier, "Binaural and monaural auditory filter bandwidths and time constants in probe tone detection experiments," *J. Acoust. Soc. Am.*, vol. 104, pp. 2412–2425, October 1998.

[34] D. M. Green, "Temporal acuity as a function of frequency," *J. Acoust. Soc. Am.*, vol. 54, no. 2, pp. 373–379, 1973.

[35] P. X. Joris, L. H. Carney, P. H. Smith, and T. C. Yin, "Enhancement of neural synchronization in the anteroventral cochlear nucleus. I. Responses to tones at the characteristic frequency," *J. Neurophysiol.*, vol. 71, no. 3, pp. 1022–1036, 1994.

[36] D. W. Grantham and F. L. Wightman, "Detectability of varying interaural temporal differences," *J. Acoust. Soc. Am.*, vol. 63, no. 2, pp. 511–523, 1978.

[37] L. R. Bernstein, C. Trahiotis, M. A. Akeroyd, and K. Hartung, "Sensitivity to brief changes of interaural time and interaural intensity," *J. Acoust. Soc. Am.*, vol. 109, no. 4, pp. 1604–1615, 2001.

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

[39] L. A. Jeffress, "A place theory of sound localization," *Journal of Comparative Physiology and Psychology*, vol. 41, no. 1, pp. 35–39, 1948.

[40] M. J. Goupell and W. M. Hartmann, "Interaural fluctuations and the detection of interaural incoherence. III. Narrowband experiments and binaural models," *J. Acoust. Soc. Am.*, vol. 122, pp. 1029–1045, August 2007.

[41] D. Griesinger, "The psychoacoustics of apparent source width, spaciousness and envelopment in performance spaces," *Acta Acustica United with Acustica*, vol. 83, pp. 721–731, 1997.

[42] K. Hiyama, S. Komiyama, and K. Hamasaki, "The minimum number of loudspeakers and its arrangement for reproducing the spatial impression of diffuse sound field," in *AES 113th Convention*, (Los Angeles, CA, USA), October 2002.

[43] S. H. Nielsen, "Auditory distance perception in different room," *J. Audio Eng. Soc.*, vol. 41, pp. 755–770, October 1993.

[44] A. D. Blumlein, "British patent specification 394,325, 1931," *J. Audio Eng. Soc.*, vol. 6, no. 2, 1958.

[45] B. B. Bauer, "Phasor analysis of some stereophonic phenomena," *J. Acoust. Soc. Am.*, vol. 33, pp. 1536–1539, November 1961.

[46] B. Bernfeld, "Attempts for better understanding of the directional stereophonic listening mechanism," in *AES 44th Convention*, (Rotterdam, The Netherlands), February 1973.

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

[48] V. Pulkki and M. Karjalainen, "Localization of amplitude-panned virtual sources I: Stereophonic panning," *J. Audio Eng. Soc.*, vol. 49, pp. 739–752, September 2001.

[49] V. Pulkki, “Virtual source positioning using vector base amplitude panning,” *J. Audio Eng. Soc.*, vol. 45, pp. 456–466, June 1997.

[50] V. Pulkki and M. Karjalainen, “Localization of amplitude-panned virtual sources II: Two- and three-dimensional panning,” *J. Audio Eng. Soc.*, vol. 49, pp. 753–767, September 2001.

[51] S. P. Lipshitz, “Stereo microphone techniques... are the purists wrong?,” *J. Audio Eng. Soc.*, vol. 34, pp. 716–744, September 1986.

[52] V. Pulkki, M. Karjalainen, and J. Huopaniemi, “Analyzing virtual sound source attributes using a binaural auditory model,” *J. Audio Eng. Soc.*, vol. 47, pp. 203–217, April 1999.

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

[54] V. Pulkki, “Microphone techniques and directional quality of sound reproduction,” in *AES 112th Convention*, (Munich, Germany), May 2002.

[55] G. Theile, “Natural 5.1 music recording based on psychoacoustic principles,” in *AES 19th International Conference*, (Schloss Elmau, Germany), June 2001.

[56] R. Kassier, H.-K. Lee, T. Brookes, and F. Rumsey, “An informal comparison between surround-sound microphone techniques,” in *AES 118th Convention*, (Barcelona, Spain), May 2005.

[57] M. A. Poletti, “Three-dimensional surround sound systems based on spherical harmonics,” *J. Audio Eng. Soc.*, vol. 53, pp. 1004–1025, November 2005.

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

[59] M. A. Gerzon, “Ambisonic in multichannel broadcasting and video,” *J. Audio Eng. Soc.*, vol. 33, pp. 859–871, November 1985.

