Capturing and reproducing realistic acoustic scenes for hearing research

Marschall, Marton; Dau, Torsten; MacDonald, Ewen; Buchholz, Jörg

Publication date:
2014

Document Version
Publisher's PDF, also known as Version of record

Technical University of Denmark

CONTRIBUTIONS TO
HEARING RESEARCH

Volume 18

Márton Marschall

Capturing and reproducing realistic acoustic scenes for hearing research

Centre for Applied Hearing Research
Department of Electrical Engineering

Capturing and reproducing realistic acoustic scenes for hearing research

PhD thesis by
Márton Marschall

Technical University of Denmark

2014

Cover photo by Nicolas Le Goff.
The defense was held on November 6, 2014.

This PhD dissertation is the result of a research project carried out at the Centre for Applied Hearing Research, Department of Electrical Engineering, Technical University of Denmark.

The project was partly financed by the Technical University of Denmark (2/3) and partly by Siemens Audiology Solutions (1/3).

Supervisors

Prof. Torsten Dau
Assoc. Prof. Ewen MacDonald
Centre for Applied Hearing Research
Department of Electrical Engineering
Technical University of Denmark
Kgs. Lyngby, Denmark

Dr. Jörg Buchholz
Department of Linguistics
Macquarie University
Sydney, Australia

Abstract

Accurate spatial audio recordings are important for a range of applications, from the creation of realistic virtual sound environments to the evaluation of communication devices, such as hearing instruments and mobile phones. Spherical microphone arrays are particularly well-suited for capturing spatial audio in three dimensions. However, practical constraints limit the number of microphones that can be used and thus the maximum spatial resolution and frequency bandwidth that can be achieved. Further, most important sound sources are near the horizontal plane, where human spatial hearing is also most accurate. This thesis therefore investigated whether the horizontal performance of spherical microphone arrays could be improved (i) through an appropriate placement of a fixed number of transducers on the sphere, and (ii) by applying mixed-order ambisonics (MOA) processing. MOA combines higher-order ambisonics (HOA) with additional, horizontally oriented spherical harmonic functions of higher orders. Simulations of a MOA array, with a higher density of microphones near the equator, and an array with a nearly uniform distribution of microphones were compared in terms of spatial resolution and robustness. A MOA array was constructed, and some of the simulation results were validated with measurements. Results showed that for MOA, the spatial resolution was improved for horizontal sources at mid to high frequencies and the robustness to noise and measurement errors was similar to that of HOA. The properties of MOA microphone layouts and processing were investigated further by considering several order combinations. It was shown that the performance for horizontal vs. elevated sources can be adjusted by varying the order combination, but that a benefit of the higher horizontal orders can only be seen at mid to high frequencies as the need for regularization limits spatial directivity at lower frequencies. Finally, the MOA array was also evaluated in terms of sound field reconstruction error in a head-sized region. Results provided a physical validation of the functioning of the MOA microphone array and further showed that the MOA approach results in a somewhat larger “sweet area” for horizontal sources than for elevated sound sources. While the focus was on the technical evaluation of the developed MOA system, potential perceptual effects concerning MOA and microphone array recordings in general are also discussed. The system developed in this work provides new possibilities for the investigation of human perception in realistic and complex acoustic environments.

Resumé

Præcise rumlige lydoptagelser er vigtige i en række af anvendelsesområder, fra skabelsen af realistiske virtuelle lydmiljøer til evalueringen af kommunikationsapparater såsom høreapparater og mobiltelefoner. Sfæriske mikrofon-arrays er særligt velegnede til at optage rumlig lyd i tre dimensioner. Praktiske problemstillinger begrænser dog antallet af mikrofoner der kan anvendes, hvilket begrænser den maksimale rumlige opløsning samt frekvens-båndbredden der kan opnås. Ydermere befinder de fleste vigtige lydkilder sig typisk nær det horisontale plan, hvor den menneskelige rumlige hørelse også er mest nøjagtig. Denne afhandling undersøgte derfor hvorvidt den horisontale ydeevne af sfæriske mikrofon-arrays kunne forbedres (i) ved en passende placering af et givent antal af mikrofoner på kuglen, og (ii) ved at anvende mixed-order ambisonics (MOA) processering. MOA kombinerer højere-ordens ambisonics (HOA) med yderligere, horisontalt orienterede sfæriske harmoniske funktioner af højere orden. Simulationer af et MOA-array med en højere densitet af mikrofoner nær ækvator, samt et array med en nær uniform fordeling af mikrofoner blev sammenlignet i form af rumlig opløsning og robusthed. Et MOA-array blev konstrueret, og en del af simulationerne blev valideret med fysiske målinger. Resultaterne for MOA viste, at den rumlige opløsning var forbedret for horisontalt placerede lydkilder ved mellem og høje frekvenser, og at robustheden overfor støj og målefejl var sammenlignelig med robustheden ved HOA. Egenskaberne ved MOA mikrofonplaceringen og processeringen blev yderligere undersøgt ved at betragte flere kombinationer af forskellige ordener. Det blev vist, at ydeevnen for horisontale lydkilder i forhold til eleverede lydkilder kan justeres ved at variere kombinationerne af ordenerne, men at fordelen ved højere horisontale ordener kun ses ved mellem og høje frekvenser, da behovet for regularisering begrænser den rumlige direktivitet ved lave frekvenser. Slutteligt blev MOA-arrayet også evalueret i forhold til rekonstruktionen af lydfelter i et område svarende til størrelsen på et hoved. Resultaterne gav en fysisk validering af funktionaliteten af MOA mikrofon-arrayet og viste yderligere, at MOA-tilgangsvinklen giver et noget større område med en god reproduktion af lydfelter for horisontale lydkilder end for lydkilder med eleverede placeringer. Selv om det primære fokus var på den tekniske evaluering af det udviklede MOA system, er potentielle perceptuelle effekter af MOA og optagelser ved hjælp af mikrofon-arrays generelt også diskuteret. Systemet, som blev udviklet i forbindelse med dette projekt, giver nye muligheder for at undersøge menneskelig perception i realistiske og komplekse akustiske miljøer.

Acknowledgments

First and foremost, I would like to thank my main supervisor, Torsten Dau, for his unwavering support, inspiration, and guidance. I'm also indebted to my co-supervisors and colleagues that I've had the chance to work closely with during this project: Jörg Buchholz, Sylvain Favrot, Jiho Chang and Ewen MacDonald. I would like to acknowledge the assistance of our project partners Siemens Audiology Solutions and Brüel & Kjær, and of Richard Schultz-Amling and Wookeun Song in particular. Finally, I would like to thank my family, friends and colleagues for all the great times in and out of office, at home and around the world.

Contents

1

General introduction

1.1 A cocktail party: from life to lab

One of the central goals in hearing research is to better understand the "cocktail-party effect": how listeners are able to extract information about their environment from a complex acoustic scene around them, and how they are able to attend to specific sources in that environment. The healthy auditory system has a remarkable ability to focus on and process information in adverse acoustic conditions. Hearing-impaired listeners, on the other hand, often report great difficulties in understanding speech in everyday conditions, where many sound sources and reverberation are encountered. These difficulties in complex, noisy situations frequently persist even when these listeners have been fitted with hearing aids, and also affect cochlear implant users. One of the biggest challenges faced by hearing instrument manufacturers is improving speech intelligibility and the users' awareness of their surroundings in dynamic, complex conditions.

The ability to test listeners in realistic acoustic environments in a repeatable and controlled manner could provide new insights into the processing strategies employed by the auditory system in adverse conditions, and could benefit the development and testing of advanced hearing instruments. However, studies of listener performance in more realistic environments have so far been held back by the complexity of running lengthy field tests and, until recently, by the difficulty of reproducing acoustic scenes in a realistic and repeatable manner in the laboratory. To address this problem, recent work at the Centre for Applied Hearing Research (CAHR) and elsewhere has focused on the development and application of virtual sound environments (VSEs) for hearing research (e.g. Minnaar et al., 2010; Seeber et al., 2010; Grimm et al., 2014). At CAHR, a loudspeaker-based virtual sound environment ("Spacelab") was installed, and a room-auralization framework, the LoRA toolbox, was developed by Favrot and Buchholz (2010), allowing a realistic playback of simulated room acoustics.

However, scenes with dynamic, spatially extended, or a large number of sound sources are generally difficult to simulate. Moreover, if the goal is to investigate a specific, existing room or space, it can be difficult and time consuming to obtain the necessary parameters and to fine-tune the acoustic model for an accurate reproduction (Bork, 2005; Cubick, 2011). Limitations of the various acoustic models, such as the number of reflections considered, or the way late reverberation is simulated can also impact the perception of the simulated rooms (Zahorik, 2009). Thus, it would be advantageous to be able to capture the acoustic space, or to record the spatial acoustic scene directly. The work presented in this thesis addresses

the development and evaluation of a microphone array system that allows the recording of dynamic, real-life acoustic scenes.

In order to record and reproduce spatial sounds, it is necessary to spatially sample the sound field, which is usually achieved with an array of microphones. Spherical microphone arrays, in particular, have the advantage that they can capture the spatial properties of the sound field in all directions around the array. Spherical arrays have seen increasing use in diverse applications from spatial audio (Abhayapala and Ward, 2002; Meyer and Elko, 2004), beamforming (Meyer and Elko, 2002; Li and Duraiswami, 2007) to room acoustic measurements (Gover et al., 2004; Park and Rafaely, 2005). As the objective is to recreate an acoustic scene around a single listener in a virtual environment, the use of such a spherical microphone array and the higher order ambisonics (Daniel, 2000; Moreau et al., 2006) technique is proposed in this work. The approach is based on a spherical harmonic decomposition of the sound field, and has the advantage that reproduction errors are minimized within the "sweet spot" around the listeners head.

With regards to the configuration of the array, a new approach has been taken. Instead of using a uniform distribution of microphones on the sphere, the microphones are arranged in a way that better matches the capabilities of human hearing: providing maximum spatial resolution in the horizontal plane, while still recording vertical (height) information. This is achieved by placing more microphones near the equator, and employing a matching mixed-order ambisonics scheme. Mixed-order ambisonics allows for a fully three-dimensional sound field to be reproduced around the listener, but with more horizontal detail where most real-life sound sources are situated, and where spatial hearing is most acute (Blauert, 1997b).

The term "acoustic scene" is used in this context to denote the physical sound field around the listener created by the various sound sources and the acoustic space, as opposed to the "auditory scene", which already implies analysis (i.e. source segregation, streaming) by the auditory system (Bregman, 1994). It follows that the main aim in this work has been to preserve the physical properties of the recorded scene as closely as possible.

In order to investigate spatial perception itself, ideally one must produce a sound field that is physically "correct", and does not rely on perceptual effects to position sound sources. Any departure from a purely physically-based reproduction introduces a new variable, whose potential effect on the measurement in question must be carefully evaluated. Similarly, technical devices, such as hearing aids or mobile phones with multiple microphones and beamforming algorithms, can only be expected to behave as designed in realistic sound fields. However, it is clear that with current technologies, it is not possible to capture and reproduce a sound field with a high enough accuracy and a large enough bandwidth that perceptual effects can be neglected (Spors et al., 2013). Although this comprises future work, these limitations underscore the importance of perceptually evaluating spatial audio systems. Nonetheless, accurate reproduction can be obtained for a more restricted range in frequency and space, and it is therefore essential before any perceptual evaluation to thoroughly characterize the physical limitations of the applied methods. These physical limitations are investigated in detail in this work.

1.2 Overview of the thesis

The thesis comprises 6 chapters and an appendix. Some of the chapters are based on already published or submitted manuscripts, as indicated in each chapter. An overview of the topics covered by the various chapters is presented below.

Chapter 2 provides a brief overview of spatial hearing, as well as spatial audio recording and reproduction techniques applicable to virtual sound environments, such as binarual synthesis, wave field synthesis (WFS), higher-order ambisonics (HOA) and directional audio coding (DirAC). An emphasis is placed on HOA, which serves as a basis for the methods applied in this work.

Chapter 3 introduces the concept of mixed-order ambisonics (MOA), and presents the design and evaluation of a MOA microphone array. The goal set out in this study was to design and evaluate a microphone array capable of capturing a fully three-dimensional representation of the sound field, but with more detail in the horizontal plane. Simulations of a MOA array, featuring an uneven distribution of microphones over the sphere, were compared with an HOA array (e.g. as proposed by Daniel et al., 2003) with a nearly-uniform microphone layout. The comparison was made in terms of physical measures related to the array output, adapted from literature on beamforming (e.g. Li and Duraiswami, 2007), as well as in terms of measures of the reproduced sound field introduced by Gerzon (1992). Some of the existing metrics had to be adapted to the mixed-order approach due to the elevation dependence of array performance. An effort was made to provide realistic simulations of transducer self-noise and characteristic variations, in order to predict the real-life performance of the microphone arrays. Finally, a prototype mixed-order array was constructed, and some of the simulated performance metrics were validated with measurements.The study demonstrates the feasibility of using MOA for microphone arrays, but also highlights that the advantages of the technique are restricted to higher frequencies, due to limitations imposed by the need for regularization.

The study presented in **Chapter 4** follows the approach presented in the previous chapter, but now investigates the impact of using different order combinations in mixed-order ambisonics. An appropriate microphone layout for each order combination is derived using a method that distributes rings of microphones on a sphere. The aim of this study was to investigate how performance for horizontal vs. elevated sources can be adjusted by changing the order combination and applying a matching array layout.

So far, the focus has been on differentiating between horizontal and vertical performance characteristics of the microphone array. The size of the “sweet area”, in which the sound field is reproduced with low error, was not explicitly considered. In **Chapter 5**, this is investigated by evaluating the error of the reproduced sound field as a function of the distance from the origin, using both simulations and measurements of the MOA microphone array. The effects of deviations from ideal transducer characteristics are also considered. The error of the reproduced sound field is of interest, as the listener’s head or the device under test should ideally be encompassed by the “sweet area” in the virtual environment. It is a property of ambisonics that the error is a function of the product of the radius and the frequency. For a

fixed radius, like the size of a listener's head, the error increases with frequency, and provides an upper frequency limit for physically correct reconstruction.

In **Chapter 6**, the main findings of the thesis are summarized. The advantages and limitations of the proposed array system, as well as directions for further improvement are considered, with a focus on potential perceptual effects. The perspectives for the application of the array are discussed.

An additional project concerned with a more fundamental aspect of auditory perception, namely frequency selectivity, was also carried out and forms a part of this thesis. In addition to spatial cues, the frequency selective properties of the auditory system play a crucial role in its ability to segregate one sound source from another, and thus in making sense of a complex acoustic environment. Specifically, the nonlinear behavior of the cochlea may play a key role in this respect. However, since the results of this investigation were not directly related to the main topic of this thesis, they are presented in the appendix.

Chapter A investigates two approaches of estimating the bandwidth of nonlinear auditory filters using a very simple, schematic auditory model. It is shown that, depending on the structure of the model, the interaction between a compressive nonlinearity and the bandwidth estimation method can lead to different estimates of the auditory filter bandwidth. This has implications on the interpretation of psychophysical estimates of human auditory frequency selectivity derived from simultaneous and forward masking experiments. In **Chapter B** a more complete auditory model was applied to investigate and further analyze differences in frequency selectivity between the two masking paradigms that have been reported in the literature. Specifically, it was tested in model whether the nonlinear phenomenon of suppression can explain the observed frequency selectivity differences.

2 Background

2.1 Virtual sound environments

A virtual sound environment attempts to create an acoustic scene that the listener perceives as a convincing (i.e. authentic or plausible) auditory scene. The exact methods employed can vary greatly, but the basic approach is that a representation of the acoustic scene (whether simulated or recorded) is processed and presented to the listener, either through loudspeakers or headphones, eliciting some of the auditory cues related to spatial hearing. The methods can be differentiated broadly based on (i) how the acoustic scene is represented, (ii) the techniques with which the scene is presented to the listener, and (iii) the overall goals in terms of fidelity, i.e whether the focus is authenticity, plausibility or artistic effect.

The first point is most closely related to how the acoustic scene itself is constructed. If the scene is simulated, it may be represented as a collection of virtual sources in a virtual acoustic space, with various parameters, such as the dimensions of the room and the location of the sources. If the scene is recorded directly, it is usually represented as multi-channel sound data, where the channels may correspond directly to loudspeaker signals, or to elements in a mathematical representation of the sound field, such as higher-order ambisonics.

The distinction between simulation and recording is not exclusive, as the acoustic scene can also be constructed by combining simulated and recorded elements, such as when anechoic recordings of real sources are placed in a simulated acoustic space, or when a recording in a real space is modified by the recording engineer by mixing various recorded signals. The latter is the case for many recordings produced for artistic or entertainment purposes.

The second point refers to the techniques used to reconstruct and deliver the acoustic signal to the ears of the listener. The main options here are delivery through headphones, requiring only two channels, and loudspeaker-based systems using anything from two up to tens or even hundreds of channels. The specific spatial audio techniques, such as binaural synthesis, wave-field synthesis or higher-order ambisonics determine the signals that the headphones or loudspeakers need to be driven with to recreate the desired scene.

Regarding the final point, it is important to distinguish between different possible goals of a virtual sound environment in terms of fidelity with regard to a real or simulated reference. At one extreme, full physical fidelity, the goal is to have the same acoustic signal enter the listener's ears as if he or she had been present in the reference environment. This is the general aim of sound field synthesis techniques like WFS and HOA. In practice, full physical fidelity is only realizable with significant limitations, either regarding the scene itself (i.e. anechoic

conditions, limited source positions) or with respect to the frequency range. In these cases, the potential perceptual effects of the errors in the acoustic signals need to be evaluated.

Another possibility is to give up on the demand of reproducing the acoustic signal exactly, and instead attempt to recreate a scene that is perceptually indistinguishable from, or at least similar to, the reference scene. A downside of this approach is that, even if human perception is not changed substantially in a given scenario, technical devices may not behave in the expected way in the virtual environment, as the physical sound field can be significantly different. Directional Audio Coding (DirAC; Pulkki, 1997) is an example of a technique that aims at perceptual fidelity.

Further down the spectrum, approaches exist without the goal of matching an explicit reference scene, rather aiming at presenting a plausible auditory scene to the listener. Most spatial audio systems used for music and entertainment, such as typical cinema surround systems, fall under this category. Finally, at the other extreme, it may not be desirable to present realistic or plausible auditory scenes at all, if the goal is, for example, to convey non-speech information through auditory means (e. g. data sonification), or to present imagined or artificial sounds for artistic effect.

In practice it is often difficult to evaluate perceptually-based methods in terms of fidelity, due to the difficulty of providing a direct comparison with an appropriate reference scene. The definition of quality attributes and attribute identification techniques (e. g. Rumsey, 2002; Berg and Rumsey, 2006) can aid in this case the subjective quality evaluation of spatial audio systems, without the need for an explicit reference. In contrast, comparing repeatable technical measures of the reference and reproduced sound fields is much more straightforward, and has been – until recently – the primary method of evaluating sound field synthesis systems, whose primary goal has been full physical fidelity. Nonetheless, successful approaches exist aiming both at physical (e.g. WFS, HOA) and perceptual (e.g. DirAC) fidelity.

The aim of the approach in this work has primarily been to preserve physical fidelity, while attempting to optimize the impact of physical limitations in a way that better matches human spatial hearing. As the “receiver” is ultimately a human listener, in the following, a short overview of the properties of human spatial hearing is provided.

2.2 Spatial hearing

Although vision is arguably our primary sense used for navigating in space, the spatial information gathered by the auditory system is crucial in directing attention and providing a sense of awareness of our surroundings, especially in directions outside of the visual field, whether in real or virtual environments (Blauert, 1997a; Shilling and Shinn-Cunningham, 2002). Additionally, spatial information has been shown to enhance the ability to focus on a specific source among a multitude of sound sources, termed the “cocktail-party” effect (e. g. Hawley et al., 2004).

Thus, one of the primary goals of a convincing virtual sound environment is to present spatial information to the listener, through the activation of one or more of the spatial hearing

mechanisms of the auditory system. This section provides a brief overview of the major auditory cues involved in spatial hearing. A more detailed review is provided by e.g. Akeroyd (2006), and an exhaustive treatment of the the subject is given by Blauert (1997b).

The human auditory system is able to extract spatial information both from a single ear (monaural cues), and by exploiting the differences between the signals reaching the two ears (binaural cues). Localization refers to the ability to assign a spatial location to a sound source. Localization in the horizontal plane is primarily based on interaural time and level differences (ITDs and ILDs). ITDs are caused by a difference in time of arrival to the ears for sound sources that are located away from the median plane, in the range of about 0 to 600 μs, depending on the azimuth angle of the source. For continuous sounds, these time differences are translated to phase differences. For pure tones, sensitivity to ITDs diminishes above about 1.5 kHz, but for wideband signals, envelope ITDs can be detected also at higher frequencies. In contrast, ILDs are detectable in a wide frequency range, but for distant sources the head only has a significant shadowing effect above about 1 kHz (maximum of 10-20 dB, increasing with frequency). The auditory system seems to use both ITDs and ILDs for localization, with ITDs dominating at lower frequencies, at least in quiet (Wightman and Kistler, 1992). For close sources (with a distance below 1 m), however, large ILDs can occur at lower frequencies due to the large difference in distance to the source between ears (Shinn-Cunningham et al., 2000). Thus, ILDs serve as an important distance cue for close sources.

Elevation-dependent spectral cues introduced by the outer ear allow localization in the vertical direction. These cues occur as high frequency peaks and notches between about 4 and at least 17 kHz (Hebrank and Wright, 1974; Algazi et al., 2001b). The exact spectral configuration is highly dependent on the shape and size of the listener's ears and head, which introduces significant individual differences in the measured transfer functions of the outer ear (Algazi et al., 2001b). Low-frequency spectral cues (below 3 kHz) due to reflections from the shoulders and torso may also aid the auditory system in estimating elevation (Algazi et al., 2001a).

The localization performance of the auditory system varies with angle, and is best in front of the listener, with the minimum audible angle (MAA) of about 1° in azimuth and 4° in elevation (Perrott and Saberi, 1990). Localization performance is thus better in the horizontal than vertical directions for sources close to the median plane, but horizontal localization degrades towards the sides. Makous and Middlebrooks (1990) found localization errors of between 2° and 20°, with better horizontal localization in front of the subject, but better vertical localization for more peripheral source positions.

In contrast to localization, distance perception in humans is relatively inaccurate, with listeners typically underestimating distance (Zahorik et al., 2005). The main acoustic cue in anechoic environments appears to be intensity and high-frequency attenuation, whereas in reverberant environments the direct-to-reverberant energy ratio provides a robust distance cue. An attribute related to distance perception is externalization, or the perception of a sound source being outside the head. Normally, real-life sound sources are perceived as externalized, but in certain cases, especially when listening through headphones, sounds can be perceived

as being inside the head. In reverberant environments, dynamic ILD fluctuations seem to be involved in the perception of externalization and distance (Catic et al., 2013).

Other spatial attributes that have been linked to specific cues include spaciousness (the perceived size of the environment) and apparent source width (the spatial extent of a sound source), both of which appear to be associated with the amount of correlation between the signals reaching the two ears (interaural cross-correlation; IACC) (Blauert, 1997b).

Overall, the localization ability of the auditory system is quite robust, and it will use whatever cues or combination of cues that are available in noisy conditions (Akeroyd, 2006). This is an advantage for spatial audio applications, as many systems will work "well enough" in practice even if coherent spatial cues are only provided in a limited frequency range (Spors et al., 2013).

In addition to providing information about the acoustic environment, binaural hearing can aid in the separation of competing sound sources. Detection levels of signals masked by a noise are lowered by the introduction of one or more cues corresponding to a spatial separation between them. This effect is known as binaural masking level difference (BMLD) or spatial release from masking, and can provide a gain in detectability of up to about 15 dB (Akeroyd, 2006). For speech intelligibility tasks, improvements of up to 6–15 dB in speech reception thresholds have been reported, with higher improvement in cases where the non-spatial features of the target and masker are otherwise similar (Shinn-Cunningham, 2002).

2.3 The acoustic scene

2.3.1 Representing the acoustic scenes

Acoustic scenes can be represented in a number of ways, combining both measured and simulated elements. This section provides a brief overview of the various ways in which an acoustic scene can be captured or constructed, and highlights the advantages and drawbacks of the different approaches.

The acoustic scene is defined here as a collection of sound sources in an acoustic environment. The acoustic environment represents the transmission path from the source to the listener, which is typically a room, but can also be an outdoor location, or a smaller enclosure, such as a car.

First, the representation of the sound source will be considered. To help differentiate between approaches, it is useful to separate the concept of a sound source further into a combination of (i) an audio signal, and (ii) the spatial attributes of the source. The audio signal corresponds to the sound emitted by the source, but without spatial information. This can be, for example, a time signal from a physical sound generation model, like a physical model of a plucked string or a speech synthesizer. More often, a single-channel anechoic recording of a sound source is used as the audio signal. The spatial attributes of the source in turn determine how the audio signal is emitted and interacts with the acoustic environment, potentially in a frequency-dependent way. The list of spatial attributes depend on the complexity of the

model, but include the position (possibly changing) and the spatial directivity of the source. The directivity can be specified by a simple source model, such as a plane wave or point source, or specified explicitly, based on a more complex source model or measured data. For instance, surrounding microphone arrays can capture sound sources with complex spatial directivities (Zotter, 2009).

For most practical purposes, the acoustic environment as a transmission path can be considered as a linear, time invariant system (Jacobsen and Juhl, 2013). This means that the effect of the acoustic environment on the sound source at a receiver position is described by the impulse response for the source-receiver pair. This impulse response can be obtained from a model, or through a measurement of the acoustic space. The time invariance does not hold for a moving source (or a moving receiver), because the transmission path changes at each time instant. This introduces an additional difficulty in handling such sources, as the impulse response needs to be constantly updated.

Accurate simulations of the acoustic environment are usually based on a geometrical model of the space, including absorption and scattering coefficients of the reflecting surfaces. Commercial room acoustic simulation programs (e. g. ODEON; Christensen and Koutsouris, 2013) can provide realistic simulations of rooms, typically using a combination of image source and ray tracing methods in order to generate an impulse response for a specific source-receiver pair (Vorländer, 2008). Methods providing real-time room acoustic simulations for virtual environments have also been presented, with newer methods also handling dynamic scenes with a moving listener or source (e.g. Noisternig et al., 2008; Schröder and Vorländer, 2011; Grimm et al., 2014). In music and entertainment applications, often much simpler mathematical or signal processing models are used to generate "good sounding" reverberation, without attempting to match a specific, physical environment (Gardner, 2002).

Impulse responses of a real room or space can also be obtained by measurement. This entails placing a sound source, typically a loudspeaker, and one or more receivers (microphones) in the acoustic space. As one impulse response only describes the transfer path between one source-receiver pair, a detailed mapping of the room may require many measurements. These impulse responses can then be convolved with an anechoic audio signal, allowing a placement of a virtual source at the measured locations in the acoustic space. An important consideration with this approach is that the directivity of the virtual source will be determined by the directivity of the physical sound source (i.e. the loudspeaker) that was used to obtain the impulse response. If the intended directivity of the virtual source (e.g. a human speaker, or a musical instrument) is very different from that of the measurement device, the results may not be realistic. Compact loudspeaker arrays that allow the radiation of arbitrary directivities can be applied (Zotter, 2009) in this case.

A second consideration is the choice of method for capturing spatial sound: the recording microphones and their positions have to be chosen in accordance with the spatial audio reproduction technique that will be applied. Standard microphone arrangements using two or more microphones exist for stereo and surround sound techniques (Huber and Runstein,

2013), while for sound field synthesis methods, microphone arrays with tens or hundreds of positions need to be used.

Instead of just measuring impulse responses in the acoustic space, a complete scene, including both the sound sources and the environment, can be recorded directly using one of the microphone techniques described above. This affords the least flexibility, but allows the capture of moving sources, as well as extended or diffuse sources that cannot be handled with impulse-response based approaches.

2.3.2 Acoustic scenes for virtual environments

In terms of constructing realistic acoustic scenes for virtual environments, the three most relevant representations of the acoustic scene are considered to be: (i) simulated acoustic scene, with a virtual source placed in a room acoustic model, and convolution with an anechoic audio signal; (ii) measured acoustics with impulse responses in several source receiver pairs, virtual sources added through convolution with anechoic signals; and (iii) direct recording of the acoustic scene.

The first approach (representing the acoustic scene as a model) is most flexible. Within the limitations of the specific acoustic model, any room or space can be simulated, and the source and receiver positions can be chosen freely. A further advantage is that the model itself is not necessarily coupled to the reproduction method, and thus the output can be adapted to suit different playback systems. The major disadvantage is the need for detailed geometrical and acoustical data, and the considerable effort required to set up and fine tune the model, especially for very complex scenes. Further, all models employ simplifications of the underlying physics in order to provide a reasonable computation time (Vorländer, 2008). Small rooms and enclosures are typically not handled well by these algorithms. Thus, these models cannot be expected to provide a physically fully accurate simulation of a given room, but rather provide a good match for various room acoustic parameters, which in many cases may be sufficient. For example, Favrot and Buchholz (2010) showed that for auralizations using their loudspeaker-based room auralization (LoRA) system, parameters such as early decay and reverberation time, clarity, speech transmission index and interaural cross correlation were generally well maintained in the reproduced virtual environment. Moreover, comparing speech intelligibility in the real room, vs. an auralization using again the LoRA system, Cubick et al. (2013) found that while absolute intelligibility scores differed between the real and virtual rooms, the speech intelligibility benefit from a beamforming hearing-aid was similar. This highlights that while simulation-based virtual scenes may not be able to provide an exact match to a specific, real environment, they can be useful for evaluating hearing instruments, for instance.

The advantage of the second approach (measuring impulse responses) is that the acoustics of the real location are directly captured. It is thus straightforward to virtualize a specific space. The response of the room can be applied to any anechoic audio signal, and thus the signal emitted by the sound source, but not its location, can freely be changed in the virtual

scene. Averaging techniques can be used to improve the signal-to-noise ratio (SNR) of the measurement, and "virtual arrays", where a single microphone is used to sample the array position in succession, can also be used (e.g. Bernschütz et al., 2010). Some disadvantages have already been mentioned, including the need for a separate measurement for each source and receiver location that is to be auralized, the inability to handle moving sources with this method, as well as the close link to the reproduction system. This approach has been primarily applied to capture, analyze and reproduce concert hall acoustics (e.g. Farina and Ayalon, 2003).

