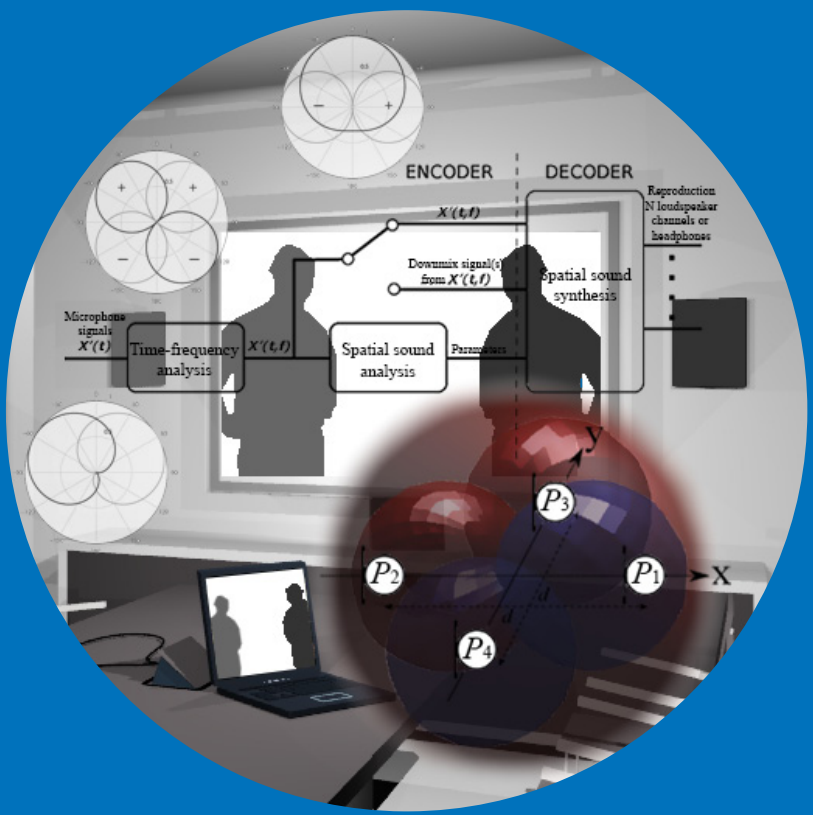
Department of Signal Processing and Acoustics

Microphone front-ends for spatial sound analysis and synthesis with Directional Audio Coding

Jukka Ahonen



Aalto University

DOCTORAL
DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 33/2013

Microphone front-ends for spatial sound analysis and synthesis with Directional Audio Coding

Jukka Ahonen

Doctoral dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the School of Electrical Engineering for public examination and debate in Auditorium S1 at the Aalto University School of Electrical Engineering (Espoo, Finland) on the 8th of March 2013, 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
Professor John Mourjopoulos, University of Patras, Greece

Professor Steven van de Par, Carl-von-Ossietzky University, Germany

Opponent
Doctor Christof Faller, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Aalto University publication series
DOCTORAL DISSERTATIONS 33/2013

ISBN 978-952-60-5035-5 (printed)
ISBN 978-952-60-5036-2 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5036-2

Unigrafia Oy
Helsinki 2013

Finland

441 697
Printed matter

Aalto University

Abstract

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

Author
Jukka Ahonen

Name of the doctoral dissertation
Microphone front-ends for spatial sound analysis and synthesis with Directional Audio Coding

Publisher School of Electrical Engineering

Unit Department of Signal Processing and Acoustics

Series Aalto University publication series DOCTORAL DISSERTATIONS 33/2013

Field of research Acoustics and Audio Signal Processing

Manuscript submitted 17 September 2012 **Date of the defence** 8 March 2013

Permission to publish granted (date) 14 December 2012 **Language** English

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

Abstract

A large number of professional and domestic audio applications utilize spatial sound reproduction. In addition to the conventional applications, such as the surround sound in movie and home theaters, spatial sound is also applied for telecommunication purposes. For instance in teleconferencing, sound emanated by talkers can be captured with multiple microphones at one end and reproduced spatially distributed with multiple loudspeakers at the other. This has benefit over a typical monophonic reproduction of the teleconference in terms of speech intelligibility and other elements of communication.

During the last decade there has been an increasing research interest in parametric spatial sound processing. Several techniques for estimating the directional parameters of a sound field from multichannel audio files or from microphone signals have been proposed. In the parametric techniques, the directional information can be efficiently transmitted and then applied to spatial sound synthesis for various purposes.

This thesis discusses Directional Audio Coding (DirAC) for capturing, transmitting and reproducing spatial sound. The perceptually motivated time-frequency processing of DirAC provides a parametric description of spatial sound, namely the arrival direction and diffuseness of sound. Direction and diffuseness, when analyzed in the time-frequency resolution of human hearing, are assumed to transmit enough information on the captured sound field for spatial hearing. DirAC has several applications of spatial audio, of which teleconferencing is mainly the focus here.

The author's research addresses the development of different microphone front-ends for DirAC. The methods to analyze a sound field with input from arrays of omnidirectional microphones and from typical directional stereo microphones were studied. A novel method for diffuseness estimation was developed as a part of this work. Microphone arrays, which exploit an acoustic shadowing between microphones, are also proposed as an acoustical front-end for DirAC, as are the methods to conduct directional analysis with such arrays. These methods overcome the issues, which occur in direction analysis with input from the conventional microphone arrays, and thus provide reliable direction estimate over the entire audio frequency range. In the thesis, DirAC processing is also applied to bilaterally-fitted hearing aids with two microphones at each ear. The use of different microphone front-ends is evaluated through measurements and listening tests.

Keywords spatial audio, microphone arrays, multichannel reproduction, teleconferencing

ISBN (printed) 978-952-60-5035-5 **ISBN (pdf)** 978-952-60-5036-2

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

Location of publisher Espoo **Location of printing** Helsinki **Year** 2013

Pages 132 **urn** http://urn.fi/URN:ISBN:978-952-60-5036-2

A! Aalto-yliopisto

Tiivistelmä

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

Tekijä
Jukka Ahonen

Väitöskirjan nimi
Mikrofonitekniikat tilaäänen analyysiin ja synteesiin Directional Audio Coding-menetelmällä

Julkaisija Sähkötekniikan korkeakoulu

Yksikkö Signaalinkäsittelyn ja akustiikan laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 33/2013

Tutkimusala Akustiikka ja äänenkäsittelytekniikka

Käsikirjoituksen pvm 17.09.2012 **Väitöspäivä** 08.03.2013

Julkaisuluvan myöntämispäivä 14.12.2012 **Kieli** Englanti

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

Tiivistelmä

Useat ammatti- ja kotikäyttöön tarkoitetut audiosovellukset hyödyntävät tilaäänentoistoa. Tavanomaisia sovelluksia ovat esimerkiksi elokuva- ja kotiteatterit. Näiden lisäksi myös telekommunikaatiosovelluksissa ääni voidaan toistaa siten, että sen tilaominaisuudet säilyvät kuulijalle. Esimerkiksi telekonferenssisovelluksessa osanottajien puhe voidaan tallentaa usealla mikrofonilla lähetyspäässä ja toistaa usealla kaiuttimella vastaanottopäässä, jolloin äänilähteet välittyvät kuulijalle eri suunnista. Tämä parantaa muun muassa puheenymmärrettävyyttä verrattuna perinteisesti käytettyyn yksikanavaiseen äänentoistoon.

Parametrisia tilaäänen prosessointimenetelmiä on tutkittu laajalti viimeisten vuosikymmenten aikana. Menetelmissä mikrofonisignaaleista tai monikanavaäänitiedostosta analysoidaan suuntatietoa äänestä. Tämä suuntatieto voidaan tallentaa ja siirtää tehokkaasti ja hyödyntää tilaäänen synteesissä eri käyttötarkoituksia varten.

Tässä väitöskirjatyössä käsitellään Directional Audio Coding (DirAC) -menetelmää, joka on tarkoitettu tilaäänen äänittämiseen, siirtoon ja toistoon. Menetelmä perustuu äänisignaalien aika-taajuusprosessointiin, jossa on huomioitu ihmisen kuulon ominaisuudet tilaäänen havaitsemiselle. Menetelmässä äänikentän parametrinen esitys koostuu äänen analysoidusta tulosuunnasta ja diffuusisuudesta. Näiden parametrien katsotaan välittävän riittävästi suunta- ja tilainformaatiota äänikentästä ihmisen suuntakuulolle, kun prosessointi suoritetaan vastaavalla aika-taajuusresoluutiolla, jolla ihmisen kuulo käsittelee ääntä. DirAC-menetelmää voidaan käyttää useissa erilaisissa audiosovelluksissa, joista tässä väitöskirjatyössä käsitellään pääasiallisesti telekonferenssisovellusta.

Väitöskirjatyössä on kehitetty useita eri mikrofonitekniikoita DirAC-menetelmälle. Työssä on tutkittu äänikentän suunta-analyysia painemikrofonihilan sekä tyypillisten stereo-suuntamikrofonien signaaleista. Osana työtä on kehitetty uusi laskentamenetelmä äänen diffuusisuusanalyysille. Työssä käsitellään myös mikrofonihiloja, joissa mikrofonien välille muodostuu akustista varjostusta. Akustista varjostusta hyödynnetään tässä työssä kehitetyissä äänen suunta-analyysimenetelmissä. Kehitetyt menetelmät ratkaisevat pääosin ongelmia, joita esiintyy perinteisillä mikrofonihiloilla, sekä mahdollistavat näin ollen luotettavan äänen tulosuunta-analyysin koko kuultavalle taajuuskaistalle. Lisäksi DirAC-menetelmää on sovellettu kuulokojeilla väitöskirjatyössä. Eri mikrofonitekniikoiden käyttöä on arvioitu mittauksin sekä kuuntelukokein.

Avainsanat tilaääni, mikrofonitekniikat, monikanavainen äänentoisto, telekonferenssi

ISBN (painettu) 978-952-60-5035-5 **ISBN (pdf)** 978-952-60-5036-2

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

Julkaisupaikka Espoo **Painopaikka** Helsinki **Vuosi** 2013

Sivumäärä 132 **urn** http://urn.fi/URN:ISBN:978-952-60-5036-2

Preface

This work has been carried out at the Department of Signal Processing and Acoustics, Aalto University, Finland, during 2006-2012, on a joint project "Enhanced Directional Audio Coding (EDAC)" with Fraunhofer Institute for Integrated Circuits IIS. During the period from August to December 2009, the work was conducted at Fraunhofer IIS, Germany.

First of all, I am thankful to Prof. Ville Pulkki, my thesis supervisor and instructor, for the opportunity to conduct this work in his spatial sound research group and for the interesting research topic. Also, I am grateful for Prof. Pulkki's excellent guidance throughout this work. Moreover, I would like to express gratitude to my former supervisor Prof. Matti Karjalainen, who passed away in May 2010, for his encouraging example and guidance to all young researchers. I also thank the pre-examiners of the thesis, Prof. John Mourjopoulos and Prof. Steven van de Par, for their comments and feedback on the manuscript of the thesis.

At the acoustics lab, I have had the privilege to work with great people. My special thanks go to the present and former co-workers of the spatial sound research group: Mikko-Ville Laitinen, Tapani Pihlajamäki, Juha Vilkamo, Marko Takanen, Olli Santala, Marko Hiipakka, Javier Gomez Bolanos, Archontis Politis, Symeon Delikaris-Manias, Teemu Koski, Olli Rummukainen, Dr. Miikka Tikander, Dr. Ville Sivonen, Dr. Toni Hirvonen, and Dr. Juha Merimaa. I also thank the rest of the personnel (present and former) at the acoustics lab. I would especially like to thank secretary Heidi Koponen, ex-secretary Lea Söderman, ex-laboratory manager Martti Rahkila, and Jussi Hynninen for their help in practical and technically related issues.

I am also thankful to my colleagues in the EDAC team at Fraunhofer IIS: Dr. Fabian Kuech, Dr. Markus Kallinger, Dr. Giovanni Del Galdo, Dr. Achim Kuntz, Oliver Thiergart, and Richard Schultz-Amling. For me, it

has been a pleasure to collaborate with such great researchers, and first of all with such great people. I also thank the personnel of my current employer, Akukon Ltd., for their encouragement and support to finalize this thesis.