[60] J. Daniel and S. Moreau, “Further study of sound field coding with higher order ambisonics,” in *AES 116th Convention*, (Berlin, Germany), May 2004.

[61] S. Moreau, J. Daniel, and S. Bertet, “3D sound field recording with higher order ambisonics – objective measurements and validation of spherical microphone,” in *AES 120th Convention*, (Paris, France), May 2006.

[62] “Soundfield ST450 portable microphone system.” Soundfield, `http://www.soundfield.com/products/st450.php`, May 2013.

[63] “the Eigenmike microphone array.” mh acoustics, `http://www.mhacoustics.com/mh_acoustics/Eigenmike_microphone_array.html`, May 2013.

[64] A. Solvang, “Spectral impairment for two-dimensional higher order ambisonics,” *J. Audio Eng. Soc.*, vol. 56, pp. 267–279, April 2008.

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

[66] S. Spors, R. Rabenstein, and J. Ahrens, “The theory of wave field synthesis revisited,” in *AES 124th Convention*, (Amsterdam, The Netherlands), May 2008.

[67] S. Zielinski, F. Rumsey, and S. Bech, "On some biases encountered in modern audio quality listening tests – a review," *J. Audio Eng. Soc.*, vol. 56, pp. 427–451, June 2008.

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

[69] E. A. G. Shaw, "Earcanal pressure generated by a free sound field," *J. Acoust. Soc. Am.*, vol. 39, no. 3, pp. 465–470, 1966.

[70] H. Møller, "Fundamentals of binaural technology," *Applied Acoustics*, vol. 36, no. 3-4, pp. 171–218, 1992.

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

[72] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *AES 108th Convention*, (Paris, France), February 2000.

[73] S. Müller and P. Massarani, "Transfer-function measurement with sweeps," *J. Audio Eng. Soc.*, vol. 49, pp. 443–471, June 2001.

[74] V. Pulkki, M.-V. Laitinen, and V. P. Sivonen, "HRTF measurements with a continuously moving loudspeaker and swept sines," in *AES 128th Convention*, (London, UK), 2010.

[75] M. Hiipakka, T. Kinnari, and V. Pulkki, "Estimating head-related transfer functions of human subjects from pressure-velocity measurements," *J. Acoust. Soc. Am.*, vol. 131, pp. 4051–4061, May 2012.

[76] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, (New Paltz, NY, USA), 2001.

[77] J. Gòmez Bolaños and V. Pulkki, "HRIR database with measured actual source direction data," in *AES 133rd Convention*, (San Francisco, CA, USA), 2012.

[78] O. Kirkeby, P. A. Nelson, and H. Hamada, "The "stereo dipole" – a virtual source imaging system using two closely spaced loudspeakers," *J. Audio Eng. Soc.*, vol. 46, pp. 387–295, May 1998.

[79] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen, "Transfer characteristics of headphones measured on human ears," *J. Audio Eng. Soc.*, vol. 43, pp. 203–217, January 1995.

[80] F. Christensen, H. Møller, P. Minnaar, J. Plogsties, and S. K. Olesen, "Interpolating between head-related transfer functions measured with low directional resolution," in *AES 107th Convention*, (New York, NY, USA), September 1999.

[81] F. L. Wightman and D. J. Kistler, "The importance of head movements for localizing virtual auditory display objects," in *2nd International Conference on Auditory Display*, (Santa Fe, NM, USA), 1994.

[82] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head-tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *J. Audio Eng. Soc.*, vol. 49, pp. 904–916, October 2001.

[83] S. Yairi, Y. Iwaya, and Y. Suzuki, "Influence of large system latency of virtual auditory display on behavior of head movement in sound localization task," *Acta Acustica United with Acustica*, vol. 94, pp. 1016–1023, 2008.

[84] T. Lentz, D. Schröder, M. Vorländer, and I. Assenmacher, "Virtual reality system with integrated sound field simulation and reproduction," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 187–187, 2007.

[85] L. Peltola, C. Erkut, P. R. Cook, and V. Välimäki, "Synthesis of hand clapping sounds," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1021–1029, March 2007.

[86] N. Tsingos, E. Gallo, and G. Drettakis, "Perceptual audio rendering of complex virtual environments," in *SIGGRAPH 31st International Conference on Computer Graphics and Interactive Techniques*, (Los Angeles, CA, USA), August 2004.

[87] C. Verron, M. Aramaki, R. Kronland-Martinet, and G. Pallone, "A 3-D immersive synthesizer for environmental sounds," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 1550–1561, August 2010.