Finally, a full recording can potentially capture all the nuances of the acoustic scene. However, the recording is subject to the artifacts and limitations of the spatial sound reproduction technique applied, such as spatial aliasing and limited SNR. Nonetheless, direct recordings of sound scenes were successfully applied by Minnaar et al. (2013) in order to investigate the preference of different hearing-aid settings in a virtual environment. Koski et al. (2013) used a combination of directly recorded background noise and spatial impulse responses to construct a virtual scene, where speech intelligibility scores closely matched those of the real environment.

In summary, while simulated virtual scenes are more flexible, reproducing multiple, potentially moving sound sources, and complex background noises is difficult. In such cases, recording the acoustic scene appears to be a better approach.

2.4 Spatial sound reproduction techniques

Given an acoustic scene that has been defined in some form, either as a set of parameters describing the sound sources and the environment, or as a set of audio signals containing spatial information, the sound scene must be rendered to the listener using one of the available spatial audio techniques. Two major presentation methods have been used for spatial audio: headphone-based and loudspeaker-based methods. In the following, a brief overview of the various methods is provided, with an emphasis on loudspeaker-based approaches.

2.4.1 Headphone-based reproduction

Headphone-based spatial audio methods are based on providing signals with (simulated or measured) binaural cues to the listener. Because the acoustic signals are generated directly at the ears, the listener's own ears and head do not provide spatial cues in this case. Consequently, the spatial hearing cues have to be included in the signal itself. These cues are typically represented as head-related transfer functions (HRTFs), which include the direction (and distance) dependent filtering effects of the head (and potentially the torso). The HRTFs can be included directly in the recording, such as when the recording is made with small microphones placed in a person's ears, or with an artificial head. Alternatively, an anechoic signal can be spatialized by filtering it with a measured or simulated HRTF corresponding to the intended direction (and distance) of the sound source (binaural synthesis). Typically,

HRTFs are measured in an anechoic environment, in which case they do not provide any information about the acoustic space. However, the binaural response can also be measured or simulated in a room, and is known as a binaural room impulse response (BRIR). Binarual technology is discussed in more detail in, e.g., Hammershøi and Møller (2005).

Headphone-based presentation allows for a precise control of the signals reaching the ears of the listener. Spatial sound presented over headphones can be very convincing, especially if the listener's own HRTFs are used. However, it is often impractical or impossible to measure HRTFs for each listener, and thus another person's or a set of general "non-individual" HRTFs must be used. Because of significant variations in head sizes and ear shapes, the applied HRTFs may be significantly different from those of the listener. This discrepancy can lead to a degradation in localization performance, front-back confusions, and lack of externalization (e.g. Hammershøi and Møller, 2005), and represents the main difficulty in applying headphone-based methods. Some improvement, especially in terms of front-back confusions, can be achieved if head tracking is used, and the HRTFs are dynamically updated in order to allow head-rotation (Begault et al., 2001).

From the perspective of virtual environment applications, the main advantages of headphone-based presentation over loudspeaker-based methods can be summarized as (i) relatively simple technical setup including only two audio channels, and (ii) precise control over acoustic cues. The primary disadvantages are (i) problems created by non-individual HRTFs, (ii) the requirement for head-tracking for virtual environments and (iii) the fact that external communication devices, e.g. phones or hearing aids, cannot be used easily in conjunction with headphones.

2.4.2 Loudspeaker-based reproduction

In contrast to headphone-based methods, in the case of presentation over loudspeakers, the listeners' own ears are exposed to the sound field, enabling the use of their own pinna cues. In addition, listeners can easily wear hearing aids or headsets in the playback environment. Several techniques are available to create spatial sound around the listener with loudspeakers. These techniques differ mainly in the supported loudspeaker geometries, the employed signal processing and in their aims regarding physically-based or perception-based reproduction.

The simplest spatial loudspeaker system, the stereo setup, employs only two loudspeakers with about 60° separation, in front of the listener. With corresponding microphone and panning techniques, it is possible to position sounds fairly accurately (often called "phantom sources" in this context) between the loudspeakers, at least if the listener is seated equidistantly from the loudspeakers. Despite the fact that the sound field generated in this manner differs significantly from that of a real source, valid ITD and ILD cues are created at the listener's ears (e.g. Spors et al., 2013). The effect breaks down if the listener moves closer to one of the speakers. Two loudspeakers using cross-talk cancellation methods can also be used to deliver binaural audio that would normally require playback through headphones. In this case the effect of the listener's own ears are compensated for, and the spatial cues are again included in

the signal itself (see Sec. 2.4.1). Various extensions of the stereo concept have been used for cinema surround systems, but these systems are typically aimed at providing well-localizable sounds only in the front, and ambiance in the other directions.

More advanced approaches use physical or perceptual criteria to (re)synthesize the desired sound field using from anywhere between a few to hundreds of loudspeakers, and are generally capable of reproducing virtual sources in arbitrary directions (at least in the horizontal plane). Some of the techniques applicable to virtual environments are described below. A more detailed review of spatial sound reproduction over loudspeakers is given in Spors et al. (2013).

2.4.3 Sound field synthesis methods

In the following, two sound field synthesis methods will be presented that aim at physically reconstructing the desired (measured or simulated) sound field, using a distribution of loudspeakers (termed "secondary sources" in the theoretical context). These two techniques are not exclusively *reproduction* techniques, as they also enable a physically-based representation of the acoustic scene.

Wave-field synthesis

Wave-field synthesis (WFS) is a physically-based sound field synthesis technique based on the principle of acoustic holography (or "holophony"), originally proposed by Berkhout (1988). It is based on an approximation of the Kirchoff-Helmholtz integral, which states that at any point within a source-free region, the sound pressure is completely determined by the sound pressure and its directional gradient on the surface of the region boundary (e.g. Spors et al., 2008). The synthesis of a sound field would then strictly require knowledge of both the sound pressure and its directional gradient, as well as a set of monopole and dipole secondary sources (Daniel et al., 2003). In the derivation of WFS, approximations are made such that only monopoles are used as secondary sources, and knowledge of only one field quantity, the directional gradient of the pressure, is needed. In practice, this implies that a closely spaced set of loudspeakers can recreate the desired sound field, if the directional gradient of the desired sound field is known at the points corresponding to the loudspeakers. WFS setups most often use simulated sound fields due to practical difficulties in creating the microphone configurations that would be required (Daniel et al., 2003). In addition, for curved secondary source geometries, an appropriate subset of secondary sources, radiating in the propagation direction of the virtual wave, needs to be selected (Spors et al., 2008). This secondary source selection is less straightforward with measured sound fields.

WFS has the advantage that the reproduction area is large. The listener can move in this area with the apparent position of the virtual sound sources staying constant, just like they would with real sound sources. However, aliasing errors will also be present in the entire listening area above the aliasing frequency, given by the frequency above which the loudspeaker spacing exceeds half a wavelength (Spors et al., 2013). Covering the full audible bandwidth would require a loudspeaker spacing of less than 2 cm, which is not feasible with current loudspeaker

designs. However, it has been reported that good localization performance is maintained for up to 20 cm loudspeaker spacing, due to the fact that low-frequency ITD cues are well preserved, and that localization is indeed independent of the listening position (e.g. Spors et al., 2013). Another limitation, due to the number of loudspeakers required, is that current implementations of WFS are practically limited to a horizontal line of loudspeakers, and thus operate only in the horizontal plane. On the other hand, the layout of loudspeakers is generally more flexible than with the main alternative, higher order ambisonics. Commercial as well as open-source implementations of the technique are readily available (Geier and Spors, 2012), and research into the perception of sound fields synthesized with WFS is ongoing (Spors et al., 2013), especially with regards to coloration artifacts.

Higher-order ambisonics

Higher-order ambisonics (HOA) is a technique based on a spherical (or circular) harmonics decomposition of the sound field. Much like the Fourier transform in the time domain, spherical harmonics represent a set of orthogonal basis functions with which the sound field on a spherical surface (or circle) can be described. The basic concept of ambisonics was introduced by Gerzon (1973), as a way to represent, record, and play back spatial audio. Classical ambisonics used only zero and first order spherical harmonics, which correspond to physically realizable, omnidirectional and figure-of-eight microphone characteristics. This allowed the development of a microphone to directly record first order ambisonics (Farrar, 1979). Significant work was done by Daniel (2000) and Daniel et al. (2003) to extend the technique to using higher orders. Daniel (2003) explicitly considered the radius of the loudspeaker array by considering the secondary sources as monopoles (and not as plane wave sources). This led to the most general formulation of ambisonics, near-field compensated higher-order ambisonics (NFC-HOA). NFC-HOA, like WFS, aims at reproducing a physically accurate sound field inside the listening area. In fact, Spors et al. (2008) showed that the two techniques are related, and that NFC-HOA can also be derived from the Kirchoff-Helmholtz integral for spherical (or circular) geometries, but with a slightly different set of assumptions than WFS. The behaviors of HOA and WFS, however, differ markedly in some aspects.

HOA exhibits a pronounced “sweet-area” at the center of the reproduction array. The error in the reproduced sound field increases with both frequency and distance from the center, and is a result of the properties of the spherical harmonics expansion, in that the contribution of high-order components vanishes towards the origin. In HOA reproduction, the upper frequency limit is not determined directly by the transducer spacing, but rather by the maximum order of the expansion, and the considered reproduction area. Ward and Abhayapala (2001) provided a rule of thumb regarding the spherical harmonics order M required to obtain a reproduction error of less than 4 % within a radius r as

$$M = \left\lceil \frac{2\pi f r}{c} \right\rceil, \tag{2.1}$$

from which the upper frequency limit of physically correct reconstruction, f_{lim}, assuming a

radius of $r = 9$ cm evaluates to approx. $M \cdot 600$ Hz. In terms of single-user virtual environments, this behavior can be considered as an advantage, because for a limited (e.g. head sized) reproduction area, a higher upper frequency limit can be obtained than the aliasing frequency for a comparable WFS setup. This comes at the cost of higher errors outside of the sweet spot.

One of the main advantages of HOA over WFS is that the playback of recorded sound fields is relatively straightforward. While no microphones exist that realize spatial directivities of higher-order spherical harmonics, Abhayapala and Ward (2002) and Meyer and Agnello (2003) proposed spherical microphone arrays that can be used to estimate these higher order components. Further work investigated the properties of spherical arrays specifically in the context of spatial sound (e.g. Poletti, 2005; Moreau et al., 2006).

With HOA, three dimensional recording and reproduction are feasible, and 3D HOA systems are available in several laboratories. However, the approach still suffers from the need for a high number of microphones and loudspeakers. The maximum order of the spherical harmonics expansion depends on the number of transducers used. For a 3D and a horizontal-only setup, respectively, the minimum number of required transducers is $(M+1)^2$ and $2M+1$ (Ward and Abhayapala, 2001). From Eq. 2.1, it can be seen that a physically correct reproduction of up to 20 kHz would require an order of 34, which would in turn require 1225 loudspeakers for 3D, and 69 for horizontal-only reproduction. It is likely that even if that many loudspeakers were installed, other error sources like reflections, noise, non-ideal loudspeaker characteristics, etc., would not allow accurate reconstruction at very high frequencies. Both HOA and WFS assume anechoic conditions in the playback environment, and thus strong reflections in the reproduction room can be detrimental to reproduction quality. In practice, however, similarly to WFS, HOA systems with a reasonable number of loudspeakers show fair localization and distance perception (e.g. Bertet et al., 2007; Favrot and Buchholz, 2009; Braun and Frank, 2011).

2.4.4 Directional audio coding

Directional audio coding (DirAC) (Pulkki, 2007) is a technique that specifically aims at a perceptually valid, rather than a physically valid, reconstruction of the target sound scene. It is based on a time-frequency analysis of the sound field, where a direction and diffuseness analysis is performed for each time-frequency bin. The input to DirAC in its original formulation is a recording from a first-order ambisonic microphone. Based on the diffuseness analysis, the audio signal is divided into a diffuse and a non-diffuse stream. In the synthesis part, the non-diffuse stream is reproduced using the estimated direction over a loudspeaker array with vector-base amplitude panning (VBAP) (Pulkki, 1997), which itself is an extension of the stereophonic panning principle to three dimensions. For synthesizing the diffuse part of the signal, a first-order virtual microphone is formed in the direction of each loudspeaker by a weighted sum of the first-order ambisonic components. But unlike in ambisonics, the signals are then decorrelated for each loudspeaker. This is intended to remove coloration and "phasing" artifacts resulting from highly correlated signals being presented over multiple loudspeakers.

While DirAC clearly does not produce a physically accurate sound field and involves non-linear processing, subjective evaluation suggests that the perceived quality of the reproduction is good for a variety of loudspeaker layouts (Vilkamo et al., 2009). A recent study (Koski et al., 2013) compared speech reception thresholds (SRTs) for one particular reference scene and the same scene reproduced using DirAC, and showed that for the best matching conditions, SRTs were within 2 dB for both normal-hearing and hearing-impaired listeners. It is, however, unclear how the nonlinear signal processing employed would affect other percepts, and how technical devices would perform in sound fields generated with DirAC.

2.4.5 Other approaches

Alternatives to the approaches mentioned so far include methods based on an inversion of measured transfer functions between the microphones and the loudspeakers in the listening environment (e.g. Kirkeby et al., 1996). Such an approach was presented by Minnaar et al. (2013) for the playback of recordings made with a spherical microphone array. Inversion techniques are practical as they are relatively easy to implement, and can also equalize microphone and loudspeaker responses, as well as the room response in one step. However, a new set of filters need to be measured using the same microphone array for each new playback location, or if any modifications are made to the playback environment. Therefore, such a method is not easily applicable for spatial audio recordings that are to be widely shared.

Another approach specifically aimed at spatial hearing research was presented by Seeber et al. (2010), termed the “simulated open-field environment”. In this approach, the focus is on a fine control of the individual room reflections. The reflections are computed using a room model, and each reflection is presented from individual, or pairs of loudspeakers. This allows for an accurate simulation particularly of the early reverberation in small rooms, but the approach is not well-suited for the reproduction of arbitrary sound scenes, including more diffuse, busy environments. A similar approach, using single loudspeakers for individual room reflections, is available as one of the rendering methods in the LoRA system (Favrot and Buchholz, 2010).

2.5 Summary and choice of method

In this chapter, an overview of the techniques applicable to the representation and reproduction of virtual sound scenes was presented. In this work, it was decided to investigate the recording of real scenes, in order to capture busy and dynamic environments. To this end, a technique based on HOA was chosen as it allows both recording and playback of 3D spatial audio, using spherical arrays of microphones and loudspeakers. HOA provides a sweet spot at the center of the reproduction array, which was deemed appropriate for use by a single person seated in the center. It also aims at a physically accurate reproduction of the sound field, which was desired in order to minimize psychophysical assumptions, and to be able to investigate technical devices, like hearing aids.

3

A mixed-order ambisonics microphone array[a]

Abstract

Spherical microphone arrays can be used to capture the spatial characteristics of acoustic scenes for analysis or reproduction, but practical constraints limit the number of microphones that can be used, and thus the maximum spatial resolution and frequency bandwidth that can be achieved. In this paper, a mixed-order ambisonics (MOA) approach is proposed to improve the horizontal spatial resolution of microphone arrays with a given number of transducers. A MOA array was realized, and its performance and robustness to variations in microphone characteristics and self-noise were investigated through both simulations and measurements. Results showed that higher horizontal resolution is achieved at mid to high frequencies, with a small increase in usable bandwidth. Robustness to various errors was similar to that of higher-order ambisonics (HOA) arrays.

3.1 Introduction

Advances in technology in the last decade have increasingly enabled the use of microphone array techniques in spatial audio recording and reproduction. There has also been increasing interest in applying spatial audio reproduction techniques, such as wavefield synthesis and higher-order ambisonics (HOA), to create virtual sound environments (VSEs) for research purposes. Such environments can create complex and realistic acoustic scenes around a listener, facilitating basic research into spatial hearing, or aiding the development and evaluation of hearing instruments and other communication devices (Minnaar et al., 2010).

Spherical microphone arrays are particularly well-suited for spatial audio applications, as they can capture the spatial characteristics of the sound field in all directions. Consequently, there have been a number of studies investigating the design of spherical arrays and their application to sound field analysis and reproduction (e.g., Abhayapala and Ward, 2002; Meyer and Elko, 2004; Rafaely, 2005; Poletti, 2005; Moreau et al., 2006; Li and Duraiswami, 2007). A spherical-harmonics-based decomposition of the sound field fits naturally with the spherical geometry, and was employed in much of the related work, although other processing strategies exist as well (e.g. DirAC, Pulkki, 2007). Recent arrays have used a rigid-sphere configuration

[a] Portions of this manuscript were previously presented in Marschall et al. (2012).

to avoid singularities at certain frequencies (see e.g. Rafaely, 2005), and evenly distributed microphones on the sphere for uniform sampling of the sound field.

To reproduce acoustic scenes with realism and high quality, high resolution and wide bandwidth, as well as low noise are needed, which places stringent requirements on the recording array. Therefore, it is important to evaluate the expected performance of such an array in the face of various errors and noise sources.

In particular, to achieve a high resolution in three dimensions, a high number of microphones are required. The spatial resolution and the usable frequency range are proportional to the order M of the spherical harmonics decomposition employed. The number of transducers required to capture spherical harmonics up to order M is at least $(M+1)^2$ (Abhayapala and Ward, 2002). This means that for higher orders, each additional order entails a large increase in the number of required transducers.

However, the most important sound sources are generally situated near the horizontal plane, and this is also where human auditory localization is most accurate, at least for frontal directions (Blauert, 1997b). Further, increasing spatial resolution in only two dimensions requires far fewer additional transducers than for three dimensions (Daniel, 2000). On the reproduction side, typical loudspeaker layouts are thus often restricted to the horizontal plane, or only use a few loudspeakers for height reproduction. In order to accommodate loudspeaker arrays with a greater number of speakers placed near the horizontal plane, Daniel (2000) proposed mixed-order ambisonics (MOA), and mixed-order playback was investigated by several authors (Travis, 2009; Trevino et al., 2010; Käsbach et al., 2011). Using MOA, a given number of transducers may be arranged in a way that better matches human perception: providing maximum resolution in the horizontal plane, while still retaining vertical spatial information. Thus, layouts with a non-uniform distribution of transducers could help bridge the gap between horizontal-only and three-dimensional spatial sound recording and reproduction.

In contrast to MOA playback however, recording in mixed order has not yet been studied in detail. Preliminary work by the present authors suggested that horizontal directivity may be improved by using mixed-order processing and a corresponding array layout for spherical microphone arrays (Favrot et al., 2011; Favrot and Marschall, 2012). This paper, which is an extended version of the study presented in Marschall et al. (2012), aims to provide (i) a more thorough introduction to the MOA concept for microphone arrays, (ii) a detailed comparison of MOA and HOA through the evaluation of two example microphone layouts, as well as (iii) an experimental validation of MOA through measurements on an array prototype.

The advantages and limitations of the mixed-order approach were evaluated by comparing an example 52-channel MOA array to a HOA array that uses the same number of transducers, but with a nearly uniform distribution of microphones over the sphere. The arrays' responses to plane waves were simulated, taking into account the characteristics of a commercially available array microphone, including self-noise, as well as amplitude and phase response variations. A set of beamforming measures were considered to quantify the performance limits of the array itself. White noise gain, directivity index, -3 dB beamwidth, and maximum sidelobe levels

were calculated for both arrays. In order to investigate the reproduced sound field, measures introduced by Gerzon were applied (Gerzon, 1992; Craven, 2003) to a simulated reproduction over an ideal loudspeaker array. Finally, validation measurements were performed on a prototype MOA array in free field, using a single loudspeaker as a sound source.

The paper begins with a summary of spherical harmonics decomposition, as well as higher-order and mixed-order ambisonics processing in Section 3.2. The details of the simulation and experimental setup are presented in Section 3.3. The applied objective measures are defined and the simulation results are described in Section 3.4, while the measurement results are detailed in Section 3.5. Finally, the findings are discussed and summarized in Sections 3.6 and 3.7.

3.2 Background

In this section, the principle of spherical arrays, as well as the mixed-order ambisonics approach are briefly summarized. For a more detailed introduction into spherical arrays for sound recording, refer to e.g. Meyer and Elko (2004) or Moreau et al. (2006).

In the following, the "ambisonics notation" as in e.g. Daniel et al. (2003) and Moreau et al. (2006) is followed. In the spherical coordinate system used, a point is described by its radius r, azimuth angle θ ($-\pi \leq \theta \leq \pi$), and elevation from the horizontal plane δ ($-\pi/2 \leq \delta \leq \pi/2$).

3.2.1 Pressure on a rigid sphere

The pressure on the surface of a rigid sphere of radius R at point (R, θ, δ), using spherical harmonics expansion, and omitting the implied time dependence of $e^{+i\omega t}$, is given as (Moreau et al., 2006):

$$p(kR, \theta, \delta) = \sum_{m=0}^{\infty} W_m(kR) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} B_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta, \delta), \tag{3.1}$$

where k is the wavenumber, $W_m(kR)$ is the radial function, describing the radial and frequency dependence of the pressure, B_{mn}^{σ} are the Fourier-Bessel series coefficients or ambisonics components, and $Y_{mn}^{\sigma}(\theta, \delta)$ are the real-valued spherical harmonics functions, defined as

$$\begin{aligned} Y_{mn}^{\sigma}(\theta, \delta) = & \\ & \sqrt{(2m+1)(2-\delta_{0,n})\frac{(m-n)!}{(m+n)!}} P_{mn}(\sin\delta) \\ & \times \begin{cases} \cos n\theta \text{ if } \sigma = +1 \\ \sin n\theta \text{ if } \sigma = -1 \end{cases}, \end{aligned} \tag{3.2}$$

and where $P_{mn}(\sin\delta)$ are the associated Legendre functions, and $\delta_{0,n} = 1$ if $n = 0$ and is 0 otherwise.

For a rigid sphere, the radial function at $r = R$ is given as (Meyer and Elko, 2004):

$$W_m(kR) = i^m \left(j_m(kR) - \frac{j'_m(kR)}{h^{(2)\prime}_m(kR)} h^{(2)}_m(kR) \right), \tag{3.3}$$

where j_m are spherical Bessel functions of the first kind, and $h^{(2)}_m$ are spherical Hankel functions of the second kind, both of order m, and the primes indicate derivatives with respect to the argument.

3.2.2 Encoding

The encoding process equates to determining the coefficients B^{σ}_{mn} from a set of pressures p_q measured at Q discrete positions (θ_q, δ_q). The discrete spatial sampling means that the series in Eq. (3.1) has to be truncated at a finite order M in practice, and the expression becomes an approximation of the pressure. The total number of spherical harmonics functions (K) included in the series up to order M is $K = (M+1)^2$ (Moreau et al., 2006).

With the above, Eq. (3.1) may be represented in matrix form as follows, noting that the equality only holds strictly if the considered sound field is spatially band limited to orders below M:

$$\mathbf{p} = \mathbf{Y} \cdot \mathrm{diag}[W_m(kR)] \cdot \mathbf{b}, \tag{3.4}$$

where $\mathbf{p} = [p_1, p_2, \ldots, p_Q]^T$ is a column vector of measured pressure values, $\mathbf{Y}$ is a $Q \times K$ matrix, where each row contains K spherical harmonics functions sampled at the angle corresponding to the q-th microphone:

$$\mathbf{Y} = \begin{bmatrix} Y^{+1}_{00}(\theta_1, \delta_1) & \cdots & Y^{-1}_{MM}(\theta_1, \delta_1) \\ \vdots & \ddots & \vdots \\ Y^{+1}_{00}(\theta_Q, \delta_Q) & \cdots & Y^{-1}_{MM}(\theta_Q, \delta_Q) \end{bmatrix}, \tag{3.5}$$

$\mathrm{diag}[W_m(kR)]$ is a diagonal matrix with K elements, containing the appropriate value of the radial function $W_m(kR)$ for each k-th spherical harmonic, and $\mathbf{b} = [B^{+1}_{00}, \ldots, B^{\sigma}_{mn}, \ldots, B^{-1}_{MM}]^T$ is a column vector of K series coefficients (ambisonics signals).

The encoding, or the estimation of the coefficients $\widetilde{\mathbf{b}} \simeq \mathbf{b}$ from the measured pressures $\mathbf{p}$ can be accomplished by inverting Eq. (3.4), providing a least-squares solution for $\widetilde{\mathbf{b}}$:

$$\widetilde{\mathbf{b}} = \mathrm{diag}\left[\frac{1}{W_m(kR)} \right] \cdot \mathbf{Y}^{+} \cdot \mathbf{p}, \tag{3.6}$$

where $\mathbf{Y}^{+}$ indicates the Moore-Penrose pseudo-inverse of $\mathbf{Y}$.

3.2.3 Limitations of a spherical array

Eq. (3.6) is also termed the encoding equation. Two limitations of the array processing are already apparent from this equation. First, the encoding involves inverting the radial function $W_m(kR)$, which has low magnitudes at low frequencies and high orders (Meyer and Elko, 2004; Moreau et al., 2006). In other words, the contribution of high order spherical harmonics at

low frequencies (relative to the size of the array) is low, and thus trying to estimate these harmonics from the measured pressures leads to a very high amplification of the transducer signals. The amplification in practice has to be limited due to the limited signal-to-noise ratio (SNR) of the microphone signals. This is achieved by applying regularization to $1/W_m(kR)$ (see Sec. 3.2.5), which leads to a reduction of the effective spherical harmonics order at low frequencies (Moreau et al., 2006). The selection of the regularization parameter is a tradeoff between acceptable noise amplification and array directivity at low frequencies.

A second limitation is that the pseudo-inverse operation on matrix $\mathbf{Y}$ in the encoding can lead to high errors if $\mathbf{Y}$ is not well-conditioned. The condition number of $\mathbf{Y}$ depends on the array layout (the sampling scheme) and the maximum order M. If the spherical harmonics functions are properly sampled, they form an approximately orthonormal basis. In this case $\mathbf{Y}$ is invertible with low error, and the condition number will be close to one (Moreau et al., 2006). For a quasi-regular sampling scheme, there need to be $Q \geq K = (M+1)^2$ transducers to capture spherical harmonics of order M (Rafaely, 2005). The maximum order determines the maximum directivity or spatial resolution achievable with the array.

At high frequencies, when the sampling distance exceeds half the wavelength, aliasing error becomes dominant and the directive properties of the array break down. More specifically, as the spatial frequency content of the sound field is generally not bandlimited, higher-order spherical harmonics are aliased into lower orders when the sampling scheme is insufficient to capture them. The exact pattern of aliasing is determined by the sampling scheme (Rafaely et al., 2007). Aliasing error is only significant at high frequencies, because the contribution of the aliased higher order components, as described by the radial function $W_m(kR)$, is low at low frequencies (i.e. low kR).

3.2.4 Mixed-order ambisonics

The idea behind MOA is to optimize the performance of a system with a given number of transducers or channels, by using only a subset of a higher-order set of spherical harmonics functions. As it was mentioned in Sec. 3.2.3, for 3D HOA, at least $(M+1)^2$ transducers are needed to capture or reproduce spherical harmonics components up to the M-th order. However, if only horizontal reproduction is desired, the number of transducers needed is only $(2M+1)$ (Daniel et al., 2003). One way to view MOA is as a combination of a higher-order planar representation with a lower-order periphonic (3D) representation. This leads to the mixed-order scheme considered in this paper, but alternate schemes have also been proposed (Travis, 2009).

The mixed-order scheme applied here uses a full set of spherical harmonic functions (SHFs) Y_{mn}^{σ} up to an order M_{3D}. Additionally, horizontal SHFs (with indices $n = m$) are selected for orders $M_{2D} > M_{3D}$.

With MOA, Eq. (3.1) is then approximated by:

$$\begin{aligned} p(kR,\theta,\delta) \simeq & \sum_{m=0}^{M_{3D}} W_m(kR) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} B_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta,\delta) \\ & + \sum_{m=M_{3D}+1}^{M_{2D}} W_m(kR) \sum_{\sigma=\pm 1} B_{mm}^{\sigma} Y_{mm}^{\sigma}(\theta,\delta). \end{aligned} \tag{3.7}$$

The MOA representation can also be described in matrix form by Equations (3.4)–(3.6), but with a mixed-order SHF matrix $\mathbf{Y}$, and with the corresponding K-long vector of mixed-order coefficients:

$$\mathbf{b} = \left[B_{00}^{+1} \dots B_{mn}^{\sigma} \dots B_{M_{3D}M_{3D}}^{\sigma} \dots B_{mm}^{\sigma} \dots B_{M_{2D}M_{2D}}^{-1} \right]^T. \tag{3.8}$$

The number of mixed-order coefficients K for an order combination of M_{2D}/M_{3D} is given by

$$K = (M_{3D}+1)^2 + 2(M_{2D} - M_{3D}). \tag{3.9}$$

3.2.5 Array processing

To derive the ambisonics signals from the measured pressures in practice, the regularized filtering approach described in e.g. Moreau et al. (2006) was used. This is obtained by the addition of regularization filters F_m to Eq. (3.6). The coefficients $\widetilde{\mathbf{b}}$ can then be obtained from the sampled pressures $\mathbf{p}$ as using the encoding matrix $\mathbf{E}$ as

$$\widetilde{\mathbf{b}} = \underbrace{\operatorname{diag}\left[\frac{F_m(kR)}{W_m(kR)} \right] \cdot \mathbf{Y}^{+}}_{\mathbf{E}} \cdot \mathbf{p}. \tag{3.10}$$

where $F_m(kR)$ are defined as

$$F_m(kR,\lambda) = \frac{|W_m(kR)|^2}{|W_m(kR)|^2 + \lambda^2}, \tag{3.11}$$

and where λ is the regularization parameter. The regularization filters limit excessive amplification when the magnitude of $W_m(kR)$ is small (see Sec. 3.2.3).

3.2.6 Decoding

Decoding refers to the process of obtaining the L-long vector of desired signals $\mathbf{s}$ from the set of ambisonics signals $\mathbf{b}$ with

$$\mathbf{s} = \mathbf{D} \cdot \mathbf{b}, \tag{3.12}$$

where the $L \times K$ matrix $\mathbf{D}$ is the decoding matrix, and $\mathbf{s}$ and $\mathbf{b}$ may be functions of time or frequency.