This work has been supported financially by Fraunhofer IIS, the Academy of Finland (project No. 105780), Tekniikan Edistämissäätiö, Walter Ahlström foundation, and by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 240453. I wish to thank all financial supporters.

Last, but not least, I am grateful to my family for their love and support. My parents sparked my interest in music, which had a major impact on my studies in acoustics and furthermore on this thesis. Finally, I thank my wife Sanna Ahonen for her love and support and my children Alex and Emma, the brightest stars in my life.

Espoo, February 7, 2013,

Jukka Ahonen

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 Jukka Ahonen and Ville Pulkki. Speech Intelligibility in Teleconference Application of Directional Audio Coding. In *AES 40th International Conference on Spatial Audio*, Tokyo, Japan, October 2010.

II Giovanni Del Galdo, Maja Taseska, Oliver Thiergart, Jukka Ahonen and Ville Pulkki. The Diffuse Sound Field in Energetic Analysis. *Journal of the Acoustical Society of America (JASA)*, **131**(3), pp. 2141-2151, March 2012.

III Jukka Ahonen. Microphone Configurations for Teleconference Application of Directional Audio Coding and Subjective Evaluation. In *AES 40th International Conference on Spatial Audio*, Tokyo, Japan, October 2010.

IV Jukka Ahonen, Giovanni Del Galdo, Fabian Kuech, and Ville Pulkki. Directional Analysis with Microphone Array Mounted on Rigid Cylinder for Directional Audio Coding. *Journal of the Audio Engineering Society (JAES)*, Vol. 60, No. 5, pp. 311-324, May 2012.

V Jukka Ahonen and Ville Pulkki. Broadband Direction Estimation Method utilizing Combined Pressure and Energy Gradients from Optimized Microphone Array. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 2011.

VI Jukka Ahonen, Ville Sivonen and Ville Pulkki. Parametric Spatial Sound Processing Applied to Bilateral Hearing Aids. In *AES 45th Conference on Applications of Time-Frequency Processing*, Helsinki, Finland, March 2012.

Author's Contribution

Publication I: "Speech Intelligibility in Teleconference Application of Directional Audio Coding"

DirAC teleconferencing with 1-D and 2-D arrays of two and four omnidirectional microphones, respectively, is presented in this paper. Two alternative methods to estimate the arrival direction of sound using the 1-D array are proposed. One of the methods was invented completely and the other partly by the present author. The paper also presents a listening test to measure speech intelligibility in DirAC teleconferencing. The author conducted the listening test and wrote the paper.

Publication II: "The Diffuse Sound Field in Energetic Analysis"

This paper presents three different methods for computing diffuseness, which are also evaluated under various conditions. The present author is involved with the coefficient of variation method by developing and formulating it together with the last author of the paper. Additionally, the present author suggested some of the evaluation methods.

Publication III: "Microphone Configurations for Teleconference Application of Directional Audio Coding and Subjective Evaluation"

The paper introduces various microphone configurations for DirAC processing and presents a listening test to assess audio quality in DirAC teleconferencing. The present author contributed to developing DirAC analysis and synthesis methods applied to stereo pairs of directional microphones, to conducting the listening test, and to writing the paper.

Publication IV: "Directional Analysis with Microphone Array Mounted on Rigid Cylinder for Directional Audio Coding"

The paper presents a 2-D microphone array on the surface of a rigid cylinder casting an acoustic shadow of an arriving sound. The idea for such a cylinder microphone array was suggested by the fourth author. The present author contributed together with the second and fourth authors to developing two alternative methods for broadband direction estimation of sound with the cylinder array. Additionally, the present author conducted the listening test to evaluate the cylinder microphone array in DirAC and wrote the paper.

Publication V: "Broadband Direction Estimation Method utilizing Combined Pressure and Energy Gradients from Optimized Microphone Array"

The general idea of the proposed method for the direction estimation is the same as in Publication IV and it came from the second author. The omnidirectional microphones with enough large housing to cast an acoustic shadow are used to estimate the direction of sound. The method for optimizing the microphone array for the direction estimation is proposed by the present author, who also conducted the evaluations of the methods and wrote the paper.

Publication VI: "Parametric Spatial Sound Processing Applied to Bilateral Hearing Aids"

A variation of the DirAC processing applied to bilaterally-fitted hearing aids is presented and evaluated in the paper. The present author contributed to modifying the DirAC processing for the hearing aid microphone signals, to performing objective evaluations, and also to writing the initial draft of the paper. Listening tests to measure the speech reception threshold (SRT) with DirAC applied to hearing aids were designed by all authors. The present and the second authors arranged the test.

List of Abbreviations

1-D	one-dimensional
2-D	two-dimensional
3-D	three-dimensional
BCC	binaural cue coding
BMLD	binaural masking level difference
BTE	behind-the-ear hearing aid
DirAC	directional audio coding
ERB	equivalent rectangular bandwidth
HRTF	head-related transfer function
IC	interaural coherence
ICC	inter-channel coherence
ICLD	inter-channel level difference
ILD	interaural level difference
ITD	interaural time difference
JND	just noticeable difference
MS	mid-side
NOS	Nederlandse omroep stichting
ORTF	office de radiodiffusion television Francaise
PEG	pressure-energy gradient
SAC	spatial audio coder
SIRR	spatial impulse response rendering
SRT	speech reception threshold
STFT	short-time Fourier transform
VBAP	vector base amplitude panning
VQ	vector quantization

List of Symbols

β	directivity parameter
θ	azimuth angle
ϕ	elevation angle
ψ	diffuseness
c	speed of sound
d	distance between microphones
f_c	center frequency
f_sa	spatial aliasing frequency
p	sound pressure as a function of time
$\boldsymbol{u}$	particle velocity vector as a function of time
t	time
v	first-order virtual microphone signal
w	omnidirectional signal of B-format
x, y, z	dipole signals of B-format

1. Introduction

Spatial hearing [1] provides efficiently information of our surroundings. Physical objects producing audible sound can be perceived from all directions, also at the rear, which obviously is not true with vision. Based either on the comparison of the signals at left and right ears (binaural listening) or on the signal at one ear only (monaural listening), the human auditory system produces the space perception [2]. This includes, among others, the perception of the direction of a sound source and also the sensation of the surrounding space.

Different audio applications serve the human spatial hearing by creating auditory spatial attributes to the reproduced sound. For instance, in public and home movie theaters with a multichannel loudspeaker system, the video picture is reproduced with spatial audio providing spatial attributes for a listener, such as discrete auditory events and surrounding ambient sound. At best, the reproduced sound creates an immersive auditory environment where a listener perceives to be in some other environment than in the listening space.

As discussed in this thesis, the auditory spatial attributes can be created to the reproduced sound with parametric spatial-sound processing methods, where the directional parameters of the sound field are estimated from multichannel audio signals. Typically only the features of the sound field that are assumed to be relevant for human spatial hearing are estimated and transmitted to a listener. That is, a complete reconstruction of the sound field is not aimed. Consequently, the parametric methods provide computationally efficient processing. In addition to a high-quality reproduction of spatial audio, the parametric methods are utilized in real-time telecommunication applications. For instance, in teleconferencing with spatial audio the auditory events are reproduced at directions that coincide with the talkers at the remote location and with the repro-

duced video picture. One benefit of the reproduction of spatial audio in teleconferencing is the increase in speech intelligibility when compared to monophonic reproduction common in current systems. Additionally, plausible spatial audio reproduction may at its best create a sensation of being present at the remote location (telepresence).

1.1 Scope of the Thesis

The primary topic of this thesis is the development of microphone front-ends for spatial sound analysis and synthesis using a parametric method, Directional Audio Coding (DirAC), mainly for telecommunication applications. The front-ends include arrays of multiple low-cost miniature microphones and also conventional stereo microphones in typical audio devices. A version of the DirAC processing is also developed for real-time teleconferencing purposes in the thesis work.

Additionally, the thesis presents a method for broadband direction estimation from microphone signals. This means to overcome the so-called spatial-aliasing issue, which typically occurs with microphone arrays and impairs directional analysis in DirAC. A microphone array, which introduces acoustic shadowing and scattering between microphones, and a novel directional analysis method utilizing these effects are proposed and evaluated to provide reliable direction estimation over a broad audio frequency range. Besides the telecommunication applications, a variation of the DirAC processing with the novel directional analysis method is applied to bilaterally-fitted hearing aids and is also presented in this thesis.

1.2 Organization of the Thesis

This thesis consists of the introduction and six publications presenting the author's research. The introduction is organized as follows: Section 2 gives an overview of the aspects of the spatial hearing, which relate to a parametric spatial sound processing. Conventional microphone techniques for capturing spatial sound are discussed in Section 3. Parametric spatial-sound processing techniques with microphone inputs, including DirAC, are presented in Section 4. Section 5 presents the contribution of this work on microphone front-ends in DirAC. Conclusions and the main

results are drawn in Section 6.

2. Spatial hearing

Parametric spatial-sound processing techniques discussed in this thesis are motivated by the human hearing. This section shortly reviews the frequency resolution of the ear, the principles of the sound source localization, and also spatial hearing in complex acoustic environments.

2.1 Frequency resolution of hearing

The audible frequency range, a.k.a the audio frequency range, is roughly from 20 Hz to 20 kHz and varies between individuals. For instance, hearing impairments, which typically occur with age, reduce this range. The frequency resolution of the human ear depends on the spectral properties of sound. Our capability to perceive changes in frequency is at best 3 Hz with single tones measured with the just noticeable differences (JND) method [2]. However, this frequency resolution changes significantly with complex sounds.

Human frequency resolution for a wideband sound is explained with bandpass auditory filters, critical bands, the center frequencies of which are associated with different positions in the cochlea. The ear processes sound frequency components inside each critical band as one entity, since the same part on the basilar membrane of the cochlea is stimulated for closely related frequencies. Inside the critical band, stronger frequency components mask and reduce the audibility of the weaker components [3, 4]. In experiments conducted externally, such a masking effect has been exploited to define psychophysically the effective width of the critical bands and their shapes. The detection threshold for a single tone with a noise-notched masker around the tone is measured by chancing the width of the masker [5]. Additionally, loudness summation with a group of tones

relates to the critical bands [6]. That is, loudness increases with tones located in different critical bands compared to tones located in the same band. These experiments indicate that the equivalent rectangular bandwidth (ERB) scale defines the frequency widths of the critical bands most accurately when the ERB band as a function of the center frequency f_c is computed as

$$\mathrm{ERB}(f_c) = 24.7 + 0.108 f_c. \tag{2.1}$$

The frequency responses of the bandpass filters on the ERB scale are shown in Fig. 2.1 for frequencies between 200 Hz and 2 kHz. The implementation of the filters is based on the gammatone filterbank by Patterson et al. [7], and it is implemented using Slaney's approximation [8] with the Matlab HUTear 2.0 Toolbox [9].

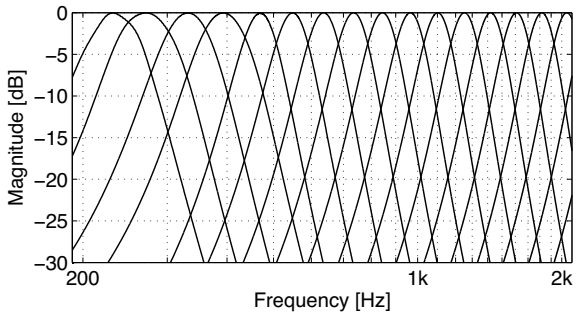


Figure 2.1. Frequency responses of bandpass filters corresponding to the ERB scale in the range from 200 Hz to 2 kHz.

The frequency resolution of the ERB scale is used by DirAC to process a frequency channel as explained in more detail in Publications I, III, IV and VI. Direction and diffuseness parameters are analyzed from the captured sound at the auditory ERB-scale frequency channels. Subsequently, the parameters are applied in spatial sound synthesis. The use of such a frequency resolution in DirAC comes from the assumption that humans cannot detect multiple directions within the same critical band. Some evidence for this assumption is given in [10].