[88] K. Brandenburg, "MP3 and AAC explained," in *AES 17th International Conference*, (Florence, Italy), August 1999.

[89] C. Faller and F. Baumgarte, "Efficient representation of spatial audio using perceptual parametrization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, (New Paltz, NY, USA), October 2001.

[90] E. Schuijers, W. Oomen, B. den Brinker, and J. Breebaart, "Advances in parametric coding for high-quality audio," in *AES 114th Convention*, (Amsterdam, The Netherlands), March 2003.

[91] J. Merimaa and V. Pulkki, "Perceptually-based processing of directional room responses for multichannel loudspeaker reproduction," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, (New Paltz, NY, USA), October 2003.

[92] J. Herre, C. Faller, S. Disch, C. Ertel, J. Hilpert, A. Hoelzer, K. Linzmeier, C. Spenger, and P. Kroon, "Spatial audio coding: Next-generation efficient and compatible coding of multi-channel audio," in *AES 117th Convention*, (San Francisco, CA, USA), October 2004.

[93] M. M. Goodwin and J.-M. Jot, "A frequency-domain framework for spatial audio coding based on universal spatial cues," in *AES 120th Convention*, (Paris, France), May 2006.

[94] J. Ahonen, "Microphone configurations for teleconference application of directional audio coding and subjective evaluation," in *AES 40th International Conference*, (Tokyo, Japan), October 2010.

[95] O. Thiergart, M. Kratschmer, M. Kallinger, and G. Del Galdo, "Parameter estimation in directional audio coding using linear microphone arrays," in *AES 130th Convention*, (London, UK), May 2011.

[96] A. Politis and V. Pulkki, "Broadband analysis and synthesis for directional audio coding using A-format input signals," in *AES 131st Convention*, (New York, NY, USA), October 2011.

[97] V. Pulkki and C. Faller, "Directional audio coding: Filterbank and STFT-based design," in *AES 120th Convention*, (Paris, France), May 2006.

[98] E. Vickers, "Frequency-domain implementation of time-varying FIR filters," in *AES 133rd Convention*, (San Francisco, CA, USA), October 2012.

[99] J. I. Marín-Hurtado and D. V. Anderson, "FFT-based block processing in speech enhancement: Potential artifacts and solutions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 2527–2537, November 2011.

[100] G. Del Galdo, M. Taseska, O. Thiergart, J. Ahonen, and V. Pulkki, "The diffuse sound field in energetic analysis," *J. Acoust. Soc. Am.*, vol. 131, pp. 2141–2151, March 2012.

[101] J. Eargle, *The Microphone Book*. Focal Press, 2001.

[102] J. Vilkamo, T. Lokki, and V. Pulkki, "Directional audio coding: Virtual microphone based synthesis and subjective evaluation," *J. Audio Eng. Soc.*, vol. 57, p. 709:724, September 2009.

[103] T. Hirvonen, J. Ahonen, and V. Pulkki, "Perceptual compression methods for metadata in directional audio coding applied to audiovisual teleconference," in *AES 126th Convention*, (Munich, Germany), May 2009.

[104] J. Ahonen, V. Pulkki, and T. Lokki, "Teleconference application and B-format microphone array for directional audio coding," in *AES 30th International Conference*, (Saariselkä, Finland), March 2007.

[105] J. Ahonen and V. Pulkki, "Speech intelligibility in teleconference application of directional audio coding," in *AES 40th International Conference*, (Tokyo, Japan), October 2010.

[106] T. Pihlajamäki and V. Pulkki, "Projecting simulated or recorded spatial sound onto 3D-surfaces," in *AES 45th International Conference*, (Helsinki, Finland), March 2012.

[107] T. Pihlajamäki and M.-V. Laitinen, "Plausible mono-to-surround sound synthesis in virtual-world parametric spatial audio," in *AES 49th International Conference*, (London, UK), February 2013.

[108] M. Kallinger, H. Ochsenfeld, G. Del Galdo, F. Kuech, D. Mahne, R. Schultz-Amling, and O. Thiergart, "A spatial filtering approach for directional audio coding," in *AES 126th Convention*, (Munich, Germany), May 2009.

[109] O. Thiergart, R. Schultz-Amling, G. Del Galdo, D. Mahne, and F. Kuech, "Localization of sound sources in reverberant environments based on directional audio coding parameters," in *AES 127th Convention*, (New Paltz, NY, USA), October 2009.

[110] A. Politis, T. Pihlajamäki, and V. Pulkki, "Parametric spatial audio effects," in *15th Int. Conference on Digital Audio Effects*, (York, UK), September 2012.