If the goal is to reproduce the sound field described by $\mathbf{b}$ over a set of loudspeakers, then desired signals are the appropriate set of driving signals for those loudspeakers. For a regular layout of L loudspeakers, assumed to be in the far field and located at (θ_l, δ_l), the rows of the

decoding matrix (for "basic" decoding) are simply given as (Daniel et al., 2003)

$$\begin{aligned}\mathbf{d}_l &= \frac{1}{L}\left[Y_{00}^{+1}(\theta_l,\delta_l)\cdots Y_{MM}^{-1}(\theta_l,\delta_l)\right] \\ &= \frac{1}{L}\mathbf{y}_l\,.\end{aligned} \tag{3.13}$$

Thus, the signal for loudspeaker l is

$$s(\theta_l,\delta_l) = \frac{1}{L}\mathbf{y}_l\cdot\mathbf{b}\,, \tag{3.14}$$

which states that the driving signal for each loudspeaker is given by the weighted sum of the ambisonics signals, and where the weights are obtained by sampling the spherical harmonics functions in the directions of the loudspeakers. Looked at another way, Eq. 3.14 defines a plane-wave decomposition beamformer (Rafaely, 2004; Rafaely, 2005), thus showing that for a regular loudspeaker layout and assuming plane wave sources, the loudspeaker signal is given by a beam formed in the direction of the loudspeaker. The properties of the beam therefore also relate to the characteristics of the reproduced sound field. With this in mind, well-established performance measures for beamforming can be applied to quantify the performance of the array output directly.

3.3 Methods

3.3.1 Microphone array design

A MOA microphone array with 52 channels was designed by the present authors and custom built by Brüel & Kjær. This array was applied in this study as an example array both for the simulation-based analysis described in Sec. 3.3.2 and the corresponding measurement-based analysis and verification described in Sec. 3.3.3.

The array consists of rings at different elevation angles, as shown in Figures 3.1 and 3.2 (left panel). There are seven rings with 2×2, 2×6, 2×10, and 1×16 microphones, at elevation angles ±80°, ±55°, ±29°, and 0°, respectively. The radius was 5 cm, which provided a good compromise between low-frequency microphone noise amplification and spatial aliasing (see Sec. 3.2.3). The layout was designed based on a slightly modified version of the method presented in Chapter 4, generating a suitable layout for an order combination of $M_{2D}=7$ and $M_{3D}=5$. An optimization of the elevation angle of the rings was performed, minimizing the condition number of the matrix $\mathbf{Y}$ (Li and Duraiswami, 2007). The obtained 52-transducer MOA layout has a condition number of 1.7 for a 7/5 order combination.

To allow a fair comparison between the MOA and HOA approach, a HOA layout with the same number of transducers (52) was simulated (see Figure 3.2, right panel), but with a nearly uniform distribution of microphones based on a t-design (Hardin and Sloane, 1996).

As this layout did not allow the capture of 6th order components with sufficiently low error (condition number of 4.7 for 6th order, versus 1.4 for 5th order), only a HOA order of $M=5$

Figure 3.1: The realized 52-channel mixed-order ambisonics microphone array.

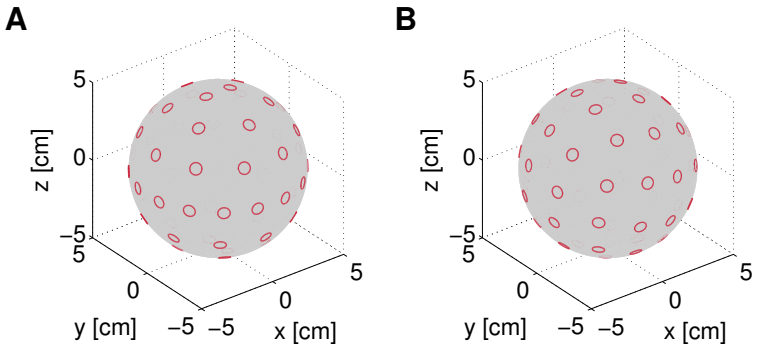


Figure 3.2: Microphone layout of the 52-channel, MOA array (left panel), and HOA array (right panel).

was used. In fact, the 52-channel t-design is given by Hardin and Sloane (1996) as a 9-design, which means that spherical harmonics of only up to 4th order satisfy the sampling condition with negligible error (see Sec. III.C. in Rafaely, 2005). However, the condition number and orthonormality error were deemed sufficiently low to use 5th order in practice.

3.3.2 Simulation framework

A simulation framework was developed in MATLAB that implements the complete signal processing chain: (i) incoming plane or spherical waves (see Sec. 3.2.1), (ii) the array processing based on the spherical harmonics decomposition of the sound field (Sec. 3.2.2 and 3.2.5), and

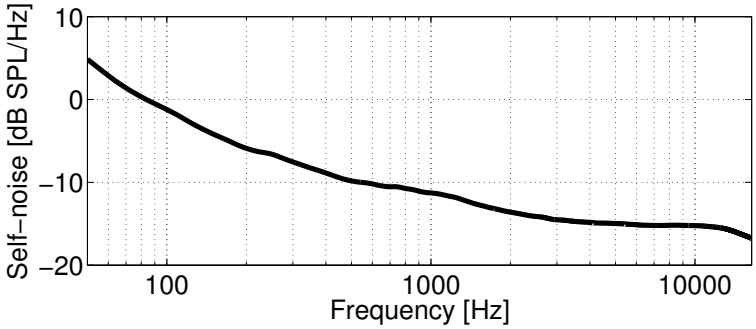


Figure 3.3: Self-noise power spectrum of the simulated array microphone .

Magnitude response	100 Hz – 5 kHz: ±2 dB
	50 Hz – 10 kHz: ±3 dB
	10 kHz – 20 kHz: +5/−3 dB
Phase response	100 Hz – 3 kHz: $< \pm 5°$
	3 kHz – 10 kHz: $< \pm 10°$

Table 3.1: Amplitude and phase characteristics of the simulated microphone.

(iii) the decoding and reproduction by a loudspeaker array (see Sec. 3.2.6). The processing was implemented in the frequency domain, and performed for each frequency bin.

Different system limitations that need to be taken into account when realizing a microphone array were also considered. Microphone self-noise as well as variations in the amplitude and phase characteristics of the individual transducers were simulated based on the specifications of the B&K type 4959 microphone used in the MOA array (Brüel & Kjær, 2012).

Each simulated pressure measurement $\hat{p}_q$ (e.g. the simulated complex pressures in $\mathbf{p}$ in Eq. 3.10) was perturbed by:

$$\widetilde{p}_q = Ae^{i\phi} \cdot \hat{p}_q + \hat{n}\,. \tag{3.15}$$

Microphone self-noise was simulated by adding the complex noise term $\hat{n}$, whose magnitude corresponds to the self-noise spectrum shown in Figure 3.3, with random phase. To simulate microphone sensitivity and phase variations, each transducer was assigned a frequency-dependent amplitude characteristic A and phase characteristic ϕ. These characteristics were defined as a relative deviations, and were drawn randomly from a set of example characteristics conforming to the specifications, which are detailed in Table 3.1.

All simulated sound sources were plane waves, with a flat spectrum and a total amplitude of 82 dB SPL (corresponding to a spectral density of 40 dB SPL/Hz). The regularization parameter was set to $\lambda = 0.01$.

3.3.3 Measurement setup

To verify the simulation-based analysis detailed above, measurements were made in the small anechoic chamber of the Technical University of Denmark (with free space volume of 60m^3, and a lower limiting frequency of 100 Hz). Since only the MOA microphone array described in Sec. 3.3.1 was physically available, measurements were only performed on this array. A single Dynaudio BM6P loudspeaker was used as a sound source, mounted either in the horizontal plane, or at 20° elevation as seen from the center of the array. The distance from the base of the loudspeaker to the base of the microphone array was 2.5 m. Similarly to the simulations described in Sec. 3.3.2, the loudspeaker was driven with white noise, and the level adjusted to measure about 82 dB SPL (lin. weighting) at the position of the array. The magnitude response of the loudspeaker was compensated for using a reference microphone placed on-axis, 1 m from the loudspeaker, measuring ±3 dB in the range 100 Hz–10 kHz at the reference location after compensation. For the elevated position, the loudspeaker had to be positioned off-axis to the microphone array position due to mounting restrictions, resulting in a trough in the magnitude response at the loudspeaker's crossover frequency. Frequency response compensation was not employed in this case in order to avoid excessive gain at this frequency.

The measurements were made using the B&K Pulse platform, recording the 52-channel microphone signal with a sampling frequency of 32 kHz. The data were then transferred to MATLAB where all the array signal processing was performed. Welch's method was applied to 10 s segments of the processed signals to obtain an estimate of the power spectral density, which then enabled the application of the same performance measures as used for the simulation study.

3.3.4 Performance measures

An inherent difficulty in MOA microphone array design is selecting appropriate measures to objectively evaluate array performance, as the properties of the array vary with elevation angle. Here, a set of measures were considered to investigate errors both at the array output and in the reproduced sound field.

For a look direction (θ_0, δ_0), the output of the beamformer in the frequency domain can be written as (cf. Eq. 3.14, omitting the $\frac{1}{L}$ normalization)

$$s(\theta_0, \delta_0, f) = \mathbf{y}_0 \mathbf{b}(f) \tag{3.16}$$

with $\mathbf{y}_0$ being the K-long row vector of SHFs evaluated at (θ_0, δ_0), and $\mathbf{b}$ the obtained ambisonics signals after array encoding.

For the beamforming measures, which are further described in Sections 3.4.1, an equiangle sampling scheme was used, assuming a regular grid of 128 points along the azimuth θ and 64 points along the elevation angle δ, where the output of the array at each point was given by Eq. 3.16.

For measures of the reproduced sound field considered in Sections 3.4.2 and 3.5.2, decoding

was simulated over a loudspeaker array consisting of 204 plane-wave sources, arranged according to a t-design (Hardin and Sloane, 1996). A large number of virtual loudspeakers were chosen to reduce reproduction error for $kr < M$ (i.e., inside the "sweet area") (Solvang, 2008) in order to ensure that the observed errors were primarily due to the processing by the microphone array. Various psychoacoustic decoder optimizations, such as "max r_E" and "in-phase" decoding (Daniel, 2000) were not considered here, as the goal was to simulate the reconstruction of the sound field captured by the array.

3.4 Simulation results

3.4.1 Beamformer-based measures

Beam pattern

The beam pattern describes the output of the array to a plane wave from a fixed direction as a function of the look direction (θ_0, δ_0) according to Eq. 3.16, and is a direct illustration of the spatial selectivity realized by the array.

Figure 3.4 shows typical beam patterns for the MOA array (panels A and B) and the HOA array (panel C) plotted for a single plane wave, with a cylindrical projection of the spherical coordinate system. For the MOA array and a source in the horizontal plane (panel A), the main lobe of the beam pattern is narrower along the azimuth than along the elevation. This highlights the higher azimuthal directivity of the MOA array for horizontal sources. Panel B displays the beam pattern for a source at the zenith, with the projection rotated so that the main lobe can easily be seen, showing that it is equally wide in both directions, and demonstrating that the beamwidth changes with elevation angle. The beam pattern for the HOA array (Panel C) is independent of the look direction and is similar to that of the MOA array for the elevated source.

White noise gain

The white noise gain (WNG) is a common measure for estimating the robustness of microphone arrays to microphone self-noise, amplitude and phase variations as well as position error (Meyer and Elko, 2004). Assuming uncorrelated white noise on each transducer, the WNG shows how much this noise is reduced (or amplified) by the array processing. A higher WNG means the array is more robust. The WNG can be interpreted as the signal power at the output of the beamformer over the sensor self-noise power (Meyer and Elko, 2004), and can be calculated as

$$\mathrm{WNG} = 10\log_{10}\left(\frac{|\mathbf{y}_0\mathbf{b}_0|^2}{(\mathbf{y}_0\mathbf{E})^H(\mathbf{y}_0\mathbf{E})}\right), \tag{3.17}$$

with the nominator being the square of the array output for a unit amplitude plane wave from the look direction (θ_0, δ_0) (cf. Eq. 3.16), and $\mathbf{E}$ being the encoding matrix (cf. Eq. 3.6).

WNG was simulated for a horizontal $(0,0)$ and an elevated $(0,+\pi/2)$ beamformer look direction (Figure 3.5) for both the MOA (top panel) and the HOA array (bottom panel). As

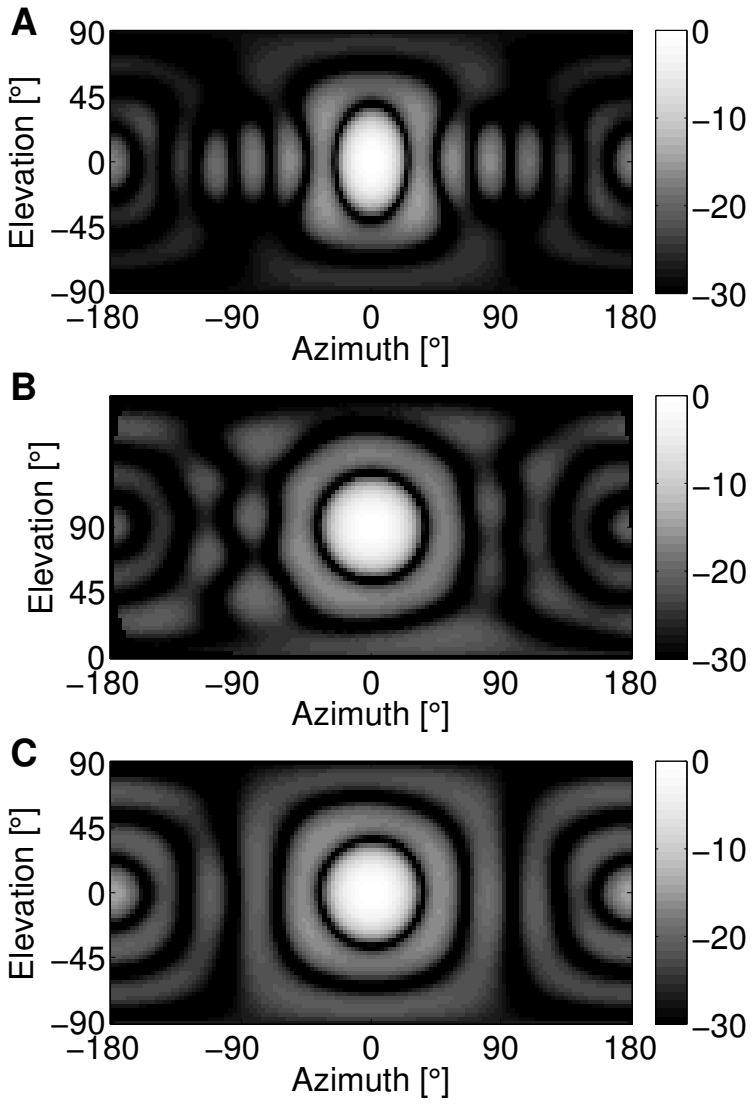


Figure 3.4: Beam patterns for the MOA array at 5 kHz for a horizontal $(\theta, \delta) = (0, 0)$ (panel A) and a vertical $(0, \pi/2)$ source (panel B). The beam pattern for the HOA array for a horizontal source is shown in panel C. The color scale indicates the normalized response magnitude in dB.

expected, the WNG shows a bandpass characteristic centered at an optimum frequency defined as (Park and Rafaely, 2005)

$$f_{\text{opt}} = \frac{cM}{2\pi R} \tag{3.18}$$

where c is the sound velocity. The effect of regularization in limiting noise amplification can be seen below about 3 kHz for both arrays: a WNG of above -10 dB is maintained even at low frequencies, meaning that sensor noise is not amplified by more than 10 dB. Without regularization, the noise amplification would quickly exceed acceptable levels and severely limit the usable bandwidth.

For the HOA array, both elevated and horizontal directions display the same bandpass characteristic around $f_{\text{opt}} = 5500$ Hz, corresponding to $M = 5$. For the MOA array, the WNG for the elevated look direction is similar to that of the 5th-order HOA array. For the horizontal

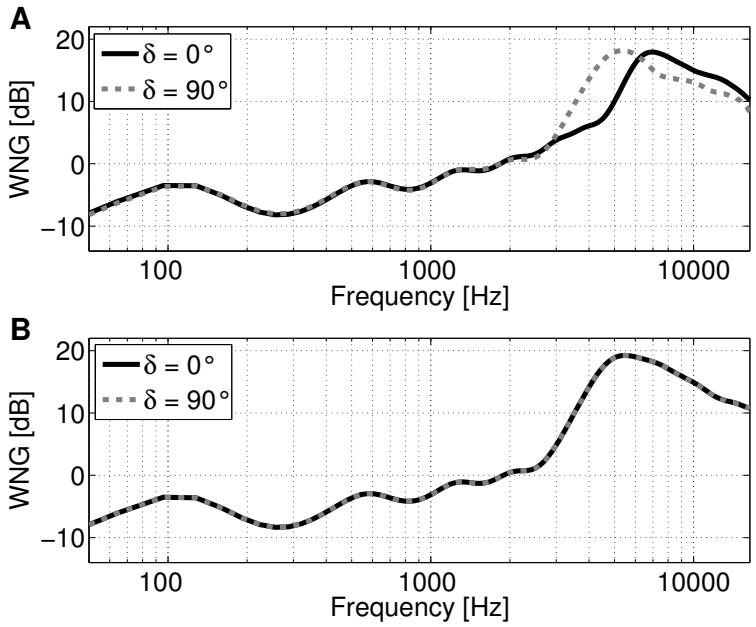


Figure 3.5: White noise gain for the MOA (top panel) and HOA (bottom panel) array, for a horizontal and an elevated source.

direction the optimum frequency is shifted upward to $f_{\text{opt}} = 7600$ Hz, corresponding to M = 7, due to the inclusion of additional spherical harmonic functions. This in turn leads to a reduction of the WNG as compared to the elevated source for frequencies between 3 and 6 kHz. Thus, the robustness of MOA encoding is similar to the robustness of an M_{3D} order HOA encoding for elevated sources, and to that of an M_{2D} order HOA encoding for horizontal sources.

Directivity Index

The directivity index (DI) describes the spatial directivity of the microphone array and can be defined as the ratio of the beamformer output in the plane wave incidence direction relative to the average of the output of the same beamformer for all look directions:

$$\text{DI} = 10\log_{10}\left(\frac{|\mathbf{y}_0\mathbf{b}_0|^2}{\frac{1}{L}\sum_{n=1}^{L}|\mathbf{y}_n\mathbf{b}_0|^2}\right) \tag{3.19}$$

A regular grid of $L = 350$ points was assumed around the array from which the look and incidence directions were selected. In order to analyze the directivity of horizontal vs. elevated sources, DIs were averaged for incidence directions above $|\delta| > 30°$ (“elevated sources") and for directions $|\delta| < 10°$ (“horizontal sources").

Figure 3.6 shows elevated and horizontal averaged DIs for the MOA (top) and HOA (bottom) arrays. Dashed horizontal lines represent the theoretical maximum achievable DI for orders $M = 1$ to 7, given as $20\log_{10}(M+1)$ (Meyer and Elko, 2004). At low frequencies, for both arrays, the DI increases with frequency, highlighting the effect of regularization, which has the effect

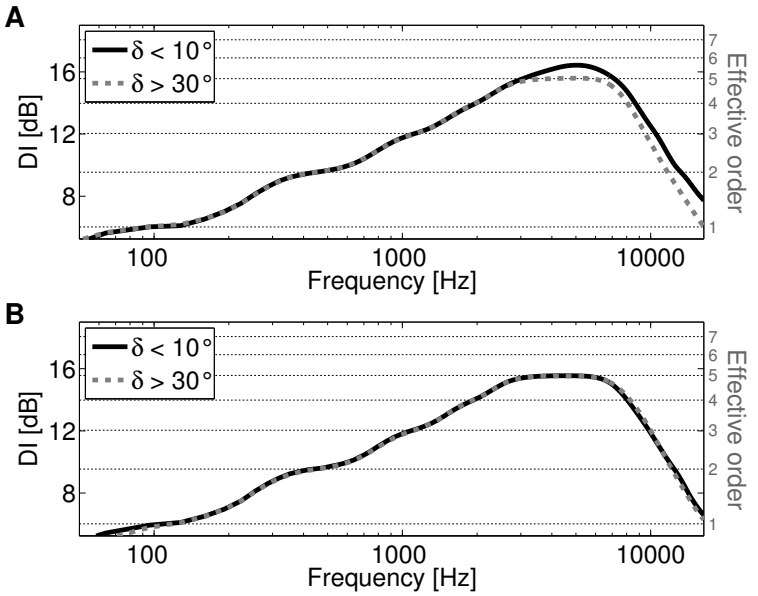


Figure 3.6: Directivity index for the MOA (top) and HOA (bottom) arrays, for horizontal (solid lines) and for elevated sources (dotted lines).

of attenuating high orders at low frequencies, and thus decreasing spatial resolution. At high frequencies, directivity is reduced again due to spatial aliasing.

For the 5th order HOA array, DIs for both elevated and horizontal source directions reach the theoretical value corresponding to $M = 5$. For the 7/5 MOA array, the elevated DI reaches the value for $M = 5$, and is very similar to the DI obtained for HOA. The horizontal DI, however, only reaches a maximum value corresponding to just below $M = 6$, but well below the applied horizontal order $M = 7$. This means that the 3D directivity of encoded horizontal sources is only slightly higher than for HOA 5th order. However, MOA beam patterns have different horizontal and vertical widths (cf. Fig. 3.4) and therefore the 3D directivity index does not fully reveal the potential benefit of MOA vs. HOA for horizontal sources.

Beamwidth and sidelobe level

The 3 dB beamwidth is a commonly used measure of beamformer resolution, and describes the angular width of the main lobe of the beam pattern (see Figure 3.4) at the −3 dB points relative to the maximum. Due to the expectation that the resolution of the MOA array differs in the horizontal and vertical directions, the beamwidth was evaluated both along the azimuth and along the elevation angle. In addition to the beamwidth, which only carries information about the main lobe, the level of the side lobes also influence the total spatial selectivity of the array. The maximum sidelobe level (MSL) is defined as the difference in level between the main lobe and the second highest peak (sidelobe) in the beam pattern, and describes the minimum attenuation the array provides in directions outside of the main lobe. However, the ambisonics formulation used here does not seek to minimize sidelobes; instead, it aims at

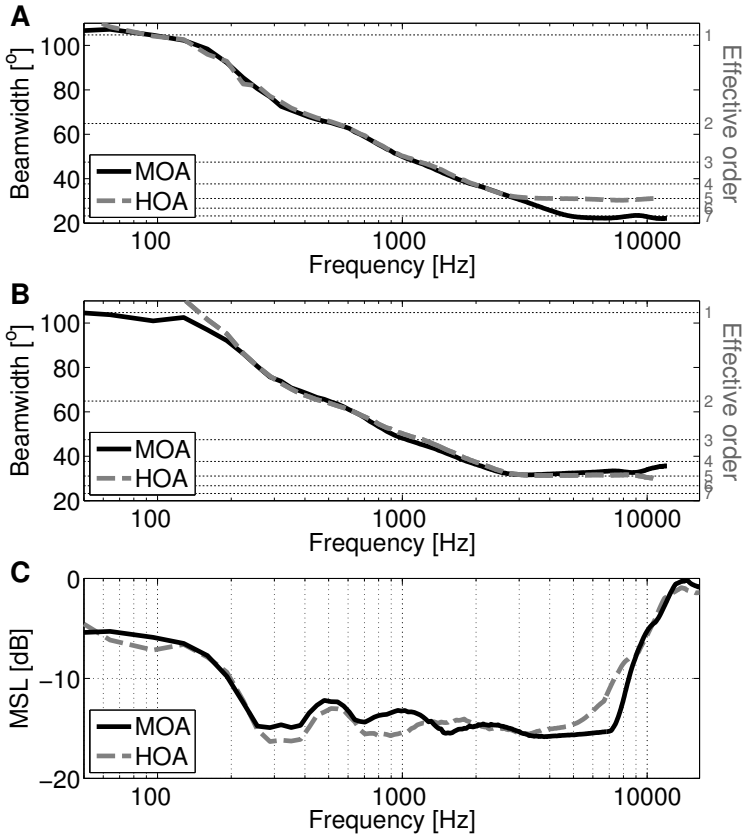


Figure 3.7: 3 dB beamwidth for the MOA and HOA arrays, along the azimuth (panel A) and along the elevation (panel B). The maximum sidelobe level (MSL) is shown in panel C. Source direction is horizontal.

providing an accurate reproduction of the sound pressure. The sidelobes levels were therefore evaluated as a diagnostic measure rather than a performance measure.

Figure 3.7 shows the simulation results for the MOA and HOA arrays. The beamwidth and MSL were calculated for 3 horizontal source directions and averaged. Panel A shows the beamwidth measured along the azimuth for both arrays. The beamwidths are identical up to about 3 kHz, decreasing with frequency. The gradual decrease, mirroring increasing DI, again shows the effect of regularization. Above 3 kHz, it can be seen that the HOA beamwidth (dashed lines) stays constant up to the aliasing frequency as the maximum resolution for order M_{3D} is reached. The MOA beamwidth (solid lines) continues to decrease until the corresponding resolution for M_{2D} is reached. This illustrates that the MOA array provides a narrower beam along the azimuth, but also that due to regularization this beam pattern is only reached at higher frequencies.

Panel B in the same figure shows that the vertical beamwidth is essentially identical between MOA and HOA for the entire frequency range, and reaches a maximum corresponding to an order of 5. This again shows that, as expected, HOA provides a beam pattern that has a circular cross-section, thus the resolution is the same in the horizontal and vertical directions, and that with MOA the vertical beamwidth is limited by the periphonic order M_{3D}. Elevated sources

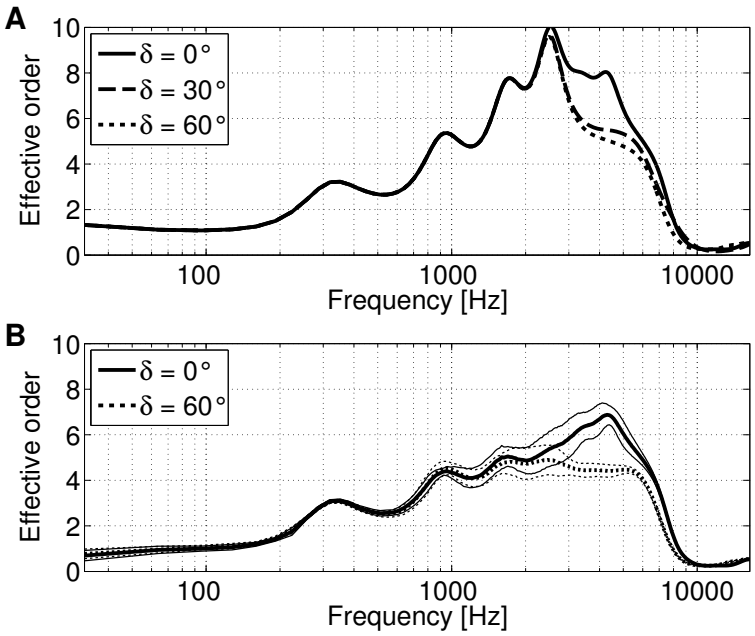


Figure 3.8: Effective order (based on r_E) for the MOA array for different source elevation angles. Ideal (top panel) and noisy simulation (bottom panel) is shown. Thick lines represent the mean, thin lines one standard deviation from the mean for the noisy simulation.

were not investigated with this measure, as it was already shown that the beam pattern for elevated sources with MOA is similar to that with HOA.

Looking at the MSL shown in panel C, it can be seen that it is generally high at very low frequencies, indicating a prominent back lobe. For most of the operating frequency range the MSL is around −15 dB. At higher frequencies (above 8 kHz) the MSL increases again due to aliasing error becoming more and more dominant. Above 10 kHz, the source direction becomes ambiguous, the beam pattern splits up into several lobes and thus the MSL reaches 0 dB. The MOA and HOA arrays behave similarly except for frequencies between 5–8 kHz, where the MOA array has slightly lower sidelobe levels. This is might be due to the more densely located transducers on the equator contributing to a better sampling of horizontal sound sources.

3.4.2 Sound-field-based measures

Effective order

The simulated plane waves encoded by the array were decoded onto a virtual, 204-element regular loudspeaker array (see Secs. 3.2.6 and 3.3.4). At the center of the virtual array, the r_E measure , the sum of all loudspeaker signals $G = \sum_i g_i$, as well as the sum of the squared loudspeaker signals $E = \sum_i g_i^2$ were calculated (Gerzon, 1992; Craven, 2003), where g_i is the driving signal for loudspeaker i. The r_E measure is the magnitude of the "energy vector" $\vec{E}$, defined as

$$\vec{E} = \frac{\sum_i g_i^2 \cdot \vec{u}_i}{\sum_i g_i^2}, \tag{3.20}$$

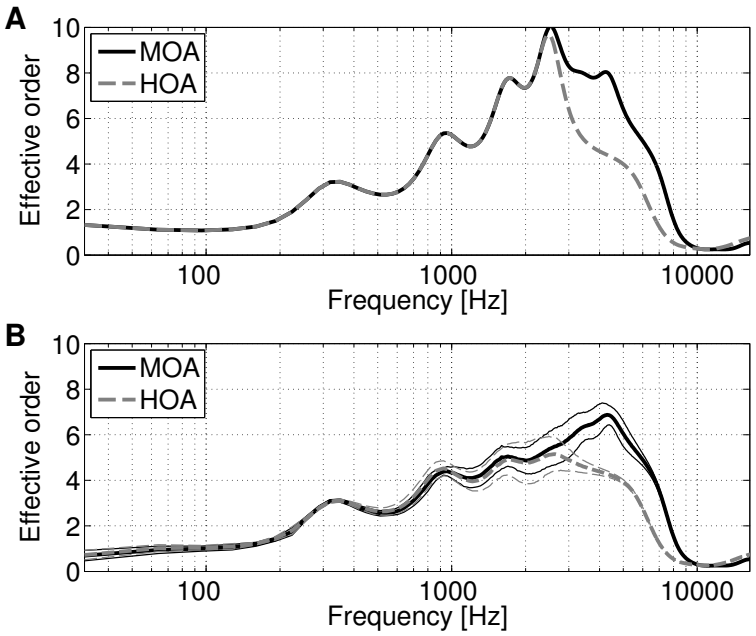


Figure 3.9: Effective order (based on r_E). Comparison of MOA (solid lines) and HOA (dashed lines) for a horizontal source. Ideal (top panel) and noisy simulation (bottom panel) is shown. Thick lines represent the mean, thin lines one standard deviation from the mean for the noisy simulation.

where $\vec{u}_i$ is a unit vector pointing in the direction of the i-th loudspeaker.

The ideal value of r_E for 3D ambisonics of order M (i.e. for ideal ambisonics components and a regular loudspeaker layout with a sufficient number of loudspeakers) is given by Daniel (2000) as

$$r_E = \frac{M}{M+1}, \tag{3.21}$$

from which the effective order may be expressed as

$$M_{\text{eff}} = \frac{r_E}{1-r_E}. \tag{3.22}$$

This measure is used here as an indication of the directionality of the reproduced sound field. For a high effective order (and r_E) sound energy originates from a small portion of the array (so it is highly directional), whereas a low effective order means that the sound energy is produced by several loudspeakers, and the original plane wave is “blurred". The above measure is considered both in ideal and noisy form, in order to evaluate the impact of errors introduced by the microphones.

Figure 3.8 shows the effective order simulated for the MOA array for different elevation angles with and without adding noise and microphone characteristic variations. For the noise-free condition (top panel), the effective order increases with frequency for all three elevation angles up to 2–3 kHz. Again, the maximum effective order in this frequency region is limited by regularization, which was applied even in the noise-free condition. As can be seen, the applied regularization scheme does not result in a completely smooth change of r_E with frequency, and a step-like behavior with peaks is seen as higher orders are introduced. Above 2 kHz, the

effective order is higher for the horizontal source than for the elevated sources, showing the result of the MOA processing. Finally, the effective order decreases sharply from above 7–8 kHz due to increasing spatial aliasing error.

The bottom panel shows the mean of 20 simulation runs with noise, where for each run, a new amplitude and phase characteristic was assigned to each microphone. In this more realistic configuration, the effective order is not affected strongly below 1 kHz, but above that frequency the effective order is reduced compared to the ideal case. For the elevated source, an effective order of between 4 and 5 is maintained in the 1–6 kHz range. For the horizontal source, the effective order exceeds that of the horizontal source above about 2 kHz, albeit peaking just below the chosen M_{2D} order of 7. The addition of self-noise by itself was not found to have a marked effect on the simulations above about 200 Hz, in accordance with previous findings (Rafaely, 2005), but not shown here. Thus the variations in array performance show the effect of the amplitude and phase characteristic variations. Between about 1–6 kHz, the effective order (and r_E) is reduced and varies more strongly with changes in transducer characteristics.

Figure 3.9 shows the same measure, but now comparing MOA (solid line) and HOA (dashed line) for a horizontal source, both with and without noise and characteristic variations. The HOA array behaves very similarly to an elevated source position in the previous comparison (cf. Figure 3.8). Differences between MOA and HOA are seen above 2 kHz, where MOA reaches higher effective orders. It can also be seen from the figure that the usable bandwidth of the MOA array is slightly extended as compared to the HOA array.

Array response and background noise

Figure 3.10 shows the sum of all loudspeaker signals G as well as the sum of the squared loudspeaker signals E (thick lines), for three plane wave sources incident on the MOA array. G corresponds to the sound pressure level at the very center of the reproduction array, and shows a flat response at 40 dB/Hz, as expected, up to the aliasing frequency of about 8 kHz for all three elevation angles. The measure G considers a phase-coherent addition of the loudspeaker signals, which in practice occurs only at low frequencies. In fact, the response at the very central point of the listening area is only dependent on the 0th order (omnidirectional) component, as the directional components are canceled (i.e. their integral over the sphere is 0).

A closer approximation of the high frequency behavior in the reproduced sound field, especially at a slightly off-center position, is given by the energetic sum of the loudspeaker signals E. Using "basic" decoding, the system tries to achieve a flat pressure response. This in turn means that towards low frequencies where coherent addition occurs, the energetic sum E is reduced, as can be seen in Figure 3.10 (bottom panel). It can be seen, however, that the reproduced energy of sound sources is essentially independent of the elevation angle. Fluctuations in G, as well a sharp increase of E are evident above about 8 kHz due to spatial aliasing.

In the same figure, thin lines show the G and E measures with only the microphone self-

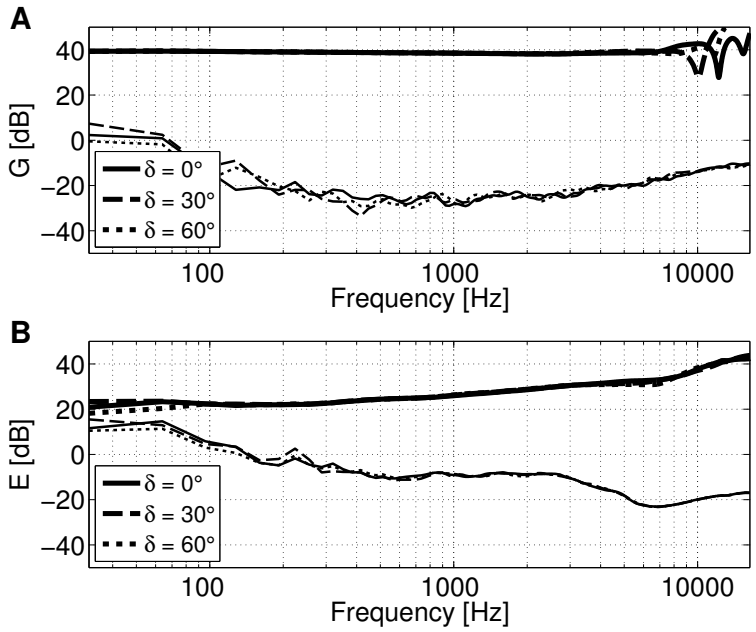


Figure 3.10: Power spectral density of the sum (G) and the energetic sum (E) of the loudspeaker signals, for a source at three different elevation angles, for the MOA array. Thick lines show the response for a plane wave of 40 dB SPL/Hz. The smoothed spectrum of the array output with no signal applied to the microphones (i.e. the predicted noise floor) is also plotted with thin lines.

noise applied, indicating the predicted noise floor in the reproduction system. Looking at G, the simulated noise floor is highest at low frequencies, and the lowest in the range between 500 and 1000 Hz, varying in the range of about 30 dB. For a signal spectrum level of about 40 dB SPL/Hz (roughly the spectrum level of loud speech at 1 m up to about 1 kHz; Olsen, 1998), the signal-to-noise ratio varies between 40 and 60 dB. The SNR predicted by the E measure is markedly smaller, especially at frequencies below 3 kHz, which reveals potential noise problems at more off-center positions, and may indicate the need for applying more regularization in practice. At the very lowest frequencies this measure likely underestimates the SNR due to the coherent addition of loudspeaker signals, as mentioned before.

The HOA array showed a small decrease (6 dB max.) in background noise power between 3 and 6 kHz, as expected from the higher WNG of HOA at these frequencies (cf. Fig. 3.5). However, as results were otherwise almost identical to those of the MOA array, they are not shown here.

3.5 Measurement results

Selected results from the simulation-based microphone array analysis presented in the previous section were validated with measurements on the physical MOA array, set up as described in Sec. 3.3.3. In order to obtain the beam patterns to calculate the beamformer-based measures, the same grid of array look directions as described in Sec. 3.3.4 was taken. Similarly, for the sound-field-based analysis, the measured signals were decoded onto the virtual loudspeaker array described in Sec. 3.3.4.

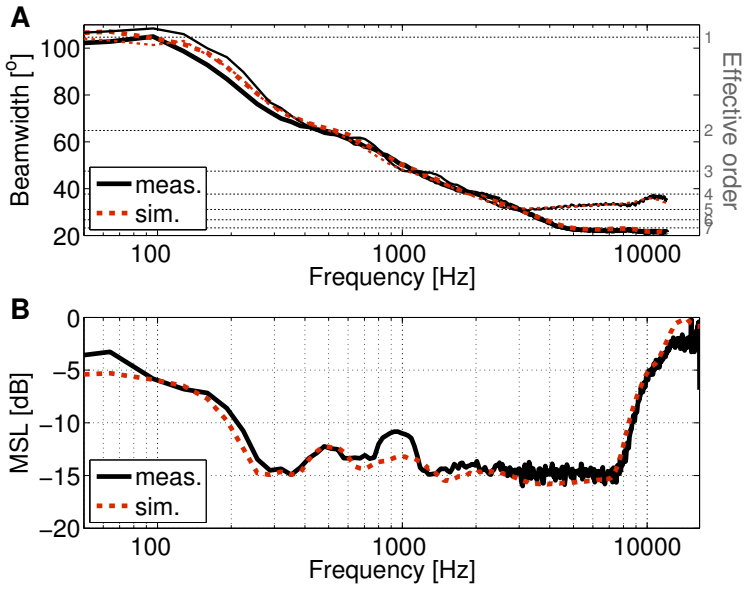


Figure 3.11: Comparison of measured (solid lines) and simulated (dotted lines) beamwidth and MSL for the MOA array. The top panel shows the 3 dB beamwidth in the azimuth (thick lines) and elevation directions (thin lines). The bottom panel shows the MSL (thick lines only). Results are for a source in the horizontal plane.

3.5.1 Beamformer-based measures

Beamwidth and sidelobe level

Figure 3.11 shows a comparison of measured and simulated beamwidth and MSL data. Identically to Section 3.4.1, the beamwidth is calculated along both the elevation and the azimuth, but is now displayed in one figure with thin and thick lines, respectively. It can be seen that both the measured beamwidth and MSL data closely match the simulations over the whole frequency range tested, suggesting that the simulation parameters chosen can accurately predict the actual behavior of the physical array.

3.5.2 Sound-field-based measures

Array response and background noise

Figure 3.12 shows the microphone array response at the center of a virtual loudspeaker array, both with a signal applied (a horizontal source), and in quiet. Similarly to Figure 3.10, both the sum and the energetic sum are displayed. Measured results are plotted with solid, simulations with dotted lines. Looking at the array response with the signal applied (thick lines), it can be seen that the measured spectrum matches the simulations well, although response variations exist in the measured result. These are partly due to the spectral fluctuations of the white noise signal itself, and any measurement errors or errors introduced by the microphone array may contribute as well.

Although overall the noise floor (thin lines) follows the predictions quite well, the background noise in the measurement is somewhat higher than predicted above about 300 Hz.

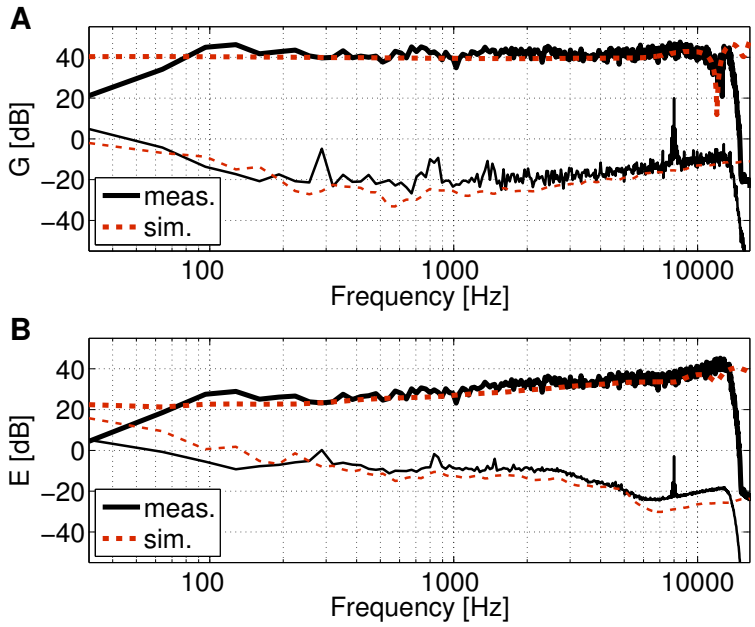


Figure 3.12: Comparison of measured and simulated array response (thick lines) and background noise (thin lines). The top panel shows the power spectral density of the sum (G), the bottom panel the power spectral density of the energetic sum (E) of the loudspeaker signals for a source in the horizontal plane.

Some tonal noise components are apparent (e.g. a strong peak at 8 kHz), which are due to noise in the electronics somewhere along the signal path, as they were not present in the acoustic field. Such noise components can be especially detrimental if they affect groups of channels, as they introduce correlated noise, which is amplified more than given by the white noise gain. Aside from these peaks, dynamic range is as predicted and discussed in Sec. 3.4.2.

Effective order

The effective order for a measured source situated in the horizontal plane, as well as for a source at 20° elevation is displayed in Figure 3.13. The effective order for both sources is similar up to about 2 kHz, being limited by regularization, as seen previously in the simulated analysis in Sec. 3.4.2. However, above this frequency, the horizontal source achieves a higher effective order, reaching up to order 6. This shows the desired result of MOA processing, although it also demonstrates that an effective order of 7 is not reached under real-world conditions (cf. Fig 3.8). A large drop in effective order is seen for the elevated source just under 2 kHz. This, however, was confirmed to be a result of the loudspeaker frequency response at the position of the array. Due to mounting restrictions, the loudspeaker was not facing the microphone array directly at the elevated position, thus introducing a trough in the frequency response at the crossover frequency. Aside from this, the measured power spectrum at the microphone position for the elevated source was not lower than for the horizontal source, indicating that the decreased effective order at other frequencies was not due to a lower SNR, but rather to the effect of MOA processing.

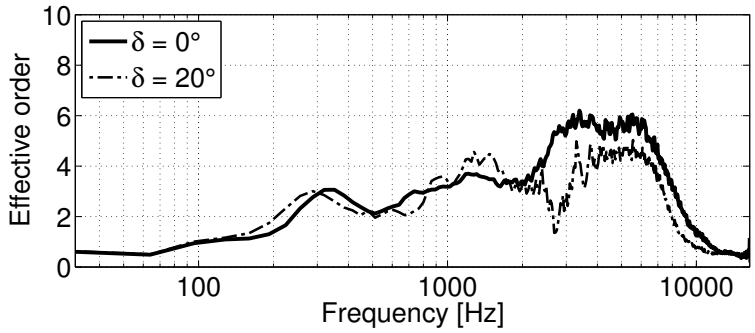


Figure 3.13: Comparison of effective order (based on r_E) for a measured horizontal and elevated source.

3.6 Discussion

3.6.1 MOA vs. HOA

One of the main goals of this paper has been to investigate the applicability of the MOA approach to microphone arrays. The aim of the MOA approach itself has been to combine a 3D (periphonic) representation of a certain order with a higher-order horizontal representation, taking advantage of the fact that (i) an increase in resolution in only two dimensions may be achieved by a smaller increase in the number of transducers, and (ii) that the properties of human hearing make it desirable to provide a better resolution in the horizontal plane when a limited number of transducers are available. The desired properties of a MOA system therefore depend on the direction of the source. For an arbitrary source direction, it should provide at least the performance of an HOA system of order M_{3D}. For horizontal sources, it should approach the performance of a higher, M_{2D}-order system.

Based on the results, the MOA microphone array presented here meets these criteria partially. The DI, beamwidth and r_E measures all show that performance for horizontal sources is improved, with a small increase in the usable bandwidth over the HOA array. While the beamwidth approaches the levels corresponding to order M_{2D}, the r_E and DI measures only show a smaller improvement. Whereas the beamwidth can be separated into horizontal and vertical components, the DI and the r_E measures are affected by horizontal and vertical spatial directivity concurrently. Therefore these measures cannot be expected to reach levels corresponding to order M_{2D} with the current mixed-order approach, which essentially only affects horizontal directivity.

It is important to note, that improved performance for horizontal sources is only seen for mid to high frequencies, above about 3 kHz. The increased spatial resolution for horizontal sources has the cost of decreased robustness (i.e. more noise amplification) at mid frequencies, as evidenced by the WNG measure. At lower frequencies the regularization scheme limits performance, and the MOA array behaves similarly to an HOA array of order M_{3D}. Unlike playback systems, microphone arrays can only take advantage of the higher orders in the frequency range where these orders can be captured without excessive amplification of the

microphone signals. As pointed out by Daniel (2009), this frequency range decreases as the order increases, and is a major limitation in building arrays of very high orders.

3.6.2 The effect of regularization and noise

Although it was not investigated here explicitly, the above suggests that it is important to evaluate the role of regularization in limiting performance in the MOA recording setup. The regularization parameter λ needs to be chosen carefully, considering the available signal-to-noise ratio. Further, alternate regularization schemes that may be better suited for MOA need to be investigated. Nonetheless, MOA seems to provide a viable extension of HOA that is also applicable to microphone arrays. The utility of the extra horizontal orders will ultimately have to be evaluated in accordance with the exact application. While microphone self-noise does not seem to affect the considered performance measures directly at higher frequencies because of adequate SNR, the noise may nonetheless become audible, and perceptual aspects will need to be considered. Listening tests can help determine the optimal trade-off between increased spatial acuity and background noise during playback.

Variations in the amplitude and phase characteristics of the transducers showed a marked effect on array performance above about 1 kHz, directly influencing the directionality of the reproduced sound field. These variations can – at least partly – be compensated for if the microphone characteristics are known. For the transducers used in the array, only the sensitivity was matched (at 250 Hz); therefore applying frequency response correction filters may offer a way to improve performance.

The results have also shown that correlated noise across microphone channels may not be negligible in practice, and needs to be considered in microphone array design.

3.6.3 Alternate approaches

An alternate MOA scheme has been proposed by Travis (2009), which may be better suited for more unequal order combinations. However, as seen from the effective order and array response measures (see Figs. 3.8 and 3.10), the problems put forward in the aforementioned study were not observed here, likely due to the already high periphonic (3D) order used in this work. It is clear, however, that highly unequal order combinations in the current MOA scheme show undesirable behavior, and potential advantages with alternate schemes should be investigated in future work.

Although spatial audio reproduction is the main focus of this paper, the MOA approach could also be utilized in beamforming applications where increased discrimination of horizontal sources is desired, for example.

3.7 Conclusions

This paper investigated applying the mixed-order ambisonics approach to microphone arrays. Specifically, two example arrays using 52 transducers were simulated: one using a quasi-regular

distribution of microphones and HOA processing, and the second using a higher density of microphones on the equator and MOA processing. Further, measurements on a physical realization of the MOA array were used to validate the simulations.

It was shown that a higher horizontal spatial directivity is obtained for the MOA array, but only above 3 kHz for the considered radius of 5 cm. For lower frequencies the regularization applied limited directivity, and the HOA and MOA arrays behaved similarly. In terms of robustness, the WNG was shown to be elevation dependent for the MOA array, with the peak of the WNG shifted up in frequency for a horizontal source. The WNG was otherwise comparable to that of the HOA array. Measurements of the MOA array prototype verified that the array performance can accurately be simulated if the transducer characteristics are taken into account, and confirmed that higher horizontal directivity with adequate SNR can be attained in practice.

It is clear that the trade-off between background noise and spatial resolution is a crucial parameter in microphone array processing for audio applications. Further, the consideration of a more realistic playback environment, various decoding strategies, and the handling of frequencies above the aliasing limit present questions that must be addressed with subjective tests in future work.

4

Performance assessment of mixed-order ambisonics for spherical microphone arrays[b]

Abstract

Mixed-order ambisonics (MOA) combines planar (2D) higher-order ambisonics (HOA) with lower-order periphonic (3D) ambisonics. MOA encoding from spherical microphone arrays has the potential to provide versatile recordings that can be played back using 2D, 3D or mixed systems. A procedure to generate suitable layouts for a given MOA order combination is introduced, consisting of rings of microphones at several elevation angles. Robustness and directivity measures were evaluated for four MOA layouts, each optimized for a planar order of 7, and a periphonic order of 1, 3, 5 and 7. Results showed that for non-horizontal directions, the MOA arrays behaved similarly to a HOA array of the corresponding, periphonic order. The inclusion of the higher-order, horizontally oriented SHFs led to increased directivity for horizontal sources, as a result of decreased horizontal beamwidth.

4.1 Introduction

Sound field recording techniques have received increasing attention in the last two decades. Several applications, e.g. in psychoacoustics and hearing instrument testing, require a realistic reproduction of these sound fields, i.e., of the spatial characteristics of the recorded scene. For these applications, high-quality recordings, scalable to playback setups of different sizes, either planar (2D) or periphonic (3D), are desirable.

Higher-order ambisonics (HOA) is a technique for either 2D or 3D systems (Moreau et al., 2006) that can process recordings from spherical microphone arrays for playback on arrays with various numbers of loudspeakers. More recently, mixed-order ambisonics (MOA) was investigated, which combines horizontal 2D HOA with lower order 3D ambisonics (Favrot et al., 2011). MOA for spherical microphone arrays can improve, compared to HOA, the directivity of horizontal sources while retaining some directivity for elevated sources (see Chapter 3). MOA recordings are very versatile and compatible with HOA playback. A MOA recording of combination order M_{2D}/M_{3D} could be played back (i) on either regular 3D or 2D loudspeaker arrays (using up to M_{3D} order and M_{2D} order HOA, respectively) or (ii) on 3D arrays with a

[b] This chapter is revised version of Favrot and Marschall (2012).

higher density of loudspeakers on the horizontal plane (using MOA with a combination order of up to M_{2D}/M_{3D}). It is desired that the encoded MOA signals be of similar quality than HOA signals of corresponding orders. The term "quality" here refers to (i) robustness to sensor noise and amplitude and phase mismatches and (ii) spatial resolution, i.e., the directivity of the array.

This study investigates the directivity and robustness of MOA encoding from spherical microphone arrays for different order combinations, for a fixed 2D order paired with various 3D orders. Because of the hybrid nature of MOA, a set of performance measures or metrics need to be evaluated, separately considering horizontal and vertical characteristics. First, a procedure to generate suitable microphone layouts for a given MOA order combination is introduced. Second, standard metrics are evaluated for the proposed MOA layouts in both horizontal and vertical directions. The metrics are compared to corresponding values for 2D and 3D HOA. Finally, the effect of the regularization on MOA encoding is discussed.

4.2 Background

First, the principle of MOA encoding using spherical arrays is briefly described here. The notations and nomenclature follow Moreau et al. (2006) and use spherical coordinates where a point in space is described by its radius r, azimuth θ ($-\pi \leq \theta \leq \pi$) and elevation δ ($-\pi/2 \leq \delta \leq \pi/2$) in relation to the origin O (the center of the spherical array).

4.2.1 Pressure on a sphere

The pressure p at a point (R, θ, δ) on the surface of a solid sphere can be approximated by (Moreau et al., 2006):

$$p(kR, \theta, \delta) = \sum_{m=0}^{M} W_m(kR) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} B_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta, \delta), \tag{4.1}$$

with k being the wave number, $W_m(kR)$ the weighting factor for the rigid sphere as described in Moreau et al. (2006), B_{mn}^{σ} the Fourier-Bessel series coefficients or ambisonics components of the sound field, and $Y_{mn}^{\sigma}(\theta, \delta)$ the real-valued spherical harmonic functions (SHFs) (Daniel, 2000) defined as

$$\begin{aligned} Y_{mn}^{\sigma}(\theta, \delta) = & \sqrt{(2m+1)(2-\delta_{0,n})\frac{(m-n)!}{(m+n)!}} P_{mn}(\sin\delta) \\ & \times \begin{cases} \cos n\theta & \text{if } \sigma = +1 \\ \sin n\theta & \text{if } \sigma = -1 \end{cases}, \end{aligned} \tag{4.2}$$

with $\delta_{0,n} = 1$ for $n = 0$, and 0 for $n > 0$, and where P_{mn} are the Schmidt semi-normalized associated Legendre functions of degree m and order n. The approximation gets more precise with increasing M. Considering the Q pressure signals captured by microphones

flush-mounted on the surface of the sphere, Eq. 4.1 can be written in matrix form as

$$\mathbf{s} = \mathbf{T} \cdot \mathbf{b}, \tag{4.3}$$

where $\mathbf{b}$ is the column vector of the $K = (M+1)^2$ ambisonics components B_{mn}^{σ}, $\mathbf{s}$ is a column vector of Q microphone pressure signals, and T represents the transfer matrix of size $Q \times K$ written as

$$\mathbf{T} = \mathbf{Y} \cdot \mathrm{diag}[W_m(kR)], \tag{4.4}$$

with the columns $\mathbf{Y}$ containing the SHFs evaluated at each microphone position (θ_q, δ_q).

4.2.2 Mixed-order ambisonics

The MOA scheme (Favrot et al., 2011) relies on a selection of SHFs. The MOA harmonic functions for an order combination of M_{2D}/M_{3D} consist of all SHFs up to order M_{3D}, and horizontally-oriented functions (with indices $n = m$) from order $m = M_{3D}+1$ to M_{2D}. The number of MOA harmonics K is then:

$$K = (M_{3D}+1)^2 + 2(M_{2D} - M_{3D}). \tag{4.5}$$

The matrix of MOA SHFs will be denoted $\widetilde{\mathbf{Y}}$.

4.2.3 Encoding

MOA components are encoded from the Q microphone signals using the ambisonic method (see Chapter 3). The frequency-dependent array encoding matrix $\mathbf{E}(f)$ $(K \times Q)$ derives the coefficients $\mathbf{b}$ (ambisonics signals) from the sampled pressures $\mathbf{p}$ as

$$\mathbf{b}(f) = \mathbf{E}(f)\mathbf{p}(f), \tag{4.6}$$

and is obtained by inverting Eq. 4.3. Using the regularized filtering approach described in e.g. Moreau et al. (2006), the encoding matrix $\mathbf{E}$ is approximated by

$$\mathbf{E}(f) \approx \mathrm{diag}\left[\frac{W_m^*(kR)}{|W_m(kR)|^2 + \lambda^2}\right]\widetilde{\mathbf{Y}}^{+}, \tag{4.7}$$

where $\widetilde{\mathbf{Y}}^{+}$ is the pseudo-inverse of $\widetilde{\mathbf{Y}}$ and λ is a regularization parameter. The regularization prevents the classical problem of excessive amplification of high orders at low frequencies, which, in practice, would lead to high noise levels at low frequencies (e.g. Moreau et al., 2006).

4.3 Methods

In order to investigate the performance of MOA spherical arrays for different mixed-order combinations, suitable sensor layouts first need to be derived for a given order combination.

i	δ_i	Q_i
0	0	$2M_{2D}+1$
$1\ldots N_r/2$	$\pm\frac{\pi}{2}\frac{i}{N_r/2+\epsilon}$	$\left\lceil\left(1-\frac{2\delta_i}{\pi}\right)\frac{Q_e}{2}+1\right\rceil$

Table 4.1: Specification of the investigated MOA layouts. Elevation angles δ_i and number of transducers Q_i for ring i.

4.3.1 Generating ring layouts

Similarly to HOA, the encoding of MOA signals relies on a least-squares minimization operation (the pseudo-inverse in Eq. 4.6). Therefore, a low condition number $\kappa(\widetilde{\mathbf{Y}})$ of the SHF matrix is necessary for a robust encoding of MOA signals. A prerequisite is that the number of transducers Q is greater than or equal to the number of MOA harmonics K (cf. Eq. 4.5), otherwise the system of equations in Eq. 4.3 is underdetermined. In practice, a higher number of sensors is needed to obtain good robustness, especially if the layout is irregular. In addition, the SHFs evaluated at the sensor positions should form an orthonormal basis. For MOA, one straightforward way to achieve the orthonormality of the horizontal harmonics is to sample the horizontal ring (the equator) with equiangular spacing.

The following procedure describes the generation of example layouts that were used in this study for a given M_{2D} and M_{3D} order. The layouts consist of rings of Q_i transducers with an equiangular spacing in azimuth and with an elevation angle of δ_i. First, each layout includes a horizontal ring (i.e. with $\delta_0=0$) of $Q_0=2M_{2D}+1$ transducers, in order to sample the horizontal SHFs up to order M_{2D}. Second, the total number of transducers on the rest of the rings, in order to fulfill $Q>K$, should at least be equal to

$$K-(2M_{2D}+1)=M_{3D}^2\,. \tag{4.8}$$

Here, more transducers than the minimum are chosen such that

$$Q_e=\begin{cases} M_{3D}^2+2 & \text{for even } M_{3D} \\ M_{3D}^2 & \text{for odd } M_{3D} \end{cases}. \tag{4.9}$$

Then, the number of elevated ($\delta\neq 0$) rings N_r is chosen as

$$N_r=\begin{cases} M_{3D} & \text{for even } M_{3D} \\ M_{3D}+1 & \text{for odd } M_{3D} \end{cases}. \tag{4.10}$$

Table 4.1 describes the elevation angle δ_i of ring i consisting of Q_i transducers. The $\lceil . \rceil$ brackets represent the ceiling function and $\epsilon=1$ for even M_{3D} and $\epsilon=0$ for odd M_{3D}.

In order to verify that this procedure provides suitable example layouts, the condition number $\kappa(\widetilde{\mathbf{Y}})$ was evaluated for $M_{2D}=7$ and $M_{3D}=1\ldots 7$, and listed in Table 4.2. None of the condition numbers are $\gg 1$, which indicates that the matrix $\widetilde{\mathbf{Y}}$ is not ill-conditioned.

M_{3D}	1	2	3	4	5	6	7
Q	17	21	25	39	49	81	101
K	16	19	24	31	40	51	64
$\kappa(\tilde{\mathbf{Y}})$	2.80	2.54	2.11	1.87	1.63	1.51	1.71

Table 4.2: Condition number of matrix $\tilde{\mathbf{Y}}$ with layouts for various order combinations. The number of transducers Q, and the number of spherical harmonic components K are also given for each order combination.

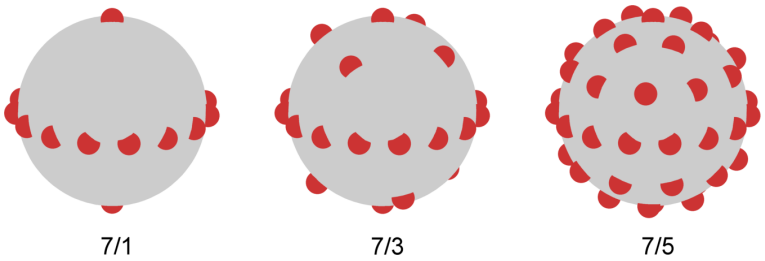


Figure 4.1: Example layouts for order combinations 7/1, 7/3 and 7/5.

As an example, the layouts obtained for combination orders 7/1, 7/3 and 7/5 are shown in Fig. 4.1. Sensors locations are indicated by red circles.

4.3.2 Metrics

Since MOA introduces an elevation dependence and considers horizontal and vertical directions separately, array performance metrics should be chosen and evaluated accordingly.

Using spherical microphone arrays for sound field reproduction entails similar signal processing to their use in beamforming, as the loudspeaker feeds for a regular loudspeaker array can be obtained by beams formed in the direction of the loudspeakers (e.g. Moreau et al., 2006). Established performance measures for spherical array beamforming (Meyer and Elko, 2004) are therefore relevant for the present study and are described in the following section. For a look direction (θ_0, δ_0) and for a regular beam pattern, the output of the beamformer can be written as (Meyer and Elko, 2004)

$$y(\theta_0, \delta_0, f) = \mathbf{y}_0 \mathbf{b}(f), \tag{4.11}$$

with $\mathbf{y}_0$ being the vector of K SHFs evaluated at (θ_0, δ_0), and $\mathbf{b}$ the obtained ambisonics signals after array processing. Fig 4.2 shows beam patterns, i.e., the beamformer output plotted as a function incoming plane wave direction, for a look direction of $(\theta_0 = 0, \delta_0 = 0)$, and for $M_{2D} = 7$ and $M_{3D} = 1, 3, 5$ and 7. A sphere radius of 5 cm and a frequency of $f = 5$ kHz was used for these simulations. The beampatterns in the left column are shown from the top, revealing the behavior of the beam in the horizontal plane, while in the right column a side-view is shown, displaying the beam in the vertical plane. It can be seen that the beam patterns consist of a main lobe in the intended direction, and several sidelobes in other directions.

In order to analyze the beampatterns, four established measures for beamforming arrays

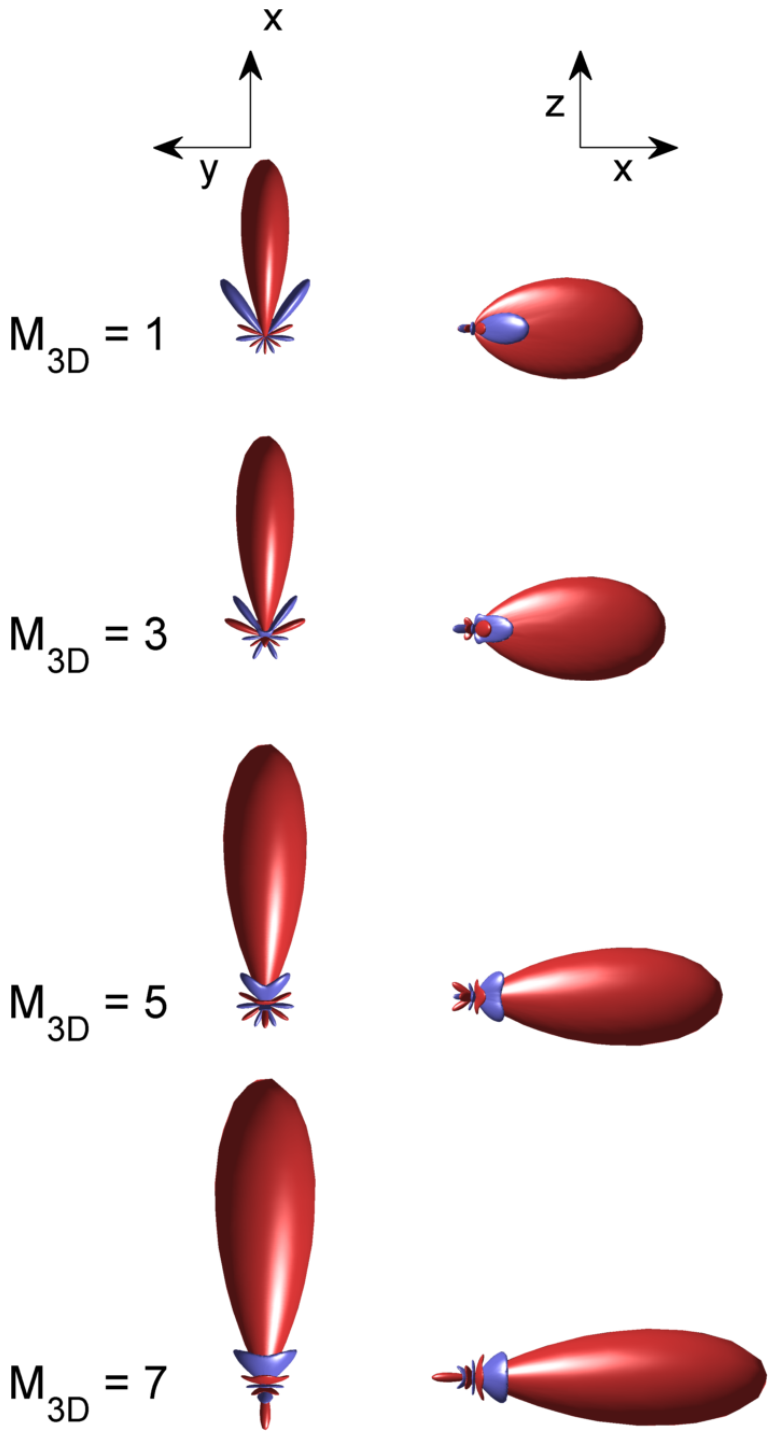


Figure 4.2: Mixed-order beampatterns for a look direction ($\theta_0 = 0, \delta_0 = 0$), $M_{2D} = 7$ and $f = 5$ kHz.

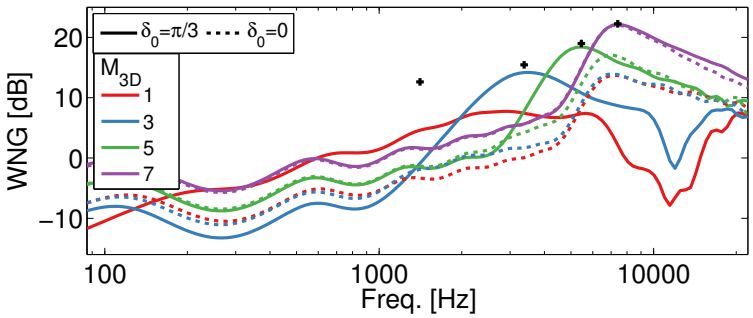


Figure 4.3: White noise gain for a horizontally-incident (dashed-lines) and an elevated (solid-lines) plane wave, for different MOA combinations.

were calculated, namely white noise gain (WNG), directivity index (DI), beamwidth, and maximum sidelobe level (MSL), and are described in the following section.

Another group of metrics relates to the characteristics of the sound field reproduced by an array of loudspeakers. The reproduced sound field is also influenced by the loudspeaker layout. Therefore, for these metrics, an ideal, sufficiently large loudspeaker array was considered in order to focus on the effects of the microphone array processing.

4.4 Metrics and results

In this section, the proposed metrics for the evaluation of MOA are described and results are presented for four hard-sphere arrays, with layouts generated using the procedure described in Sec. 4.3.1, with $M_{2D} = 7$ and $M_{3D} = 1, 3, 5$ and 7. The last case corresponds to 7th order HOA, and was used as a reference. A regularization parameter of $\lambda = 0.01$ was applied (cf. Eq. 4.7) in order to provide a more realistic prediction of MOA performance, and to demonstrate the impact of regularization. This value of the regularization parameter was found to give good results with simulated microphone self-noise.

4.4.1 White noise gain

The white noise gain (WNG) is a commonly used measure for estimating the robustness of beamforming microphone arrays against transducer self-noise, characteristic variations, and position errors (Meyer and Elko, 2004). WNG represents the signal power at the output of the beamformer over the sensor self-noise power, assuming spatially uncorrelated white noise. The inverse of the WNG shows how much this noise is amplified by the array processing. Thus, a higher WNG means more robust processing. For a unit-amplitude plane wave arriving from the look direction, the WNG can be calculated as (cf. Eqs. 4.11 and 4.7)

$$\mathrm{WNG} = 10\log_{10}\left(\frac{|\mathbf{y}_0\mathbf{b}_0|^2}{(\mathbf{y}_0\mathbf{E})^H(\mathbf{y}_0\mathbf{E})}\right). \tag{4.12}$$

Figure 4.3 shows the WNG for a horizontal ($\delta_0 = 0$) and an elevated ($\delta_0 = \frac{pi}{3}$) look direction for the four considered MOA layouts. For comparison, black crosses indicate the maximum of

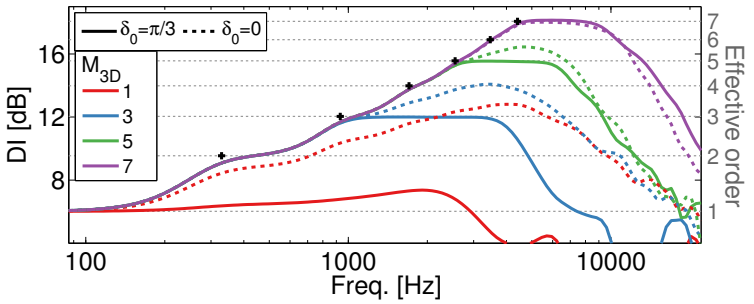


Figure 4.4: Directivity index for horizontal and elevated beams, for different MOA orders.

the theoretical WNG with HOA for each layout. The theoretical WNG is calculated as (Park and Rafaely, 2005)

$$\mathrm{WNG_t} = 10\log 10\left(\frac{QK^2}{\sum_{m=0}^{M}\frac{2m+1}{|W_m(kR)|^2}}\right). \tag{4.13}$$

$\mathrm{WNG_t}$ presents a typical band-pass characteristic centered at an optimum frequency defined as $f_{\mathrm{opt}}(m) = c\,m/2\pi R$ (Park and Rafaely, 2005), where c is the speed of sound, and m is the spherical harmonics order.

For MOA arrays, the WNG for the horizontal look direction shows a bandpass characteristic (dashed lines), centered at the optimum frequency for 7th order HOA, albeit with lower maximum values. This shows that for the MOA layout and processing, while the optimum frequency is shifted higher, robustness for horizontal sources is lower than for full 7th order HOA. For the elevated look direction, the WNG peaks are close to the theoretical WNG for the corresponding M_{3D} orders, except for $M_{3D} = 1$. The behavior of the MOA arrays in terms of WNG for elevated sources is thus more similar to that expected for the corresponding lower, M_{3D} order. Comparing WNGs for the horizontal and elevated look directions for a given order combination reveals that the extension to higher planar orders comes at the cost of decreased robustness for mid frequencies. At lower frequencies, the WNG is limited by regularization and is similar for horizontal and elevated look directions.

4.4.2 Directivity index

The directivity index (DI) indicates how directive a beamformer is, which is directly linked to the spherical harmonics order used. The DI is defined as the ratio of the beamformer output for an incoming plane wave in the look direction, relative to the average output of the beamformer for all incidence directions (e.g. Meyer and Elko, 2004):

$$\mathrm{DI} = 10\log_{10}\left(\frac{|\mathbf{y}_0\mathbf{b}_0|^2}{\sum_{l=1}^{L}|\mathbf{y}_0\mathbf{b}_l|^2}\right), \tag{4.14}$$

where $\mathbf{b}_l$ are the ambisonics signals for the l-th incoming plane wave out of L waves distributed evenly around the sphere. DIs were computed after the simulation of $L = 150$ plane waves for a horizontal and an elevated look direction, and are shown in Figure 4.4.

Horizontal gray dashed lines, labeled "effective order", represent the maximum achievable DI ($\mathrm{DI}_{\max}$) values for a regular beamformer of order $m = 1$ to 7, given as (Meyer and Elko, 2004)

$$\mathrm{DI}_{\max}(m) = 20\log_{10}(m+1). \tag{4.15}$$

As mentioned above, the regularization parameter attenuates the contribution of higher order components at low frequencies. Thus, the spatial directivity of the arrays is expected to be frequency dependent. The frequency $f_a(m)$ at which the order m is "activated" by the regularization parameter λ (cf. Eq. 4.7) is defined as the frequency for which $\lambda = |W_m(kR)|$. The black crosses in the figure indicate points at $(f_a(m), \mathrm{DI}_{\max}(m))$, for $m = 2\ldots7$.

For the elevated beam direction, DIs increase with frequency until the maximum DI for the given M_{3D} order is reached at about $f_a(M_{3D})$, as indicated by the black crosses. The DI below these frequencies is limited by the applied regularization, which reduces the effective order of the array in order to avoid excessive noise amplification. The DI starts to drop above the optimum frequency $f_{\mathrm{opt}}(M_{3D})$, where directivity is reduced due to the increasing contribution of spatial aliasing errors (Rafaely, 2005). For horizontal beams, DI values above f_a are higher than for elevated beams, but do not reach $\mathrm{DI}_{\max}(M_{2D} = 7)$. Instead, maximum horizontal DIs lie below the mean of $\mathrm{DI}_{\max}(M_{3D})$ and $\mathrm{DI}_{\max}(M_{2D})$. The frequency above which the DI starts decreasing again is also higher for horizontal than for elevated beams, but does not reach $f_{\mathrm{opt}}(M_{2D})$. The directivity of the MOA array is thus similar to the directivity of an M_{3D} order HOA array for elevated directions, and in between the directivity of an M_{2D} and an M_{3D} order HOA array for horizontal directions.

4.4.3 Beamwidth

The beamwidth is a measure of the spatial resolution of the beamformer. It is commonly defined as the angular width of the main lobe of the beampattern at the -3 dB points relative to the maximum. As this measure focuses only on the main lobe, and does not consider potential sidelobes, it only partially describes the total directivity of the beam pattern. The measure can, however, be conveniently evaluated both along the azimuth and the elevation, making it a relevant metric for MOA.

Beamwidths were computed for a horizontal look direction $(0, 0)$ for the four arrays, and are shown in Figure 4.5.

For comparison, "effective orders", representing the ideal beamwidth for HOA of orders 2 to 8, are plotted as dashed horizontal lines. Black crosses again mark the "activation frequency" $f_a(m)$ for each order. In general, beams get narrower with increasing frequency, as the beamwidth is limited by regularization at low frequencies. The beamwidth along the azimuth decreases up to $f_a(M_{2D})$, above which the minimum beamwidth is reached, corresponding to an effective order of about 7 for $M_{3D} = 5$ and 7, and 8 for $M_{3D} = 1$ and 3. The beamwidth along the elevation decreases up to the activation frequency of the periphonic order $f_a(M_{3D})$, with the minimum beamwidth roughly corresponding to the periphonic order, except for $M_{3D} = 1$. Thus, the beamwidth for horizontal sources is closely linked to the chosen periphonic (M_{3D})

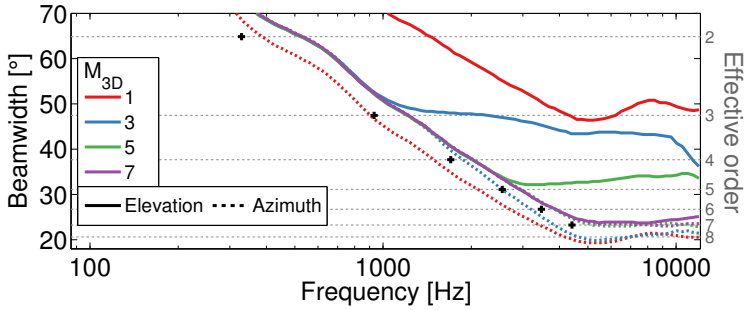


Figure 4.5: Beamwidth along the azimuth and elevation for a horizontal look direction, for different MOA orders.

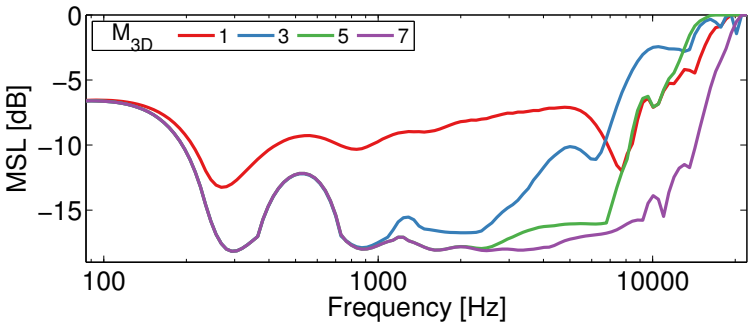


Figure 4.6: Maximum side lobe levels for a horizontal look direction, for different MOA orders.

order along the elevation, and to the planar (M_{2D}) along the azimuth. This behavior can also be seen directly in Figure 4.2.

4.4.4 Maximum side lobe level

The beamwidth alone does not fully describe array directivity, as any sidelobes will introduce sensitivity in directions other than that of the main lobe. High sidelobe levels will impair overall spatial directivity, and must be considered alongside the main lobe width. Maximum side lobe levels (MSLs) are defined as the level of the second highest peak in the beampattern relative to the level of the main lobe, and describe the minimum attenuation of sounds from directions outside the main lobe. MSLs were calculated for the four arrays and are displayed in Figure 4.6.

Comparing the MSLs for the different orders above 1 kHz reveals that the loss of directivity for MOA for horizontal sources at higher frequencies occurs due to increasing sidelobe levels, and not due to an increase in the width of the main lobe (cf. Figure 4.5). This in turn reflects the increasing contribution of spatial aliasing for higher frequencies. The highest sidelobe levels are seen for the $M_{3D} = 1$ array, which indicates that this layout and order combination may not be optimal.

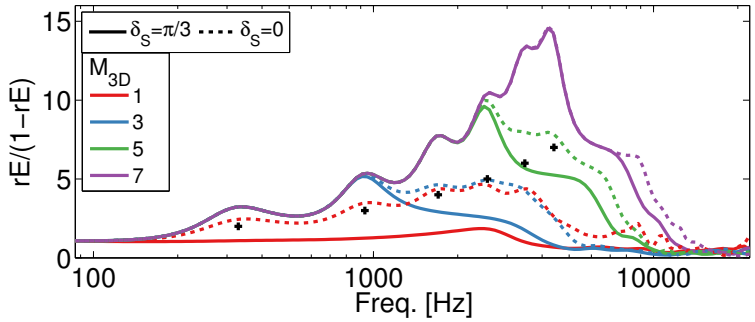


Figure 4.7: Effective order transformed from r_E, for a horizontal and an elevated sound source, and for different MOA orders.

4.4.5 Sound field reproduction

Sound fields captured by microphone arrays can in turn be reproduced by playback over arrays of loudspeakers. Deriving the appropriate loudspeaker driving signals from HOA or MOA signals is termed "decoding". Comparison of the reconstructed sound fields can provide a measure of array encoding performance when using a sufficiently large loudspeaker array, such that any errors are primarily due to the microphone array. This is easily achieved by applying a virtual loudspeaker array, where each element is considered to emit ideal plane waves. The comparison used here further relies on the norm of the "energy vector" r_E, a concept proposed by Gerzon (1992). This measure quantifies, on the playback side, the spatial distribution of energy. r_E is 1 if all energy is from one specific direction, and 0 if it is distributed evenly, and thus indicates the directivity of the reproduced sound field. Ideally, for 3D HOA, $r_E = M/(M+1)$ according to Daniel (2000). The transform $M_{\text{eff}} = r_E/(1-r_E)$ can then be used to relate r_E to the effective order M_{eff}. The transformation was used here to ease the comparison between different MOA order combinations.

A virtual, 204-element loudspeaker array, based on a spherical t-design (Hardin and Sloane, 1996) was used to decode the MOA signals encoded by the four microphone arrays considered. The r_E values were calculated after the simulation of a horizontal ($\delta_s = 0$) and an elevated ($\delta_s = \pi/3$) sound source. The corresponding effective orders are shown in Figure 4.7. The black crosses indicate points at which order $m = 2\ldots7$ is activated $(f_a(m), m)$. It can be seen in general that a higher effective order is reached for horizontal sources than for elevated sources above the frequency $f_a(M_{3D})$, demonstrating the effect of mixed-order processing. Prominent peaks occur in the transformed r_E values at the activation frequencies f_a, where higher effective orders than the highest applied order of 7 are seen.

To verify the impact of the regularization scheme, another simulation was performed without applying regularization. Since only an ideal plane wave was considered without any additional noise sources in the simulations, excessive signal gains did not occur. The r_E results obtained in this ideal case are shown in Figure 4.8. It can be seen that this time, effective orders were limited to the applied range of orders, not exceeding 7. This suggests that the observed peaks in Figure 4.7 are an artifact of the regularization scheme used, with which r_E does not change smoothly with frequency. Further, for the horizontal source and $M_{3D} = 5$, the effective

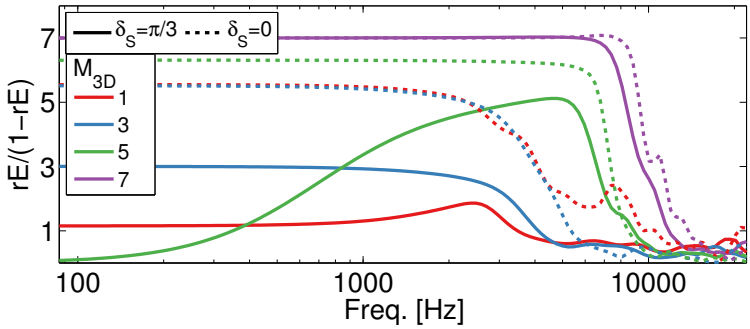


Figure 4.8: Effective order transformed from r_E, without regularization, shown for a horizontal and an elevated sound source, and for different MOA orders.

order decreases towards low frequencies. Although not shown explicitly here, it was confirmed that with the specific layout, some 6th order spherical harmonic components were not well captured, which, without regularization, introduced substantial errors at low frequencies.

4.5 Discussion

Various metrics were evaluated for the set of MOA layouts considered here. Separate evaluation along horizontal and vertical dimensions provided a tool to assess the effects of MOA processing in terms of robustness and directivity. For the latter characteristic, the directivity index provided an aggregated view of the results from the beamwidth and side lobe level measures. On the other hand, the beamwidth allowed a separation of horizontal and vertical spatial resolution for a single beam. For the reproduction side, the norm of the energy vector r_E provided another measure of directivity.

In general, the measures showed that various order combinations in MOA can be used to control planar versus periphonic performance. However, benefits from the higher planar order (M_{2D}) could only be seen above the activation frequency for the periphonic order (M_{3D}). Further, due to the increasing slope of $W_m(kR)$ with order (cf. Eq. 4.7; see e.g. Moreau et al., 2006), the frequency range in which SHFs of higher orders contribute significantly decreases with increasing order. The beamwidth measures also showed that for horizontal beams, the increased directivity is provided by a decrease in beamwidth along the azimuth, but not along the elevation.

The regularization scheme applied introduced peaks in the r_E response around the activation frequencies of each spherical harmonic order. Such an impact was not observed with the directivity index. Whether it is more desirable to obtain a smooth r_E response with frequency needs further investigation. The amount and type of regularization applied controls the frequency-dependent trade-off between spatial directivity and background noise. These parameters are expected to have a pronounced impact on the perception of the reproduced sound field, and thus subjective tests will likely be required to obtain the optimal parameters for audio applications.

The layout generation algorithm presented here serves only as an example, with further

optimization needed to generate practically applicable MOA layouts. In particular, layouts for low periphonic orders had somewhat higher condition numbers, which does not pose a numerical problem, but indicates that the layout may not be optimal. Layouts for high periphonic orders (i.e. 6 and 7) on the other hand appear to be using more transducers than necessary. A more advanced layout generation procedure could consider the orthonormality error of the sampled SHFs (e.g. Li and Duraiswami, 2007), or the spatial error between the reproduced and desired SHFs (Moreau et al., 2006; Favrot et al., 2011).

The investigated MOA scheme may not be well-suited for highly unequal order combinations, as evidenced by the very large difference in performance characteristics for horizontal and elevated sources for the 7/1 order combination. Similar observations were made by Travis (2009) in connection with mixed-order loudspeaker reproduction. Different mixed-order schemes, such as the one presented in the aforementioned paper could be investigated in future work.

4.6 Summary and conclusions

The aim of MOA is to combine a higher-order planar representation of the sound field with a lower-order periphonic representation. This study investigated the properties of MOA spherical microphone arrays with various performance metrics. A procedure to generate layouts for a given order combination was described. To highlight the properties of MOA, four example layouts were generated, and encoding performance was assessed by (i) evaluating beamformer metrics for a horizontal and an elevated beam separately, and by (ii) evaluating the reproduced sound field for a horizontal and an elevated sound source. Results showed that for non-horizontal directions, the MOA arrays behaved similarly to a HOA array of the corresponding, periphonic order. The inclusion of the higher-order, horizontally oriented SHFs led to increased directivity for horizontal sources, as a result of decreased horizontal beamwidth. However, overall directivity was lower than that of HOA of the corresponding, planar order. Further, improvements were restricted to frequencies above the frequency where the applied regularization scheme activated the higher orders. Nonetheless, adjusting MOA order combinations provided a way to control horizontal vs. vertical performance characteristics of microphone arrays.

5

Sound field reconstruction performance of a mixed-order ambisonics microphone array[c]

Abstract

Accurate spatial audio recordings are important for a range of applications, from virtual sound environments for hearing research to the evaluation of communication devices, such as hearing instruments and mobile phones. Spherical microphone arrays present one method, whereby accurate spatial recordings can be made. Recently, a mixed-order ambisonics (MOA) approach was proposed to improve the horizontal spatial resolution of spherical arrays. This was achieved by increasing the number of microphones near the horizontal plane while keeping the total number of transducers fixed. The approach is motivated by the fact that human spatial hearing is most acute in the horizontal plane. This study investigates the performance of a MOA rigid-sphere microphone array in terms of sound field reconstruction error, and the impact of variations in microphone characteristics. Specifications of a commercially available microphone were used to simulate self-noise, sensitivity, and phase response variations between the microphones. To quantify the reconstruction error and the "sweet area", the reconstructed sound field based on a simulated as well as an actual array measurement was compared with the reference sound field for both horizontal and elevated sources. It was found that at mid to high frequencies, the MOA approach results in a larger sweet area for horizontal sources than for elevated sources.

5.1 Introduction

There has been increasing interest in the past years in applying spherical microphone arrays for sound field capture, with the aim of analyzing (e.g. Meyer and Elko, 2002; Rafaely, 2005; Park and Rafaely, 2005; Rafaely et al., 2007; Li and Duraiswami, 2007; Williams and Takashima, 2010; Jacobsen et al., 2011) or reproducing the sound field using a loudspeaker array (Abhayapala and Ward, 2002; Meyer and Agnello, 2003; Moreau et al., 2006). From the perspective of hearing research, accurate recording and reproduction of spatial audio is important for the design of virtual environments, where complex and realistic acoustic scenes may be presented to the listener in a controlled manner. Such virtual environments can also aid the development and

[c] This chapter is a revised version of Marschall and Chang (2013).

evaluation of communication devices, such as hearing instruments or mobile phones. To the extent that the sound field is reproduced accurately, technical devices should also perform in the virtual space as they would in a real environment. To achieve a realistic and high quality reproduction, high spatial resolution, wide bandwidth and low noise are needed. It is therefore important to analyze the impact of various noise and error sources in microphone arrays for sound field reproduction.

A major limitation in the accurate recording of sound fields at higher frequencies is spatial aliasing (Rafaely et al., 2007), which stems from the discrete sampling realized when arrays of microphones are used. Practical considerations limit the number and arrangement of the microphones, and these in turn limit the maximum directivity and operating frequency of the array. In an effort to improve the performance of rigid-sphere arrays in the horizontal plane, previous work investigated arrays with a higher density of microphones on the equator (Favrot et al., 2011; Favrot and Marschall, 2012; Marschall et al., 2012). A mixed-order ambisonics (MOA) approach was employed, which, similarly to the approaches proposed by Daniel et al. (2003) and Travis (2009), combines a subset of higher-order spherical harmonic functions with a complete lower-order set. This is done in order to increase spatial resolution in the desired direction, usually the horizontal plane. The approach is motivated by the fact that the most important sound sources are usually in or near the horizontal plane, where human spatial hearing is generally most accurate (Blauert, 1997b).

While the properties of MOA microphone arrays have been investigated in previous work, sound fields recorded and reconstructed using MOA have not yet been investigated in detail. The properties of the reconstructed sound field in a head-sized region around the center are of particular interest, as the system's upper frequency limit is given by the frequency at which errors exceed a specific threshold at the listener's ears.

In this study, simulations as well as measurements of a MOA rigid-sphere microphone array were evaluated in terms of sound field reconstruction error in the horizontal plane. A reference sound field consisting of a single sound source in anechoic conditions was considered, either in the horizontal plane or at an elevated position. In order to evaluate the error introduced by the spherical microphone array, the sound field was reconstructed from the measurements made with the spherical array, and compared with the reference sound field. Figure 5.1 illustrates this concept. The sound field reconstruction error was quantified by evaluating the mean-square error, the magnitude error, as well as the normalized cross-correlation between the reference and reconstructed sound fields. With this approach, no assumptions were made regarding the loudspeaker setup; the reconstructed field was derived directly from the measured (or simulated) spherical harmonic coefficients, reflecting error-contributions from the microphone array only.

Besides array geometry, the characteristics of the microphones also affect the performance of the microphone array. Self-noise, as well as mismatch between the individual transducers can both degrade array performance (Moreau et al., 2006). Based on the specifications of a commercially available array microphone, realistic self-noise, as well as sensitivity and phase response variations between the microphones were considered.

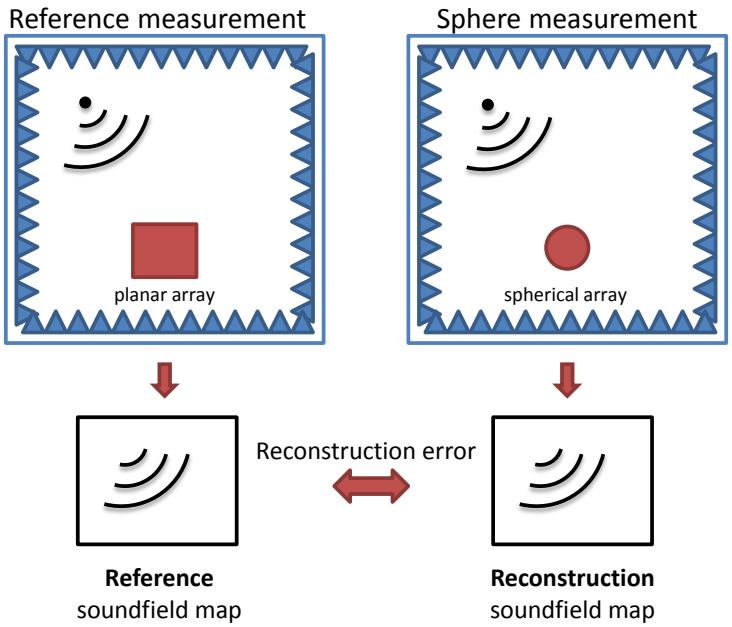


Figure 5.1: Illustration of the approach used to quantify the sound field reconstruction error of the spherical microphone array. The reconstruction error is evaluated by comparing the reference sound field (left) with the sound field reconstructed from a spherical array measurement of the same field (right). The applied error measures and further details are described in the text.

5.2 Methods

5.2.1 Mixed-order ambisonics

Here, a brief description of the MOA scheme that was used in the present study is provided. More detailed information can be found in Chapter 3. The pressure at a point with radius r, azimuth θ, and elevation δ, using mixed-order spherical harmonics expansion, can be approximated by

$$\begin{aligned} P(r,\theta,\delta) \simeq & \sum_{m=0}^{M_{3D}} W_m(kr) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} B_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta,\delta) \\ & + \sum_{m=M_{3D}+1}^{M_{2D}} W_m(kr) \sum_{\sigma=\pm 1} B_{mm}^{\sigma} Y_{mm}^{\sigma}(\theta,\delta), \end{aligned} \tag{5.1}$$

where k is the wavenumber, B_{mn}^{σ} are the expansion coefficients or ambisonics components, Y_{mn}^{σ} are the real-valued spherical harmonics functions , and $W_m(kr)$ is the radial function, which for a rigid sphere is given as (Meyer and Elko, 2002)

$$W_m(kr) = i^m \left(j_m(kr) - \frac{j_m'(kR)}{h_m'(kR)} h_m(kr) \right), \tag{5.2}$$

where R is the radius of the sphere, j_m and h_m are spherical Bessel and Hankel functions, and the primes indicate derivatives with respect to the argument.

In the mixed-order scheme used here, as shown in Eq. 5.2, all spherical harmonics functions

Figure 5.2: The 52-channel, mixed-order ambisonics microphone array investigated in this study.

Y_{mn}^{σ} are included up to the periphonic order M_{3D}. For higher orders $M_{2D} > M_{3D}$, only the horizontal functions (with indices $n = m$) are included in the summation.

5.2.2 Array design

For the measurements as well as the simulations, the hard-sphere MOA microphone array developed previously at the Technical University of Denmark was considered (Marschall et al., 2012), shown in Figure 5.2. The array consists of 7 rings, from top to bottom, of 2, 6, 10, 16, 10, 6, and 2 microphones for a total of 52 transducers, and uses B&K Type 4959 microphones. A radius of 5 cm was chosen, which is still a practically realizable size considering the number of microphones, and was also shown to be a suitable size for spatial audio recordings by Weller et al. (2011). The layout was optimized for an M_{2D} / M_{3D} order combination of 7/5. These orders were also used for all array processing.

5.2.3 Simulation framework

A simulation framework was developed that implements a reconstruction of the incident sound field based on a simulated measurement with a MOA microphone array in the frequency domain. A reference sound field P_{inc} consisting of an incoming plane wave was generated, and the sound pressure on the surface of the rigid sphere $P(R,\theta,\delta)$ was calculated at the measurement points (Williams, 1999). The incoming plane wave $\widetilde{P}_{\text{inc}}$ was reconstructed from the surface pressure on the sphere by first deriving the expansion coefficients $\widetilde{B}_{mn}^{\sigma}$, which can be expressed from Eq. 5.1 (see Chapter 3). Regularization was applied to avoid excessive amplification of noise (see Chapter 3) , with the same regularization parameter of $\lambda = 0.01$. Then, by removing the second term in Eq. 5.2 that represents the scattering effect due to the

existence of the rigid sphere, the reconstructed pressure can be expressed as

$$\begin{aligned}\widetilde{P}_{\text{inc}}(r,\theta,\delta) \simeq &\sum_{m=0}^{M_{3D}} i^m j_m(kr) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} \widetilde{B}_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta,\delta) \\ &+ \sum_{m=M_{3D}+1}^{M_{2D}} i^m j_m(kr) \sum_{\sigma=\pm 1} \widetilde{B}_{mm}^{\sigma} Y_{mm}^{\sigma}(\theta,\delta).\end{aligned} \tag{5.3}$$

This reconstructed sound field was compared with the reference field, and the error between these fields was used as performance indicator. The exact error measures used are described in Sec. 5.2.5.

Measurement errors were simulated and added to the surface pressure on the sphere. Errors due to self-noise and variations in transducer characteristics were simulated based on the specifications of the Brüel & Kjær type 4959 array microphone (Brüel & Kjær, 2012). Microphone self-noise was simulated by adding a value to the simulated pressure measurement, whose magnitude for each frequency band was set to correspond to a typical self-noise magnitude spectrum for the transducer type, with random phase. As the noise is defined in the frequency domain as a spectral density, the sound pressure levels given in the following are normalized to a bandwidth of 1 Hz, as the overall pressure depends on the bandwidth considered. At the two frequencies used in the simulations, 400 Hz and 4000 Hz, the self-noise levels were −8 dB SPL and −14 dB SPL, respectively. The simulated plane wave had a sound pressure level of 42 dB SPL at each frequency, to match the measurements, unless otherwise stated. This resulted in a signal-to-noise ratio (SNR) of 50 dB and 56 dB at 400 Hz and 4000 Hz, respectively.

To simulate sensitivity and phase response variations, each transducer was assigned a frequency dependent amplitude and phase characteristic (described as a relative amplitude and phase deviation for each frequency) randomly selected from a set of example characteristics conforming to the specifications.

5.2.4 Measurement setup

The measurements were carried out in the small anechoic chamber at the Technical University of Denmark, which has a lower limiting frequency of about 100 Hz. The measurement setup is shown in Figure 5.3. The sound scene consisted of a single loudspeaker (Dynaudio BM6P), 2.5 m from the center of the measurement area. The loudspeaker was mounted level with the measurement plane. For the measurements with the spherical array, the array was mounted such that the center of the sphere coincided with the center of the measurement area. To measure the reference sound field, a line array with 7 equally spaced microphones (B&K type 4958) was mounted on a turntable, as shown in Panel A of Figure 5.3. Using several successive measurements with different positions of the line array, the sound field was sampled in the horizontal plane every 1.5 cm from the center along the radius, up to 10.5 cm, and every 10° in azimuth. The loudspeaker was driven with a white noise signal and was adjusted to measure 84 dB SPL (lin. weighting) at the microphone array. The recordings were made with

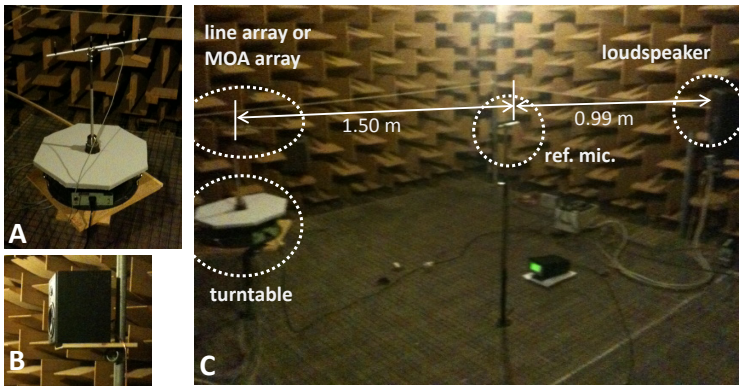


Figure 5.3: Measurement setup. Panel A shows the rotating line array used for the reference measurement. Panel B shows the loudspeaker, and Panel C provides an overview of the measurement setup.

the B&K Pulse platform, sampled at 32 kHz, and then processed in the frequency domain, with a frequency resolution of 1 Hz.

5.2.5 Error measures

Based on the simulated or measured spherical harmonic coefficients, the sound field was reconstructed using Eq. 5.3 in the horizontal plane ($\delta = 0$) in 36 positions (every 10° in θ), at several radii r. In the same 36 positions at each radius, the reference sound field was either directly measured, or, for the simulations, calculated analytically, assuming an acoustic plane wave from the direction of the loudspeaker. Hence, the reference and the reconstructed sound fields at each radius r can be expressed as 36-element vectors, $\mathbf{P}$ and $\widetilde{\mathbf{P}}$. For example, $\mathbf{P}$ is defined as

$$\mathbf{P}(r) = \left[P_{\text{inc}}(r, \theta_1, \delta = 0) \quad P_{\text{inc}}(r, \theta_2, \delta = 0) \quad \cdots \quad P_{\text{inc}}(r, \theta_{36}, \delta = 0) \right]^{\text{T}}. \tag{5.4}$$

The reconstruction error was calculated between $\mathbf{P}(r)$ and $\widetilde{\mathbf{P}}(r)$ on a ring for each r. The considered error measures are described in the following.

First, the mean-square error was calculated to show the overall error, given as

$$e_{\text{mse}}(r) = \sqrt{\frac{\sum_{i=1}^{36} |P_i(r) - \widetilde{P}_i(r)|^2}{\sum_{i=1}^{36} |P_i(r)|^2}}. \tag{5.5}$$

This measure is sensitive to errors in both magnitude and phase. For spatial audio applications, absolute phase error (such as a constant phase shift) may be less problematic than a disturbance in the relative phase relationships in the sound field, because the latter could affect auditory localization (e.g. Blauert, 1997b). In an effort to also consider the effects of magnitude error separately, a magnitude error and a correlation coefficient were also calculated between

the reference and reconstructed field. The magnitude error is defined as

$$e_{\text{mag}}(r) = \sqrt{\frac{\sum_{i=1}^{36} \left[|P_i(r)| - |\widetilde{P}_i(r)| \right]^2}{\sum_{i=1}^{36} |P_i(r)|^2}} \tag{5.6}$$

and is insensitive to phase differences.

The spatial correlation coefficient CC is the normalized cross-correlation between the (complex) reference and reconstructed fields, defined as

$$CC(r) = \frac{|\mathbf{P}(r)^* \cdot \widetilde{\mathbf{P}}(r)|^2}{(\mathbf{P}(r)^* \cdot \mathbf{P}(r))(\widetilde{\mathbf{P}}(r)^* \cdot \widetilde{\mathbf{P}}(r))}. \tag{5.7}$$

This measure is similar to the modal assurance criterion (Allemang, 2003) that has been used extensively in structural acoustics for the evaluation of mode shapes. The correlation coefficient provides a value between 0 and 1, indicating the degree of similarity between the magnitude and phase response shapes of the two sound fields. Values close to 1 indicate similar responses. The measure is insensitive to constant phase shifts and to amplitude scaling, thus providing another view of the reconstruction error that may be more appropriate for spatial audio applications.

5.3 Results and discussion

5.3.1 Sound field reconstruction

Simulations

The sound field is considered on a disc around the origin in the horizontal plane, in several discrete points as described previously. Figure 5.4 shows the magnitude and phase of the reference and reconstructed sound fields, for a 42 dB SPL plane wave at 4 kHz. Looking at the reference magnitude (top left panel), it can be seen that the magnitude is constant over the displayed area, as expected for a plane wave. The reference phase (bottom left panel) shows the phase decreasing in the $+x$ direction, which, with the sign convention used here, indicates a propagating wave along the x axis, from left to right.

Both the magnitude and phase of the reconstructed field (middle and right panels) show some differences compared to the reference field. Considering the magnitudes (top), it can be seen that the target level is only reached up to about $r = 10$ cm distance, and for higher radii, there is a decrease in level along the y axis, and an increase along the x axis. Similarly, the phase responses (bottom) correspond well with the reference for radii smaller than about 10 cm, but diverge for higher r. Comparing the reconstructions with simulated noise (right panels) and without simulated noise (middle panels), very little difference can be seen between them.

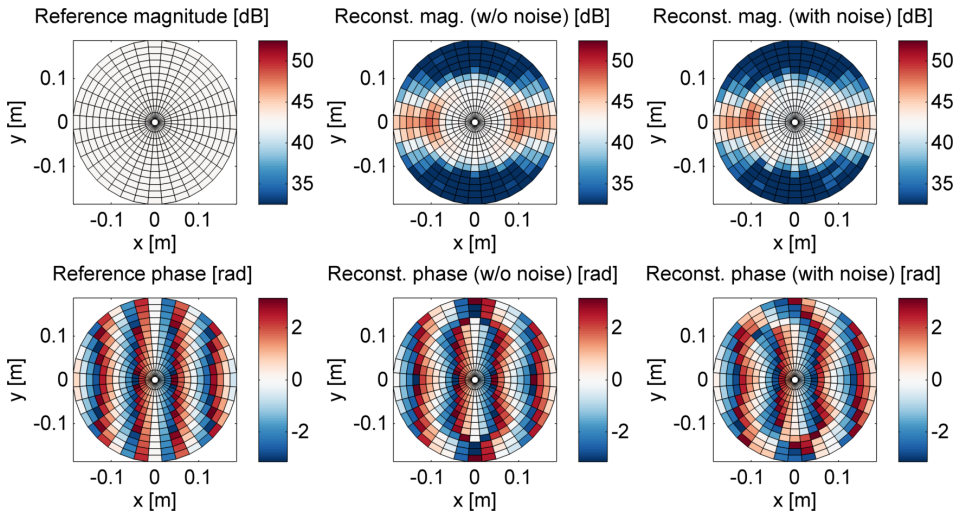


Figure 5.4: Magnitude and phase of the reference sound field (left), reconstructed sound field without noise simulation (middle), and reconstructed sound field with noise simulation (right). The reference field is a 4 kHz plane wave, with amplitude 42 dB SPL, propagating in the $+x$ direction. The displayed area is a horizontal disc, with radius $r = 20$ cm.

These results demonstrate a fundamental property of arrays using spherical harmonics, namely that the error of the reconstruction is lowest at the origin and increases with the radius r. More precisely, the contribution of higher orders for small kr, as given by the radial function (Eq. 5.2), is low. Thus, when the reconstruction is limited to order M, the error due to the truncation of the spherical harmonic series will be lower closer to the origin and for lower frequencies. An estimate for the order M that is needed for the representation of a sound field with wavenumber k within a sphere with radius r is usually given as $M \geq kr$ (Rafaely, 2005; Moreau et al., 2006). Expressing r results in

$$r \leq \frac{cM}{2\pi f} \tag{5.8}$$

which, for a horizontal order of $M_{2D} = 7$ and $f = 4$ kHz, results in a radius of approximately 10 cm. This roughly corresponds to the radius in which the reconstructed sound field matches the reference in Figure 5.4. The similarity of the results with and without noise simulation suggests that, at this frequency, limitations imposed by the array geometry, i.e. the aliasing error, dominate. Errors introduced by the transducers are secondary.

Measurements

Figure 5.5 shows direct measurements of the reference sound field in the anechoic chamber (left) and the reconstructed field based on the spherical array measurement (right) for 4 kHz and a horizontal sound source. Note that the displayed area is smaller than in Figure 5.4, as the measurement extended only to $r = 10.5$ cm from the center. The measured results are very similar to the simulations presented previously, and show that the plane wave was well realized in the experimental setup, and that the reconstructed field appears to correspond well

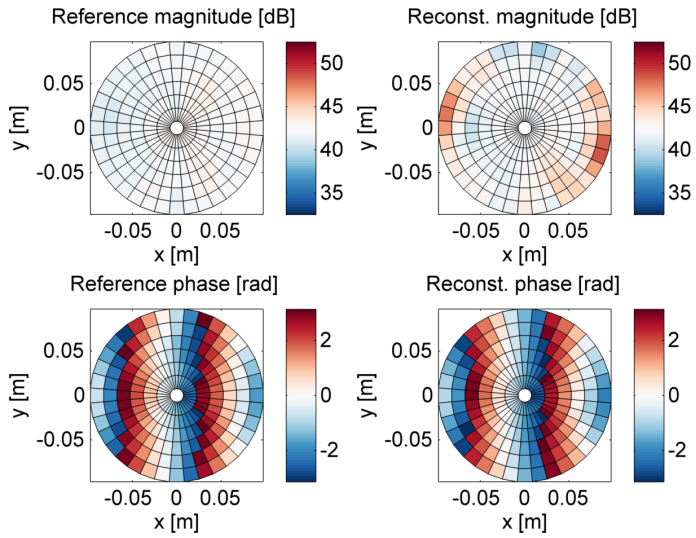


Figure 5.5: Magnitude and phase of the measured reference sound field (left), and the sound field reconstructed from the spherical array measurements (right), at 4 kHz. The displayed area is a horizontal disc, with radius $r = 10$ cm.

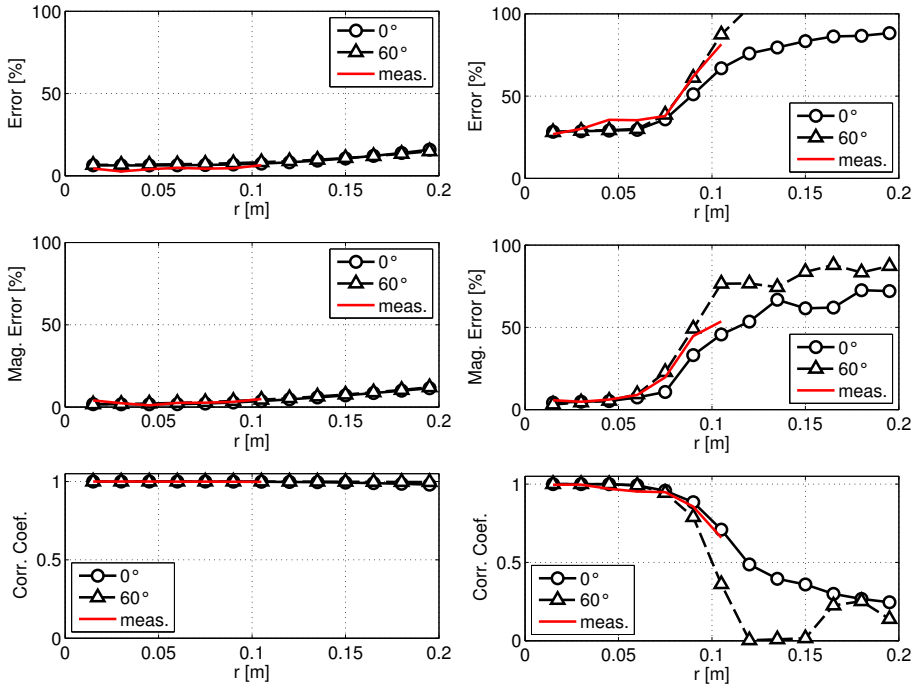
with the reference. Some magnitude errors are apparent at the very edges of the 10 cm disc, as was also seen in the simulations at that radius.

So far, only a qualitative comparison between the reference and reconstructed sound fields was presented. In order to quantify the reconstruction error, the error measures described in Sec.5.2.5 are presented in the following.

5.3.2 Effect of elevation angle

One of the desired outcomes of applying a mixed-order processing is to improve the performance of the array for horizontal sources. In order to investigate the effect of source elevation, the sound field reconstruction error was considered for a horizontal and an elevated source at two frequencies. Figure 5.6 shows the three error measures for a relatively low frequency of 400 Hz (left panels) and a higher frequency of 4000 Hz. Black curves indicate simulation results, while red curves indicate measured data for a horizontal source. Regarding the low-frequency results (left panels) it can be seen that both the total error and the magnitude error remain below about 20%, and that the correlation coefficient is also close to 1 for the whole range of radii investigated. This suggests that the sound field is well reconstructed at this frequency in the investigated area. Further, no marked differences are observed between the horizontal and the elevated source directions in terms of reconstruction error.

Conversely, at the higher frequency, both the total and the magnitude error are at or above about 50% for a radius of 10 cm for both incidence angles. The correlation coefficient is also sharply reduced above $r = 9$ cm, indicating that the sound field reconstruction fails. For the horizontal source, the results are in line with the estimation of the maximum reconstruction radius of about 10 cm. Comparing the horizontal and the elevated source angle, it can be observed that the errors are higher for the elevated source. Thus, for a horizontal source, the radius for a fixed error and frequency is increased, or alternatively (not explicitly shown here),



(a) 400 Hz (b) 4000 Hz

Figure 5.6: Sound field reconstruction error as a function of distance r from the origin, shown for two source elevation angles (0° and 60°), and two frequencies (400 Hz and 4000 Hz). Simulation results shown in black, measurements for a horizontal source shown in red.

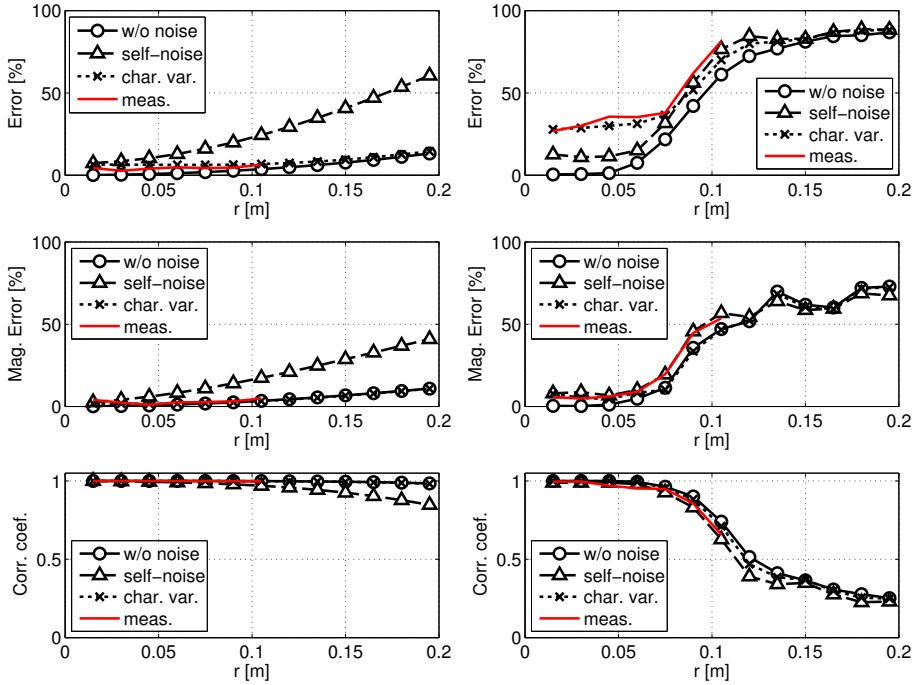
the frequency range for a fixed radius and fixed error is extended. This demonstrates the expected result of the MOA processing in terms of sound field reconstruction error.

The measured sound field reconstruction errors (red curves) for a horizontal source are generally well-matched to the simulations, although the measured reconstruction error is slightly higher than predicted for the 4 kHz condition. This implies that the transducer noise and error model accurately reflects the real-life behavior of the microphones, although there may be a small effect of additional noise or error sources that were not considered, such as electrical noise or transducer positioning errors.

It is interesting to note that the level of the total error (top right panel) is relatively high (30-35%) for small r, up to about 6 cm, but that both the magnitude error and the correlation coefficient show a low error for the same region. This suggests that there is an absolute phase shift introduced somewhere in the recording chain, which affects the total error, but not the magnitude error or the correlation coefficient, because the latter two measures are insensitive to absolute phase. The source of the error is thus postulated to be the microphone phase response, and this is further investigated in the next section.

5.3.3 Contribution of error sources

The effect of microphone self-noise and the variations in microphone characteristics (differences in magnitude and phase responses) were also considered separately. Simulations with no added noise or error sources were also run. The latter condition indicates errors due to



(a) 400 Hz (b) 4000 Hz

Figure 5.7: Sound field reconstruction error as a function of distance r from the origin, shown for two frequencies (400 Hz and 4000 Hz), for a horizontal sound source. Three simulations are shown in each graph: no noise simulation applied (circles), only microphone self-noise simulated (triangles), and only amplitude and phase characteristic variations simulated (crosses). Measured sound field reconstruction errors are again shown in red. The simulated plane wave amplitude was reduced by 20 dB compared to the measurement.

the truncation of the spherical harmonic series and spatial aliasing. Similarly to the previous section, three error measures are displayed in Figure 5.7 for 400 Hz (left) and 4000 Hz (right). For reference, the measured sound field reconstruction errors from Figure 5.6 are replotted (red curves). The simulated plane wave amplitude was reduced here by 20 dB compared to the measurement (and the simulations of the previous section) in order to show the effect of self-noise more clearly. The SNR was thus reduced to 30 dB at 400 Hz, and 36 dB for 4000 Hz.

Regarding the low frequency results (left panels) for the simulations with no noise (circles) and with only characteristic variations (crosses), the errors are low across all measures. At this frequency, self-noise (triangles) is the dominant error source, its effect increasing with radius. Naturally, the contribution of self-noise error is dependent on the source amplitude. In comparison to the left panels of Figure 5.6, where the total error is shown for a higher source amplitude, the error is reduced.

At the higher frequency (right panels), note the elevated total error with only the characteristic variations simulated (crosses), coupled with a low magnitude error and high correlation coefficient for $r < 7$ cm. The simulations with self-noise only are closer to those without noise, and neither error source results in a marked change from the no-noise condition for $r > 8$ cm. This suggests that the increased total error also observed in Figure 5.6 for small radii is indeed caused by the phase responses of the microphones. Aside from that, at high frequencies, aliasing error appears to dominate. Further, self-noise simulations with a lower

SNR seem to be a better match to the measured data than with the matching source amplitude shown previously in Figure 5.6. This points to simulated self-noise being underestimated at 4 kHz.

5.4 Conclusions and outlook

Summarizing the results, it can be said that MOA processing exhibits the desired behavior under realistic conditions, providing improved performance for horizontal sound sources. More specifically, the sound field reconstruction error is reduced for horizontal sources as compared to elevated sources, at least for higher frequencies. The results correspond well with the findings of our previous studies, where beamforming measures were used to evaluate the benefits of MOA for microphone arrays (see Chapters 3 and 4). There, it was found that the benefit of MOA in terms of spatial resolution appears mainly at mid to high frequencies, corresponding to the frequencies at which the higher orders are "activated" by the regularization scheme used.

It should be noted that the performance of the array, especially at low frequencies and with regard to self and background noise, depends on the regularization parameter λ (Moreau et al., 2006).The effect of the regularization parameter was not investigated here explicitly, but its present value was found to be a good compromise between noise amplification and its impact on performance in terms of spatial resolution and sound field reconstruction error.

Measured sound field reconstruction errors matched the simulations well, confirming that a reasonably accurate reconstruction of the sound field is possible in practice, with an upper frequency limit of about 4000 Hz for a head-sized region. In terms of the effects of various noise sources, similarly to earlier studies (e.g. Rafaely, 2005), it was found that the dominant source of error at high frequencies is spatial aliasing. This means that, assuming realistic transducer characteristics, high-frequency performance is still mostly dependent on the array geometry. Using microphones with a higher variation in their magnitude and phase responses could of course result in a greater contribution of these errors. In contrast, at low frequencies and low levels, the contribution of transducer self-noise becomes significant.

The utility of considering several sound field error measures was also highlighted. While the mean-square error measure provides a view of the overall error, its sensitivity to absolute phase may overestimate the contribution of phase errors in practice for spatial audio applications. The magnitude error on the other hand does not consider phase information at all. However, relative phase relationships in the sound field serve as an important localization cue, especially at low frequencies. The spatial correlation coefficient is sensitive to changes in the relative amplitude and phase relationships, which are also the cues that would likely affect localization for a human listener. Thus, while the three error measures considered together provide a broader view of the types of errors present in the reconstructed sound field, if a single measure had to be chosen, the correlation coefficient may be the most appropriate one for predicting degradation in localization performance.

The general approach presented in this paper can be extended to include the quantification

of errors introduced by the playback system, if similar measurements (or simulations) of the sound field in a loudspeaker array are made. This way the errors introduced by the whole recording-reproduction chain, as well as the separate contribution of the microphone array and the playback system could be evaluated. Practical issues restricted the measurement of the reference sound field to the horizontal plane only, but in principle a volume could also be considered.

Finally, taking into account that the intended application of the array is sound field reproduction for human listeners, in the future, methods must be developed to define acceptable performance criteria for such arrays. This is needed in order to assign perceptual relevance to the purely physical error measures considered so far.

ACKNOWLEDGMENTS

The authors would like to acknowledge the support of Siemens Audiology Solutions and Brüel & Kjær Sound & Vibration Measurement.

6

Overall discussion

6.1 Main contributions

The main topic of the present thesis has been the development and objective evaluation of a spherical microphone array for spatial audio recording, with the primary aim of capturing realistic acoustic scenes for applications related to hearing research. As practical considerations limit the number of available transducers, it was investigated whether the microphones could be arranged in a way to better match typical acoustic scenes, as well as the properties human hearing. As the most important sound sources are typically in or near the horizontal plane, it was a design goal to obtain a fully three-dimensional representation of the sound field, but with more detail in the horizontal plane.

The main contributions of this thesis can be summarized as

(1) Introduction and evaluation of a mixed-order ambisonics (MOA) approach for spatial audio recordings (Chapters 3 and 4);

(2) Design and development of a microphone array based on MOA principles (Chapters 3 and 4);

(3) Objective evaluation of array performance through various performance metrics, including beamforming-based measures and technical measures of the reproduced sound field (Chapters 3 and 4), as well as the sound field reconstruction error (Chapter 5).

6.2 Mixed-order ambisonics microphone arrays

6.2.1 Mixed-order microphone layouts

The starting point in Chapter 3 was to investigate whether a recording system with anisotropic properties could be realized (i) through an appropriate placement of transducers on the sphere, and (ii) by applying a matching, mixed-order spherical harmonics processing. Previously described spherical microphone arrays generally aimed at providing a direction independent spatial resolution, whether in two (Tiana-Roig et al., 2011; Weller et al., 2011) or three dimensions (e.g. Meyer and Elko, 2004; Moreau et al., 2006), and featured a uniform or nearly uniform distribution of microphones. In this work, a ring-based layout was considered, with horizontal rings of microphones at several elevation angles, and the smallest transducer spacing on the equator. The layout was optimized such that in addition to a complete set of

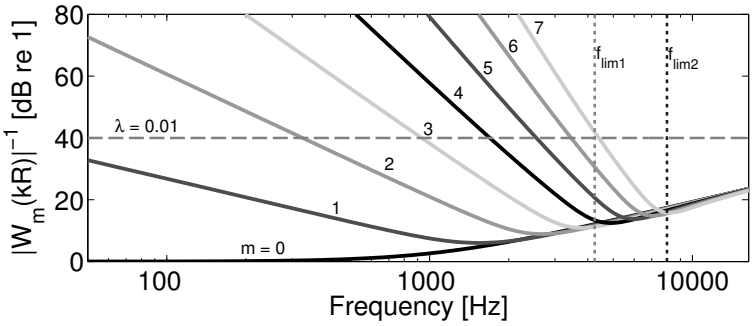


Figure 6.1: The reciprocal of the radial function $W_m(kR)$ for oders 0 to 7, for a rigid sphere of 5 cm radius. f_{lim1} indicates the upper frequency limit for 7th order sph. harm. representation in a head-sized region, while f_{lim2} indicates the approximate aliasing frequency of the MOA array.

spherical harmonic functions up to 5th order, it allowed the capture of additional, horizontally-oriented SHFs up to 7th order. Thus, a mixed-order spherical harmonics expansion with a higher horizontal order was applied, termed mixed-order ambisonics (MOA). MOA was first proposed by Daniel (2000) to accommodate loudspeaker layouts with a larger number of loudspeakers placed horizontally, but its application to microphone arrays had not been investigated previously.

In Chapter 3, two specific array layouts were considered, both with 52 transducers: a ring-based layout as described above (MOA array), and a layout featuring a nearly uniform distribution of microphones, suitable for 5th-order ambisonics (HOA array). It was shown through simulations of both arrays that with MOA, a direction-dependent spatial resolution is achieved. Specifically, with MOA, a higher spatial directivity was observed for sound sources in the horizontal plane, both in comparison to the HOA array, and in comparison to elevated sound sources with the MOA array. However, for the considered array radius of 5 cm, the improved spatial directivity was only attained above about 3 kHz, up to the aliasing frequency of about 8 kHz. At frequencies below 3 kHz, the regularization applied limited directivity, and the HOA and MOA arrays showed similar performance characteristics.

6.2.2 Limitations in MOA processing

The effects of MOA processing were investigated more closely in Chapter 4, where several order combinations were considered. It was shown that performance for horizontal vs. elevated sources can be adjusted by changing the order combination, but that a benefit of the higher horizontal orders was only seen above the frequency at which these orders were "activated" by the regularization. This roughly corresponds to the frequency at which the magnitude of the radial function $W_m(kr)$ for that order equals the regularization parameter λ. Unlike with higher-order playback systems, microphone arrays can only take advantage of higher orders in the frequency range where these orders can be captured without excessive amplification of the microphone signals. Without regularization, the operating bandwidth of higher-order arrays is quite limited.

To illustrate this further, the magnitude of $1/W_m(kR)$, which indicates the encoding gain

for order m, is plotted in Figure 6.1 for a 5 cm sphere. The regularization parameter of $\lambda = 0.01$, which was used throughout this thesis, corresponds to a maximum amplification of 40 dB (indicated by the gray dashed line). The activation frequency for each order is thus given by its intersection with the maximum amplification level. Below this frequency, the given order is attenuated. The approximate aliasing frequency of the array is indicated as f_{lim2}. It can be seen that the nominal operating frequency range for a 7th order array (without regularization) would be around an octave, whereas with regularization, the frequency range can be extended to about two decades, albeit at the cost of utilizing only first-order components at the lowest frequencies. It is also apparent that the higher the order, the smaller its usable frequency range, due to the increasing slope of the radial function towards low frequencies. This presents a fundamental limitation in the design of microphone arrays of very high orders. These same limitations apply to the additional horizontal orders in MOA, and limit the frequency range where they can contribute to the array output. Allowing for more or less signal amplification (i.e. adjusting λ) also increases or decreases the frequency above which higher horizontal spatial resolution is achieved. In other words, adequate SNR is required to utilize the higher horizontal orders.

Another consideration regarding mixed order combinations stems from the fact that the chosen array radius is smaller than the target listening area, which is taken to be about the size of a human head. The limiting frequency up to which the sound field in this larger area is well represented by SHFs up to order M can be estimated from the "rule of thumb" $M \approx kr$. For 7th order this frequency is indicated in Figure 6.1 as f_{lim1}. Any orders that are activated above this frequency will contribute little to a physically accurate sound field in this listening area. In the present case, it can be seen that the 7th order components reach the maximum amplification around this frequency, thus they still contribute to the sound field around the ears. The frequency range between f_{lim1} and f_{lim2} represents the region where spatial information is captured by the array, but a physically-based (holographic) reconstruction is no longer possible in the target area. Here, an alternative, non-holographic reproduction strategy can be employed (e.g. "energetic" or "max r_E" decoding, see Moreau et al., 2006; Zotter and Frank, 2012), that may still provide valid localization cues.

The size of the area of physically correct reproduction, or the "sweet area" was investigated explicitly in Chapter 5 by considering the sound field reconstruction error as a function of distance from the origin. It was also confirmed here that the MOA approach results in a lower sound field reproduction error for horizontal sound sources, or, a slightly larger sweet area for a fixed error level.

The ring-based microphone layout provided a fairly straightforward way of generating microphone layouts suitable for MOA by varying the number and density of transducer rings. However, as shown in Chapter 4 and by Travis (2009), mixed order combinations with a low periphonic and high horizontal order have some undesirable properties, stemming from a very large difference between vertical and horizontal directivity for horizontal sources. Further, for array layouts with a low number of microphones outside of the horizontal plane, spatial aliasing for elevated sound sources will occur from much lower frequencies than for horizontal

sources. In realistic sound fields, where the sound incidence direction cannot be controlled, aliasing from elevated sources will likely cause spectral coloration and may impair localization of horizontal sources, negating the benefit of higher horizontal directivity. The MOA approach is thus best applied with a balanced choice of both the order combination, as well as with regards to the distribution of microphones on the sphere. Based on the objective measures considered as part of this work, for the current MOA scheme, a periphonic order of at least 3 is recommended.

6.3 Validation of the simulation framework

A large part of this study was based on computer simulations, as this allowed for flexibility in studying and comparing various array configurations. The measurements in Chapters 3 and 5 verified that the performance of the realized MOA array was well matched by the simulations, and demonstrated that array performance can be simulated accurately if the characteristics of the transducers are taken into account. The simulation framework developed in this work may thus prove useful for designing future microphone arrays. Finally, the measurements confirmed that with MOA the desired higher horizontal directivity can also be attained under real-world conditions.

6.4 Performance metrics

A number of performance metrics were investigated to objectively quantify array performance, based on measures of the beam pattern, technical measures of the reproduced sound field, as well as an explicit quantification of the sound field reconstruction error in a plane. Measures of the beam pattern provided a tool to directly evaluate spatial directivity, which was valuable during the development process of the array. The beamwidth proved especially useful in assessing the impact of MOA processing, as it allowed for a separation of the spatial resolution along the horizontal and vertical dimensions.

These beamforming measures are related to, but do not provide a direct view of the reproduced sound field. To assess the intended application of the array more directly, measures of the recorded sound field reproduced over a simulated loudspeaker array were also considered. The magnitude of the "energy vector" r_E was evaluated, which was proposed by Gerzon (1992) as a technical measure that may be related to ILD-based localization at higher frequencies. This measure was nonetheless applied from a purely technical perspective in order to gauge the effective order (i.e. the effective spatial resolution) of the reproduced sound field. Results from this measure further revealed the impact of the regularization applied. The simulations of the reproduced field also allowed an estimation of the expected dynamic range at playback.

Finally, in Chapter 5, direct measures of the error of the reconstructed sound field were applied. A distinction is made here between the terms *reproduction*, which implies playback over a loudspeaker system, and *reconstruction*, which refers to the sound field as represented

by the captured spherical harmonic coefficients. In other words, the reproduction error includes errors introduced by the playback system, whereas the reconstruction error only includes errors introduced by the microphone array. In Chapter 5, only the reconstruction error was considered, but the technique can easily be extended to quantify the errors introduced by a loudspeaker array, if measurements are made in the playback environment. Such an approach could be applied to validate the complete recording and reproduction chain. From the error measures considered in this chapter, the spatial correlation coefficient seemed most well-suited for spatial audio applications, as it is sensitive to relative amplitude and phase relationships in the sound field, which are also cues that would likely affect localization for a human listener.

6.5 Perceptual effects

One of the main limitations of this study is that only a technical evaluation of MOA recording has been presented, and a perceptual evaluation of the system has not yet been undertaken. As the results of this thesis also demonstrate, a physically accurate reproduction of the sound field in the entire audible frequency range cannot practically be achieved, and the perceptual significance of various errors will have to be evaluated in future work. As stated before, how critical specific errors are will also depend on the specific application. Audiological applications, for example, might have a different set of requirements than virtual reality systems.

In particular, there are four main areas related to the work presented where perceptual tests will be needed to find optimal parameters or drive future development.

6.5.1 Effects of high-frequency aliasing

As the recorded sound field is not bandlimited, the microphone array will normally be exposed to audio frequencies above its spatial aliasing frequency, which will cause these frequencies to appear from undesired directions, and may also cause spectral coloration effects or other annoying artifacts (Avni et al., 2013). For 3D systems such as the one considered here, localization of elevated sound sources may be particularly affected by aliasing, as spectral localization cues extend well above typical aliasing frequencies of such systems. Signal processing strategies to mitigate the disturbing effects of aliasing and to provide the most acceptable or pleasing sound need to be investigated. One approach could be to bandlimit the spherical harmonics-based processing to below the aliasing frequency, and at higher frequencies utilize the shadowing effect of the rigid sphere, which, in essence, makes the individual microphones directive at high frequencies.

6.5.2 Effects of the playback environment

To focus the discussion on the microphone array, only ideal playback environments with a large number of loudspeakers were considered in this study. Reproduction on a real-life

loudspeaker array will naturally include additional error sources related to the characteristics of the loudspeakers and the playback room. As it was mentioned before, the frequency limit for a physically accurate reproduction in a head-sized region is lower than the aliasing frequency of the microphone array itself. Various decoding strategies have been proposed that allow the presentation of the captured spatial information in a perceptually meaningful way (e.g. Daniel, 2000; Zotter and Frank, 2012; Zotter et al., 2012), and could be applied at higher frequencies. For applications where physical accuracy is not paramount, it may be appropriate to use such a perceptually optimized reproduction strategy in the entire frequency range.

6.5.3 Effects of regularization

The trade-off between background noise and spatial resolution is a crucial parameter in microphone array signal processing, and is related to the applied regularization scheme as well as to the value of the regularization parameter. Technical measures alone cannot resolve whether increased spatial acuity or a reduction in audible background noise is more beneficial in a given context. Especially at higher frequencies, the considered objective measures may not be affected by microphone noise due to adequate SNR, but the noise may still be audible. Subjective tests can help choose between alternative regularization methods and determine the optimal value of the regularization parameter.

6.5.4 Perceptual benefits of MOA

The technical measures considered in this work showed a benefit of MOA over HOA recording in terms of increased spatial resolution for horizontal sources. However, whether this additional spatial information is perceptually relevant, or whether any unexpected artifacts are introduced needs to be evaluated with subjective tests. In terms of localization, Bertet et al. (2013) showed improvements in localization on a horizontal-only playback system for increasing microphone orders from 1 through 4. It was furthermore reported that localization accuracy for frontal sources for the 4th order system approached the range given for natural sound sources, suggesting that a 4th order system already provides close to sufficient information to the listener for natural localization. However, other spatial attributes and subjective quality may still benefit from a higher-order representation. Avni et al. (2013) showed that the judgment of a range of attributes improved with increasing order for orders well above 5. Technical applications, such as the evaluation of advanced hearing instruments or communication headsets, could also benefit from the extended frequency range of physically correct reproduction.

6.6 Perspectives

6.6.1 Towards an acoustic scene library

The ultimate goal of the system developed in this thesis is to enable the recording and reproduction of real-life acoustic scenes, and to develop a spatial acoustic scene library for hearing research. With this in mind, a set of recordings was already collected with the aim of validating methods to construct "cocktail-party" scenes. Two acoustic spaces with different degrees of reverberation were sampled: a library and a medium-sized break room. In each space, realistic background noise, as well as spatial room impulse responses for several source positions were recorded at a single listener position. In addition, acoustic scenes with and without close talkers were recorded in a large cafeteria. These recorded scenes will also be useful in applying a perceptual evaluation of the microphone array itself, and to address the issues outlined in the previous section.

6.6.2 Potential applications

This project presented a first step towards applying virtual environment technologies in hearing research for the presentation of realistic, dynamic scenes. The tools developed here could be used in the future to create new psychophysical tests that utilize more natural, complex auditory stimuli. In particular, speech material from existing speech tests (e.g. Nilsson et al., 1994; Nielsen and Dau, 2011) could be combined with spatial audio recordings in order to create speech intelligibility tests with real-life, dynamic backgrounds. Listening tests based on realistic acoustic scenes with spatially distributed, dynamic interferers may provide a better understanding of the limitations of the auditory system. Such tests could also provide better predictions regarding the real-life performance of hearing-impaired listeners, as well as regarding the potential benefits of hearing-aid use. Further, the spatial audio recordings could be used to investigate not only localization, but other spatial-hearing related attributes as well, such as distance perception, externalization and apparent source width. Finally, the recordings could be combined with visual stimuli in order to provide a combined audio-visual environment for perceptual testing and the evaluation of communication devices.

Bibliography

Abhayapala, T. D. and D. B. Ward (2002). "Theory and design of high order sound field microphones using spherical microphone array". In: *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. Vol. 2. IEEE, pp. II–1949.

Akeroyd, M. A. (2006). "The psychoacoustics of binaural hearing". In: *International journal of audiology* 45.S1, pp. 25–33.

Algazi, V. R., C. Avendano, and R. O. Duda (2001a). "Elevation localization and head-related transfer function analysis at low frequencies". In: *The Journal of the Acoustical Society of America* 109.3, pp. 1110–1122.

Algazi, V. R., R. O. Duda, D. M. Thompson, and C. Avendano (2001b). "The CIPIC HRTF database". In: *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop*. IEEE, pp. 99–102.

Allemang, R. J. (2003). "The modal assurance criterion–twenty years of use and abuse". In: *Sound and Vibration* 37.8, pp. 14–23.

Avni, A., J. Ahrens, M. Geier, S. Spors, H. Wierstorf, and B. Rafaely (2013). "Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution". In: *The Journal of the Acoustical Society of America* 133.5, pp. 2711–2721.

Begault, D. R., E. M. Wenzel, and M. R. Anderson (2001). "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source". In: *Journal of the Audio Engineering Society* 49.10, pp. 904–916.

Berg, J. and F. Rumsey (2006). "Identification of quality attributes of spatial audio by repertory grid technique". In: *Journal of the Audio Engineering Society* 54.5, pp. 365–379.

Berkhout, A. J. (1988). "A holographic approach to acoustic control". In: *Journal of the Audio Engineering Society* 36.12, pp. 977–995.

Bernschütz, B., C. Pörschmann, S. Spors, and S. Weinzierl (2010). "Entwurf und Aufbau eines variablen sphärischen Mikrofonarrays für Forschungsanwendungen in Raumakustik und Virtual Audio". In: *Fortschritte der Akustik*.

Bertet, S., J. Daniel, L. Gros, E. Parizet, and O. Warusfel (2007). "Investigation of the perceived spatial resolution of higher order ambisonics sound fields: A subjective evaluation involving virtual and real 3D microphones". In: *Audio Engineering Society 30th International Conference: Intelligent Audio Environments*. Audio Engineering Society.

Bertet, S., J. Daniel, E. Parizet, and O. Warusfel (2013). "Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources". In: *Acta Acustica united with Acustica* 99.4, pp. 642–657.

Blauert, J. (1997a). “An introduction to binaural technology.” In: *Binaural and spatial hearing in real and virtual environments*. Ed. by R. H. Gilkey and T. R. Anderson. Lawrence Erlbaum Associates, Inc.

Blauert, J. (1997b). *Spatial hearing: the psychophysics of human sound localization*. MIT press.

Bork, I. (2005). “Report on the 3rd round robin on room acoustical computer simulation. Part II: Calculations”. In: *Acta Acustica united with Acustica* 91.4, pp. 753–763.

Braun, S. and M. Frank (2011). “Localization of 3D ambisonic recordings and ambisonic virtual sources”. In: *International Conference on Spatial Audio*.

Bregman, A. S. (1994). *Auditory scene analysis: the perceptual organization of sound*. MIT press.

Brüel & Kjær. *Short 20 kHz Array Microphone – Type 4959 Product data*. Tech. rep. retrieved 15 Jan 2012.

Catic, J., S. Santurette, J. M. Buchholz, F. Gran, and T. Dau (2013). “The effect of interaural-level-difference fluctuations on the externalization of sound”. In: *The Journal of the Acoustical Society of America* 134.2, pp. 1232–1241.

Christensen, C. L. and G. Koutsouris (2013). *Odeon Room Acoustics Software Version 12 – User manual*. Odeon A/S.

Craven, P. G. (2003). “Continuous surround panning for 5-speaker reproduction”. In: *Audio Engineering Society 24th International Conference: Multichannel Audio, The New Reality*. Audio Engineering Society.

Cubick, J. (2011). “Validation of a virtual sound environment system for the testing of hearing aids”. MA thesis. Technical University of Denmark.

Cubick, J., S. Favrot, P. Minnaar, and T. Dau (2013). “Validation of a virtual sound environment system for hearing aid testing”. In: *Proceedings of AIA-DAGA 2013, Merano*.

Daniel, J. (2000). “Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia”. PhD thesis. University of Paris VI, France.

Daniel, J. (2003). “Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format”. In: *Audio Engineering Society 23rd International Conference: Signal Processing in Audio Recording and Reproduction*. Audio Engineering Society.

Daniel, J. (2009). “Evolving views on HOA: From technological to pragmatic concerns”. In: *Ambisonics Symposium, Graz, Austria*.

Daniel, J., S. Moreau, and R. Nicol (2003). “Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging”. In: *Audio Engineering Society Convention 114*. Audio Engineering Society.

Farina, A. and R. Ayalon (2003). “Recording concert hall acoustics for posterity”. In: *Audio Engineering Society 24th International Conference: Multichannel Audio, The New Reality*. Audio Engineering Society.

Farrar, K. (1979). “Soundfield microphone”. In: *Wireless World* 85.1526, pp. 48–50.

Favrot, S. E. and M. Marschall (2012). “Metrics for performance assessment of mixed-order Ambisonics spherical microphone arrays”. In: *AES 25th UK conference and 4th International Symposium on Ambisonics and Spherical Acoustics*.

Favrot, S. and J. M. Buchholz (2009). “Distance perception in loudspeaker-based room auralization”. In: *Audio Engineering Society Convention 127*. Audio Engineering Society.

Favrot, S., M. Marschall, J. Käsbach, J. Buchholz, and T. Weller (2011). “Mixed-order Ambisonics recording and playback for improving horizontal directionality”. In: *Audio Engineering Society Convention 131*. Audio Engineering Society.

Favrot, S. and J. M. Buchholz (2010). “LoRA: A loudspeaker-based room auralization system”. In: *Acta Acustica united with Acustica* 96.2, pp. 364–375.

Gardner, W. G. (2002). “Reverberation algorithms”. In: *Applications of digital signal processing to audio and acoustics*. Springer, pp. 85–131.

Geier, M. and S. Spors (2012). “Spatial Audio with the SoundScape Renderer”. In: *27th Tonmeistertagung - VDT International Convention*.

Gerzon, M. A. (1973). “Periphony: with-height sound reproduction”. In: *Journal of the Audio Engineering Society* 21.1, pp. 2–10.

Gerzon, M. A. (1992). “General metatheory of auditory localisation”. In: *Audio Engineering Society Convention 92*. Audio Engineering Society.

Gover, B. N., J. G. Ryan, and M. R. Stinson (2004). “Measurements of directional properties of reverberant sound fields in rooms using a spherical microphone array”. In: *The Journal of the Acoustical Society of America* 116.4, pp. 2138–2148.

Grimm, G., T. Wendt, V. Hohmann, and S. D. Ewert (2014). “Implementation and perceptual evaluation of a simulation method for coupled rooms in higher order ambisonics”. In: *Proc. of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, 2014*.

Hammershøi, D. and H. Møller (2005). “Binaural technique – Basic methods for recording, synthesis, and reproduction”. In: *Communication Acoustics*. Ed. by J. Blauert. Springer-Verlag.

Hardin, R. H. and N. J. Sloane (1996). “McLaren’s improved snub cube and other new spherical designs in three dimensions”. In: *Discrete & Computational Geometry* 15.4, pp. 429–441.

Hawley, M. L., R. Y. Litovsky, and J. F. Culling (2004). “The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer”. In: *The Journal of the Acoustical Society of America* 115.2, pp. 833–843.

Hebrank, J. and D Wright (1974). “Spectral cues used in the localization of sound sources on the median plane”. In: *The Journal of the Acoustical Society of America* 56.6, pp. 1829–1834.

Huber, D. and R. Runstein (2013). *Modern Recording Techniques*. Taylor & Francis.

Jacobsen, F. and P. M. Juhl (2013). *Fundamentals of General Linear Acoustics*. Wiley.

Jacobsen, F., G. Moreno-Pescador, E. Fernandez-Grande, and J. Hald (2011). “Near field acoustic holography with microphones on a rigid sphere”. In: *The Journal of the Acoustical Society of America* 129.6, pp. 3461–3464.

Käsbach, J., S. Favrot, and J. Buchholz (2011). “Evaluation of a mixed-order planar and periphonic ambisonics playback implementation”. In: *Forum Acusticum*.

Kirkeby, O., P. A. Nelson, F. Orduna-Bustamante, and H. Hamada (1996). “Local sound field reproduction using digital signal processing”. In: *The Journal of the Acoustical Society of America* 100.3, pp. 1584–1593.

Koski, T., V. Sivonen, and V. Pulkki (2013). “Measuring speech intelligibility in noisy environments reproduced with parametric spatial audio”. In: *Audio Engineering Society Convention 135*. Audio Engineering Society.

Li, Z. and R. Duraiswami (2007). “Flexible and optimal design of spherical microphone arrays for beamforming”. In: *Audio, Speech, and Language Processing, IEEE Transactions on* 15.2, pp. 702–714.

Makous, J. C. and J. C. Middlebrooks (1990). “Two-dimensional sound localization by human listeners”. In: *The journal of the Acoustical Society of America* 87.5, pp. 2188–2200.

Marschall, M. and J. Chang (2013). “Sound-field reconstruction performance of a mixed-order Ambisonics microphone array”. In: *Proceedings of Meetings on Acoustics*. Vol. 19. 1. Acoustical Society of America.

Marschall, M., S. Favrot, and J. Buchholz (2012). “Robustness of a mixed-order Ambisonics microphone array for sound field reproduction”. In: *Audio Engineering Society Convention 132*. Audio Engineering Society.

Meyer, J. and T. Agnello (2003). “Spherical microphone array for spatial sound recording”. In: *Audio Engineering Society Convention 115*. Audio Engineering Society.

Meyer, J. and G. W. Elko (2004). “Spherical microphone arrays for 3D sound recording”. In: *Audio signal processing for next-generation multimedia communication systems*. Springer, pp. 67–89.

Meyer, J. and G. Elko (2002). “A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield”. In: *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. Vol. 2. IEEE.

Minnaar, P., S. Favrot, and J. M. Buchholz (2010). “Improving hearing aids through listening tests in a virtual sound environment”. In: *The Hearing Journal* 63.10, pp. 40–42.

Minnaar, P., S. F. Albeck, C. S. Simonsen, B. Søndersted, S. A. D. Oakley, and J. Bennedbæk (2013). “Reproducing real-life listening situations in the laboratory for testing hearing aids”. In: *Audio Engineering Society Convention 135*. Audio Engineering Society.

Moreau, S., J. Daniel, and S. Bertet (2006). “3D sound field recording with higher order ambisonics – Objective measurements and validation of spherical microphone”. In: *Audio Engineering Society Convention 120*. Audio Engineering Society.

Nielsen, J. B. and T. Dau (2011). “The Danish hearing in noise test”. In: *International Journal of Audiology* 50.3, pp. 202–208.

Nilsson, M., S. D. Soli, and J. A. Sullivan (1994). “Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise”. In: *The Journal of the Acoustical Society of America* 95.2, pp. 1085–1099.

Noisternig, M., B. F. Katz, S. Siltanen, and L. Savioja (2008). “Framework for real-time auralization in architectural acoustics”. In: *Acta Acustica united with Acustica* 94.6, pp. 1000–1015.

Olsen, W. O. (1998). “Average speech levels and spectra in various speaking/listening conditions. A summary of the Pearson, Bennett, & Fidell (1977) report”. In: *American Journal of Audiology* 7.2, pp. 21–25.

Park, M. and B. Rafaely (2005). “Sound-field analysis by plane-wave decomposition using spherical microphone array”. In: *The Journal of the Acoustical Society of America* 118.5, pp. 3094–3103.

Perrott, D. R. and K. Saberi (1990). “Minimum audible angle thresholds for sources varying in both elevation and azimuth”. In: *The Journal of the Acoustical Society of America* 87.4, pp. 1728–1731.

Poletti, M. A. (2005). “Three-dimensional surround sound systems based on spherical harmonics”. In: *Journal of the Audio Engineering Society* 53.11, pp. 1004–1025.

Pulkki, V. (1997). “Virtual sound source positioning using vector base amplitude panning”. In: *Journal of the Audio Engineering Society* 45.6, pp. 456–466.

Pulkki, V. (2007). “Spatial sound reproduction with directional audio coding”. In: *Journal of the Audio Engineering Society* 55.6, pp. 503–516.

Rafaely, B. (2004). “Plane-wave decomposition of the sound field on a sphere by spherical convolution”. In: *The Journal of the Acoustical Society of America* 116.4, pp. 2149–2157.

Rafaely, B. (2005). “Analysis and design of spherical microphone arrays”. In: *Speech and Audio Processing, IEEE Transactions on* 13.1, pp. 135–143.

Rafaely, B., B. Weiss, and E. Bachmat (2007). “Spatial aliasing in spherical microphone arrays”. In: *Signal Processing, IEEE Transactions on* 55.3, pp. 1003–1010.

Rumsey, F. (2002). “Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm”. In: *Journal of the Audio Engineering Society* 50.9, pp. 651–666.

Schröder, D. and M. Vorländer (2011). “RAVEN: A real-time framework for the auralization of interactive virtual environments”. In: *Forum Acusticum*.

Seeber, B. U., S. Kerber, and E. R. Hafter (2010). “A system to simulate and reproduce audio–visual environments for spatial hearing research”. In: *Hearing research* 260.1, pp. 1–10.

Shilling, R. D. and B. Shinn-Cunningham (2002). “Virtual auditory displays”. In: *Handbook of Virtual Environments: Design, Implementation, and Applications*. Ed. by K. M. Stanney. CRC Press, pp. 65–92.

Shinn-Cunningham, B. (2002). “Speech intelligibility, spatial unmasking, and realism in reverberant spatial auditory displays”. In: *Proc of International Conference on Auditory Displays, Atlanta, GA*.

Shinn-Cunningham, B. G., S. Santarelli, and N. Kopco (2000). “Tori of confusion: Binaural localization cues for sources within reach of a listener”. In: *The Journal of the Acoustical Society of America* 107.3, pp. 1627–1636.

Solvang, A. (2008). “Spectral impairment of two-dimensional higher order Ambisonics”. In: *Journal of the Audio Engineering Society* 56.4, pp. 267–279.

Spors, S., R. Rabenstein, and J. Ahrens (2008). “The theory of wave field synthesis revisited”. In: *Audio Engineering Society Convention 124*. Audio Engineering Society, pp. 17–20.

Spors, S., H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter (2013). “Spatial sound with loudspeakers and its perception: A review of the current state”. In: *Proceedings of the IEEE* 101.9, pp. 1920–1938.

Tiana-Roig, E., F. Jacobsen, and E. Fernandez-Grande (2011). “Beamforming with a circular array of microphones mounted on a rigid sphere”. In: *The Journal of the Acoustical Society of America* 130.3, pp. 1095–1098.

Travis, C. (2009). “A new mixed-order scheme for Ambisonic signals”. In: *Ambisonics Symposium 2009*.

Trevino, J., T. Okamoto, Y. Iwaya, and Y. Suzuki (2010). “High order Ambisonic decoding method for irregular loudspeaker arrays”. In: *Proceedings of the 20th International Congress on Acoustics*.

Vilkamo, J., T. Lokki, and V. Pulkki (2009). “Directional audio coding: Virtual microphone-based synthesis and subjective evaluation”. In: *Journal of the Audio Engineering Society* 57.9, pp. 709–724.

Vorländer, M. (2008). *Auralization: Fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*. Springer.

Ward, D. B. and T. D. Abhayapala (2001). “Reproduction of a plane-wave sound field using an array of loudspeakers”. In: *Speech and Audio Processing, IEEE Transactions on* 9.6, pp. 697–707.

Weller, T., S. Favrot, and J. Buchholz (2011). “Application of a circular 2D hard-sphere microphone array for higher-order ambisonics auralization”. In: *Forum Acusticum*.

Wightman, F. L. and D. J. Kistler (1992). “The dominant role of low-frequency interaural time differences in sound localization”. In: *The Journal of the Acoustical Society of America* 91.3, pp. 1648–1661.

Williams, E. G. (1999). *Fourier acoustics: Sound radiation and nearfield acoustical holography*. Academic Press.

Williams, E. G. and K. Takashima (2010). “Vector intensity reconstructions in a volume surrounding a rigid spherical microphone array”. In: *The Journal of the Acoustical Society of America* 127.2, pp. 773–783.

Zahorik, P. (2009). “Perceptually relevant parameters for virtual listening simulation of small room acoustics”. In: *The Journal of the Acoustical Society of America* 126.2, pp. 776–791.

Zahorik, P., D. S. Brungart, and A. W. Bronkhorst (2005). “Auditory distance perception in humans: A summary of past and present research”. In: *Acta Acustica united with Acustica* 91.3, pp. 409–420.

Zotter, F., H. Pomberger, and M. Noisternig (2012). “Energy-preserving ambisonic decoding”. In: *Acta Acustica united with Acustica* 98.1, pp. 37–47.

Zotter, F. (2009). “Analysis and synthesis of sound-radiation with spherical arrays”. PhD thesis. Insitute of Electronic Music, Acoustics, University of Music, and Performing Arts, Graz, Austria.

Zotter, F. and M. Frank (2012). “All-round ambisonic panning and decoding”. In: *Journal of the Audio Engineering Society* 60.10, pp. 807–820.

A

The effect of compression on tuning estimates in a simple nonlinear auditory filter model[d]

Abstract

Behavioral experiments using auditory masking have been used to characterize frequency selectivity, one of the basic properties of the auditory system. However, due to the nonlinear response of the basilar membrane, the interpretation of these experiments may not be straightforward. Specifically, there is evidence that human frequency-selectivity estimates depend on whether an iso-input or an iso-response measurement paradigm is used (Eustaquio-Martin and Lopez-Poveda, 2011). This study presents simulated tuning estimates using a simple compressive auditory filter model, the bandpass nonlinearity (BPNL), which consists of a compressor between two bandpass filters. The BPNL forms the basis of the dual-resonance nonlinear (DRNL) filter that has been used in a number of modeling studies. The location of the nonlinear element and its effect on estimated tuning in the two measurement paradigms was investigated. The results show that compression leads to (i) a narrower tuning estimate in the iso-response paradigm when a compressor precedes a filter, and (ii) a wider tuning estimate in the iso-input paradigm when a compressor follows a filter. The results imply that if the DRNL presents a valid cochlear model, then compression alone may explain a large part of the behaviorally observed differences in tuning between simultaneous and forward-masking conditions.

A.1 Introduction

Frequency selectivity is one of the fundamental properties of the auditory system and describes the ability to separate frequency components of complex stimuli. This property of hearing in humans can be characterized behaviorally using masking experiments. However, it is known that the response of the basilar membrane of the inner ear exhibits a compressive response to stimuli at medium sound pressure levels, and that frequency selectivity at low levels is aided by an active mechanism in the cochlea (Pickles, 1986). Consequently, the application of linear analysis techniques to estimate frequency selectivity is not straightforward, as the involved nonlinearities need to be taken into account. This is especially relevant when comparing results

[d] This chapter was originally published as Marschall et al. (2013). Figures were updated for this version.

from frequency selectivity measures obtained through different measurement paradigms. In particular, behavioral estimates of frequency tuning have been shown to depend on the temporal configuration of the stimulus, i.e., whether simultaneous or non-simultaneous masking is used (Moore et al., 1984; Oxenham and Shera, 2003); on the sound pressure level of the stimulus (Patterson and Moore, 1986); and on whether the input level ("iso-input" tuning) or the output level ("iso-response" tuning) of the filter is held constant in the measurement paradigm (Eustaquio-Martin and Lopez-Poveda, 2011).

Estimates of tuning derived from non-simultaneous masking conditions, when the signal and the masker do not overlap in time, tend to show sharper tuning than those derived from simultaneous masking conditions (Moore et al., 1984; Oxenham and Shera, 2003). Non-simultaneous masking conditions include forward masking, where the masker precedes the signal, and the pulsation threshold task, where the signal and the masker are alternated in time. It has been suggested that the difference between simultaneous and non-simultaneous estimates of frequency selectivity may be mostly due to effects of suppression, but the exact mechanism and the extent of suppressive contributions is under debate (see Moore and O'Loughlin, 1986, for a review). Suppression here refers to the nonlinear phenomenon whereby the auditory system's response to a sound can, under certain conditions, be decreased by the presence of another sound. In animal studies, suppression has been observed as two-tone rate suppression in auditory-nerve fibers (Sachs and Kiang, 1968) and also in the mechanical response of the basilar membrane (Ruggero et al., 1992). Houtgast (1972) found psychophysical evidence of two-tone suppression in humans where a decrease in the pulsation threshold of a tone was observed as a result of an added suppressor.

In this paper, we present an alternative explanation for the observed tuning differences, based on compression. For a linear system, tuning estimates measured using either an iso-input or iso-response method will be identical. However, for a nonlinear filter, the tuning estimates derived from each method may differ. Here, we explore how these tuning estimates differ depending on filter structure and the implication this has on behavioral estimates of frequency tuning in the auditory system.

A.2 Iso-input and iso-output tuning estimates of nonlinear filter structures

In an iso-input paradigm, a constant signal input power is maintained for the frequency range of interest, and the tuning characteristics of the system are described by the output power as a function of the input frequency. Conversely, in an iso-response paradigm, the signal input power is adjusted instead, so that the output power (response) of the system remains constant at each frequency. The tuning in the system is then described by the input signal power required to achieve constant output, as a function of frequency. For a linear system, these two methods lead to the same result. This is illustrated in Panel A of Figure A.1.

Now consider the case where a simple compressive non-linear element is added to before

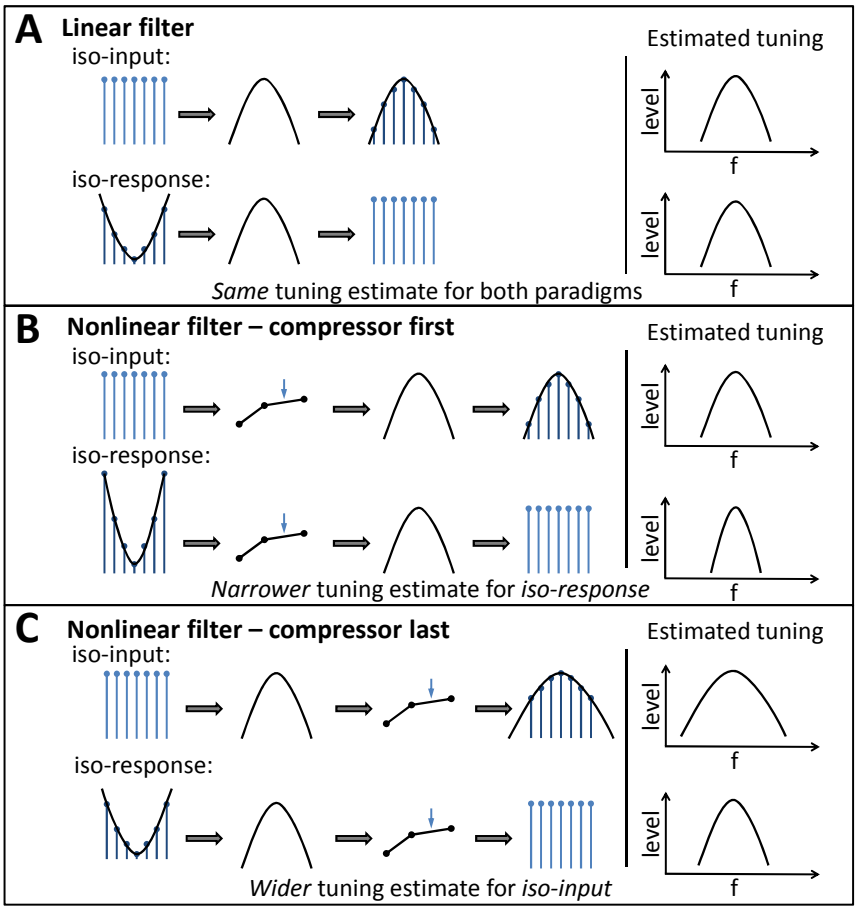


Figure A.1: Schematic of tuning estimates from iso-input and iso-response methods when applied to a linear filter (A), a compressor followed by a linear filter (B), and a linear filter followed by a compressor (C).

the filter (see Figure A.1, Panel B). For an iso-input paradigm, the signal power as a function of frequency is still constant after compression. Therefore, the output levels reflect the underlying tuning of the filter and the tuning estimate remains unchanged. However, if an iso-response paradigm is used, this is not the case. For frequencies that are attenuated by the filter, a larger change in input level is required due to compression. Thus, the addition of the compressor before the filter leads to tuning estimates that are sharper than the underlying filter tuning when an iso-response method is used.

Conversely, consider the case where a compressor is added after the filter (see Figure A.1, Panel C). If an iso-response method is used, the compressor has no effect as the output level of the filter is already constant. However, if an iso-input method is used, the compressor following the filter reduces the difference of the filter output across frequency. This leads to an estimate of tuning that is wider or less sharp than the underlying filter.

Now consider the case of a bandpass nonlinearity, where the compressor is level dependent and sandwiched between two bandpass filters. Tuning estimates from simple simulations of both iso-input and iso-response methods are plotted in Figure A.2. In the simulation, the compressor was set to be linear at low levels, compressive at medium levels (5:1 compression ratio), and linear again at high levels, to mimic the compressive behavior of the basilar

membrane. For simplicity, triangular filters, as well as dimensionless, logarithmic input and output values were assumed. The input level for the iso-input condition was varied from 20 to 80 dB in 10 dB increments. The reference level for the iso-response condition was varied from 10 to 40 dB in 5 dB increments.

Due to the properties of the nonlinearity, at low and at high levels the behavior of the system is linear. Therefore the tuning estimate with both paradigms gives the same filter shape, and corresponds to the filtering produced by the two filters applied in succession. However, at medium levels, where the compressive function is active, the differences between the two paradigms become apparent. When the signal level at the nonlinearity reaches the compression threshold, the slopes of the estimated filter function are affected. A large level difference between two frequency points at the input of the compressor is transformed into a smaller one at the output. For the iso-input paradigm, the amplitude changes arising as a result of the first filter are compressed, while those resulting from the second filter are unaffected. Thus, in the region of compression, the estimated filter slope will be shallower than in the linear case. For the iso-response paradigm, due to the compression, larger differences are needed at the input of the compressor to counteract the attenuation of off-center frequencies by the second filter. This leads to a steeper estimated filter slope for the whole system.

A further consequence of the two filter arrangement, specifically that of a filter preceding the nonlinearity, is that the onset of compression is frequency dependent. More off-frequency components require a higher level at the input than on-frequency components to be processed compressively. This effect can also be seen in Figure A.2. Changes in the filter slopes, indicating the onset and offset of compression, appear at different levels for different frequencies.

To summarize, when the bandpass nonlinearity is investigated with an iso-input paradigm, the estimated tuning is wider for the compressive region than in the linear case. Conversely, when an iso-response paradigm is used, the estimated tuning is narrower than in the linear case. This implies that the measurement paradigm has to be carefully considered when estimating tuning of nonlinear systems.

A.3 The sharpening of tuning in forward masking

As mentioned previously, tuning estimates derived from non-simultaneous masking paradigms have been observed to be sharper than those derived from simultaneous masking paradigms. While this phenomenon has been attributed to suppression, it may also be a result of compression. So far, only single sinusoids have been considered. However, when investigating frequency selectivity using a psychophysical task, additional signals are required as it is not possible to access the output of the auditory filters directly. In a typical forward masking paradigm, a probe tone is used to gauge the excitation from a masker at the (frequency) place of the probe. Thus, the level of the probe tone is held constant and the level of the off-frequency masker is varied such that the probe tone is just audible. This corresponds to an iso-response paradigm where we assume that the probe tone will become audible when the signal-to-noise ratio (SNR) at the nonlinear filter output reaches some fixed level.

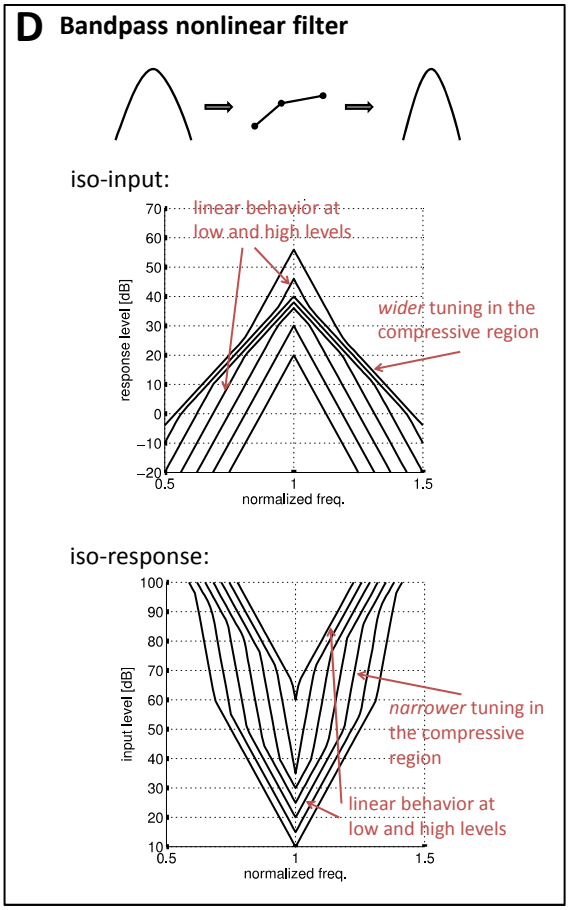


Figure A.2: Simple simulation of tuning estimates at different levels from iso-input (top) and iso-response (bottom) methods when applied to a bandpass nonlinear filter. In the compressive region, a shallower tuning is observed with the iso-input method. In contrast, with the iso-response method, a sharper tuning is seen in the compressive region.

Consider the behavior of a bandpass nonlinearity in a forward masking paradigm. Assuming that the impulse responses of the filters in the bandpass nonlinearity are short compared to the temporal separation of the masker and probe, then the masker and signal are processed independently. If the masker level is sufficiently high, the masker is compressed but the probe tone is not. As the masker moves further off-frequency, a greater change in masker level is needed at the input to the compressor in order to achieve a fixed SNR at the compressor output. This will result in a sharpened tuning estimate. In contrast, in a simultaneous masking paradigm, the masker and the signal are processed together. Thus, any compression of the masker also reduces the signal level, such that the relative levels of the signal and the masker do not change. Therefore, the sharpened tuning observed in forward masking vs. simultaneous masking experiments could be explained directly by cochlear compression, without considering suppression explicitly. Here, it is further assumed that the bandpass nonlinearity presents a valid functional model of the nonlinear behavior of the basilar membrane. This assumption, however, is supported by a number of studies having successfully used models based on the bandpass nonlinearity (e.g. Lopez-Poveda and

Meddis, 2001; Plack et al., 2002; Jepsen et al., 2008) to account for a wide range of human psychophysical data.

A.4 Summary

In this paper we have demonstrated that tuning estimates of nonlinear filters can vary significantly depending on whether an iso-input or iso-response measurement paradigm is used. This suggests that the measurement paradigm used to estimate tuning of nonlinear systems needs to be considered carefully. Further, given a small set of plausible assumptions, we have demonstrated that cochlear compression can explain the sharpening of tuning observed in non-simultaneous masking paradigms.

A.5 Acknowledgments

The authors would like to acknowledge the support of Siemens Audiology Solutions.

A.6 References

Eustaquio-Martin, A. and E. Lopez-Poveda (2011). "Isoresponse Versus Isoinput Estimates of Cochlear Filter Tuning". In: *J. Assoc. Res. Otolaryngol.* 12 (3). 10.1007/s10162-010-0252-1, pp. 281–299.

Houtgast, T. (1972). "Psychophysical Evidence for Lateral Inhibition in Hearing". In: *The Journal of the Acoustical Society of America* 51.6B, pp. 1885–1894.

Jepsen, M. L., S. D. Ewert, and T. Dau (2008). "A computational model of human auditory signal processing and perception". In: *The Journal of the Acoustical Society of America* 124, pp. 422–438.

Lopez-Poveda, E. A. and R. Meddis (2001). "A human nonlinear cochlear filterbank". In: *The Journal of the Acoustical Society of America* 110, pp. 3107–3118.

Marschall, M., E. MacDonald, and T. Dau (2013). "The effect of compression on tuning estimates in a simple nonlinear auditory filter model". In: *Proceedings of Meetings on Acoustics*. Vol. 19.

Moore, B. C. J. and B. J. O'Loughlin (1986). "The use of nonsimultaneous masking to measure frequency selectivity and suppression". In: *Frequency selectivity in hearing*. Ed. by B. C. J. Moore. Academic Press, London. Chap. 4.

Moore, B. C. J., B. R. Glasberg, and B. Roberts (1984). "Refining the measurement of psychophysical tuning curves". In: *The Journal of the Acoustical Society of America* 76.4, pp. 1057–1066.

Oxenham, A. J. and C. A. Shera (2003). "Estimates of Human Cochlear Tuning at Low Levels Using Forward and Simultaneous Masking". In: *Journal of the Association for Research in Otolaryngology* 4, pp. 541–554.

Patterson, R. D. and B. C. J. Moore (1986). "Auditory filters and excitation patterns as representations of frequency resolution". In: *Frequency selectivity in hearing*. Ed. by B. C. J. Moore. Academic Press, London. Chap. 3.

Pickles, J. O. (1986). “The neurophysiological basis of frequency selectivity”. In: *Frequency selectivity in hearing*. Ed. by B. C. J. Moore. Academic Press, London. Chap. 2.

Plack, C. J., A. J. Oxenham, and V. Drga (2002). “Linear and nonlinear processes in temporal masking”. In: *Acta Acustica united with Acustica* 88.3, pp. 348–358.

Ruggero, M. A., L. Robles, and N. C. Rich (1992). “Two-tone suppression in the basilar membrane of the cochlea: mechanical basis of auditory-nerve rate suppression”. In: *Journal of Neurophysiology* 68, pp. 1087–1099.

Sachs, M. B. and N. Y. S. Kiang (1968). “Two-Tone Inhibition in Auditory-Nerve Fibers”. In: *The Journal of the Acoustical Society of America* 43.5, pp. 1120–1128.

B

Modeling the effects of compression and suppression on estimates of auditory frequency selectivity[e]

Abstract

Estimates of tuning derived from non-simultaneous masking experiments tend to show sharper tuning than those derived from simultaneous masking experiments. Previous studies have suggested that the wider tuning observed in simultaneous masking is due to the more widely tuned influence of suppression that only occurs when the masker and the signal are present at the same time. This study, using a modeling approach, investigates an alternative explanation, involving the effect of compression alone. To this end, a computational model of auditory signal processing and perception was used to simulate behavioral measures of frequency selectivity in both simultaneous and forward masking. The model included the dual-resonance nonlinear (DRNL) filter as the frequency selective stage, which simulates key aspects of nonlinear cochlear processing, such as compression and two-tone suppression. The modeling results show that the effect of compression may directly lead to a narrower tuning estimate in forward masking if a filter structure with a nonlinearity sandwiched between two bandpass filters is assumed.

B.1 Introduction

Frequency selectivity is one of the basic properties of the auditory system: it describes its ability to partly separate the frequency components of complex stimuli. This property of hearing in humans can be characterized behaviorally using masking experiments. Estimates of tuning derived from nonsimultaneous masking conditions, when the signal and the masker do not overlap in time, tend to show sharper tuning than those derived from simultaneous masking conditions. It has been suggested that a major part of the difference between estimates of frequency selectivity in the two masking conditions may be due to effects of suppression (see Moore and O'Loughlin, 1986, for a review). Suppression refers to the phenomenon that the auditory system's response to a sound can be decreased by the presence of another sound.

The present study investigates, using a modeling approach, an alternative explanation

[e] This chapter was originally published as Marschall et al. (2011).

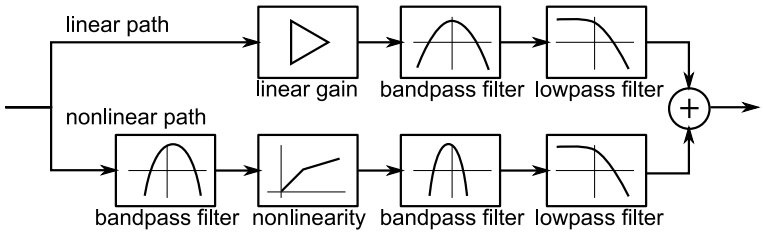


Figure B.1: The basic structure of the DRNL filter (see text).

based on the effect of peripheral compression alone. To this end, a computational auditory model was used to simulate two behavioral measures of frequency selectivity: psychophysical tuning curves (PTCs) and notched-noise thresholds.

As the underlying tuning of the auditory system is of interest, it is relevant to consider which estimates reflect this tuning better, those obtained from simultaneous masking or those from nonsimultaneous masking experiments. Further, as the temporal position of the signal causes an apparent change in frequency selectivity, the processing of dynamic signals, such as speech, may be affected if the auditory filters change with time. It is therefore important to understand these effects of frequency selectivity and their relation to the active processes in the cochlea.

B.2 Model and method

The model used in this study was the computational auditory signal processing and perception model (CASP, Jepsen et al., 2008), with some modifications. Instead of using a modulation filterbank stage, a modulation lowpass filter with a cutoff frequency of 8 Hz was considered. The model uses the dual-resonance nonlinear (DRNL) filterbank as the frequency selective stage (Meddis et al., 2001). The basic structure of the DRNL filter is shown in Figure B.1. The filter consists of two paths: a linear path representing the high-amplitude linear response of the basilar membrane, and a nonlinear path accounting for the low-amplitude linear and medium amplitude compressive response. The nonlinear element in the filter is an instantaneous "broken-stick" function that is composed of a linear and a compressive part.

The parameters of the DRNL were modified based on suggestions by Plack et al. (2002) to better account for suppression-related data. Specifically, the two bandpass filters in the nonlinear path had different bandwidths and center frequencies (CF). The first filter had a somewhat higher bandwidth and was centered slightly above CF, whereas the second filter had a somewhat narrower bandwidth and was centered slightly below CF. The exact filter bandwidths were adjusted to get a reasonable match with measured pilot notched-noise and PTC data.

In order to investigate the effect of peripheral compression on the estimates of tuning, two versions of the model were used to simulate each experiment. A "nonlinear" version, as

described above, and a “linear” version, in which the broken-stick function was replaced by a linear function with all other parameters unchanged.

All simulations were performed at 1 kHz, with only one frequency channel. The nominal center frequency of the filter corresponded to the frequency of the test tone. The configuration of the stimuli for the PTCs followed Moore et al. (1984), with the exception that a low-level notched noise was not used. For the notched-noise simulation, the setup was as in Oxenham and Bacon (2003).

B.3 Modeling suppression

The model can simulate some aspects of two-tone suppression, including suppression-areas that are broadly similar to human psychophysical data (Plack et al., 2002, not shown here, but see). This is because it includes a *bandpass nonlinearity*, a structure composed of a compressive nonlinearity between two bandpass filters, which has previously been shown to be a simple model of two-tone suppression (Duifhuis, 1976). In the context of the model, suppression refers to the reduction of the level of the signal at the output of the DRNL stage due to the addition of another (masker) tone or noise. If the masker level is sufficiently high and off-frequency to the signal, the masker and the signal are compressed together, but subsequently most of the masker energy is removed by the second filter. This leaves a lower signal level at the output than in the case without the masker due to the compression of the signal. It is clear that this interaction between the masker and the signal can only occur if they are present at the same time, i.e. in simultaneous masking, but not in forward masking.

B.4 Results

The results of the simulations are shown in Figure B.2. The nonlinear model (upper panels) predicts a sharper tuning for forward masking, as evidenced by the narrower tuning curve and the steeper slope of the notched-noise function in forward masking (marked by crosses). The differences in tuning are similar to those observed in human data (Moore et al., 1984; Oxenham and Bacon, 2003). The linear model (lower panels) shows similar tuning for both masking conditions, and this tuning is close to the simultaneous-masking prediction of the nonlinear model.

B.5 Discussion

Suppression and frequency selectivity

To explain the link between changes in frequency selectivity and suppression, one hypothesis has been that the influence of suppression is more widely tuned than that of excitation. Then, in simultaneous masking, a combination of suppression and excitation produces masking,

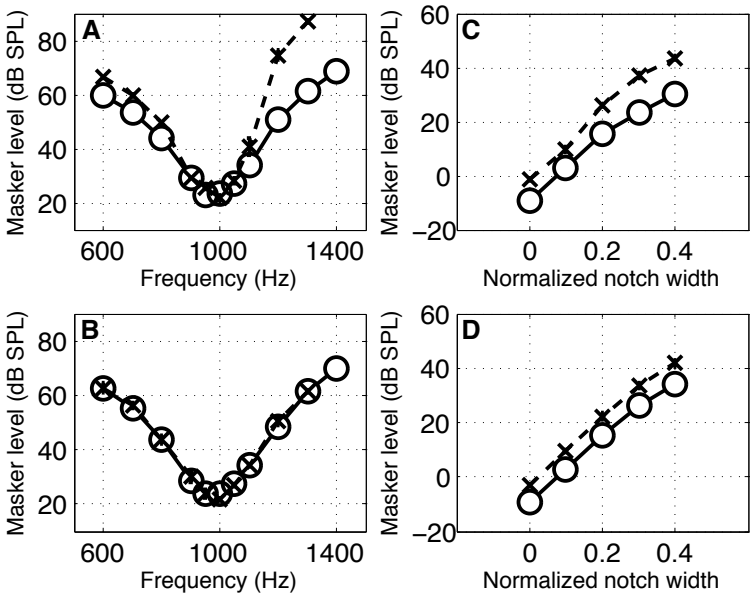


Figure B.2: Simulation results. Panels A and B show simulated PTCs, and panels C and D show simulated notched-noise thresholds. The top panels (A and C) show the predictions of the nonlinear model, the bottom panels (B and D) show predictions of the linear model. Circles indicate simultaneous, crosses forward masking thresholds.

whereas in forward masking, this additional, more widely-tuned influence of the masker on the signal is absent. Hence, the measured tuning in forward masking is sharper.

This hypothesis can be tested in the framework of the model by comparing the frequency-selectivity estimates for the nonlinear and the linear model versions. In the model, suppression arises as a result of compression; therefore both suppression and compression are absent in the linear version. If suppression would cause a broadening of tuning in simultaneous masking in the nonlinear model, then with suppression removed, the tuning should be narrower in the linear model.

The simulation results from the present study are clearly not consistent with this hypothesis. In fact, the opposite was observed: the simultaneous tuning estimate remains the same between the model versions, and it is the forward masking estimate that is broader in the linear model.

The effect of compression

Consider that in forward masking the masker is processed independently from the signal because of the temporal gap between them. If the masker level is sufficiently high, the masker is compressed by itself. As a result, a greater change in masker level is needed before the compressor to achieve the same change in level as when the masker is processed linearly. If a filter follows the nonlinearity, as in the current model, the compressive function before the second filter effectively sharpens the second filter (viewed from the input). In simultaneous masking, the masker and the signal are processed together, and any compression of the masker also reduces the signal level, such that the relative levels of the signal and the masker do not

change. Therefore, this sharpening caused by the compressive function only occurs in forward masking.

From this reasoning, it follows that in the linear model, where compression is removed, the tuning in forward masking should be wider than in the nonlinear model. This is consistent with the simulation results.

B.6 Summary and conclusions

In summary, two possible mechanisms affecting tuning estimates in the different masking paradigms have been identified: (1) suppression, as a result of compression, causing a *widening* of tuning in *simultaneous* masking; and (2) compression directly leading to a *narrower* tuning in *forward* masking. In the framework of the model, explanation (2) seems to be dominant, based on a comparison of the linear and nonlinear model versions. Consequently, the simultaneous masking estimates reflect the underlying tuning of the model more closely. The critical assumption here is that some sort of filtering follows the compressive function, and that the bandpass nonlinearity structure used in the DRNL filter is an appropriate model of the behavior of the basilar membrane.

B.7 References

Duifhuis, H (1976). "Cochlear nonlinearity and second filter: Possible mechanism and implications". In: *The Journal of the Acoustical Society of America* 59.2, pp. 408–423.

Jepsen, M. L., S. D. Ewert, and T. Dau (2008). "A computational model of human auditory signal processing and perception". In: *The Journal of the Acoustical Society of America* 124, pp. 422–438.

Marschall, M., J. Buchholz, and T. Dau (2011). "Modeling the effects of compression and suppression on estimates of auditory frequency selectivity". In: *Fortschritte der Akustik*, pp. 605–606.

Meddis, R., L. P. O'Mard, and E. A. Lopez-Poveda (2001). "A computational algorithm for computing nonlinear auditory frequency selectivity". In: *The Journal of the Acoustical Society of America* 109, pp. 2852–2861.

Moore, B. C. J. and B. J. O'Loughlin (1986). "The use of nonsimultaneous masking to measure frequency selectivity and suppression". In: *Frequency selectivity in hearing*. Ed. by B. C. J. Moore. Academic Press, London. Chap. 4.

Moore, B. C. J., B. R. Glasberg, and B. Roberts (1984). "Refining the measurement of psychophysical tuning curves". In: *The Journal of the Acoustical Society of America* 76.4, pp. 1057–1066.

Oxenham, A. J. and S. P. Bacon (2003). "Cochlear Compression: Perceptual Measures and Implications for Normal and Impaired Hearing". In: *Ear & Hearing* 24, pp. 352–366.

Plack, C. J., A. J. Oxenham, and V. Drga (2002). "Linear and nonlinear processes in temporal masking". In: *Acta Acustica united with Acustica* 88.3, pp. 348–358.

Contributions to Hearing Research

Vol. 1: *Gilles Pigasse*, Deriving cochlear delays in humans using otoacoustic emissions and auditory evoked potentials, 2008.

Vol. 2: *Olaf Strelcyk*, Peripheral auditory processing and speech reception in impaired hearing, 2009.

Vol. 3: *Eric R. Thompson*, Characterizing binaural processing of amplitude-modul-ated sounds, 2009.

Vol. 4: *Tobias Piechowiak*, Spectro-temporal analysis of complex sounds in the human auditory system, 2009.

Vol. 5: *Jens Bo Nielsen*, Assessment of speech intelligibility in background noise and reverberation, 2009.

Vol. 6: *Helen Connor*, Hearing aid amplification at soft input levels, 2010.

Vol. 7: *Morten Løve Jepsen*, Modeling auditory processing and speech perception in hearing-impaired listeners, 2010.

Vol. 8: *Sarah Verhulst*, Characterizing and modeling dynamic processes in the cochlea using otoacoustic emissions, 2010.

Vol. 9: *Sylvain Favrot*, A loudspeaker-based room auralization system for auditory research, 2010.

Vol. 10: *Sébastien Santurette*, Neural coding and perception of pitch in the normal and impaired human auditory system, 2011.

Vol. 11: *Iris Arweiler*, Processing of spatial sounds in the impaired auditory system, 2011.

Vol. 12: *Filip Munch Rønne*, Modeling auditory evoked potentials to complex stimuli, 2012.

Vol. 13: *Claus Forup Corlin Jespersgaard*, Listening in adverse conditions: Masking release and effects of hearing loss, 2012.

Vol. 14: *Rémi Decorsière*, Spectrogram inversion and potential applications for hearing research, 2013.

Vol. 15: *Søren Jørgensen*, Modeling speech intelligibility based on the signal-to-noise envelope power ration, 2014.

Vol. 16: *Kasper Eskelund*, Electrophysiological assessment of audiovisual integration in speech perception, 2014.

Vol. 17: *Simon Krogholt Christiansen*, The role of temporal coherence in auditory stream segregation, 2014.

Vol. 18: *Márton Marschall*, Capturing and reproducing realistic acoustic scenes for hearing research, 2014.

The end.

To be continued...