2.2 Sound source localization

The human auditory system uses both binaural and monaural cues for sound source localization. Binaural cues are the differences in the duration of sound wave propagation from a source to the left and right ears and

in the levels between ear signals. The importance of such differences for the localization of the source was stated in Lord Rayleigh's duplex theory already in 1907 [11]. Monaural cues are based on the spectral filtering by the pinna and the torso.

Sound localization in the horizontal plane relies mainly on interaural time and level differences (ITD and ILD) in each critical band. ITD and ILD are considered as primary localization cues roughly below and above 1500 Hz, respectively. This frequency corresponds to the largest delay time between the ears, about 630 μs, which arises when sound arrives perpendicularly from the side with relation to the head [1]. The ITD decoded from the waveforms of the ear signals is no longer unambiguous at high frequencies with a cycle time shorter than 630 μs. Nevertheless, at high frequencies, the head casts a direction-dependent acoustic shadow at the ears and thus provides the ILD as a localization cue. Additionally, at considerably high frequencies, localization is found to be based also on the time differences in the envelopes of the ear signals [12] [13]. The same ITD and ILD cues are obtained for sound arriving from different points located within the so-called cone of confusion [1], having the axis of symmetry along a line through the ears. The apex of the cone is at the mid-point of this line. The ITD and ILD define a cone where a sound source is located, and furthermore a more accurate location is determined by the effect of head rotations to binaural cues and the spectral differences of the ear signals.

In the ear canals, reflections from the outer ear, the head and torso are merged with the unreflected sound that arrives directly to the ear canal entrance. Depending on the angle of arrival, the merging produces spectral differences as monaural localization cues. These have a primary meaning in the vertical sound localization and front-back separation, the latter of which is provided by the pinna effect. Spectral differences can be represented with head-related transfer functions (HRTF) from the sound source to the ear in free field. The transfer function from the ear canal entrance to the eardrum is considered to be independent of the angle of the arrival of the sound [14]. Thus, HRTFs are measured with binaural microphones placed typically at the ear canal entrances [15]. In addition to the spectral differences for sound source localization, the HRTF pair measured from the ears includes the ITD and ILD.

The accuracy for sound localization depends on direction. This can be quantified psychophysically with the absolute difference between a per-

ceived auditory event and sound event directions, or with the JND in the perception of direction. According to Blauert in [1], these measures define the localization blur[1], depending on direction, stimulus and frequency. Studies on the absolute difference using broadband white-noise pulses shown that the localization blur in the horizontal plane is $\pm 3.6^\circ$ in the front, about $\pm 10^\circ$ for the left and right sides, and $\pm 5.5^\circ$ at the rear. The JND measured with sines and narrowband white-noises is at best 1° in the front and increases to the sides. Additionally, the sound localization in the median plane is weaker than in the horizontal plane. For instance, the localization blur in the median plane is $\pm 9^\circ$ in the front measured with speech [1].

2.3 Spatial hearing in complex acoustic environments

A common listening scenario includes typically several sound sources contending for our attention. Also, reverberation in rooms forms a complex sound field affecting, for instance, sound source localization and intelligibility. Despite such possible complexities, humans have capabilities that help us to conceive a relatively accurate impression of our surroundings, and to decode relevant directional and spatial information. These capabilities are strongly based on binaural hearing.

Sound emanated from a source in a room is reflected from surfaces, and consequently similar sounds arrive at different times from different directions at the ears and are captured with different amplitudes. Despite the disinformation from the reflections, we are capable of localizing a sound source in a room relatively accurately. This capability is explained by the precedence effect [16] (also referred as the Haas effect [17]) whereby the apparent direction is one where the first sound comes from. Thus, a direct sound from a source defines mostly the perceived direction in reverberant spaces. The precedence effect requires a certain delay time between successive sounds. Experiments, where two loudspeakers at $\pm 40^\circ$ produced the same sound with different delays in an anechoic chamber, show that the precedence effect occurs if the signals differ by about 1-35 ms [1]. Smaller delays (< 1 ms) create a summing-localization, and with larger delays (> 35 ms) discrete sound sources are perceived.

[1]Localization blur expresses the smallest change in a sound event direction that produces change in the localization of the auditory event [1].

In a complex listening scenario consisting of many talkers in background noise, humans are able to selectively attend to one specific talker. Such a capability, known as the cocktail party effect, was described first by Cherry [18]. Cherry showed that the effect is much more efficient in binaural than monaural listening with simultaneous spatially-separated sound sources. This fact has also been shown with experiments, for instance, in [19]. The cocktail party effect appears to be related to the binaural masking level difference (BMLD), which measures the detection threshold of a test signal in the presence of the masking sound for diotic [2] and dichotic [3] listening [20]. In the classical BMLD experiments, the test signal is introduced into listeners's ears with various interaural phase differences via a headphone reproduction. The detection threshold for the test signal has been shown to be 12–15 dB lower in the dichotic condition than in the diotic in the presence of the diotic masking noise [1].

Besides the interaural differences, ITD and ILD, humans detect also the similarity of binaural signals. This similarity can be described with the interaural coherence (IC) [1], which appears to be related to the sensation of space. IC is defined to be the maximum of the absolute value of the cross-correlation function between the ear signals. In the experiments [1], two signals, one in each ear, were presented with various degrees of coherence to listeners. Fully coherent signals resulted in a perception of a spatially point-like auditory event. Decreasing of the coherence widened the auditory event until, finally, two separate auditory events were perceived. IC is reported also to influence ITD and ILD cues [21], which are considered to aid sound source localization only when the energy of the direct sound of a single source inside a critical band of the inner ear is sufficient.

[2] Similar sounds are presented to the listener's ears.
[3] Two different sounds are presented to a listener, one in each ear.

3. Techniques for recording and reproducing spatial sound

This section gives an overview of the techniques for reproducing and recording spatial sound. First, sound panning techniques to position monophonic virtual sources in loudspeaker listening are shortly introduced. After this, typical recording techniques to capture spatial sound are presented.

3.1 Panning techniques

Reproducing the same sound with adjacent loudspeakers creates a virtual sound source[1] between them. The virtual source moves towards the loudspeaker that emanates a stronger or an earlier sound signal in amplitude or time panning, respectively. The perceived direction of the virtual sound source (estimated with the panning direction) is related to the summing localization discussed in Section 2.3.

In amplitude panning, the sound signal is multiplied with loudspeaker-specific gain factors. Different panning laws are proposed to give an approximation for the panning direction from the gain factors. The sine law presented in [22] is

$$\frac{\sin\theta_\mathrm{S}}{\sin\theta_0} = \frac{g_1 - g_2}{g_1 + g_2}, \tag{3.1}$$

where g_1 and g_2 are the gain factors of the loudspeakers and θ_S is the panning direction. The loudspeakers are placed in the azimuths directions $\pm\theta_0$ with respect to the listener. The sine law is derived by solving a loudspeaker-gain ratio that produces the same ITD as a physical sound source in the direction of θ_S. The shadowing and diffraction of the head is ignored in the derivation of the sine law. The panning direction estimates

[1] A virtual sound source is an auditory object perceived in a direction that does not coincide with any physical sound source.

the direction of the virtual source more accurately by using the tangent law [23, 24]

$$\frac{\tan\theta_{\mathrm{T}}}{\tan\theta_0} = \frac{g_1 - g_2}{g_1 + g_2}, \tag{3.2}$$

where θ_{T} is the panning direction. The tangent law is derived by approximating the sound propagation around the head to the contralateral ear (the ear on the opposite lateral side of the loudspeaker) with a curved line. However, amplitude panning with the tangent law reproduces only the ITD cues correctly. The gain factors, which provide the panning direction θ_{T} according to the tangent law, can be computed for the adjacent loudspeakers n and m as

$$\begin{aligned} g_n &= \cos(\theta_x) \\ g_m &= \sin(\theta_x), \end{aligned} \tag{3.3}$$

where θ_x is the azimuth angle between the panning direction and the loudspeaker n. The gain factors, regardless of the panning law, need to be normalized as

$$\hat{g}_n = \frac{g_n}{\sqrt{\sum_{n=1}^{N} g_n^p}} \tag{3.4}$$

to keep the perceived loudness of the produced virtual source constant. Here, p depends on the acoustics of the listening room. In a normal room with reverberation, the value is typically chosen to be $p = 2$.

The tangent law is reformulated with the vector base amplitude panning (VBAP) [25, 26], in which the panning can be performed in a two-dimensional (2-D) plane or three-dimensional (3-D) space between adjacent loudspeakers or inside loudspeaker-triplets, respectively.

In time panning, a delay is applied to one loudspeaker signal to position the virtual sound source between two loudspeakers. However, the delay should be small enough, at most about 1 ms, that summing localization occurs. With larger delays the sound is perceived to arrive from the leading loudspeaker according to the precedence effect or as discrete sounds from both loudspeakers.

3.2 Recording techniques

The directional characteristic and arrangement of the microphones determines how the directional information is provided from the microphone signals. The microphones discussed here have either omnidirectional (zeroth-

order) or first-order directivity, the former being equally sensitive in all directions and the latter is sensitive depending on direction. General, first-order directivity is defined as a superposition of the omnidirectional and bidirectional (cosine) components of directivity as

$$e(\theta) = \beta + (1 - \beta)\cos\theta, \tag{3.5}$$

where θ is the azimuth angle and $\beta \in \{0, 1\}$ is the directivity parameter. Figure 3.1 shows some ideal first-order directional patterns. However, the patterns of the real directional microphones are frequency-dependent. That is, they are typically deformed at very low and high frequencies. When discussing recording techniques in this section, the patterns of the microphones are assumed to be ideal. Also, the directional microphone refers here to the first-order directivity as distinct from the higher-order microphones [27]. Depending on the inter-microphone distances, different stereo- and multi-microphone configurations are divided into coincident, spaced and near-coincident techniques, and their influences on sound capturing as well as on sound reproduction are discussed.

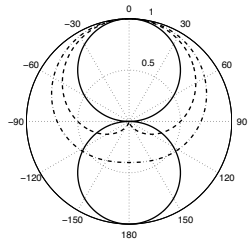


Figure 3.1. Directional patterns as a function of θ for ideal first-order microphones: dipole (solid line, β = 0), cardioid (dashed line, β = 0.5), and subcardioid (dash-dotted line, β = 0.7).

3.3 Coincident-microphone techniques

In the coincident techniques, a few directional microphones are placed as close as possible to one another with different on-axis directions. Consequently, plane waves in free field produce only differences in amplitude between microphones reflecting the directional characteristic of the sound. The signals of the stereo microphones, when fed directly to two loudspeakers, are thus equal to the signals computed with amplitude panning for stereo reproduction. With multiple coincident microphones, the microphone signals are conventionally matrixed. This includes summa-

tion, subtraction, and equalization of the microphone signals, resulting in various number of loudspeaker signals for reproduction.

3.3.1 Stereo microphone techniques

The first recording technique for stereo reproduction was proposed by Blumlein in 1931 [28]. In the Blumlein pair, two dipole microphones, a.k.a figure-of-eight microphones, are placed in a coincident position with an angle of 90° between the on-axis directions of the dipoles. The ideal directional patterns of such perpendicular dipoles are shown in Fig. 3.2 (a) and expressed in the horizontal plane as

$$\begin{aligned} e_x(\theta) &= \cos(\theta + 45^\circ) \\ e_y(\theta) &= \sin(\theta + 45^\circ). \end{aligned} \tag{3.6}$$

The Blumlein pair provides uniform signal power with all azimuths, since $\sin^2(\theta + 45^\circ) + \cos^2(\theta + 45^\circ) = 1$. Reproducing the signals of the perpendicular dipoles with loudspeakers in directions of $\pm 45^\circ$ results in a virtual sound source, the panning direction of which follows the tangent law in Eq. (3.2). The Blumlein pair provides a wide and consistent reproduction of the direct sound and a uniform spread of the reverberant sound for the frontal quadrant ($|\theta| \leq 45$) [29]. However, the sound arriving from the sides are captured with the same amplitudes, but with inverted polarities. Applying the captured microphone signals as such to the loudspeakers results in the loudspeaker signals that are out of phase with respect to one another. This causes signal cancellation at the listening position.

XY stereo techniques are realized with a pair of identical coincident microphones having any first-order directionality [29, 30]. The angle between the on-axis directions of the stereo microphones varies typically from 90° to 135° [31]. Figure 3.2 (b) shows directional patterns for two ideal cardioid microphones in a typical XY stereo pair at an angle of 120°. The directional characteristics of the microphones and their orientations result in non-uniform signal power distribution over azimuths.

XY stereo pairs relate to mid-side stereo pairs (MS) [30, 32, 33] composed of a center microphone having any directivity aimed at the center of the recording area and a side microphone with dipole directivity. Figure 3.2 (c) shows the ideal directional patterns of the MS stereo pair with dipole and subcardioid microphones. The linear combination of the XY stereo microphone signals produces certain MS stereo microphone signals and vice versa. This is also illustrated in Fig. 3.2. Subtraction and

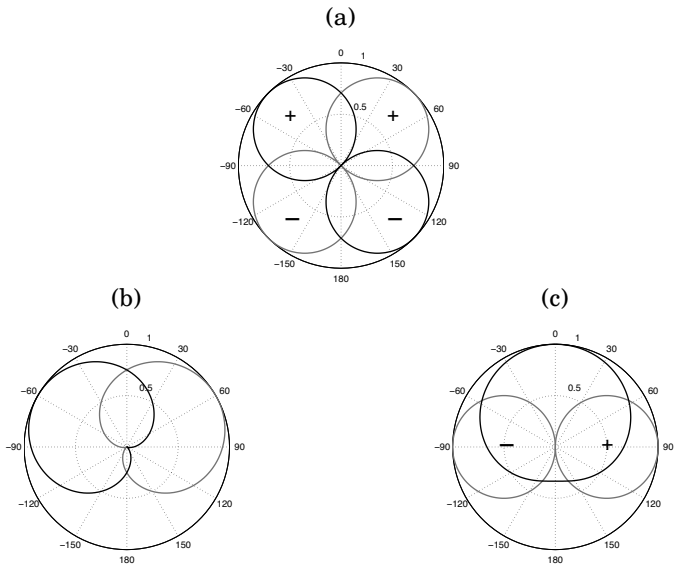


Figure 3.2. Directional patterns as a function of θ for ideal coincident stereo microphones: (a) Blumlein pair of two dipoles, (b) XY pair of two cardioids, and (c) MS pair of subcardioid and dipole.

summation of two coincident cardioids in (b) results in a dipole and a subcardioid in (c), respectively.

3.3.2 B-format signals and the Soundfield microphone

B-format signals are based on the decomposition of the sound field into spherical harmonics. In this thesis only the first-order B-format signals are utilized. They consist of the omnidirectional signal w and the orthogonal dipole signals x, y, and z (Fig. 3.3). In theory, all signals are captured at the same point in the sound field. The signal w is related to the sound pressure p and the dipoles to the approximation of the x, y, and z components of the particle velocity vector $\boldsymbol{u}$. Because of such relations first-order B-format signals represent the three-dimensional (3-D) directional sound field. For a plane wave in free field with sound pressure $p(t)$, the first-order B-format signals can be expressed as

$$\begin{aligned} w(t) &= p(t) \\ x(t) &= \sqrt{2}\, p(t) \cos(\theta) \cos(\phi) \\ y(t) &= \sqrt{2}\, p(t) \sin(\theta) \cos(\phi) \\ z(t) &= \sqrt{2}\, p(t) \sin(\phi), \end{aligned} \tag{3.7}$$

where t is time and ϕ indicates the elevation angle of the arrival direction of the plane wave. Dipole signals are amplified by 3 dB (multiplied by $\sqrt{2}$)

in order to compensate their energy with respect to the omnidirectional signal in a diffuse sound field.

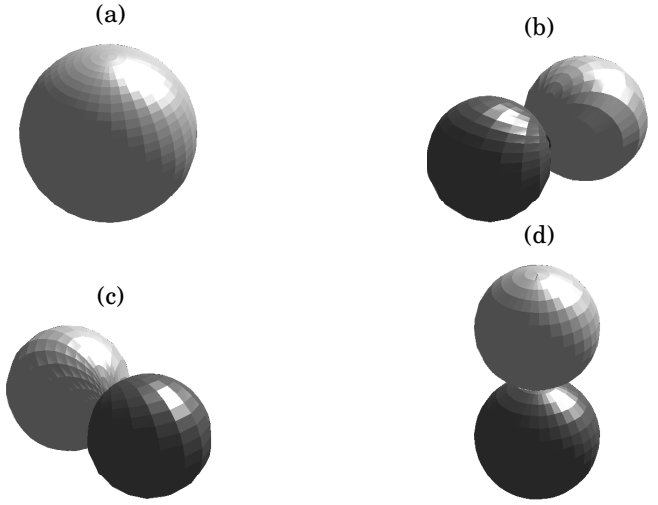


Figure 3.3. Directional patterns of the ideal B-format: (a) omnidirectional signal w, (b) dipole signal x, (c) dipole signal y, and (d) dipole signal z.

In addition to a native B-format recording with an omnidirectional microphone and three dipole microphones [34, 35], the B-format signals are often obtained from multi-capsule microphones with some other directional patterns, like the Soundfield microphone [36, 37]. The Soundfield microphone comprises four subcardioid or cardioid capsules placed at the corners of a tetrahedron. The output signals of the capsules, named A-format signals, are converted to B-format signals by summation, subtraction, and equalization. However, the capsules are not perfectly coincident. Consequently, spatial aliasing occurs and impairs the directional patterns of the B-format signals at high frequencies. This issue is discussed more in Section 3.3.3 that discusses how an array of omnidirectional microphones can be used to derive the B-format signals.

B-format recording was originally developed for Ambisonics technology [38], which provides spatial sound synthesis for any loudspeaker setup. The B-format signals are matrixed to derive virtual microphone signals, which are then applied to the loudspeakers in the reproduction. The virtual microphone signal represents a signal that would be captured with a microphone pointing to a loudspeaker direction. The signal v of the first-

order virtual microphone is obtained from the B-format signals as

$$v(t)_n = \beta w(t) + (1-\beta) \begin{bmatrix} x(t) \\ y(t) \\ z(t) \end{bmatrix} \cdot \begin{bmatrix} \cos(\theta_n)\cos(\phi_n) \\ \sin(\theta_n)\cos(\phi_n) \\ \sin(\phi_n) \end{bmatrix}^T, \qquad (3.8)$$

where θ_n and ϕ_n indicate the direction of loudspeaker n.

The directional patterns of the first-order virtual microphones are, however, relatively broad. Consequently, the virtual microphone signals applied to the adjacent loudspeakers correlate significantly with one another, especially with high loudspeaker densities. This produces coherence between loudspeaker channels and, furthermore, the reproduced sound and spatial image may be colored and blurred [39]. Coherence is partly avoided by using a sparse layout with a considerably small number of loudspeakers, for instance, using four loudspeakers at $\pm 45^\circ$ and $\pm 135^\circ$. In principal, any higher-order directivity could also be applied with the virtual microphones. Thus, the directional patterns would be narrower than with the first-order virtual microphones, decreasing the degree of coherence between the loudspeakers. Higher-order virtual microphones require a large number of real microphones in the recording. For instance, 25 omnidirectional microphones in a spherical array are used in [27] to produce third-order virtual microphone. In practice, the higher-order microphones suffer from a low signal-to-noise ratio (SNR) and a limited frequency range (340-3400 Hz with the spherical array in [27]) and, thus, typically the first-order directivity is applied in the Ambisonics.

3.3.3 Deriving B-format signals from an omnidirectional microphone array

In addition to the Soundfield microphone, arrays composed of multiple omnidirectional microphones have also been employed to derive B-format signals [40, 41]. Figure 3.4 depicts a B-format array of six microphones arranged in a regular octahedron. Microphone pairs $\{p_1, p_2\}$, $\{p_3, p_4\}$, and $\{p_5, p_6\}$ on the Cartesian axes are used to derive the corresponding dipole signals (x, y and z) as pressure gradients. That is, microphone signals of each pair are subtracted from one another. The omnidirectional signal w is computed as an average of all microphone signals.

The opposing microphones are positioned a few centimeters apart in the array. Consequently, when deriving the dipole signals, the inadequate sampling in space results in spatial aliasing, deforming the directional

patterns at high frequencies. The spatial-aliasing frequency, which gives a theoretical upper frequency limit for a dipole pattern, is

$$f_{\mathrm{sa}} = \frac{c}{2d}, \tag{3.9}$$

where c is the speed of sound and d is the inter-microphone distance. Obviously, a larger inter-microphone distance results in a lower spatial-aliasing frequency and vice versa.

The telecommunication applications of DirAC, as presented in Publications I, III, IV and [42], employ the microphone array discussed above, but typically with a smaller number of microphones. The arrays are one-dimensional (1-D) or 2-D with one or two microphone pairs, respectively. For instance, a 2-D array can be built by using only the microphone pairs $\{p_1, p_2\}$ and $\{p_3, p_4\}$ shown in Fig. 3.4. It is obvious that less dipole signals are thus obtained and, consequently, the whole 3-D directional sound field cannot be represented.

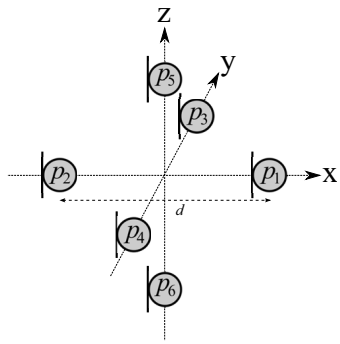


Figure 3.4. B-format array of six omnidirectional microphones. A microphone pair at each coordinate axis is used to derive the corresponding dipole signal (x, y and z) as a pressure gradient. The omnidirectional signal w is computed as an average of all microphone signals. The microphones in a pair are separated by a distance d.

3.4 Spaced and near-coincident microphone techniques

Unlike in the coincident techniques, microphones may also be placed at considerable distances from one another. Sound is thus captured at different time instants in different microphones, depending on the arrival direction, microphone spacing, and frequency. Reproducing mutually delayed microphone signals with different loudspeakers equals to the time panning, as presented in Section 3.1.

Spatially separated microphones are not typically used in parametric spatial sound processing, but in linear techniques where the captured sound signals are reproduced directly with loudspeakers. Recently, a method to apply spaced microphones with DirAC has been presented [43].

3.4.1 Stereo microphone techniques

In spaced stereo techniques, two microphones with identical directional patterns are positioned a few ten centimeters to a few meters apart [31]. Even though the microphones may have any directivity, they are typically omnidirectional, providing mainly time differences between captured signals. For instance, the technique named A/B stereo consists of two parallel, omnidirectional microphones separated 60 cm apart. The reproduced spatial image is typically inconsistent and more blurred with spaced microphones compared to the reproduction with coincident directional microphones. On the other hand, the reproduction is described as 'airy' and 'ambient' with spaced microphones [29], and sound power is captured equally from all directions, features which may be preferred for the reproduction.

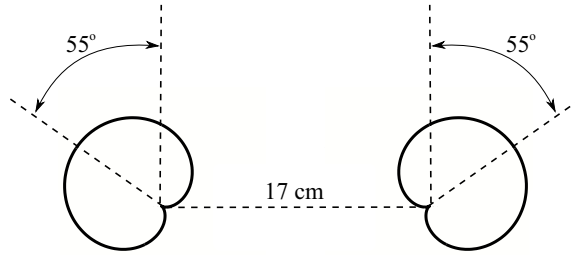


Figure 3.5. ORTF stereo configuration of two cardioid microphones in a near-coincident configuration.

In near-coincident stereo techniques, the distance between microphones is smaller than in the spaced configuration presented above. Another difference is that directional microphones are employed instead of omnidirectional. Because of the spacing of the microphones and their directional characteristics, near-coincident techniques have different properties at different frequencies. The distance between microphones may be small compared to the wavelength of the sound at low frequencies, and therefore time differences between captured signals are practically nonexistent. However, the different microphones having typically first-order di-

rectivity capture the sound with different amplitudes, a fact that makes near-coincident microphones appear like the XY stereo microphones at low frequencies. At high frequencies, the captured microphone signals differ in time, like with spaced microphones. For instance, in the common ORTF stereo configuration, two cardioid microphones at an angle of 110° between their on-axis directions are separated by 17 cm, as depicted in Fig. 3.5. Another common configuration, NOS stereo, consists of two cardioid microphones 30 cm apart with an angle of 90° between their on-axes [30].

3.4.2 Techniques with multiple spaced microphones

Instead of the multichannel reproduction from coincident microphone signals, as in the Ambisonics technology, multiple spaced microphones are most commonly used to capture spatial sound. As an example, Fig. 3.6 depicts a five-cardioid microphone array which can be applied to a standard 5.0 loudspeaker setup [44]. The spacing between microphones is based on their directional patterns and on-axis directions [45]. Furthermore, the patterns, directions, and spacing have influences on the reproduction. In addition to being directly reproduced, microphone can also be mixed with one another to obtain the desired reproduction. For instance, a common Fukada tree technique [46] uses the five-cardioid microphone array and the ORTF stereo pair (the ORTF microphones are placed to the left and right of the five-cardioid array).

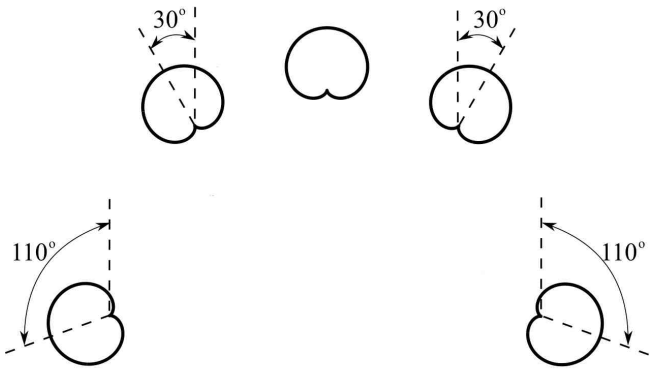


Figure 3.6. Array of five cardioid microphones for the ITU 5.1-channel reproduction.

Depending on the microphone spacing in a chosen array, multichan-

nel reproduction may suffer from different problems. In the coincident techniques, coherence between microphone signals is relatively high, especially at low frequencies and with high loudspeaker densities. This increases the coherence between loudspeaker channels and causes coloration and blurring in the reproduction. Positioning the microphones at considerably distances from one another reduces the coherence, but larger distances may delay microphone signals relative to one another more than desired. Consequently, the reproduced reverberation due to the delay is perceived to be artificial. Thus, when specifying a spacing for an array, one will typically be a compromise between the issues presented above.

3.5 Binaural recording

Figure 3.7 depicts the principles for binaural recording. The sound is captured with a pressure microphone placed in the ear canal of a dummy or a human head [47] and the binaural microphone signals from both ears are reproduced to a listener using headphones. The binaural signals replicate the ear canal signals and thus transfer all spatial information of the sound to the listener, resulting in a sound image corresponding to the recording conditions. Ideally, reproduced sounds are perceived from directions coinciding with the original sound source directions, localized outside of the head and with correct timbre.

Figure 3.7. Principles for the binaural recording.

In practice, microphone positioning has a major influence on the sound capturing and, furthermore, on the sound perception. The microphones are typically positioned at the entrance of either blocked or open ear canals [48]. The blocked-ear approach inhibits the ear canal effect, whereas in the open-ear approach, the effect needs to be removed from the bin-

aural microphone signals by filtering[2]. A disadvantage of the blocked-ear approach is it is assumed to change the acoustic impedance from the eardrum to the entrance. This must be compensated in the binaural signals [49]. The processing mentioned above may introduce some coloration to the reproduction of the signals.

[2]The ear canal effect occurs inherently in a headphone listening. Because of this, it should be removed from binaural signals recorded with microphones at open ear canals

4. Parametric spatial sound processing with microphone front-end

The signal-processing techniques for a parametric spatial sound are discussed in this section. The general processing principle to provide parametric data from microphone signals for spatial sound reproduction is shown in Fig. 4.1. In the time-frequency analysis, the microphone signals are transformed into the time-frequency domain by applying, for instance, the short-time Fourier transform (STFT). Typically, the applied time-frequency transform is designed to mimic the time-frequency resolution of human hearing. That is, for instance, performing spatial sound processing of the frequency channels that correspond to the human frequency selectivity. In the spatial sound analysis, the parametric data, which contains the directional information of the captured sound, is estimated in each time-frequency position. The parametric data and microphone signals, or their down-mixed signal(s), are transmitted to the decoder and utilized to synthesize spatial sound for loudspeaker or for headphone reproduction.

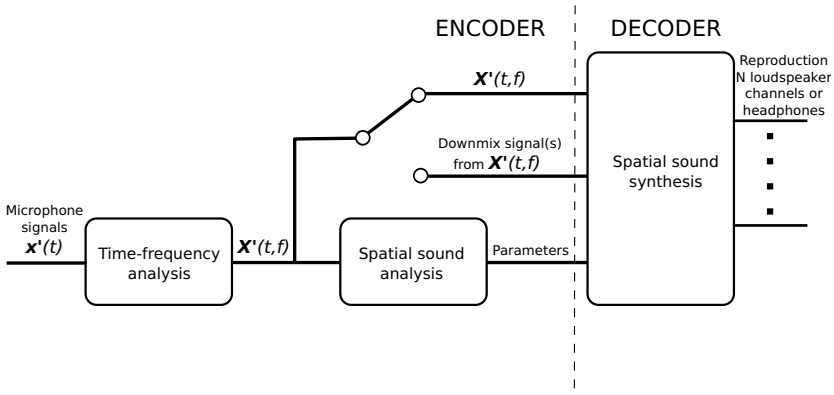


Figure 4.1. Flow diagram of the general principles for processing of the spatial sound from microphone signals.

Parametric methods that utilize multichannel audio files as input e.g. in 5.0 format, have also been proposed and used for spatial audio coding [50, 51, 52, 53], basically following the principles illustrated in the flow diagram in Fig. 4.1. However, as this thesis focuses on the parametric processing of microphone signals, such methods applying audio files are not discussed here.

4.1 Directional Audio Coding

In DirAC [54], the parametric data to describe spatial sound is obtained from B-format signals by estimating the arrival direction and diffuseness of sound inside the auditory frequency channels depending on time. The estimated parameters can be applied to audio signals in various DirAC applications, such as in high-quality reproduction of spatial sound with any loudspeaker setup or with headphones [55, 56], teleconferencing [57], and directional filtering [42]. The present author has focused his research on applying DirAC processing to the teleconference application, which is mainly discussed in this section.

DirAC shares the same spatial hearing assumptions and partly the same processing methods with Spatial Impulse Response Rendering (SIRR) [58, 59, 60], which is a technique to estimate direction and diffuseness from room impulse responses measured with a B-format microphone and applying them in convolving reverberators. The assumptions are as follows:

1. The arrival direction of sound reproduced correctly provides the localization cues (ITD, ILD, and monaural cues) to a listener.
2. Diffuseness reproduced correctly provides the IC cues to a listener.
3. Direction, diffuseness, and the short-time spectrum of sound measured at one point determine the listeners' spatial auditory perception.

Moreover, the human auditory system is assumed to decode only one direction and corresponding coherence cues at each time instant and inside one auditory frequency channel. Because of this, one direction value and one diffuseness value are estimated inside frequency channels on the ERB scale (see Section 2.1).

DirAC analysis

The arrival direction and diffuseness of sound are estimated using the energy analysis of a sound field. The analysis is based on the sound intensity [61], which is defined for a harmonic sound field as

$$\boldsymbol{I}(t) = p(t)\boldsymbol{u}(t). \tag{4.1}$$

In DirAC, the pressure p and the particle velocity vector $\boldsymbol{u}$ are measured and approximated with B-format signals (see Section 3.3.2). The signals are transformed to the time-frequency domain, and subsequently the active sound intensity vector[1] is given by

$$\boldsymbol{I}_a = \frac{1}{\sqrt{2}Z_0}\mathrm{Re}\{W^*\boldsymbol{X}\}. \tag{4.2}$$

The dependence on time and frequency are omitted for notational simplicity in the equations. Here, Z_0 is the acoustic impedance of air and $\mathrm{Re}\{\cdot\}$ expresses the real part of the complex number. The time-frequency transformed B-format signals are denoted with W (omnidirectional signal) and $\boldsymbol{X} = [X\ Y\ Z]^T$ (dipole signals in vector form). Moreover, $Z_0 = \rho c$, where ρ is the mean density of air and c is the sound velocity in air. The most prominent direction of sound (represented with azimuth θ and elevation ϕ) is obtained as a direction opposite to that of active sound intensity vector at each frequency channel.

The diffuseness is a real value between 0 and 1, and it expresses the fraction of the sound energy which occurs with multiple incoherent plane waves in different directions. The diffuseness at each frequency channel is traditionally computed as a ratio of the magnitude of the intensity vector to the energy density[2] as

$$\psi = 1 - \frac{||\mathrm{E}\{\boldsymbol{I}\}||}{c\mathrm{E}\{E\}}, \tag{4.3}$$

where $\mathrm{E}\{\cdot\}$ is the expectation operator implemented with recursive temporal integration in DirAC. The energy density E is computed from the B-format signals as

$$E = \frac{1}{2}\rho_0 Z_0^{-2}\left[|W|^2 + |\boldsymbol{X}|^2/2\right]. \tag{4.4}$$

[1]The active sound intensity is defined as the real part of the complex sound intensity vector and it represents the transport of sound energy per unit time. The imaginary part of the complex vector, namely the reactive sound intensity, represents the local, oscillating sound energy.

[2]Energy density is a scalar quantity that expresses the total sound energy per unit volume.

In addition to Eq. (4.3), another method for computing diffuseness is proposed by the present author in [62]. The method is also presented in Publication II and called coefficient of variation estimator. The diffuseness

$$\psi_{\mathrm{CV}} = \sqrt{1 - \frac{||\mathrm{E}\{\boldsymbol{I}\}||}{\mathrm{E}\{||\boldsymbol{I}||\}}} \tag{4.5}$$

is estimated from the temporal variation of the active sound intensity vectors. In a diffuse sound field, the vectors point in random directions at different time instants, whereas in a non-diffuse field with single plane wave they point in the same direction, constantly. As reported in Publication II, Eq. (4.5) provides a more accurate estimate for diffuseness than Eq. (4.3) when a limited representation of the sound field is available, for instance when using a 1-D or 2-D microphone array in DirAC, as mentioned in Section 3.3.3.

DirAC synthesis

For the DirAC synthesis, the estimated directions (azimuth θ and elevation ϕ) and diffuseness ψ at each frequency channel are transmitted along with one or more audio channels. For instance, one audio channel (the omnidirectional signal of B-format) is utilized in the DirAC synthesis in the teleconference application. The time-frequency transformed omnidirectional signal W is divided into non-diffuse and diffuse streams (W_{nd} and W_{d}) for each frequency channel from the estimated diffuseness ψ as

$$\begin{aligned} W_{\mathrm{nd}} &= W\sqrt{1-\psi} \\ W_{\mathrm{d}} &= W\sqrt{\psi/N}, \end{aligned} \tag{4.6}$$

where N is the number of the loudspeakers used in the reproduction. The non-diffuse and diffuse streams contain mainly direct and ambient sounds, respectively. Based on the estimated directions θ and ϕ, the sound of the non-diffuse stream is actively steered by reproducing it as point-like virtual sources using amplitude panning with the tangent law (see Section 3.1), that is, the sound of the non-diffuse stream is multiplied with the loudspeaker-specific gain factors at each frequency channel. Depending on the application, the sound of the diffuse stream is reproduced coherently with all loudspeakers or it is individually de-correlated for each output channel to decrease the coherence between loudspeaker signals and to increase the sensation of the surrounding sound. De-correlation can be performed, for instance, with exponentially decaying white-noise bursts or with pseudorandom delays [54, 55]. However, a real-time teleconference application of DirAC typically uses coherent reproduction instead of

de-correlation, which is computationally relatively complex. Besides, the noise-bursts or delays should be short enough to avoid unwanted artificial reverberation in teleconferencing[3], but consequently the timbre may then be changed.

DirAC teleconferencing

Applying spatial sound in teleconferencing has advantages when the directions of the many talkers at one end are reproduced to the listeners at the other end, as illustrated with the schematic of DirAC teleconferencing in Fig. 4.2. This kind of spatial distribution of the remote talkers improves aspects of communication, as speech intelligibility, recalling ability, comprehension, talker recognition and focal assurance[4] [63, 64, 65, 66]. Improvement in speech intelligibility can be linked to the cocktail party effect described in Section 2.3. DirAC can be used to provide a real-time spatial audio for teleconferencing with a full-duplex connection [57]. Moreover, various microphone configurations introduced in this thesis, and also in [62], make DirAC processing applicable and flexible for different professional and domestic telecommunication equipment. Some of these configurations are evaluated by listening tests in terms of speech intelligibility and audio quality in teleconferencing in Publications I and III.

An interesting question is the required data rate for transmitting the direction and diffuseness parameters in DirAC teleconferencing. This question and perceptual compression methods for directional metadata have been studied in [67] and are partly reviewed next. First, loudspeakers in a horizontal plane only are employed in typical teleconferencing. Thus, the azimuth θ is estimated as the direction parameter, omitting the elevation ϕ, and transmitted for the synthesis. The perceptual compression methods utilize human localization accuracy, which depends on the arrival direction of sound. The localization accuracy was already reviewed in Section 2.2. A circle can be divided into distinct sectors based on the angle subtended by human localization blur inside which the estimated azimuths angles are quantized as the center angle of the sector. Moreover, the number of the sectors is decreased assuming that remote talkers are positioned in the frontal hemisphere only. According to the usability test with an audio-visual teleconferencing in [67], a minimum of 8 distinct

[3]Excessive reverberation may reduce the speech intelligibility.

[4]The involvement of the participants is provided with focal assurance cues. For instance, the participants are able to recall afterwards the aspects/issues related/associated with the other participants.

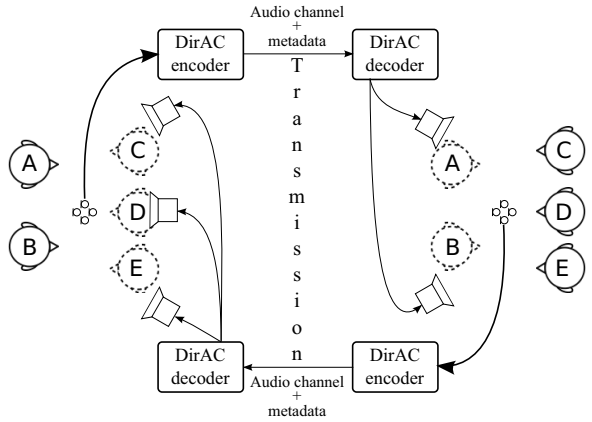


Figure 4.2. Schematic view of DirAC teleconferencing with multiple talkers (A-E). At these locations, sound emitted by talkers is captured with a microphone array. The direction and diffuseness parameters are estimated and transmitted as metadata along with one channel of audio between the locations. The remote talkers (virtual sources expressed with dashed lines) are reproduced spatially separated in directions that coincide with the talkers' directions in the sending location.

sectors was found to be sufficient for a typical small-scale teleconferencing with three talkers at one end. The direction parameter at a single frequency channel can thus be represented using 3 bits. To compress the diffuseness parameter, values from 0 to 1 can be divided, based on the usability test, into four ranges with upper-limit values 0.25, 0.5, 0.75, and 1. Within each range, the estimated diffuseness is quantized to the upper-limit values, resulting in 2 bits for the diffuseness parameter for each frequency channel. When considering the wideband telecommunication range of 50-7000 Hz [68], the number of frequency channels on the ERB scale is about 30, and with a 30-ms update rate, this results in 3-kbit/s and 2-kbit/s data rates for direction and diffuseness, respectively.

4.2 Relation between DirAC and other parametric techniques with microphone front-ends

DirAC is closely related to a number of other methods which estimate the directional parameters of the sound field and efficiently reproduce spatial audio. Although, DirAC was proposed after some such as methods as binaural cue coding (BCC) [50, 51, 69], it was the first parametric method for processing spatial sound directly from microphone signals. Recently, some other methods, which are discussed next, have also been proposed

to be used with microphone signals as acoustic front-ends.

4.2.1 Method for using stereo microphone signals in spatial audio coders

A method for providing a parametric representation of the spatial sound from coincident first-order stereo microphones is proposed in [70]. It can be applied to spatial audio coders (SAC), like MPEG surround [71], to generate multichannel audio signals. In the method, the direct-to-diffuse sound ratio and the arrival direction of direct sound are analyzed at each time-frequency position, a fact that makes the method to be almost identical to DirAC. The direct-to-diffuse sound ratio, which represents the amount of direct and ambient sounds in the microphone signals, is computed using cross-correlation. The direction estimate is based on the amplitude ratio between the signals of given microphones, utilizing the directional response information. Depending on the directional patterns of the microphones, the estimated directions from the amplitude ratios can be uniquely determined only for certain directions. Because of this, the most suitable configurations consist of cardioid or super-cardioid microphones, as reported in [70]. Furthermore, the estimated parameters are used to generate parameters compatible for the SAC decoder. In MPEG surround, these compatible parameters are the inter-channel level difference (ICLD) and the inter-channel coherence (ICC), which are applied to down-mixed signal to derive multichannel signals.

4.2.2 Sparsity-based method using relative phase differences of omnidirectional microphones

Another signal-processing method similar to DirAC has been presented in [72]. The arrival direction of sound is estimated from four near-coincident omnidirectional microphones, positioned in a tetrahedral. To yield a 3-D vector for the arrival direction, the relative phases of the microphone signals are compared to one another at each time-frequency position. Contrary to DirAC processing, the diffuseness, or a similar parameter, is not analyzed, but the diffuseness information is assumed to be inherently encoded with the direction estimation. In the synthesis phase, the direction estimation is applied to binaural sound reproduction with headphones by filtering a microphone signal with direction-dependent HRTFs at each time-frequency position. In [73], the same method with a tetrahedral mi-

crophone array is proposed to be also applied to Wave Field Synthesis (WFS) [5].

4.2.3 Decomposition of B-format signals into two plane waves and application in spatial sound reproduction

B-format signals are also employed in the method proposed in [75]. In this method, the B-format signals are decomposed into two plane waves, the directions and amplitudes of which are estimated at each time instant and frequency bin. In the synthesis phase, the plane waves are reconstructed, based on the analyzed directions and amplitudes, and reproduced with loudspeakers or with headphones [75, 76]. The differences between the proposed and DirAC methods come mainly from the analysis of two or one directions at each time-frequency position, respectively, and also from the fact that the diffuseness parameter is not utilized in the proposed method.

4.2.4 Binaural microphone signals applied in parametric spatial sound processing

As mentioned in Section 3.5, binaural signals transfer all spatial information of the sound to a listener in a headphone reproduction. In addition to this traditional approach, a parametric spatial sound processing method has also been applied to binaural signals. A method for up-mixing binaural signals into multiple loudspeaker signals is presented in [77]. In this method, the parametric data to describe spatial sound consists of the ITD values, which are determined by using cross-correlation between the ear-canal signals at the frequency channels. Moreover, the cross-correlation is computed from the signal waveforms and envelopes below and above 1.6 kHz, respectively. Based on these facts, the method to decode the ITD aims to mimic that of human spatial hearing. For spatial sound synthesis, the estimated ITD values are mapped into azimuth angles and then used in spatial sound synthesis.

Another method, proposed in [78], also uses binaural signals in direction estimation with cross-correlation. However, the ITD values are estimated only from the signal waveforms and without the frequency channel divi-

[5] WFS [74] relies on Huygens' principle assuming that the recorded sound field can be reconstructed as a superposition of spherical sound waves, which are reproduced from a large number of closely-spaced loudspeakers to a listener.

sion. Moreover, the method is not used directly for a spatial sound synthesis, but for tracking the head orientation in a horizontal plane from the binaural signals. Nevertheless, the information of the head orientation can be utilized in a binaural reproduction, as presented in [79].

Generally, estimating the direction from binaural signals resembles parametric processing in the DirAC application, proposed in Publication VI, where two microphones in each hearing aid device are used to estimate the direction and diffuseness of sound at the frequency channels. However, the direction computation is based on energetic analysis, not on computing time differences. Additionally, the direction and diffuseness estimates are utilized to filter sound depending on direction and to suppress ambient sounds, respectively, for the hearing impaired.

5. Contributions of this work to DirAC with different microphone front-ends

This section shortly reviews the microphone front-ends introduced for DirAC processing in this thesis. The detailed description is given in the publications.

5.1 Arrays of omnidirectional microphones

As already mentioned, DirAC generally utilizes the first-order B-format microphone signals (omni w and dipoles x, y and z). In some practical applications, it is feasible to use arrays of omnidirectional microphones which provide the omnidirectional signal w and only one or two dipole signals, depending on the number of the microphones.

Four omnidirectional microphones, placed in the corners of the square, enable creating the horizontal B-format signals w, x and y, as described in Section 3.3.3. In the DirAC analysis applied to such a 2-D array, the x and y components of the active sound intensity vector (I_x and I_y) are computed. Thus, the estimated azimuth angle θ alone expresses the arrival direction of sound. Nevertheless, omitting the elevation angle ϕ is tolerable, for instance, in teleconferencing where typically a horizontal loudspeaker setup is utilized.

A 1-D array of two omnidirectional microphones provides the omnidirectional signal w and dipole signal x, which allow the computation of I_x. A lateral angle from -90° to 90°, when defining the on-axis direction of the dipole x to be 90°, is estimated to express the arrival direction of sound. Consequently, the DirAC analysis cannot distinguish between sound sources in the front from those in the rear. This is tolerable in teleconferencing, since the participants and the loudspeakers (emitting the sound of the remote participants) are typically situated in the front and

rear with respect to the microphone array, as shown Fig. 4.2. Additionally, the sound from the loudspeakers captured by the microphones can be reduced by acoustic echo cancellation [80, 81]. Two alternative methods for computing the arrival direction of sound with the 1-D microphone array are proposed in Publication I, pp. 4-6.

As already mentioned in Section 3.3.3, the directional patterns of the dipoles (created by subtracting microphone signals in the 1-D and 2-D arrays) are deformed above the spatial-aliasing frequency given in Eq. (3.9). Additionally, the SNR is low in the dipole signal at low frequencies. Such an issue arises from the fact that self-noises are mutually different in the microphones. Subtracting the microphone signals actually adds the noises of the opposing microphones and increases the noise level by 3 dB. At the same time the desired signal is highly attenuated, especially, at low frequencies. Due to the low SNR and spatial aliasing at low and high and frequencies, the direction can be estimated only within a limited frequency window, as addressed in detail with the 2-D microphone array in Publication IV, pp. 4-5.

5.2 Pair of coincident cardioid microphones

DirAC can also be applied to two coincident cardioid microphones in an XY pair, as proposed in Publication III. Obtaining the signals w and x resembles a conversion from the XY microphone signals to the MS microphone signals (see Section 3.3.1). Subtraction between the left and right cardioid microphone signals results in the dipole signal x. However, summing the microphone signals results in omnidirectional only with the microphones pointed in opposing directions. Otherwise the created signal w has directivity and consequently the direction estimation is slightly incorrect in the DirAC analysis. Nevertheless, this can be partly compensated by replacing the signal w with the left or right cardioid microphone signal, depending on which has the higher signal energy. Thus, sound arriving from the left hemisphere results in a higher signal level in the left cardioid microphone than in the right and vice versa (see detailed description in Publication III, p. 5). After obtaining the required signals, the directional analysis is performed similarly to that with the 1-D array of two omnidirectional microphones. In practice, the cardioid microphones are not perfectly placed at the same point, a fact that deforms the direc-

tional pattern of the dipole signal x at high frequencies and upsetting the direction estimation.

In DirAC synthesis, loudspeaker signals can be rendered from the cardioid microphone signals using three alternative methods. The direction and diffuseness parameters can be applied to audio signals which are obtained as follows:

1. Averaging the cardioid microphone signals.
2. Transmitting both cardioid microphone signals and using the left/right microphone signal to synthesize the loudspeaker signals for the left/right hemisphere. (See the detailed description in Publication III, p. 5.)
3. Transmitting and using the left or right microphone signal depending on the estimated direction. (See the detailed description in Publication III, p. 5.)

5.3 Microphone arrays utilizing acoustic shadowing

As mentioned above, spatial aliasing and the low SNR cause an inaccurate direction estimation in the DirAC analysis with the microphone arrays above. To improve the estimation, especially at high frequencies, Publication IV proposes inserting a rigid cylinder between the omnidirectional microphones inside the 2-D array, as shown in Fig. 5.1. The rigid cylinder casts a shadow and scatters sound, depending on its arrival direction and frequency. Consequently, the microphones capture the sound with a measurable inter-microphone level and spectral differences, which are utilized in two alternative methods for direction estimation. The first one, called the pressure-energy-gradient (PEG) method, approximates the x and y components of the active sound intensity vector directly by computing energy gradients at high frequencies with prominent inter-microphone level differences. That is, the power spectra of the opposing microphone signals are subtracted from one another:

$$\widetilde{I}_x = |P_1|^2 - |P_2|^2$$
$$\widetilde{I}_y = |P_3|^2 - |P_4|^2.$$

Here, P_n is the STFT spectrum of the signal of microphone n. Moreover, $n = \{1, 2, 3, 4\}$. At low frequencies (without acoustic shadowing),

the dipole signals x and y are first created from the pressure gradient (subtracting the STFT spectra of opposing microphones from one another) and then computing the sound intensity components I_x and I_y. The detailed description of this method is given in Publication IV, pp. 315-317. The second method applies vector quantization (VQ) [82], which generally provides an optimal data classification. Here, when estimating the arrival direction of sound, an a priori array manifold of the multi-directional responses[1] is compared to the STFT spectra of the microphone signals with a normalized cross-correlation function in the VQ method. The detailed description of the method is given in Publication IV, pp. 317-318. As corroborated with anechoic chamber measurements and listening tests in Publication IV, both methods provide accurate direction estimations over the entire audio frequency range and improves the perceived audio quality in DirAC comparing to the 2-D array without the cylinder.

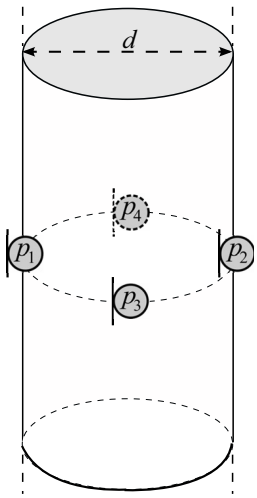


Figure 5.1. Array of four omnidirectional microphones mounted on a rigid cylinder of diameter d.

The PEG method has also been used with a 2-D microphone array composed of four omnidirectional microphones with relatively large housings, as proposed in Publication VI. A shadowing object is not used in the array, but the microphones are arranged such that their housings produce shadowing and scatter sound, depending on the arrival direction and frequency. The size of the microphones determines the frequency limit, below and above which the pressure or energy gradients, respectively, are

[1]The array manifold is derived a priori either from anechoic impulse response measurements for various directions, with respect to the microphone array, or modeling multi-directional responses of the cylinder array.

utilized to estimate the direction of the arrival of sound. Hence, the distance between opposing microphones should be optimized with respect to the microphone sizes to match the computations using both pressure and energy gradients. Such an optimizing method is proposed in Publication V, pp. 98-99.

5.4 DirAC processing applied to bilateral hearing aids

DirAC processing is also applied to bilaterally-fitted hearing aids, as proposed in Publication VI. Sound is captured with two microphones in a hearing aid device at both ears, as depicted in Fig. 5.2. The head casts an acoustic shadow and causes the sound to scatter, which is captured with the microphones at either side of the head. The direction estimation is partly based on the PEG method discussed above. The x component of the active sound intensity vector is approximated by using the pressure and energy gradients below and above the frequency limit, about 640 Hz, determined by the head's diameter. The y component of the intensity vector is computed using the dipole signal y_{L} or y_{R}, obtained as the pressure gradient from the microphones of the left or right device, respectively, depending on which one has the higher signal energy. The detailed description of the directional analysis of the hearing aid microphone signals is presented in Publication VI, pp. 6-7.

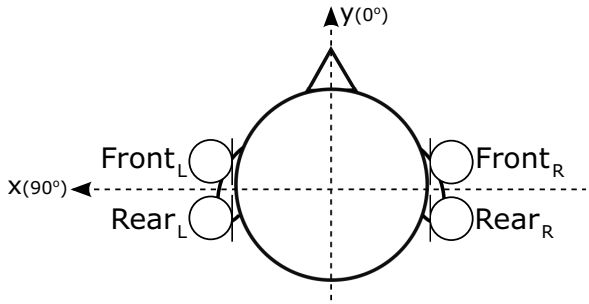


Figure 5.2. Schematic view of microphones in hearing aid devices in both ears. In DirAC analysis, the signals of the microphones ($\mathrm{Front_L}$, $\mathrm{Rear_L}$, $\mathrm{Front_R}$, $\mathrm{Rear_R}$) are used to estimate the direction and diffuseness parameters. In DirAC synthesis, the parameters are applied to the front microphone signals ($\mathrm{Front_L}$ and $\mathrm{Front_R}$) to amplify/attenuate direct sound and to suppress ambient sound.

In the synthesis phase, one microphone signal from each hearing aid device ($\mathrm{Front_L}$ and $\mathrm{Front_R}$ in Fig. 5.2) is amplified or attenuated at each time-frequency position, depending on the estimated direction. Moreover, the

diffuseness parameter, which is also estimated from the intensity vectors, is used to suppress ambient sound in the microphone signals. After this, the modified signals are reproduced to a listener. The detailed description of the synthesis phase is given in Publication VI, pp. 7-8. The DirAC processing applied to bilaterally-fitted hearing aids with behind-the-ear devices (BTE) was evaluated with a listening test to measure the speech reception threshold (SRT) for various speech signal and noise configurations. The test results corroborated that the DirAC processing provides an improvement of SRT that is comparable to other processing methods for bilateral hearing aids. However, the required data rate between the devices is relatively low with DirAC. The listening test and the results are presented in detail in Publication VI, pp. 9-11.

6. Conclusions and main results

This thesis introduced a number of microphone configurations applied to the parametric time-frequency processing of DirAC for recording, transmitting, and reproduction of spatial sound. From an application point of view, the thesis focused mainly on DirAC teleconferencing, where the sound emitted by the remote talkers at the sending end is efficiently reproduced spatially distributed at the receiving end. Applying the spatial sound to teleconferencing is shown to improve speech intelligibility. In the thesis work, directional analysis methods to estimate the direction and diffuseness of sound at frequency channels depending on time, were presented and evaluated with different microphone configurations. For subjective and objective evaluations, the listening tests and anechoic measurements were conducted. The main results of the thesis are briefly presented below.

Publication I: 1-D and 2-D arrays of two and four omnidirectional microphones, respectively, applied to DirAC teleconferencing were presented and evaluated with listening tests to measure speech intelligibility. In the tests, two spatially distributed loudspeakers reproduced different speech samples simultaneously. The sound was captured with the 1-D and 2-D microphone arrays for DirAC and with an XY pair of optimally directed dipole microphones (a.k.a the Blumlein pair). Subsequently, the sound was provided with loudspeakers to the test subjects using DirAC or reproducing directly the microphone signals of the XY pair. DirAC (with both arrays), the XY pair, and the test reference all gave the same results for speech intelligibility in the tests. The benefit with DirAC comes from the fact that only one channel of audio along with directional metadata can be transmitted with a low data rate, whereas two channels of audio are needed to transmit with the XY pair for the teleconference application. Thus, DirAC efficiently provides spatialized teleconferencing.

Publication II: A novel method to estimate diffuseness in DirAC (coefficient of variation estimator) was proposed and evaluated with simulations in the thesis. The proposed method was compared to the traditional method (energy density-based estimator) and another novel method (active intensity-based estimator). In the simulations, the methods were studied in 1-D, 2-D, and 3-D sound field analyses with comparable arrays of omnidirectional microphones. According to the simulations, all methods produced accurate diffuseness estimates in the 2-D and 3-D analyses. However, the traditional method provided a somewhat inaccurate estimate in the 1-D analysis, whereas the novel methods provided reasonably accurate estimates. This is an important result, because the typically used configuration of two microphones enables only the 1-D sound field analysis in DirAC, and thus a method to produce reliable diffuseness estimation is desired.

Publication III: DirAC processing using two cardioid microphones at a coincident position was presented and studied with listening tests. Two cardioid microphones enable the 1-D analysis of the sound field. The listening tests were conducted to assess overall audio quality in teleconferencing with spatially separated talkers. The speech sound was reproduced with loudspeakers, captured with microphones, processed, and reproduced with loudspeakers to the listeners both for sweet-spot and off-sweet-spot listening. The test reference was the XY pair in which two dipole microphones were optimally directed towards the loudspeakers for recording, and then the microphone signals as such were reproduced to the listener. DirAC-processed signal and the test reference had the same preference scores for audio quality for sweet-spot listening. For the off-sweet-spot listening, DirAC achieved a higher preference score than the test reference. Other microphone configurations, such as the 1-D and 2-D arrays of omnidirectional microphones, applied to DirAC were also studied in the listening tests. They produced about the same overall audio quality as the conventional XY pair of the cardioid microphones for the sweet-spot listening and better quality for the off-sweet-spot listening.

Publications IV and V: Introducing acoustic shadowing for omnidirectional microphones in a 2-D array was studied in this work. Four omnidirectional microphones placed on the surface of a rigid cylinder or four microphones with relatively large housings were used to create an acoustic shadow. Acoustic shadow causes level and spectral differences in the captured microphone signals, that depend on the sound direction and

its frequency. These differences are utilized in DirAC in two alternative methods to estimate the arrival direction of sound at high frequencies. The proposed PEG method uses the sound pressure and energy gradients of the microphone pairs. Another method uses VQ to estimate the direction of arrival of sound. Both methods were found to provide a reliable direction estimate even at high frequencies, where spatial aliasing affects detrimentally the direction estimate with traditional microphone arrays. This was corroborated by anechoic measurements and listening tests.

Publication VI: In this work, DirAC processing is applied to bilaterally-fitted hearing aids, where the direction and diffuseness of sound are estimated from two microphones in the hearing aid device at each ear. Subsequently, a single microphone signal from each device is amplified or attenuated at each time-frequency position to filter sound depending on the estimated direction. The estimated diffuseness is used to suppress ambient sounds. Finally, the modified microphone signal of each device is reproduced to a listener. DirAC processing with the hearing aid signals was evaluated with listening tests in which the speech reception threshold (SRT) was measured for various speech signal and noise configurations. The tests corroborated that DirAC improves significantly the SRT, when compared to the situation where the hearing aid signals were reproduced directly to the listener.

Bibliography

[1] J. Blauert, "Spatial hearing: The psychophysics of human sound localization," *The MIT press, Cambridge, MA, USA*, 1983.

[2] B.C.J. Moore, "Introduction to the psychology of hearing," *Academic Press*, 4th edition ed., 1997.

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

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

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

[6] E. Zwicker, G. Flottorp, and S.S. Stevens, "Critical band width in loudness summation," *J. Acoust. Soc. Am.*, vol. 29, no. 5, pp. 548–557, May 1957.

[7] R.D. Patterson, I. Ninno-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," *Technical report 2341. Applied Psychology Unit*, vol. Tech. Rep 2341, 1988.

[8] M. Slaney, "An efficient implementation of the Patterson-Holdsworth filter bank," *Apple Computer Inc.*, vol. Tech. Rep. 35, 1993.

[9] A. Härmä and K. Palomäki, "HUTear - a free Matlab toolbox for modeling of human auditory system," *in Proc. Matlab DSP Conf.*, pp. 96–99, November 1999, Available at: http://www.acoustics.hut.fi/software/HUTear.

[10] D.R. Perrott, "Discrimination of the spatial distribution of concurrently active sound sources: Some experiments with stereophonic arrays," *J. Acoust. Soc. Am.*, vol. 76, no. 6, pp. 1704–1712, December 1984.

[11] Lord Rayleigh, "On our perception of sound direction," *Philosophical magazine*, vol. 13, no. 74, pp. 214–232, 1907.

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

[13] T.N. Buell and E.R. Hafter, "Discrimination of interaural differences of time in the envelopes of high-frequency signals: Integration times," *J. Acoust. Soc. Am.*, vol. 84, no. 6, pp. 2063–2066, December 1988.

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

[15] H. Möller, M.F. Sörensen, D. Hammershöi, and C.B. Jensen, "Head-related transfer functions of human subjects," *J. Audio Eng. Soc*, vol. 43, no. 5, pp. 300–321, May 1995.

[16] H. Wallch, E.B. Newman, and M.R. Rosenzweig, "The precedence effect in sound localization," *The American Journal on Psychology*, vol. 62, pp. 315–336, July 1949.

[17] H. Haas, "The influence of a single echo on the audibility of speech," *J. Audio Eng. Soc*, vol. 20, no. 2, pp. 146–159, March 1972.

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

[19] A.W. Bronkhorst, "The cocktail part phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–127, January/February 2000.

[20] E. Zwicker and H. Fastl, "Psychoacoustics: Facts and models," *Springer-Verlag*, Berlin Heidelberg, 1990.

[21] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Am.*, vol. 116, no. 5, pp. 3075–3089, November 2004.

[22] B. B. Bauer, "Phasor analysis of some stereophonic phenomena," *J. Acoust. Soc. Am.*, vol. 33, no. 11, pp. 1536–1539, November 1961.

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

[24] J.C. Bennett, K. Parker, and F.O. Edeko, "A new approach to the assessment of stereophonic sound system performance," *J. Audio Eng. Soc*, vol. 33, no. 5, pp. 314–321, May 1985.

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

[26] V. Pulkki, "Spatial sound generation and perception by amplitude panning techniques," *PhD thesis, Helsinki University of Technology*, 2001, Available at: http://lib.tkk.fi/Diss/2001/isbn9512255324/.

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

[28] A. Blumlein, "British patent specification 394,325: Directional effect in sound systems," *Reprinted in:* Stereophonic techniques*, Audio Eng. Soc*, vol. 6, no. 2, pp. 91–98, April 1986.

[29] S.P. Lipshitz, "Stereo microphone techniques: Are the purists wrong?," *in Proc. AES 78th Convention*, Anaheim, CA, USA, 1985.

[30] R. Streicher and W. Dooley, "Basic stereo microphone perspectives - a review," *J. Audio Eng. Soc*, vol. 33, no. 7/8, pp. 548–556, July/August 1985.

[31] J. Eargle, "The microphone book," *Focal Press*, Boston, USA, 2001.

[32] W. Dooley and R. Streicher, "M-S stereo: A powerful technique for working in stereo," *J. Audio Eng. Soc*, vol. 30, no. 10, pp. 707–718, October 1982.

[33] M. Hibbing, "XY and MS microphone techniques in comparison," *J. Audio Eng. Soc*, vol. 37, no. 10, pp. 823–831, October 1989.

[34] E. Benjamin and T. Chen, "The native B-format microphone: Part I," *in Proc. AES 119th Convention*, New York, NY, USA, 2005.

[35] E. Benjamin and T. Chen, "The native B-format microphone: Part II," *in Proc. AES 120th Convention*, Paris, France, 2006.

[36] K. Farrar, "Soundfield microphone: Design and development of microphone and control unit," *Wireless World*, vol. 85, pp. 48–50, November 1979.

[37] K. Farrar, "Soundfield microphone: Detailed functioning of control unit," *Wireless World*, vol. 85, pp. 99–103, November 1979.

[38] M.A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc*, vol. 21, no. 1, pp. 2–10, February 1973.

[39] A. Solvang, "Spectral impairment for two-dimensional higher order Ambisonics," *J. Audio Eng. Soc*, vol. 56, no. 4, pp. 267–279, April 2008.

[40] J. Merimaa, "Measurement, analysis, and visualization of directional room responses," *in Proc. AES 111th Convention*, New York, NY, USA, 2001.

[41] J. Merimaa, "Applications of a 3-D microphone array," *in Proc. AES 112th Convention*, Munich, Germany, 2002.

[42] M. Kallinger, G. D. Galdo, F. Kuech, D. Mahne, and R. Schultz-Amling, "Spatial filtering using Directional Audio Coding parameters," *in Proc. IEEE Int. Conf. on Acoust, Speech and Signal Processing*, vol. 2, pp. 217–220, 2009.

[43] M-V. Laitinen, F. Kuech, and V. Pulkki, "Using spaced microphones with Directional Audio Coding," *in Proc. AES 130th Convention*, London, UK, 2011.

[44] ITU-R BS.775-1, "Multichannel stereophonic sound system with and without accompanying picture," *International Telecommunication Union Radiocommunication Assembly*, 1992-1994.

[45] F. Rumsey, "Spatial audio," *Focal Press*, Oxford, 2001.

[46] A. Fukada, "A challenge in multichannel music recording," *in Proc. AES 19th Int. Conf.*, 2001.

[47] S. Paul, "Binaural recording technology: A historical review and possible future developments," *Acta Acustica United with Acustica*, vol. 95, no. 5, pp. 767–788, September/October 2009.

[48] V.R. Algazi, C. Avendano, and D. Thompson, "Dependence of subject and measurement position in binaural signal acquisition," *J. Audio Eng. Soc*, vol. 47, no. 11, pp. 937–947, November 1999.

[49] H. Möller, M.F. Sörensen, C.B. Jensen, and D. Hammershöi, "Transfer characteristics of headphones measured on human ears," *J. Audio Eng. Soc*, vol. 43, no. 4, pp. 203–217, January 1995.

[50] F. Baumgarte and C. Faller, "Binaural Cue Coding - part I: Psychoacoustic fundamentals and design principles," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 509–519, November 2003.

[51] C.Faller and F. Baumgarte, "Binaural Cue Coding - part II: Schemes and applications," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 520–531, November 2003.

[52] M. Goodwin and J-M. Jot, "Analysis and synthesis for Universal Spatial Audio Coding," *in Proc. AES 121st Convention*, San Francisco, CA, USA, 2006.

[53] J. Herre, K. Kjörling, J. Breebart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Röden, 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, no. 11, pp. 932–955, November 2008.

[54] V. Pulkki, "Spatial sound reproduction with Directional Audio Coding," *J. Audio Eng. Soc*, vol. 55, no. 6, pp. 503–516, June 2007.

[55] J. Vilkamo, T. Lokki, and V. Pulkki, "Directional Audio Coding: Virtual microphone-based synthesis and subjective evaluation," *J. Audio Eng. Soc*, vol. 57, no. 9, pp. 709–724, September 2009.

[56] M-V. Laitinen and V. Pulkki, "Binaural reproduction for Directional Audio Coding," *in Proc. IEEE Workshop on Applications of Sig. Proc. to Audio and Acoust.*, New Paltz, NY, USA, 2009.

[57] J. Ahonen, V. Pulkki, and T. Lokki, "Teleconference application and B-format microphone array for Directional Audio Coding," *in Proc. AES 30rd Int. Conf.*, Saariselka, Finland, 2007.

[58] J. Merimaa and V. Pulkki, "Spatial Impulse Response Rendering I: Analysis and synthesis," *J. Audio Eng. Soc*, vol. 53, no. 12, pp. 1115–1127, December 2005.

[59] V. Pulkki and J. Merimaa, "Spatial Impulse Response Rendering II: Reproduction of diffuse sound and listening tests," *J. Audio Eng. Soc*, vol. 54, no. 1/2, pp. 3–20, February 2006.

[60] J. Merimaa, "Analysis, synthesis, and perception of spatial sound – binaural localization modeling and multichannel loudspeaker reproduction," *PhD thesis, Helsinki University of Technology*, 2006,
Available at: http://lib.tkk.fi/Diss/2006/isbn9512282917/.

[61] F.J. Fahy, "Sound intensity," *Elsevier Science Publishers Ltd.*, Essex, England, 1989.

[62] J. Ahonen, V. Pulkki, F. Kuech, G. Del Galdo, M. Kallinger, and R. Schultz-Amling, "Directional Audio Coding with stereo microphone input," *in Proc. AES 126th Convention*, Munich, Germany, 2009.

[63] M.A. Ericson and R.L. McKinley, "The intelligibility of multiple talkers separated spatially in noise," *in Binaural and Spatial Hearing in Real and Virtual Environments, edited by R.H. Gilkey and T.R. Anderson*, pp. 701–724, Lawrence Erlbaum Associates, Mahwah, NJ, 1997.

[64] D.R. Begault, "Virtual acoustic displays for teleconferencing: Intelligibility advantage for telephone-grade audio," *J. Audio Eng. Soc*, vol. 47, no. 10, pp. 824–828, October 1999.

[65] M.M. Boone and W. de Bruijn, "Improving speech intelligibility in teleconferencing by using wave field synthesis," *in Proc. AES Convention*, Amsterdam, The Netherlands, 2003.

[66] J.J. Baldis, "Effects of spatial audio on memory, comprehension, and preference during desktop conferences," *in Proc. SIGCHI conf. on Human factors in computing systems*, pp. 166–173, 2001.

[67] T. Hirvonen, J. Ahonen, and V. Pulkki, "Perceptual compression methods for metadata in Directional Audio Coding applied to audiovisual teleconference," *in Proc. AES 126th Convention*, Munich, Germany, 2009.

[68] ITU-T Recommendation, "G.722: 7 kHz audio-coding within 64 kbit/s," *International Telecommunications Union*, Geneva, Switzerland, 1988, Available at: http://www.itu.int/rec/T-REC-G.722-198811-I/en.

[69] C. Faller, "Parametric coding of spatial audio," *PhD thesis, École Polytechnique Fédérale de Lausanne*, 2004.

[70] C. Faller, "Microphone front-ends for spatial audio coders," *in Proc. AES 125th Convention*, San Francisco, CA, USA, 2008.

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

[72] M. Cobos, J. Lopez, and S. Spors, "A sparsity-based approach to 3-D binaural sound synthesis using time-frequency array processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, Article ID 415840, February 2010.

[73] M. Cobos, S. Spors, J. Ahrens, and J. Lopez, "On the use of small microphone arrays for wave field synthesis auralization," *in Proc. AES 45th Int. Conf.*, Helsinki, Finland, 2012.

[74] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *J. Acoust. Soc. Am.*, vol. 93, no. 5, pp. 2764–2778, September 1993.

[75] S. Berge and N. Barrett, "High angular resolution planewave expansion," *in Proc. 2nd Int. Symposium on Ambisonics and Spherical Acoustics*, Paris, France, 2010.

[76] S. Berge and N. Barrett, "A new method for B-format to binaural transcoding," *in Proc. AES 40th Int. Conf.*, Tokyo, Japan, 2010.

[77] J. Jakka, "Binaural to multichannel audio upmix," *Master's thesis, Helsinki University of Technology*, 2005, Available at: http://otalib.aalto.fi/en/collections/e-publications/diplomityot/1996-2005/.

[78] M. Takanen and M.Karjalainen, "Real-time tracking of speech sources using binaural audio and orientation tracking," *in Proc. AES 40th Int. Conf.*, Tokyo, Japan, 2010.

[79] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, and G. Lorho, "Augmented reality audio for mobile and wearable appliances," *J. Audio Eng.Soc.*, vol. 52, no. 6, pp. 618–639, June 2004.

[80] H. Buchner and W. Kellermann, "Acoustic echo cancellation for surround sound using perceptually motivated convergence enhancement," *in Proc. Int. Workshop on Acoust. Echo and Noise Control*, 2001.

[81] J. Herre, H. Buchner, and W. Kellermann, "Acoustic echo cancellation for two and more reproduction channels," *in Proc. IEEE Int. Conf. on Acoust, Speech and Signal Processing*, pp. 17–20, 2007.

[82] A. Gersho and R.M. Gray, "Vector quantization and signal compression," *Kluwer Academic Publishers*, 1992.

Aalto-DD 33/2013

ISBN 978-952-60-5035-5
ISBN 978-952-60-5036-2 (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