[111] J. Ahonen, V. Sivonen, and V. Pulkki, "Parametric spatial sound processing applied to bilateral hearing aids," in *AES 45th International Conference*, (Helsinki, Finland), March 2012.

[112] T. Koski, V. Sivonen, and V. Pulkki, "Measuring speech intelligibility in noisy environments reproduced with parametric spatial audio," in *AES 135th Convention*, (New York, NY, USA), October 2013.

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

[114] ITU-R BS.1534-1, *Method for the subjective assessment of intermediate quality level of coding systems*. ITU, 2003.

[115] V. Pulkki and J. Merimaa, "Spatial impulse response rendering II: Reproduction of diffuse sound and listening tests," *J. Audio Eng. Soc.*, vol. 54, pp. 3–20, January/February 2006.

[116] M.-V. Laitinen, S. Disch, and V. Pulkki, "Sensitivity of human hearing to changes in phase spectrum," *J. Audio Eng. Soc.*, vol. 61, pp. 860–877, November 2013.

[117] V. Pulkki, A. Politis, G. Del Galdo, and A. Kuntz, "Parametric spatial audio reproduction with higher-order B-format microphone input," in *AES 134th Convention*, (Rome, Italy), May 2013.

[118] F. Baumgarte and C. Faller, "Binaural cue coding—part I: Psychoacoustic fundamentals and design principles," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 509–519, November 2003.

[119] C. Faller and F. Baumgarte, "Binaural cue coding—part II: Schemes and applications," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 520–531, November 2003.

[120] C. Faller, "Parametric multichannel audio coding: Synthesis of coherence cues," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 299–310, January 2006.

[121] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegård, "Low complexity parametric stereo coding," in *AES 116th Convention*, (Berlin, Germany), May 2004.

[122] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "Parametric coding of stereo audio," *EURASIP Journal on Applied Signal Processing*, pp. 1305–1322, 2005.

[123] J. Breebaart and C. Faller, *Spatial audio processing — MPEG surround and other applications*. John Wiley and Sons, Inc., 2007.

[124] J. Breebaart, G. Hotho, J. Koppens, E. Schuijers, W. Oomen, and S. van de Par, "Background, concept, and architecture for the recent MPEG surround standard on multichannel audio compression," *J. Audio Eng. Soc.*, vol. 55, pp. 331–351, May 2007.

[125] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Rödén, W. Oomen, K. Linzmeier, and K. S. Chong, "MPEG surround—the ISO/MPEG standard for efficient and compatible multichannel audio coding," *J. Audio Eng. Soc.*, vol. 56, pp. 932–955, November 2008.

[126] J. Rödén, J. Breebaart, J. Hilpert, H. Purnhagen, E. Schuijers, J. Koppens, K. Linzmeier, and A. Hölzer, "A study of the MPEG surround quality versus bit-rate curve," in *AES 123rd Convention*, (New York, NY, USA), October 2007.

[127] J. Herre, H. Purnhagen, J. Koppens, O. Hellmuth, J. Engdegård, J. Hilpert, L. Villemoes, L. Terentiv, C. Falch, A. Hölzer, M. L. Valero, B. Resch, H. Mundt, and H.-O. Oh, "MPEG spatial audio object coding—the ISO/MPEG standard for efficient coding of interactive audio scenes," *J. Audio Eng. Soc.*, vol. 60, pp. 655–673, September 2012.

[128] C. Tournery, C. Faller, F. Kuech, and J. Herre, "Converting stereo microphone signals directly to MPEG-surround," in *AES 128th Convention*, (London, UK), May 2010.

[129] J.-M. Jot, J. Merimaa, M. M. Goodwin, A. Krishnaswamy, and J. Laroche, "Spatial audio scene coding in a universal two-channel 3-D stereo format," in *AES 123rd Convention*, (New York, NY, USA), October 2007.

[130] M. M. Goodwin and J.-M. Jot, "Spatial audio scene coding," in *AES 125th Convention*, (San Francisco, CA, USA), October 2008.

[131] M. A. Gerzon, "General metatheory of auditory localization," in *AES 92nd Convention*, (Vienna, Austria), March 1992.

[132] J. Herre, C. Falch, D. Mahne, G. Del Galdo, M. Kallinger, and O. Thiergart, "Interactive teleconferencing combining spatial audio object coding and DirAC technology," *J. Audio Eng. Soc.*, vol. 59, pp. 924–935, December 2011.

[133] G. Hotho, S. van de Par, and J. Breebaart, "Multichannel coding of applause signals," *EURASIP Journal on Advances in Signal Processing*, 2008.

[134] A. Seefeldt, M. S. Vinton, and C. Q. Robinson, "New techniques in spatial audio coding," in *AES 119th Convention*, (New York, NY, USA), October 2005.

[135] B. Cheng, C. Ritz, and I. Burnett, "Principles and analysis of the squeezing approach to low bit rate spatial audio coding," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, (Honolulu, HI, USA), 2007.

[136] C. Faller, "Modifying the directional responses of a coincident pair of microphones by postprocessing," *J. Audio Eng. Soc.*, vol. 56, pp. 810–822, October 2008.

[137] S. Berge and N. Barrett, "High angular resolution planewave expansion," in *2nd International Symposium on Ambisonics and Spherical Acoustics*, (Paris, France), May 2010.

[138] M. Cobos, J. J. Lopez, and S. Spors, "A sparsity-based approach to 3D binaural sound synthesis using time-frequency array processing," *EURASIP Journal on Advances in Signal Processing*, 2010.

[139] M. Neuendorf, M. Multrus, N. Rettelbach, G. Fuchs, J. Robilliard, J. Lecomte, S. Wilde, S. Bayer, S. Disch, C. Helmrich, R. Lefebvre, P. Gournay, B. Bessette, J. Lapierre, K. Kjörling, H. Purnhagen, L. Villemoes, W. Oomen, E. Schuijers, K. Kikuiri, T. Chinen, T. Norimatsu, C. K. Seng, E. Oh, M. Kim, S. Quackenbush, and B. Grill, "MPEG unified speech and audio coding – the ISO/MPEG standard for high-efficiency audio coding of all content types," in *AES 132nd Convention*, (Budapest, Hungary), April 2012.

[140] A. Alexandridis, A. Griffin, and A. Mouchtaris, "Capturing and reproducing spatial audio based on a circular microphone array," *Journal of Electrical and Computer Engineering*, 2013.

[141] M. Noisternig, A. Sontacchi, T. Musil, and R. Höldrich, "A 3D ambisonic based binaural sound reproduction system," in *AES 24th International Conference*, (Banff, Canada), June 2003.

[142] J. Breebaart, J. Herre, L. Villemoes, C. Jin, K. Kjörling, J. Plogsties, and J. Koppens, "Multi-channel goes mobile: MPEG surround binaural rendering," in *AES 29th International Conference*, (Seoul, Korea), September 2006.

[143] M. M. Goodwin and J.-M. Jot, "Binaural 3-D audio rendering based on spatial audio scene coding," in *AES 123rd Convention*, (New York, NY, USA), October 2007.

Errata

Publication I

Equation $\mathbf{I} = (1/\sqrt{2})\mathrm{Re}(W^*\mathbf{V}')$ should be replaced by $\mathbf{I} = -(1/\sqrt{2})\mathrm{Re}(W^*\mathbf{V}')$.

Eq. 10, and the clause preceding it, should be replaced by: The energy-weighted gain factors $\widehat{g}_i$ have to be normalized after the smoothing before being used for panning as actual gain factors $\tilde{g}_i$ so that the energy of the nondiffuse stream is preserved

$$\tilde{g}_i(k,n) = \frac{\widehat{g}_i(k,n)}{\sqrt{\sum_{i=1}^{N} \widehat{g}_i^2(k,n)}}.$$

Publication IV

Equation $\mathbf{I} = (1/\sqrt{2})\mathrm{Re}(W)^*\mathbf{V}'$ should be replaced by $\mathbf{I} = -(1/\sqrt{2})\mathrm{Re}(W^*\mathbf{V}')$.

Publication V

Equation $\mathbf{I} = (1/\sqrt{2})\mathrm{Re}(W^*\mathbf{V}')$ should be replaced by $\mathbf{I} = -(1/\sqrt{2})\mathrm{Re}(W^*\mathbf{V}')$.

Publication VI

Equation $\mathbf{I} = (1/\sqrt{2})\mathrm{Re}(W^*\mathbf{V}')$ should be replaced by $\mathbf{I} = -(1/\sqrt{2})\mathrm{Re}(W^*\mathbf{V}')$.

Publication VII

Eq. 1 should be replaced by $\mathbf{I}(k,n) = -(1/\sqrt{2})\mathrm{Re}\{W(k,n)^*\mathbf{U}(k,n)\}$.

Aalto-DD 5/2014

ISBN 978-952-60-5528-2
ISBN 978-952-60-5529-9 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

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

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS