Technical University of Denmark

CONTRIBUTIONS TO
HEARING RESEARCH

Volume 9

Sylvain Favrot

A loudspeaker-based room auralization system for auditory research

DTU Electrical Engineering
Department of Electrical Engineering

A loudspeaker-based room auralization system for auditory research

PhD thesis by
Sylvain Favrot

Technical University of Denmark
2010

Cover illustration: View of the loudspeaker array
at the SpaceLab facility in CAHR, DTU.

ISBN 978-87-92465-23-8
Printed in Denmark by Rosendahls - Schultz Grafisk a/s

Preface

This thesis was submitted to the Technical University of Denmark (DTU) as partial fulfillment of the requirements for the degree of Doctor of Philosophy (Ph.D.) in Electronics and Communication. The work presented in this thesis was completed between January 15, 2007 and April 16, 2010 at the Centre for Applied Hearing Research (CAHR), Department of Electrical Engineering, DTU. The work was done under the supervision of Assoc. Prof. Jörg M. Buchholz and Professor Torsten Dau. The project was funded by the Technical University of Denmark. All the experiments presented in this thesis were approved by the Science-Ethics Committee for the Capital Region of Denmark; reference H-KA-04149-g.

Sylvain Favrot
Kgs. Lyngby, April 16th, 2010

A public defense was held on June 29, 2010 with Prof. Michael Vorländer (Aachen University, Germany), Dr. Bernhard Seeber (MRC Nottingham, UK) and Assoc. Prof. Finn Agervist (DTU) as members of the assessment committee.

Acknowledgment

These 3 years have been full of new and rich experiences for me: playing with 29 loudspeakers, conducting listening experiments, cycling to work, going to conferences, a big news...

I would like to thank Torsten Dau for giving me the opportunity to realize this PhD project and for his interest and advises and for fruitful discussions. I would also like to thank Jörg Buchholz, my main supervisor for his dedication, his openness, and for all the valuable discussions we had. It was a pleasure to work with you on this project. This work would not have been possible without the help of Claus Lynge Christiansen and his detailed explanation about the ODEON software.

Thanks to Iris Arweiler, for her expertise and valuable discussions about the speech experiment. Thanks to Sandra Rodiño for organizing the test-subject appointments and carrying out the NFC distance perception experiment. Mange tak Morten for at oversætte mit resumé. I would like to thank all CAHR people and specially my office mates for the great working environment. Thanks to the people in Widex and Oticon, and all the other people I played a demonstration of the system to for showing interest in the project.

Big thanks to my friends here in Denmark and in France. It was great to relax and party with you. Un grand merci à ma famille pour le soutient, l'accueil. Merci aussi de m'avoir rendu visite a plusieurs reprises. Finally, merci! Sarah for your support and for being there for me when I started my PhD.

Une thèse se finit, un volcan s'éveille...

Abstract

In complex acoustic environments, such as a train station or a café, hearing-impaired people often experience difficulties to communicate even when wearing hearing instruments, whereas normal-hearing people are typically able to communicate without effort in such conditions. In order to systematically study the signal processing of realistic sounds by normal-hearing and hearing-impaired listeners, a flexible, reproducible and fully controllable auditory environment is needed.

A loudspeaker-based room auralization (LoRA) system was developed in this thesis to provide virtual auditory environments (VAEs) with an array of loudspeakers. The LoRA system combines state-of-the-art acoustic room models with sound-field reproduction techniques. Limitations of these two techniques were taken into consideration together with the limitations of the human auditory system to localize sounds in reverberant environments. Each part of the early incoming sound to the listener was auralized with either higher-order Ambisonic (HOA) or using a single loudspeaker. The late incoming sound was auralized with a specific algorithm in order to provide a diffuse reverberation with minimal coloration artifacts.

In order to assess the usability of the LoRA system, one objective and two subjective evaluations were carried out. The objective evaluation showed that the physical characteristics of the acoustic scenario were preserved by the involved signal processing of the system. The first subjective evaluation assessed the impact of the auralization technique used for the early incoming sound (HOA or single loudspeaker) on speech intelligibility. A listening test showed that speech intelligibility experiments can be reliably conducted with the LoRA system with both techniques. The second evaluation investigated the perception of distance in VAEs generated by the LoRA system. These results showed that the distance of far field sources are similarly perceived in these VAEs as in real environments. For close sources (< 1 m), a comprehensive study about the near field compensated HOA method was presented and an alternative post-processing was proposed that allowed for the perception of very close sound sources, nearly as accurately as real sources.

Beside investigating the auditory system, such virtual auditory environments (VAEs) are also relevant for evaluating and optimizing hearing instruments and communication devices.

Resumé

Hørehæmmede oplever ofte problemer med at kommunikere i komplekse akustiske miljøer, som f.eks. en togstation eller en café, også selvom de bruger høreapparat. Normalthørende har typisk intet besvær ved akustisk kommunikation under samme betingelser. For at kunne udføre systematiske undersøgelser af hørelsens signalbehandling af realistiske lyde hos normalthørende og hørehæmmede, er det nødvendigt at have et fleksibelt, reproducerbart og fuldt kontrolleret auditivt miljø.

I dette projekt blev der udviklet et højttalerbaseret rum auraliserings system (loudspeaker-based room auralization, LoRA) som skaber et virtuelt auditivt miljø (Virtual Auditory Environment, VAE) ved brug af et højttaler-array. LoRA systemet kombinerer "state-of-the-art" rum-akustiske modeller med teknikker til at reproducere lydfelter. Der blev taget højde for disse to teknikkers begrænsninger, sammenholdt med overvejelser omkring begrænsninger ved menneskets hørelse i forhold til lokalisering af lyde i akustiske miljøer med efterklang. De tidlige dele af en indkommende lyd blev auraliseret enten ved higher-order Ambisoncs (HOA) eller ved brug af en enkelt højttaler. Den senere del ad lyden blev auraliseret ved en specifik algoritme for at skabe diffus efterklang med et minimum af artefakter som f.eks. farvning af lyden.

For at vurdere brugbarheden af LoRA systemet blev der udført én objektiv og to subjektive evalueringer. Den objektive evaluering viste at de fysiske karakteristika af det akustiske scenarie blev bevaret af den involverede signalbehandling. Den ene subjektive evaluering vurderede virkningen af auraliserings-teknikken som blev brugt for de tidligt indkommende lyde (HOA eller enkelt højttaler) i forhold til taleforståelighed. Et lytteforsøg viste at pålidelige taleforståeligheds-eksperimenter kan udføres ved brug af LoRA systemet med begge teknikker. Den anden subjektive evaluering blev udført for at undersøge opfattelse af afstand i virtuelle auditive miljøer som var genereret af LoRA systemet. Resultaterne viste at afstanden til fjernfelts lydkilder opfattedes ensartet i virtuelle og virkelige miljøer. For nære lydkilder (< 1 m) blev der udført en omfattende undersøgelse af en såkaldt nærfelts kompenseret HOA metode, og efterfølgende blev der foreslået en alternativ efterbehandling af signalerne som muliggjorde afstandsopfattelse af meget nære lydkilder, næsten så præcist som virkelige kilder.

Udover muligheden for at undersøge hørelsen, er sådanne virtuelle auditive miljøer også relevante i forbindelse med evaluering og optimering af eksempelvis høreapparater and andre kommunikationsredskaber.

Contents

List of abbreviations

2D	Two dimensions
3D	Three dimensions
AWW	Angular weighting windows
B&K	Brüel & Kjær
BRIR	Binaural room impulse response
C_{80}	Clarity
EDT	Early decay time
G	Strength
HOA	Higher-order Ambisonics
HRTF	Head-related transfer function
IACC	Inter-aural cross correlation coefficient
IN	Intensity normalized
LoRA	Loudspeaker-based room auralization
mRIR	Multichannel room impulse response
NA	Natural amplitude cues
NFC	Near-field compensated
RA	Random amplitude
RF	Regularization function
RIR	Room impulse response
SNR	Signal-to-noise ratio
SPL	Sound pressure level
SRT	Speech reception threshold
STI	Speech transmission index
T_{30}	Reverberation time
VAE	Virtual auditory environment

WFS	Wave field synthesis

1 General introduction

Hearing is the most important sense for human communication. Thereby, acoustic technologies have greatly contributed to the field of communication, in particular speech communication. This area of acoustics is referred to as *communication acoustics* (Blauert, 2005) and has evolved from the advent of electrical amplification with transistors to the advent of digital signal processing with computers. Modern communication and information systems have been developed with embedded knowledge and intelligence, such as perceptual audio-coders, hearing instruments (e.g., hearing aids and cochlear implants) or automatic speech recognition systems. The performance of these systems is generally degraded in complex acoustic environments, such as reverberant (indoor or outdoor) spaces with multiple sound sources (e.g., concurrent talkers in background noise). These complex acoustic environments are encountered in every-day life, e.g., in a train station, a supermarket, a café or a busy restaurant. Hearing-impaired people often experience difficulties to communicate in these complex environments even when wearing hearing instruments, whereas normal-hearing people are typically able to communicate without effort in such conditions. Therefore, a comprehensive understanding of the signal processing principles in the normal and impaired auditory system is required to improve and optimize hearing instruments and other communication systems, particularly in complex acoustic environments.

In order to gain insight into how listeners process and perceive realistic sounds, systematic investigations in a fully versatile (i.e., modifiable) auditory environment would be valuable. Specifically, these experimental auditory environments should be reproducible (i.e., identical for all subjects) and flexible (i.e., rapidly usable by any subject with or without hearing aids). In order to provide complex environments coping with these characteristics, *virtual auditory environments* (VAEs) have been considered in which the scene perceived by the listener does not correspond to the

"real" physical listener environment (Blauert, 2005). Instead, in such VAE, acoustic scenes are recorded or simulated and the signals provided to the listeners are processed in such a way that the same auditory perception (authenticity) as that in corresponding physical scenes is created. Beside investigating the auditory system, such VAEs are also relevant for evaluating and optimizing hearing instruments and communication devices. It is therefore of great importance that spatial attributes (e.g., localization of the sound sources, awareness of the surrounding space) are perceived similarly (spatial authenticity) in a VAE as in the corresponding real physical environment. VAEs are also used in various applications such as tele-conferences, training, entertainment, computer games, sound reinforcement or augmented reality.

In this thesis, the development and evaluation of the loudspeaker-based room auralization (LoRA) system is presented. This system aims at providing controllable VAEs for (i) investigating auditory processing of realistic sounds and (ii) evaluating and optimizing hearing instruments and communication devices in real-life acoustics scenarios.

Acoustic scene simulations

VAEs can either stem from a recorded physical acoustic scenario or from a simulated one, referred to as physically-based and model-based VAE, respectively. This thesis focuses on the latter case where acoustic room models compute a room impulse response (RIR) for each source-receiver configuration in a defined space, geometrically modeled in three dimensions (3D). For a given scenario, RIRs include relevant information (such as time, direction of arrival and frequency content) about the direct sound (source-receiver direct path) and about the reflections from the surrounding space (i.e., room) at the listener's position.

Acoustic room models are used in commercially available softwares (e.g., CATT; Dalenbäck, 1996; Ramsete; Farina, 2000; ODEON; Christensen, 2007) designed for prediction of room acoustic parameters and for auralization. The models use different methods to derive RIRs. Specular reflections are generally determined with one of the following algorithms: (i) the image source method (Allen and Berkley, 1979), (ii) the ray tracing method (Krokstad *et al.*, 1968; Lehnert and Blauert, 1992) or one of its derivate, cone tracing (Dalenbäck, 1996) or pyramid tracing (Farina, 2000), or

(iii) a combination of the two, referred to as hybrid methods (Naylor, 1993). These methods are typically used for computing the direct sound and early reflections of the RIR. Some acoustic room models include scattering (due to the surface material) and diffraction (due to the surface dimension) in these algorithms. For the late reflections, the ray tracing method (or one of its derivate) is typically modified after a given order of reflection (number of encountered surfaces) by using a stochastic process, e.g., the Lambert distribution (Kuttruff, 1991) or the oblique Lambert distribution (Christensen, 2007) to modify the reflection angle. Acoustic room models thus allow for computing model-based acoustic environments that are described by a set of RIRs with different source locations but including the same receiver and room.

Auralization of single virtual sources

Once RIRs are simulated, sound signals carried by the source represented in the RIR are presented to the listener. This process is known as *auralization* and consists of processing the RIR according to the reproduction technique and then convolving it with the anechoic sound signal carried by the source (Blauert, 1997; Vorländer, 2008). Reproduction techniques for auralizing RIRs can be grouped into two categories: binaural techniques and sound field reproduction techniques. In both cases, the direct sound and each reflection of the simulated RIR are considered as a single source.

For a single virtual sound source, the *binaural* reproduction technique aims at providing signals to the listener's ears that would have resulted from the interaction of the sound wave emitted by the corresponding physical source with the listener's external ear, head and torso (Møller, 1992). This format requires a set of head-related transfer functions (HRTFs) which encode the interaction of the sound wave with a listener for a given source direction. Using individualized HRTFs requires individual listener HRTF measurements and thus a significant technical effort. Non-individualized HRTFs can be employed to avoid this technical effort. However, they often lead to poorer performance in localization tasks (e.g., Wenzel *et al.*, 1993). Non-individualized HRTFs are typically measured with a manikin (Gardner and Martin, 1994) or a "typical" listener. Binaural reproduced VAEs are either reproduced with headphones or loudspeakers. The latter case is referred to as cross-talk cancellation or transaural systems (Møller, 1992; Vorländer, 2008) which control the signals at the listener's ears with at least

two loudspeakers by canceling out the undesired acoustic paths (for a stereo set-up, the left-loudspeaker-to-right-ear and right-loudspeaker-to-left-ear path are canceled). The binaural reproduction format requires a real-time update of the HRTFs when the listener rotates his/her head. In this case, a head-tracker must be used to monitor the head position.

The other approach to reproduce a single virtual sound source aims at reproducing the *sound field* that the corresponding physical sound source would have had created. This sound field is reproduced either locally or over an extended area with the help of an array of loudspeakers. The inherent advantage of this technique is that no previous knowledge of the subject or the device is required. Three-dimensional (3D) vector-based amplitude panning (VBAP; Pulkki, 1997) creates a phantom source by appropriately panning the three closest loudspeakers to the virtual sound source direction in order to reproduce some aspects of the original sound field like interaural differences. VBAP is a simple and efficient way of spatializing virtual sources. However, the quality of phantom sources depends on their location relative to the nearby loudspeaker locations (e.g., Shirley *et al.*, 2007) which leads to an inhomogeneous quality of the VBAP reproduction over the entire 3D sphere. Higher-order Ambisonics (HOA; Gerzon, 1973; Malham and Myatt, 1995; Daniel, 2000) is based on spherical harmonics decomposition, at a given order, of the sound field which is homogeneously reproduced locally at the center of the loudspeaker array (i.e., the sweet spot). The name HOA regroups here all sound field reproduction techniques (e.g., Poletti, 2005) utilizing spherical harmonics. The accuracy of the reproduction decreases with distance from the array center, with increasing frequency and with decreasing order (which is linked to the number of loudspeakers). Wave-field synthesis (WFS; Berkhout *et al.*, 1993) is another sound field reproduction technique and is based on the Kirchhoff-Helmholtz integral. WFS aims at reproducing wave fronts over a large area. The aliasing frequency below which the reproduction is accurate depends on the spacing of the loudspeakers and therefore lies in the audible frequency range even when a very large amount of loudspeakers is used. Above this aliasing frequency, the spatial characteristics of WFS reproduced sound fields become largely inaccurate and it has been shown that, for similar practical arrays, HOA provides an overall more robust reproduction of virtual sources (Daniel *et al.*, 2003). Thus, WFS allows for a

good low-frequency reproduction within a large area (and hence can be presented to a large number of listeners) whereas HOA provides a more accurate reproduction up to higher frequencies but is locally restricted (sweet spot) and is therefore better suited for presentations to a single listener.

Loudspeaker-based room auralization (LoRA)

The loudspeaker-based room auralization (LoRA) system has been developed in this project to provide VAEs that meet the following requirements: they should be reproducible, versatile, flexible and authentic. *Reproducibility* is required to create identical conditions to every subject. *Versatility* is required for practical modifications of any element of the acoustic scenario. *Flexibility* in the use of the system is required which implies that VAEs should be independent of the listener and/or the device in order to reduce the time and technical effort, in particular for experiments with a large number of participants. *Authenticity* is important when investigating auditory processing of realistic sounds and evaluating communication devices in real-life scenarios. The LoRA system was designed to use outputs from room acoustic models, which allow for the versatility. The reproducible VAEs were provided by using sound field auralization based on HOA which allows for the flexibility of the system. Sound field auralization in this system allows for head-rotations and the possibility of wearing hearing-aids without modifying the processing of the VAE. Even though the performance of state-of-the-art room acoustic models and the accuracy of the HOA technique do not imply authenticity, these techniques were selected here (in this thesis) to provide performances approaching authenticity.

Room acoustic softwares often include HOA auralization outputs. However, their Ambisonic order, i.e., the degree of accuracy of the reproduced sound field (which depends on the number of loudspeakers available) is usually limited (e.g., 3^{rd}-order for CATT, 2^{nd}-order for ODEON) and is applied to the whole RIR. Furthermore, HOA introduces coloration artifacts (Solvang, 2008) which is in particular critical for the reproduction of the late reflections.

The novel LoRA system considers the impact of the different parts of the RIR on auditory perception together with the technical limitations of the acoustic room model and of HOA auralization. Therefore, the direct sound, the early reflections

and the late reflections are processed independently. The Ambisonic order can be independently set for the direct sound and the early reflections and a special processing that minimizes coloration artifacts is used to auralize the late reflections. The LoRA system is configurable with any spherical or circular loudspeaker array arrangements.

Loudspeaker array setup

Loudspeaker arrays for sound-field reproduction are ideally used in an anechoic room in order to avoid uncontrolled contributions from the surrounding walls. Alternatively, and for practical reasons, acoustically-damped rooms with low reverberation time can be used. In order to listen to VAEs provided by the LoRA system and to carry out experiments, a 3D spherical 29-loudspeaker array was set-up in the SpaceLab facility in the Centre for Applied Hearing Research (CAHR) at DTU (see picture on the cover of the thesis and Fig. 3.1 and 2.6). This listening room has a volume of 50 m^3 and a reverberation time of $T_{30} = 0.16$ s at 125 Hz and below 0.1 s for higher frequencies. More detailed characteristics can be found in Appendix A. Loudspeaker equalization is performed in order to flatten the frequency response of each loudspeaker at the center of the array and to compensate for possible time delays between loudspeaker channels. The loudspeaker equalization method used in the SpaceLab facility is described in Appendix B.

Evaluation of the system

The goal of the LoRA system is to provide complex VAEs for investigating the normal, impaired and aided-impaired auditory system. An evaluation of the system is thus required to assess if the characteristics of the provided VAEs are similar to that of the corresponding physical auditory environments.

HOA reproduction of single sources in anechoic conditions (without room auralization) have been objectively evaluated throughout literature: (i) in a theoretical analysis of reproduction errors by Ward and Abhayapala (2001) and Poletti (2005), (ii) in terms of spectral impairment (Solvang, 2008), (iii) in terms of localization vectors (Gerzon, 1992; Daniel, 2000) and (iv) in reproducing interaural differences (Daniel, 2000). However, in the case of the LoRA system, the early reflections are reproduced

with HOA and the late reflections with a specific algorithm. Consequently, the whole auralized RIR needs to be evaluated.

Since the LoRA system focuses on authentic auditory perception rather than an acoustically-exact reproduction of the sound field, a subjective evaluation is needed. A review of subjective evaluations of single HOA reproduced sources (without room reverberation) can be found in Bertet (2009, chap. 3). These studies (and Frank *et al.*, 2008) showed a more precise direction localization when the Ambisonic order increases. When the whole RIR is reproduced, the ideal reference case (in the simulated or recorded space) is usually not available in the test space, in contrast to anechoic sources. It is therefore difficult to directly compare quality attributes as it is done in studies with anechoic scenes (e.g., Bertet, 2009, chap. 5). While direction localization definitively represents an important aspect of the reproduced VAEs, other aspects such as distance localization and speech intelligibility are also of relevance especially in the context of the LoRA system. Thereby, these two types of evaluation were selected for the LoRA system.

Overview of the study

Chapter 2 describes the design and the detailed implementation of the LoRA system and presents an objective evaluation of the LoRA system by means of room acoustic parameters such as reverberation time, clarity, strength and interaural cross correlation coefficients, an approach also followed in Kuster (2009). These parameters represent characteristics of the RIR that can be evaluated at the room model stage as well as after the auralization stage. Therefore, their comparison provides objective insights about the signal processing of the LoRA system.

Chapter 3 investigates the impact of the reproduction format of the direct sound and the early reflections on speech intelligibility. Natural enhancement of speech intelligibility by early reflections was assessed when reproduced by Ambisonic 1^{st}- and 4^{th}-order and by a reference reproduction format using a single loudspeaker. This last format uses only the nearest loudspeaker to the direct sound and each reflection direction and represents an alternative for the reproduction of the early part of the RIR in the LoRA system. Many attributes of the reproduction format potentially influence speech intelligibility, such as spectral, spatial and temporal contents. Therefore,

speech intelligibility outcomes constitute a global indicator of the performance of the LoRA system.

Another global indicator of the performance of the LoRA system that is investigated in this thesis is distance perception. **Chapter 4** presents the results of a distance perception experiment including "far field" sources, i.e., sources further away than the loudspeaker array radius (1.8 m, for two types of simulated rooms), as well as "near field" sources. In the latter case, the near field compensating (NFC) method for HOA was utilized in a simulated classroom and in anechoic conditions. This constituted a preliminary study about the NFC method. The auditory system makes use of several cues (e.g., intensity, direct-to-reverberation ratio) when perceiving distances. This study focused on the reproduction of the direct-to-reverberation ratio cue for sources between 1 and 10 m from the listener. The reproduction of this cue is the result of the overall processing in the system, including the acoustic room model and the LoRA system.

Chapter 5 presents a comprehensive study of near field compensated HOA and considers an alternative angular weighting window for the post-processing of the technique. This novel weighting window as well as two existing windows were objectively compared in terms of spectral responses and interaural level differences. A listening experiment was also performed to evaluate if virtual sources reproduced with NFC-HOA can be localized inside the loudspeaker array radius.

Finally, **Chapter 6** present a summary of the main outcome of this work and a discussion about future perspectives.

2

LoRA: a loudspeaker-based room auralization system *

Abstract

In order to study basic human perception in reverberant environments, a novel loudspeaker-based room auralization (LoRA) system is proposed in this paper. The LoRA system efficiently combines modern room acoustic models with high-order Ambisonic auralization. An objective evaluation has been carried out demonstrating the applicability of the LoRA system. Room acoustic parameters (reverberation time, clarity, speech transmission index and inter-aural cross correlation coefficients) of room impulse responses were compared at the input and the simulated output of the LoRA system. Results show that the involved signal processing preserves the temporal, spectral and spatial properties of the room impulse response captured by these parameters. This flexible research platform will be useful for studying auditory processing and perception in normal-hearing and hearing-impaired listeners in fully controlled and realistic environments.

2.1 Introduction

Speech communication in realistic environments (i.e., multiple sound sources in a reverberant room) is often problematic for hearing-impaired people, even though their decreased sensitivity to sound might have been accounted for by amplification through hearing aids. Since normal-hearing people perform well in such complex situations, differences between the normal and impaired auditory system have been widely studied. However, this auditory research has often been carried out by using synthetic

* This chapter was published as Favrot and Buchholz (2010).

stimuli (such as clicks, tones, or white noise) or stimuli presented in a poorly controlled environment (e.g., using a small number of loudspeaker-sources in an arbitrary laboratory space).

In order to systematically study the signal processing of realistic sounds by normal-hearing and hearing-impaired listeners, a flexible and fully controllable auditory environment is needed. In order to provide such environment, a loudspeaker-based room auralization (LoRA) toolbox has been developed in this study. The LoRA toolbox combines state-of-the-art acoustic room models with loudspeaker-based auralization techniques. Limitations of these two techniques have been taken into consideration together with the limitations of the human spatial localization in reverberant environments for an efficient combination. Besides studying basic human perception, such auditory environment is also of interest for evaluating room acoustical qualities as well as for assessing and optimizing the performance of modern speech and audio technologies, such as hearing aids, cochlear implants, (perceptual) audio-coders, or automatic speech recognition systems.

Virtual auditory environments (VAE) ideally create the percept of acoustic events that are generated by sound sources in a given space, whereas neither the sound sources nor the space are physically present in the actual listening environment (Begault, 1994; Blauert, 1997, 2005). A number of VAE systems (recent reviews can be found in Kleiner *et al.*, 1993; Blauert, 2005; Lokki, 2002) have been developed over the years mainly for applications such as vehicle simulators (training and assistance for pilots, Begault, 1994), tele-conferences, computer games or entertainment (Pulkki, 2007). Some of these VAEs are generated by binaural room simulation (BRS), which combines geometrical acoustic room models (Allen and Berkley, 1979) with binaural auralization techniques (Kleiner *et al.*, 1993; Farina, 1993) to derive binaural room impulse responses (BRIRs) for specific rooms, sources and listener positions. These BRIRs are then convolved with an anechoic sound sample and presented via headphones.

The main advantage of binaural systems is that the acoustic scenes are fully controllable (e.g., the number and positions of sound sources, the type of room). Moreover, if listener-specific BRIRs are used, these systems achieve almost "authentic" reproduction of the scene, which even allows the application in rigorous psychoa-

coustical experiments (Azzali *et al.*, 2005; Djelani *et al.*, 2000; Lindau *et al.*, 2007). The term "authentic" utilized here refers to the definition by P. Novo in Blauert (2005) who states that the objective of such a reproduction "is to evoke in the listener the same percepts that would have been evoked in the corresponding auditory real environments".

Although BRS systems have already been used for auditory perception research, their usefulness for investigating the auditory processing of realistic sounds is limited. First, in order to achieve the authenticity required for application in psychoacoustical experiments, the BRS system needs to be tailored to the individual listener. Therefore, listener-specific BRIRs need to be derived, which requires the measurement of individual head-related transfer-functions (HRTFs). This introduces a significant technical effort, particularly if many subjects are involved in the considered experiment. This problem is even increased when the effects of the processing by hearing aids or cochlear implants are assessed. In such a case, the binaural system also needs to consider the device-specific characteristics and limitations. Second, listener's head movements or rotations, which provides an important cue for sound source localization (Blauert, 1997), need to be incorporated (Minnaar *et al.*, 2001). In order to address this issue, advanced BRS systems include real-time interaction of the listener with the virtual world, i.e., the listener's body and head movements modify the auditory perception as it would have been the case in the real world. Real-time BRS systems require a "head-tracker" which monitors the head position in order to update the BRIR. Although this method eliminates front-back reversals, the externalization of the sound event, i.e., inside-the-head localization, is not significantly improved (Begault *et al.*, 2001, at least for speech) which is a drawback for authentic reproduction.

In order to by-pass these two difficulties inherent to BRS systems, a loudspeaker-based room auralization (LoRA) system is utilized in the present study. The VAE generation apply similar acoustic room models as used in BRS systems, but uses an array of loudspeakers instead of binaural technology and headphones for auralization of the acoustic scene.

Loudspeaker-based auralization aims at reproducing the sound field at the listener's location (in the virtual room) in the center of the loudspeaker array and thus does not consider individual-listener specific information such as individual HRTFs.

The effect of the presence of the listener on the reproduced sound field is thereby the same as it would be on the original (real) sound field. Listener's head rotations and the use of hearing devices are thus possible without any modification of the system. In addition, the possibility of off-line signal processing allows the use of very sophisticated room models, which in turn enables the reproduction of very complex and realistic acoustical scenes. The disadvantage of such systems is that: (1) a large amount of loudspeakers is required for a precise sound-field reproduction, (2) the listening room needs to be acoustically treated to avoid reverberation from the playback room to interfere with the virtual acoustic scene and (3) listener's head translations are not accounted for.

A number of techniques have been previously proposed for reproducing arbitrary sound fields via loudspeakers; the most widely used being wave-field synthesis (WFS; Berkhout *et al.*, 1993), vector-based amplitude panning (VBAP; Pulkki, 1997) and Ambisonics (Gerzon, 1973). The present LoRA toolbox uses Ambisonics and its extension, higher-order Ambisonics (HOA; Malham and Myatt, 1995; Daniel, 2000). HOA is based on spherical harmonics decomposition of three-dimensional sound fields. The reachable order of the decomposition depends on the number of available loudspeakers and determines the directionality of the reproduced sound sources.

Although modern room acoustic modeling programs often include loudspeaker-based auralization modules (Dalenbäck, 1996; Rindel, 2000), these modules are typically realized by simply applying Ambisonic coding to the entire room impulse response which in particular leads to coherent loudspeaker signals. Hence, these programs do not take into account that different parts of the room impulse response have very different impact on human perception. Within the LoRA system, the advantages and limitations of the different techniques (i.e., acoustic room models and higher order Ambisonics) are considered together with knowledge of human perception of reverberant sounds (i.e., utilizing aspects of the precedence effect, Blauert, 1997; Zurek, 1987; Litovsky *et al.*, 1999) to approach the authenticity of the simulated acoustic scenes.

In order to evaluate the performance of the proposed VAE, an objective evaluation was performed on the simulated output of the system, including standard room measures such as reverberation time (T_{30}), early decay time (EDT), clarity (C_{80}), speech

transmission index (STI), and the inter-aural cross correlation coefficient ($IACC$). These objective measures provide a first evaluation step, and ultimately, listening experiments need to be conducted, such as speech intelligibility experiments, localization (discrimination) experiments, or distance perception experiments. However, such perceptual analysis was out of the scope of the present study and will be considered in a separate investigation.

The strategies and techniques implemented in the LoRA toolbox are described in detail in section 2. The objective evaluation of this system is carried out for an example loudspeaker set-up and is presented in section 3. The results of the study are discussed in section 4.

2.2 Methods

The LoRA toolbox combines acoustic room models with loudspeaker-based auralization. In the following, relevant details and limitations of these two techniques and the strategy for combining the two are described.

2.2.1 Acoustic room models

Acoustic room models are used to simulate the properties of the sound field for a specific room and source-receiver configuration. First, the geometry of the room is defined by a three-dimensional (3D) room model representing the coordinates of the prominent surfaces. A frequency-dependent acoustic absorption coefficient is assigned to each surface according to its material. Then, the position and characteristic of the source and the receiver are defined. With this information, a room impulse response (RIR) is simulated at the receiver location. Such RIR typically contains temporal, spectral and directional information of the room's response. Several professional programs with built-in acoustic room models have been developed over the past years (e.g., CATT, Dalenbäck, 1996; Ramsete, Farina, 2000; ODEON, Rindel, 2000). A number of round robin studies (Vorlander, 1995; Bork, 2000, 2005) have shown that these state-of-the-art programs are able to estimate accurately the characteristics

of the impulse response for conventional rooms down to a lower frequency limit of about 200 Hz (Bork, 2005). Throughout the present study, ODEON has been used.

	Direct Sound	**Early reflections**	**Late reflections**
(i) Acoustic room model interface	Discrete part: time and direction of arrival, attenuation for each reflection		Reverberation part: energy and vectorial intensity curves
(ii) Precedence effect	Very precise localisation (localisation dominance)	Reduced localisation (lag discrimination suppression)	Very poor localisation, limited audibility
(iii) Auralisation	High order Ambisonics or single loudspeaker	High order Ambisonics	Ambisonic 1st order envelopes multiplied by uncorrelated noise

Table 2.1: Summarized characteristics of each part of the room impulse response regarding (i) the input of the LoRA processing, (ii) the relevant auditory localization aspects and (iii) the auralization strategy.

The output of such models (i.e., the room impulse response) is used as input for the loudspeaker-based auralization. In order to be able to process RIRs calculated with different programs, it is important here to define a RIR input format for the LoRA toolbox. This interface considers two components of the RIR (see (i) in Tab. 2.1): (1) the direct sound and the early reflections (the discrete component) and (2) the late reflections (the reverberation component). This decomposition is in general agreement with the way RIRs are computed in most programs. The discrete component is defined by the direct sound and each single early reflection's frequency-dependent amplitude, time of arrival, and direction of arrival (i.e., azimuth and elevation). The diffuse component is defined by the frequency-dependent envelopes of the energy and the vectorial intensity (norm and direction) of the late reflections (with 10 ms resolution). For these two components of the RIR, spectral information is provided in eight octave bands (from 63 Hz to 8 kHz, see Tab. 2.2), as it is usually done in room acoustics (Dalenbäck, 1996; Farina, 2000; Rindel, 2000). The discrete and reverberant components typically overlap in time for conventional rooms.

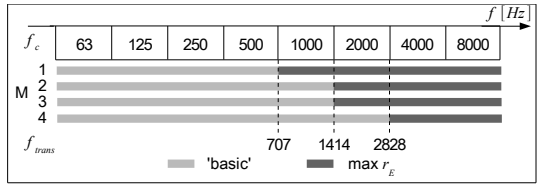

Table 2.2: Octave bands with center frequencies f_c from 63 Hz to 8 kHz. The lower part of the figure shows the frequency channels for which "basic" (light gray bars) and "max r_E" with energy normalization (dark gray bars) Ambisonic decoding is applied considering Ambisonic orders 1 to 4. The transition frequency f_{trans} between these two decoding techniques corresponds to the higher cut-off frequency of one of the octave bands.

Matlab routines required to interface data exported from ODEON are included in the LoRA toolbox as an example. Similar routines can be written for other room simulation programs. For simple room configurations, public domain programs can alternatively be used for simulating the early part of the RIR (e.g., Allen and Berkley, 1979; Miller, 2001). Late reflections energy envelopes could then be derived from analytical formulas as, for instance, described by Kuttruff (1991).

2.2.2 Ambisonics

The LoRA toolbox aims at reproducing the sound field corresponding to each part of the room impulse response (RIR) at the center of a loudspeaker array. The elements composing each part of the RIR are considered as sources far enough to produce plane-wave sound fields at the receiver location. Ambisonics is used here to reproduce these sound fields. This method is based on a spherical harmonics decomposition of a given sound field truncated at a given order (Gerzon, 1973; Malham and Myatt, 1995; Daniel, 2000). In the encoding phase, the sound field is decomposed into spherical harmonics components. Then, in the decoding phase, the sound field is reconstructed by assigning a linear combination of the spherical harmonics components to each loudspeaker according to its position. Hence for the reproduction of a plane wave, for example, a simple gain is applied to each loudspeaker channel, which is dependent on: (1) the loudspeaker layout (i.e., the location of the loudspeaker and the total number of loudspeakers), (2) the direction of the original plane wave, and (3) the applied Ambisonic order.

The order of the decomposition (i.e., the Ambisonic order) determines the spatial directionality of the reproduced sound field. Ambisonic directivity patterns are examplarily shown in Fig. 2.1 for an Ambisonic order M from 1 to 4; the surface represents loudspeaker gains in any direction to reproduce a frontal incident plane wave (similar to a far sound source in the same direction). It can be seen that the beamwidth of a sound source decreases with increasing Ambisonic order M. Hence, for a spatially precise reproduction of individual sound sources, a high Ambisonic order is required. This relation between localization accuracy and Ambisonic order have been shown in perception studies (Bertet *et al.*, 2007; Frank *et al.*, 2008). However, for a homogeneous reproduction in all directions, the maximum order M is limited by the spatial distribution of the loudspeakers, i.e., the placement and number N of the loudspeakers. For regular loudspeaker layouts, this limitation is given by the relation:

$$N \geq (M+1)^2 \tag{2.1}$$

for a three-dimensional (3-D) reproduction and

$$N \geq 2M+1 \tag{2.2}$$

for a two-dimensional (2-D) reproduction (Ward and Abhayapala, 2001).

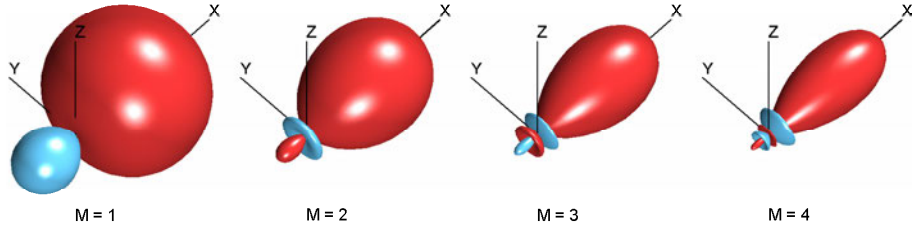

Figure 2.1: Three-dimensional Ambisonic directivity patterns for orders M = 1-4. The surface represents loudspeaker gains in every direction to reproduce a source in positive x direction. Dark surfaces represent positive gains; bright surfaces represent negative gains.

A known drawback of Ambisonics is that the sound field reproduction is done by coherent loudspeaker signals which can lead to perceptual artefacts such as coloration (Daniel, 2000; Solvang, 2008). This is particularly critical when the number

of loudspeakers is well above the limit imposed by the Ambisonic order (Eq. 2.1) as was shown in Solvang (2008) for 2-D Ambisonic reproduction. This is also a problem for the reproduction of diffuse sound fields where the energy in all directions should be equal. Since the reverberation component of the RIR is typically a diffuse sound field, its reproduction is therefore critical with Ambisonics.

Another limitation of Ambisonic is the fact that it is only possible to theoretically reconstruct the exact sound field at the center of the loudspeaker array. The reproduction error increases with: (1) increasing distance r from the center, (2) increasing wave number $k = 2\pi f/c$ and (3) decreasing Ambisonic order M. A simple rule of thumb indicates that, for an Ambisonic order M equal to the product kr, the reproduction error is around 4 % (Ward and Abhayapala, 2001; Poletti, 2005). Hence, a frequency limit can be defined up to which a satisfactory sound field reproduction is achieved for a given area around the center of the loudspeaker array (i.e., the sweet spot) and a given Ambisonic order. This frequency limit f_{lim} is defined by:

$$f_{lim} = \frac{Mc}{2\pi r} \tag{2.3}$$

where c stands for the speed of sound in air. For example, for forth order Ambisonics and for a sweet spot of the size of the head ($r = 10\ cm$), the sound field can be reconstructed with a reasonable accuracy for frequencies up to about $f_{lim} = 2183\ Hz$. As the reconstruction error increases strongly for higher frequencies ($f > f_{lim}$), a different decoding method is typically used. This method maximizes the energy in the expected direction by weighting each spherical harmonics component by an order-dependent gain and is known as "max r_E" decoding. This decoding method is associated with an energy normalization: an overall gain is applied to all harmonic components to preserve the total energy in the sweet spot (Daniel, 2000). The frequency limit f_{lim} (obtained from Eq. (2.3)) gives an indication of the transition frequency f_{trans} between the non-weighted ("basic") and weighted ("max r_E") decoding.

Within the LoRA toolbox, each element (i.e., the direct sound, each early reflection, and each sample of the energy and intensity curves of the late reflections) of the RIR provided by the room model (section 2.2.1) is realized with Ambisonics as an individual virtual source located beyond the radius of the loudspeaker array. Follow-

ing this far-field assumption, and moreover assuming a listening position in the center of the loudspeaker array, near-field compensation (NFC, Daniel, 2003) has not been applied in the HOA implementation. Amplitudes are handled in eight octave bands and the transition frequency is approximated by the nearest cut-off frequency of the octave filter bank as shown in Tab. 2.2.

2.2.3 Combining room model and loudspeaker-based auralization

The main aim of the LoRA toolbox is to efficiently combine acoustic room models (section 2.2.1) and high-order Ambisonics-based auralization (section 2.2.2) to realize a virtual auditory environment that approaches authentic reproduction. In order to accomplish this goal, aspects of human spatial perception associated with the precedence effect (Blauert, 1997; Zurek, 1987; Litovsky *et al.*, 1999) are considered in the implementation of the LoRA toolbox.

Strategy

Within the LoRA toolbox, each part of the room impulse response (i.e., the direct sound, early reflections and late reflections) provided by the acoustic room model is processed separately, depending on the role of this part for the spatial perception of the whole response (see (ii) in Tab. 2.1).

The precedence effect considers human localization performance in reverberant environments (for recent reviews see Zurek, 1987; Blauert, 1997; Litovsky *et al.*, 1999). Different aspects of the precedence effect seem to be important for developing a highly authentic VAE and are therefore considered here. The localization dominance or the law of the first wave front indicates that the direction of the direct sound dominates overall localization. Hence, particular care has to be taken when auralizing the direct sound. Lag discrimination suppression indicates that the auditory system's ability to localize individual reflections is significantly deteriorated by the presence of the direct sound as well as preceding reflections. Moreover, reflection masking (Buchholz *et al.*, 2001) indicates that a large number of individual reflections, in particular within the late part of the RIR, are inaudible. Hence, the requirements for auralizing reflec-

tions, in particular late reflections, are significantly less stringent than for the direct sound.

Depending on the loudspeaker layout available, the LoRA toolbox processes each component of the RIR (see (iii) in Tab. 2.1) according to the above considerations.

As the *direct sound* requires a very precise spatial reproduction high-order Ambisonics is used for its auralization. As described in section 2.2.2, the order is limited by the number of loudspeakers and therefore the highest Ambisonic order M_{max} according to Eq. 2.1 is chosen to provide the highest localization accuracy.

In contrast to the direct sound, rendering of reflections does not necessarily require a very high precision. The auralization of the *early reflections* thus might be done with a lower Ambisonic order than the one used for the direct sound. However, because high-order Ambisonics processing is not computational demanding, the same order for the direct sound and the early reflections is applied.

Individual *late reflections* have even a lesser impact on localization indicating that the spatial precision of their reproduction is not very critical. Moreover, only limited information, i.e., the envelopes of the energy and the vectorial intensity of the late reflections, is available from the acoustic room model. Within the LoRA toolbox, the energy and intensity envelopes are interpreted as encoded spherical harmonics components at the first Ambisonic order. These spherical harmonics components are then decoded at the first order to derive an envelope of the reverberation for each loudspeaker. These envelopes are multiplied with uncorrelated noises to recreate the late part of the RIR while preserving the diffuseness of the reverberation and limiting coloration effects (see Merimaa and Pulkki, 2005 for a similar strategy).

Detailed information on the implementation of the different RIR components within the LoRA toolbox is provided in the following section.

Implementation

For a given loudspeaker layout, the LoRA toolbox processes off-line the RIR provided by the acoustic room model (section 2.2.1). This results in a set of impulse responses (one for each loudspeaker), referred to as a multi-channel room impulse response (mRIR). The mRIR is derived separately for the direct sound, the early reflections and the late reflections, according to the strategies developed in section 2.2.3 and

summarized in Tab. 2.1. The overall mRIR is then calculated by adding the three mRIR components.

For the *direct sound*, the acoustic room model provides detailed information, i.e., the time τ and the direction of arrival (i.e., azimuth θ and elevation δ) and the amplitude $g^{attn}[k]$ in eight octave bands (indexed with the letter k).

Ambisonic gains $g_i^{Amb}[k]$ for each loudspeaker i depend on the direction (θ and δ) of the direct sound, the Ambisonic order M_{max} and the loudspeaker layout (see section 2.2.2). For the low frequency bands $k \leq k_{trans}$, these Ambisonic gains are obtained using a "basic" decoding scheme whereas for the high frequency bands $k > k_{trans}$, a "max r_E" decoding with energy normalization is used. The decoding schemes applied to the different frequency bands are shown in Tab. 2.2 for Ambisonic order M = 1-4.

In order to obtain the direct sound response at a desired sampling frequency f_s for each loudspeaker channel, the LoRA toolbox makes use of the impulse response $h_k[n]$ of a linear-phase Kaiser window octave-band filterbank (Orfanidis, 1995) (1727 points at 44.1 kHz). To consider the entire auditory frequency range, within the auralization process the lowest and highest channel of the filterbank were replaced by a lowpass and highpass filter with corresponding lower and upper cut-off frequencies. For each loudspeaker, the filter impulse response $h_k[n]$ of each band k is multiplied by the corresponding Ambisonic gain $g_i^{Amb}[k]$ and by the amplitude $g^{attn}[k]$ of the direct sound. The sum of these weighted band impulse responses is delayed by the time of arrival of the direct sound τ to obtain the direct sound part of the RIR for each loudspeaker $L_i^{DS}[n]$ as given by:

$$L_i^{DS}[n] = \sum_{k=1}^{8} g^{attn}[k]\, g_i^{Amb}[k]\, h_k[n-\tau] \tag{2.4}$$

The LoRA toolbox can also reproduce the direct sound using only the nearest available loudspeaker. In this case, the same method is applied but the Ambisonic gains $g_i^{Amb}[k]$ are replaced by the so-called "nearest loudspeaker" gains $g_i^{nl}[k]$. These gains equal 1 for the nearest loudspeaker to the direction of the direct sound for all bands and 0 for all other loudspeakers. This is of particular interest for arrays with few loudspeakers when, within the acoustic room model, the direction of the virtual source

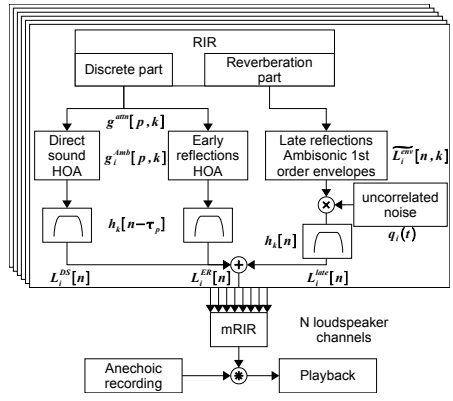

Figure 2.2: Implementation diagram of the LoRA processing. The multichannel RIR is derived in eight frequency bands and for each part of the input RIR. A virtual auditory scene is finally simulated by the playback of an anechoic signal convolved with the mRIR.

is made coincident with the direction of a loudspeaker. Thus, the multi-channel room impulse response $L_i^{DS}[n]$ for the direct sound is obtained by one of these methods.

Within the LoRA toolbox, each discrete *early reflection* is realized in the same way as specified for the direct sound in Eq. (2.4) (see also section 2.2.3). The responses for each of the P early reflections are then added to obtain the multi-channel room impulse response $L_i^{ER}[n]$ for the early reflections, i.e.,:

$$L_i^{ER}[n] = \sum_{p=1}^{P} \sum_{k=1}^{8} g^{attn}[p,k]\; g_i^{Amb}[p,k]\; h_k[n-\tau_p] \tag{2.5}$$

The envelopes of the energy and of the vectorial intensity of the *late reflections* in the eight octave bands are provided by the room acoustic model. As mentioned in section 2.2.3, they are decoded at the first Ambisonic order for each sample T of the envelope and each frequency band k. Due to the diffuseness of the late reflections, a "Basic" Ambisonic decoding is applied to all frequency bands. Loudspeaker specific envelopes of the late reflections $L_i^{env}[T,k]$ are thus obtained for the eight octave bands.

In order to recreate the natural high temporal density of late reflections, the LoRA toolbox multiplies the obtained envelopes with Gaussian noise. For each band k, the

loudspeaker specific envelopes are interpolated and smoothed $\widetilde{L}_i^{env}[n,k]$ to match the required sampling frequency f_s. The envelopes multiplied by the Gaussian noise $q_i[n]$ are then filtered by the corresponding Kaiser window design filter (with impulse response $h_k[n]$, see above). The level of the resulting reverberation for each loudspeaker channel i is adjusted according to the level of the envelopes in order to minimize the non-deterministic characteristic of the Gaussian noise. For each loudspeaker i, the Gaussian noise $q_i[n]$ is chosen uncorrelated from the noises used for the other loudspeakers in order to preserve the diffuseness of the reverberation component (see section 2.2.3). Thereby, the multi-channel RIR for the late reflection $L_i^{late}[n]$ is obtained by summing the filtered product of envelopes and noises as described by:

$$L_i^{late}[n] = \sum_{k=1}^{8} (\widetilde{L}_i^{env}[n,k]\; q_i[n]) \otimes h_k[n], \tag{2.6}$$

where the $\otimes$ symbol describes the convolution operator.

A signal flow diagram illustrating the implementation of the overall RIR is shown in Fig. 2.2. A temporal signal with sampling frequency f_s is derived for each loudspeaker channel for the direct sound $L_i^{DS}[n]$, the early reflections $L_i^{ER}[n]$ and the late reflections $L_i^{late}[n]$. The addition of these three parts provides the final multi-channel room impulse response (mRIR) for a given source receiver configuration in one room. A reverberant signal is auralized by convolving an anechoic signal with the mRIR, resulting in a mutli-channel sound signal. In order to create more complex environments, different multi-channel sound signals using different mRIRs (with the same receiver location) and anechoic source signals can be added.

2.3 Objective evaluation

In the following, an evaluation of the performance of the LoRA processing on a physical basis is presented. The aim of this objective evaluation is to assess the RIR distortions potentially introduced by the involved signal processing and playback methods.

2.3.1 Evaluation method

The present evaluation of the LoRA system is based on the comparison of room acoustic parameters extracted from: (1) the RIR provided by the room acoustic model (ODEON), which serves as input to the LoRA system and is therefore used as reference and (2) the simulated IR (monaural or binaural) at the listener's location derived from the mRIR which is the output of the LoRA system. The ODEON program was one of the three programs to be judged 'Unquestionably reliable in the prediction of room acoustical parameters' in the first round robin (Vorlander, 1995; Rindel *et al.*, 2009) and was part of the third round robin (Bork, 2005) where all programs show good agreement with the measured parameters. Therefore, ODEON provides a relevant reference for this study. Two categories of room acoustic parameters are considered here: monaural parameters to assess spectral and temporal aspects of the processing and binaural parameters to assess spatial aspects of the processing.

The monaural parameters considered here are the reverberation time T_{30}, the early decay time EDT, the clarity C_{80} and the strength G in seven octave bands (from $125\ Hz$ to $8\ kHz$) as well as the speech transmission index STI. The parameters are computed for both the input RIR and the output mRIR according to ISO 3382 standard (1997). A monaural RIR at the center of the loudspeaker array (i.e., the listener location) is obtained by summing up all channels of the output mRIR. By applying appropriate delays to the channels of the mRIR before the summation, monaural RIRs are also simulated at different positions inside the array. The parameters EDT and T_{30} are expressed in seconds, therefore relative errors (δEDT and δT_{30}) are calculated between the output mRIR parameter and the reference parameter. For the early-to-late ratio C_{80}, the strength G and the STI, the absolute deviations (ΔC_{80}, ΔG and ΔSTI) of the output from the reference parameters are computed.

The binaural parameter considered here is the inter-aural cross correlation coefficient (IACC) in seven octave bands (from $125\ Hz$ to $8\ kHz$). The IACC is computed separately for the early part $IACC_{0,t}$ and the late part $IACC_{t,+}$ of each binaural RIR according to ISO 3382 (1997). The value for the time t separating the early and the late part is typically $50\ ms$ for small rooms and $80\ ms$ for large rooms. The binaural evaluation of the LoRA system requires the derivation of binaural RIRs from the input RIR and from the output mRIR. ODEON automatically computes the binaural RIR

using a given set of head-related transfer functions (HRTFs). The reference IACCs of the input RIR are then calculated from this binaural RIR. The output mRIR provides a RIR for each loudspeaker according to its direction. The same set of HRTFs as used in the reference system is applied to calculate a binaural RIR for each individual loudspeaker's RIR. These binaural RIRs are summed up to obtain a single binaural RIR at the center of the loudspeaker array which is then used to compute the IACC at the output of the LoRA system. Delays are added to the individual loudspeaker channels to derive the IACC at different off-center positions. The absolute deviations $\Delta IACC_{0,t}$ and $\Delta IACC_{t,+}$ between the output and the reference IACC are finally computed both for the early and the late part of the response.

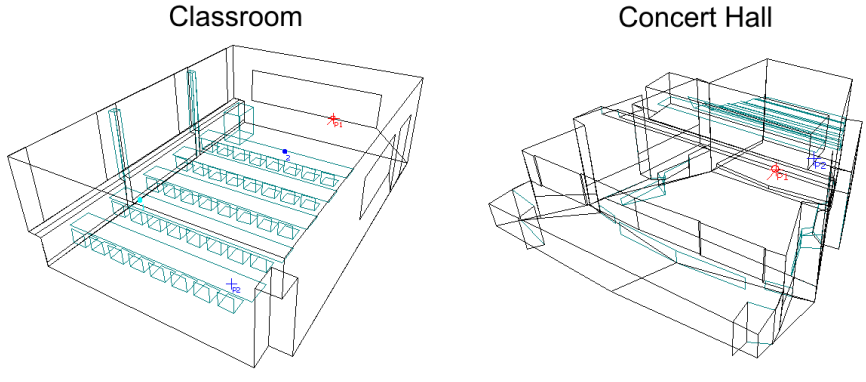

Figure 2.3: Geometrical 3D model of the classroom (40 seats, 170 m^3) and the concert hall (1033 seats, 12400 m^3) used in the evaluation.

Two example rooms were simulated with ODEON: a classroom and a concert hall (see Fig. 2.3). For each room, eight RIRs corresponding to eight source-receiver configurations were computed. Median monaural and binaural parameters of these input RIRs are shown in Fig. 2.4 and Fig. 2.5 for both rooms. The multi-channel room impulse responses were derived by the LoRA toolbox at a sampling rate of 44.1 kHz for a loudspeaker array with an example layout of 29 loudspeakers (see Fig. 2.6), which allowed a maximum available Ambisonic order of 4 (see Eq. (2.1)). Monaural and binaural parameter errors were computed at four listening positions: at the center of the loudspeaker array (0 cm) and off-center positions with a distance of 7, 15 and

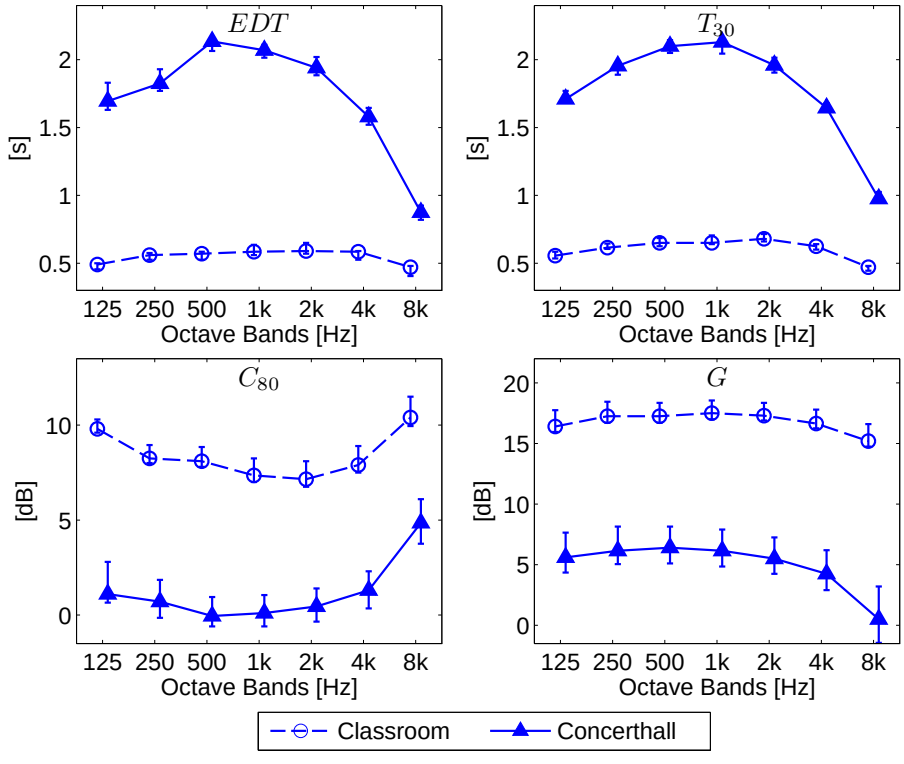

Figure 2.4: Monaural median room acoustic parameters (T_{30}, EDT, C_{80}, G) derived from the input RIRs for the classroom (dashed curve) and the concert hall (solid curve).

19 cm (see Fig. 2.7, gray circles). In order to evaluate the overall performance of the LoRA processing with reference to a specific room, the median and the interquartile of the errors were calculated over the eight source-receiver configurations for each listening positions.

Since it is difficult to interpret the error introduced by the LoRA toolbox, two different indicators were considered: the head movement limen and the subjective limen. For each individual parameter, this head movement limen consisted of the range of parameter deviation produced by small shifts in the receiver location in the virtual room. Shifts within an area of 27 $\times$ 27 cm (see Fig. 2.7) were considered and correspond to small head movements which are not expected to significantly modify the

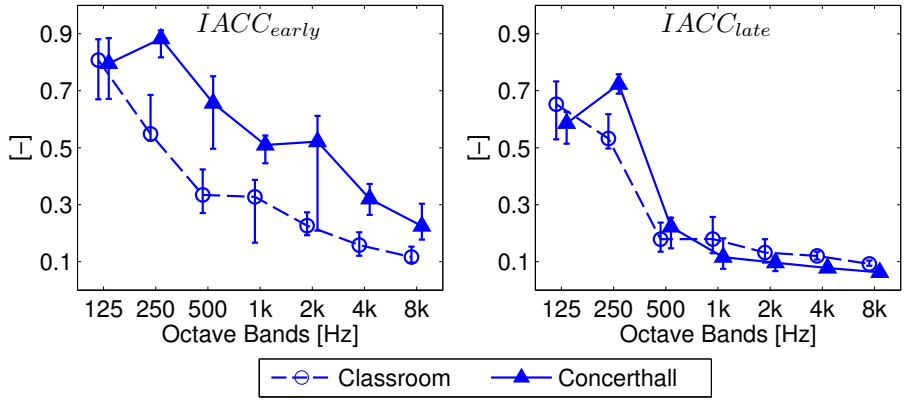

Figure 2.5: Binaural median room acoustic parameters ($IACC_{early}$ and $IACC_{late}$) derived from the input RIRs for the classroom (dashed curve) and the concert hall (solid curve).

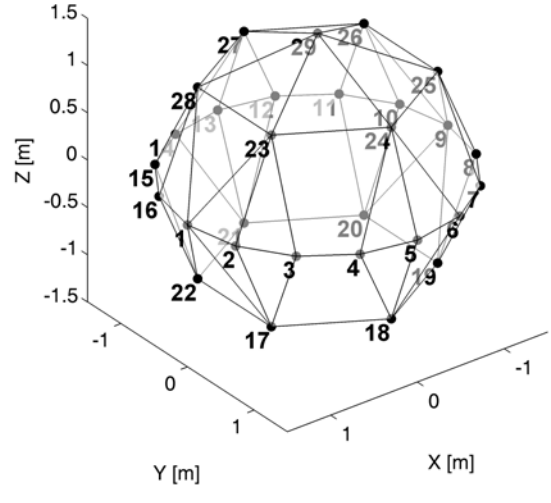

Figure 2.6: Positions of the 29-loudspeaker array used in the objective evaluation.

overall percept of the considered room. Therefore, this range is considered to be relevant to evaluate parameter errors introduced by the LoRA processing. Room acoustics parameters were computed with ODEON for all points of the grid around each of the eight receiver locations in the two rooms. For each room acoustic parameter and for each room, the acceptable range of error was derived as the interval between the minimum and the maximum deviation from the parameter at the center point of the grid for the eight positions. The subjective limen which is frequency-independent corresponds to a very rough estimation of the just noticeable difference (JND) for the different parameters as proposed by Cox *et al.* (1993); Bradley *et al.* (1998) and typically used in round robin studies on room acoustic computer simulations (Vorlander, 1995; Bork, 2000, 2005). Similarly to these studies, the single and the double tolerance of this subjective limen were considered here.

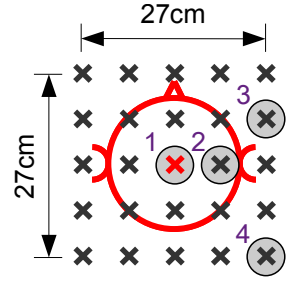

Figure 2.7: Location shifts of the receiver in the virtual room together with the typical size of the human head (20 cm diameter). The range of the room acoustic parameters computed for each point of the grid defines the head movement limen used as a basis for the evaluation of parameter errors. The four gray circles represent the locations inside the loudspeaker array where the parameter error was computed.

2.3.2 Results

Monaural parameters

For each listener position, median and interquartile (error bars) of the monaural parameter errors for the eight source-receiver configurations are shown in Fig. 2.8 for the classroom and in Fig. 2.9 for the concert hall. Median δEDT, δT_{30}, ΔC_{80} and ΔG errors for octave bands from $125\ Hz$ to $8\ kHz$ are plotted together with the head

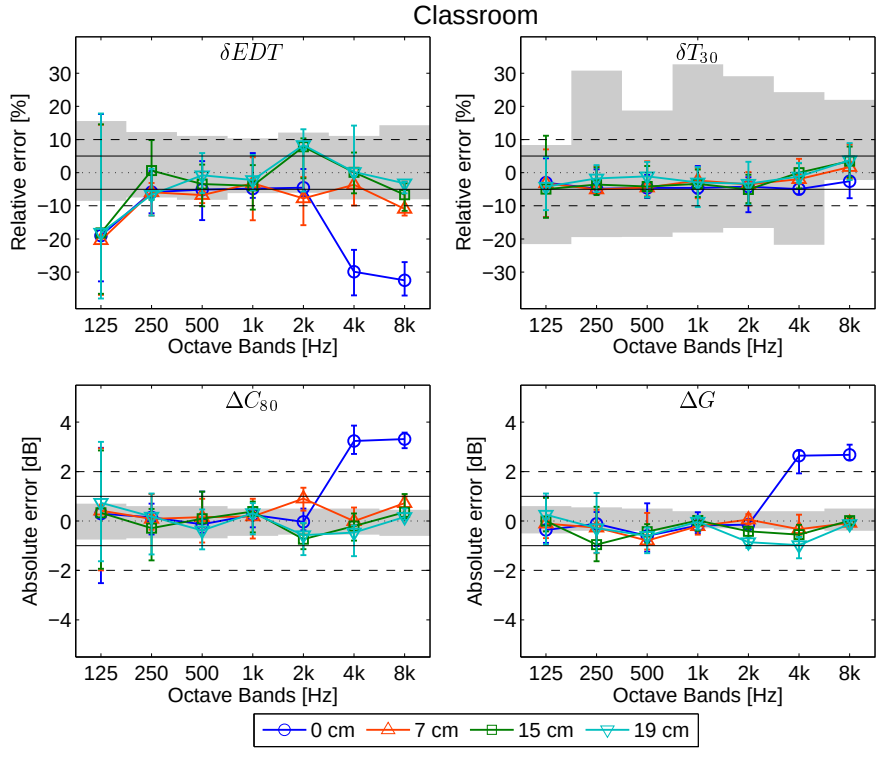

Figure 2.8: Monaural parameter errors (δEDT, δT_{30}, ΔC_{80} and ΔG) median for the classroom as a function of the considered frequency bands for each listener position. The shaded areas represents the head movement limen and the plain and dashed horizontal lines represents the single and double tolerance subjective limen respectively.

movement limen (shaded area) and the single and double tolerance of the subjective limen (plain and dashed horizontal lines). Generally, the non-negligible interquartile values show that the parameter errors fluctuate with source-receiver configurations.

For the classroom, the median monaural parameters errors for the octave bands at 125 Hz to 2 kHz are roughly within the head movement limen and mainly within the single tolerance subjective limen for all listening positions. For higher octave bands (4 and 8 kHz), these parameters are also within the two limens for listening positions outside the center of the loudspeaker array. For the center position (‘0 cm’), δEDT,

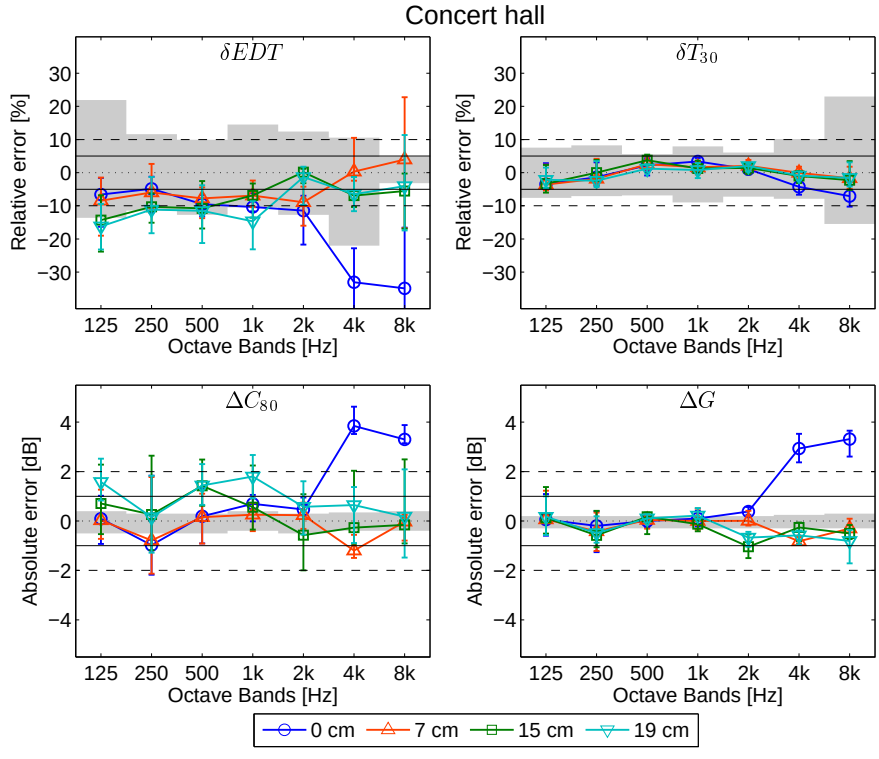

Figure 2.9: Monaural parameter errors median for the concert hall as a function of the considered frequency bands for each listener position. The two types of limens are plotted in the same way as for Fig. 2.8.

ΔC_{80} and ΔG errors are noticeably outside both limens. Similar observations can be made for the concert hall, with a generally larger negative δEDT and a stronger varying error across listening positions for ΔC_{80}.

The large error observed in the center of the array for some parameters is due to the energy normalization at the higher frequency bands for the discrete (early) part of the RIR. For these frequency bands (4 and 8 kHz), the energy normalization provides an increase in energy to compensate for the energy loss at high frequencies due to the imperfect reconstruction of the sound field outside the exact center of the array (see section 2.2.2). Therefore, at the center of the array, the energy increase of the

direct sound and early reflections leads to large positive ΔC_{80} and ΔG errors. In addition, the early decay time (EDT) is computed over the first 10 dB of the energy decay curve which includes mainly the early part of the RIR. Therefore, the energy normalization of the early part of the RIR leads to smaller EDT values as shown by the large negative δEDT errors. The estimation of the reverberation time T_{30} is not significantly affected by the energy normalization of the direct sound and the early reflections, because the T_{30} is mainly determined by the late part of the RIR. It should be emphasized that the large errors at the exact center of the array at high frequencies are not expected to be of any perceptual relevance since the listener will always cover a certain area, his or her ears being separated by at least 18 cm. The small parameter errors for listening locations outside the center indicates that, in practice, the reproduced sound will exhibit temporal and spectral characteristics that are similar to the corresponding room acoustic simulations.

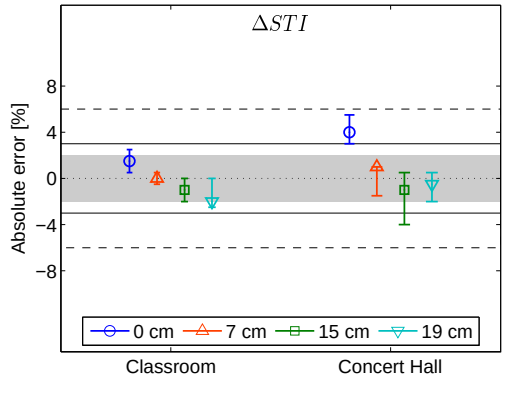

Figure 2.10: Median and interquartile speech transmission index error ΔSTI together with the acceptable error range for the classroom and the concert hall for each listener position. The two types of limens are plotted in the same way as for Fig. 2.8.

The speech transmission index errors ΔSTI are shown in Fig. 2.10. The median errors for positions outside the center of the array are within the head movement and the single tolerance subjective limen. Results at the center of the array show errors within this two limens for the classroom but only within the double tolerance subjec-

tive limen for the concert hall. This larger ΔSTI error for the center position is in line with the above mentioned large errors (δEDT, ΔC_{80} and ΔG) for the higher frequency bands.

In summary, the objective evaluation of the LoRA system showed that the temporal and spectral characteristics of the RIR are not significantly distorted by the LoRA processing, because no significant change is introduced to the room acoustic parameters considered.

Binaural parameter

For each listening position, the median errors and interquartiles of the inter-aural cross correlation coefficients are plotted in Fig. 2.11 where the upper panels represent data for the classroom and the bottom panels for the concert hall. The left panels show the IACC error for the early part of the room impulse response and the right panels show the late part. The splitting time between early and late part was 50 ms for the classroom and 80 ms for the concert hall. Each panel was plotted in the same way as in Fig. 2.8 and Fig. 2.9.

For the classroom, the early $\Delta IACC_{0,50}$ median errors lie roughly within the head movement limen and the single tolerance subjective limen for all considered listening positions. For the concert hall, the same is observaed for the two lowest octave bands at 125 and 250 Hz. For octave bands from 500 Hz to 1 kHz, the $\Delta IACC_{0,80}$ median errors are lower than both limens. This indicates that the LoRA processing increased the diffuseness (lower IACC) of the early part of the RIR at these frequencies. Since this part of the response is dominated by the discrete components (i.e., direct sound and early reflections) the increase in diffuseness might be linked to the spatial spread of energy introduced by the applied Ambisonic auralization (see Fig. 2.1). For the octave bands at 4 and 8 kHz, the $\Delta IACC_{0,80}$ median errors are larger than both limens for the center position and within these limens for all off-center positions. This difference between center and off-center positions is mostly due to the "max r_E" Ambisonic decoding and the corresponding energy normalization at these frequency bands (see section 2.3.2).

For the classroom, late $\Delta IACC_{50,+}$ median errors are within the head movement limen and the single tolerance subjective limen for all listening positions. A similar

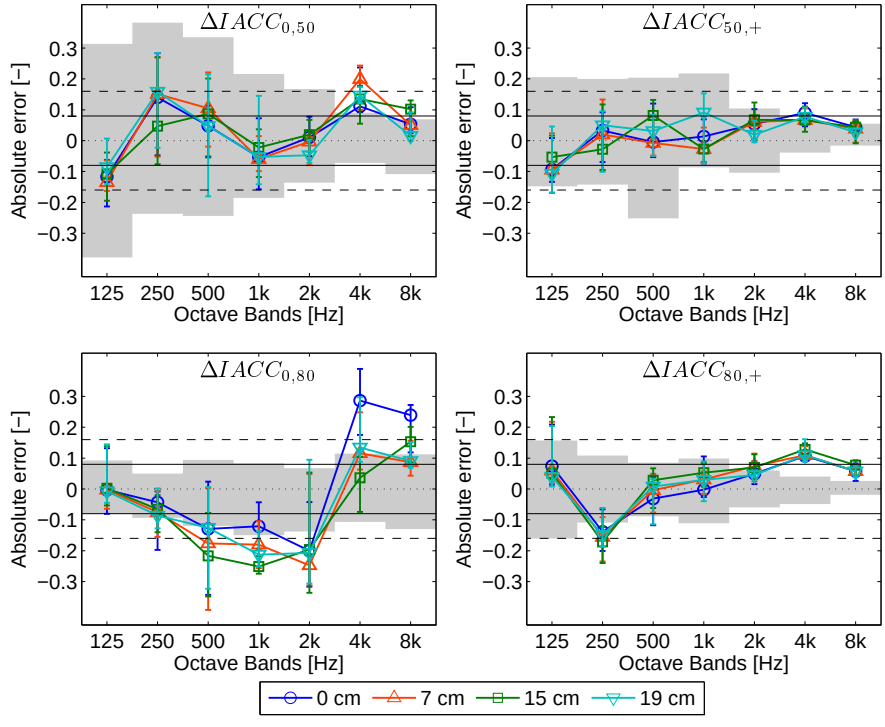

Figure 2.11: IACC errors median for the classroom (top panels) and the concert hall (bottom panels) as a function of the considered frequency bands.

observation is made for the concert hall for octave bands up to 2 kHz. For higher frequency bands, median $\Delta IACC_{80,+}$ errors are above the head movement limen, and interquartile values are very small. This indicates that late IACCs are consistently higher at the output of the LoRA system than at the input where they have rather low (< 0.1) values (see Fig. 2.4, lower right panel). The part of the room response where the late IACCs were computed is dominated by the reverberation process that is implemented in the LoRA system. This reverberation process is realized by modulated noise that is uncorrelated between loudspeakers (see section 2.2.3). Each of these 29 loudspeaker signals leads to cross-talk between signals arriving at the left and right

ear. This induces non-negligible correlations between the overall signals arriving at the left and right ear, even though the 29 loudspeaker signals are uncorrelated. This problem has also been highlighted by Hirst *et al.* (2006) for first order Ambisonic systems.

To analyze the influence of the number of loudspeakers on late IACC at high frequencies, late IACC errors were derived at the center of different three-dimensional loudspeaker arrays, comprising 8, 20, 29, 42 and 92 loudspeakers, and are plotted in Fig. 2.12. For both rooms, when the number of loudspeakers increases, the late IACC error for bands 2-8 kHz decreases from large positive errors (8 loudspeakers) to very small errors (92 loudspeakers). Less loudspeakers are necessary to obtain sufficiently low IACC for the classroom (short reverberation time and absolute IACC above 0.1) than for the concert hall (long reverberation time and absolute IACC below 0.1). These results suggest that the number of 29 loudspeakers provides a good compromise between array size and IACC errors for large rooms.

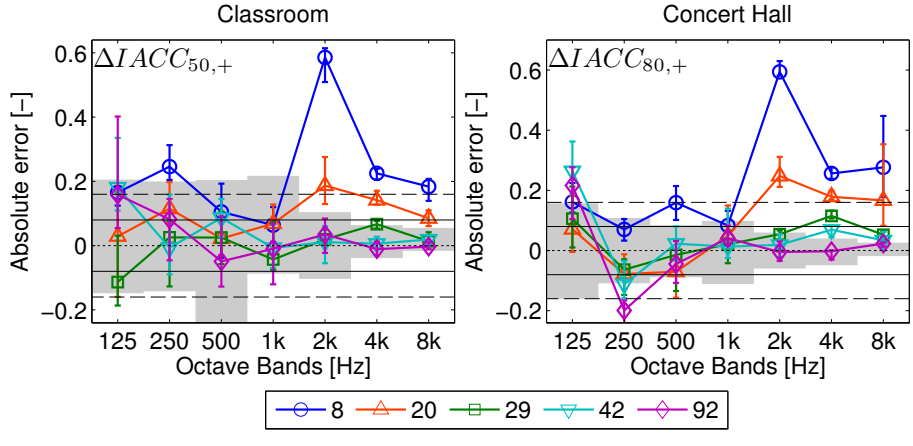

Figure 2.12: Late IACC median errors for the position in the center of arrays with different numbers of loudspeakers (8, 20, 29, 42 and 92, see legend) for the classroom (left panel) and the concert hall (right panel).

The inter-aural cross correlation coefficient analysis indicates that the LoRA system preserves the spatial properties of the room impulse response. However, a rather

large number (at least 29) of loudspeakers is required to reach low late IACCs in the high frequency bands for conventional rooms.

2.4 Summary and conclusions

A novel loudspeaker-based room auralization (LoRA) system has been proposed to generate realistic virtual auditory environments (VAEs) which are, for example, suitable for investigating human auditory perception. This system effectively combines acoustic room models and loudspeaker-based auralization to provide an accurate perception of the simulated acoustic scene. Since auditory perception, as well as the output of the acoustic room models, is very different for the direct sound, early reflections and late reflections, these components are processed independently in the LoRA system.

First, the direct sound dominates auditory localization of reverberant sounds and the room acoustic model provides very detailed information to the LoRA toolbox. Therefore, high-order Ambisonics (HOA) is used for auralizing the direct sound, providing very accurate localization cues. The localization accuracy is thereby limited by the applied Ambisonic order (Bertet *et al.*, 2007), which in turn is limited by the available number of loudspeakers. Second, individual early reflections can hardly be localized by the auditory system. They are nevertheless auralized with the same method as the direct sound since each early reflection is individually described by the room acoustic model as for the direct sound. Finally, auditory localization is very poor for individual late reflections and moreover, a large amount of them is masked (or partially masked) by surrounding reflections as well as the direct sound. Hence, accurate reproduction of the directional characteristics of individual reflection is not needed. Therefore, only energy and intensity envelopes of the late part of the room response have been considered from the room acoustic models and decoded as a simple first-order Ambisonic signal. Nevertheless, late reflections represent the diffuse reverberation of the room which typically contains significant energy and carries important information for distance perception and listener envelopment (Bradley and Soulodre, 1995). Reproduction of diffuse reverberation with a limited number of loudspeaker is challenging as conventional techniques, like Ambisonics, often lead to coloration ef-

fects. In order to realize perceptually convincing reverberation, the LoRA system multiplies the first-order Ambisonic representation of the late-reflection envelopes with noise that is uncorrelated for each loudspeaker signal. The development strategies and realization aspects described above ensured that the LoRA system: (i) delivers highly accurate localization cues for the direct sound, (ii) provides convincing diffuse reverberation, and (iii) minimizes artifacts such as sound coloration. These characteristics are essential for generating auditory environments that provide highly realistic acoustic scenes to the listener.

Objective evaluations of the reproduction of a single source in an anechoic environment with HOA have already been reported (e.g., Daniel, 2000; Ward and Abhayapala, 2001; Poletti, 2005; Solvang, 2008), showing that HOA can deliver highly accurate localization cues. In contrast to these studies, the LoRA system reproduces a sound source (or several of them) in a reverberant environment and thus reproduces very complex sound fields. In consequence, an evaluation of the whole room impulse response was necessary to objectively assess the processing of the overlapping discrete and reverberant (diffuse) components of the RIR. The objective evaluation carried out in the present study consisted of extracting and comparing different room acoustic parameters from the input and the output of the LoRA toolbox when different listening positions were simulated within the loudspeaker array. The derived parameter errors were then used to assess the LoRA processing in terms of temporal and spectral distortion (reverberation time, early-to-late ratio, strength and speech transmission index) and of spatial distortion (inter-aural cross correlation coefficients).

Monaural and binaural parameter errors showed that the parameter variation introduced by the LoRA system for several positions in the sweet spot (7 to 19 cm from the center of the array) were (i) of similar magnitude as variations introduced by small location shifts in the virtual room (head movement limen) and (ii) in the single and double tolerance interval of the subjective limen. In this respect, the evaluation demonstrated that the proposed signal processing preserves the general temporal, spectral, and spatial properties of the room impulse response at the listener's ear positions in the loudspeaker array. Moreover, an additional analysis showed that when a large number of loudspeakers is used (29), very diffuse reverberation (i.e., $IACC_{late} < 0.1$) can be achieved. This means that the LoRA system is able to suc-

cessfully reproduce the required diffuseness of the reverberation part, which is usually problematic in loudspeaker-based auralization systems.

The results of this parameter error study were obtained with an example Ambisonic order of 4 for the reproduction of the early part of the response. Lowering the Ambisonic order of the early part will potentially affect the room acoustic parameter errors at high frequencies, where the energy normalization is applied (Tab. 2.2), and at listeners positions further away from the center of the loudspeaker array. Localization accuracy in anechoic environments has been shown to decrease with the Ambisonic order (Bertet *et al.*, 2007; Frank *et al.*, 2008). It is unclear how these findings in anechoic environments translate to the auralization of reverberant environments realized by the LoRA system. According to the precedence effect the direct sound will dominate overall localization (Zurek, 1987; Blauert, 1997; Litovsky *et al.*, 1999) and the Ambisonics order will have some impact on the localization accuracy of the direct sound. However, the existence of early reflections (and reverberation) will increase the apparent source width and will reduce auditory localization accuracy. In consequence, a lower Ambisonics order might be sufficient for the auralization of rooms than it is required for auralizing anechoic sound sources.

The flexibility and modularity of the LoRA system provides the opportunity to systematically study the influence of the different components of the system (room acoustic model, auralization technique) on the reproduction of the auditory environment both in the physical and perceptual domain. In this paper, the objective evaluation through room acoustic parameters provided the first step towards rigorous evaluation of the LoRA system. In order to assess the accuracy of the system from a perception point of view, a subjective evaluation of the generated VAE is required. Moreover, in order to evaluate the applicability of the LoRA system to study the auditory processing of reverberant sounds, psychoacoustical experiments need to be conducted, considering localization performance, distance perception and speech intelligibility in rooms. All these tests need to be part of a conclusive evaluation of the LoRA toolbox and will be object of future research.

2.5 Acknowledgment

The authors thank Torsten Dau for very useful discussions and comments on an earlier version of this paper. The study was supported by a stipend from the Technical University of Denmark.

3

Validation of a loudspeaker-based room auralization system using speech intelligibility measures †

Abstract

A novel loudspeaker-based room auralization (LoRA) system has been proposed to generate versatile and realistic virtual auditory environments (VAEs) for investigating human auditory perception. This system efficiently combines modern room acoustic models with loudspeaker auralization using either single loudspeaker or high-order Ambisonics (HOA) auralization. The LoRA signal processing of the direct sound and the early reflections was investigated by measuring the speech intelligibility enhancement by early reflections in diffuse background noise. Danish sentences were simulated in a classroom and the direct sound and each early reflection were either auralized with a single loudspeaker, HOA or first-order Ambisonics. Results indicated that (i) absolute intelligibility scores are significantly dependent on the reproduced technique and that (ii) early reflections reproduced with HOA provide a similar benefit on intelligibility as when reproduced with a single loudspeaker. It is concluded that speech intelligibility experiments can be carried out with the LoRA system either with the single loudspeaker or HOA technique.

† This chapter was published as Favrot and Buchholz (2009b). Figures were updated for this thesis.

3.1 Introduction

Recently, a novel loudspeaker-based room auralisation (LoRA) system has been proposed (Favrot and Buchholz, 2010), which aims at generating fully controllable and highly realistic virtual auditory environments (VAEs) that are suitable for investigating human auditory perception as well as assessing and optimizing the performance of modern hearing devices.

The LoRA system effectively combines state-of-the art acoustic room models and loudspeaker-based auralization. Each component of the discrete part of the room's response (i.e., the direct sound and the early reflections) is auralized individually as a single source. Different reproduction techniques can be chosen to auralize these discrete components using either single loudspeakers (the closest from the component's incoming direction) or Ambisonics (first-order; Gerzon, 1973 or high-order, HOA; Daniel, 2000). The reproduction of the (late) diffuse reverberation part is realized by multiplying (directional) intensity envelopes of the room's response with noise that is uncorrelated across loudspeakers.

When a large number of loudspeakers is available, then the single loudspeaker technique is the most accurate one for reproducing the direct sound and the early reflections. Ambisonics allows the reproduction of sound events from any direction, whereby the localization accuracy depends on the applied Ambisonic order which itself depends on the available number of loudspeakers (Daniel, 2000). Ambisonics reproduces the sound field accurately inside a sweet spot (e.g., in the center of a loudspeaker array) up to a certain frequency f_{max}. This frequency f_{max} and the size of the sweet spot are determined by the applied Ambisonic order (e.g., $f_{max} = 2.2$ kHz for forth-order Ambisonics within a sweet spot of 20 cm diameter). Apart from reproducing simulated reflections, this technique can be used to reproduce a whole auditory scene which has been recorded either with a standard soundfield microphone (first-order Ambisonics; Gerzon, 1973) or with a more advanced microphone array (higher-order Ambisonics, HOA; Moreau *et al.*, 2006). Ambisonics is also suitable for reproducing moving sound sources as it allows for smooth source direction changes. Hence, although the single loudspeaker technique might in principle provide the best

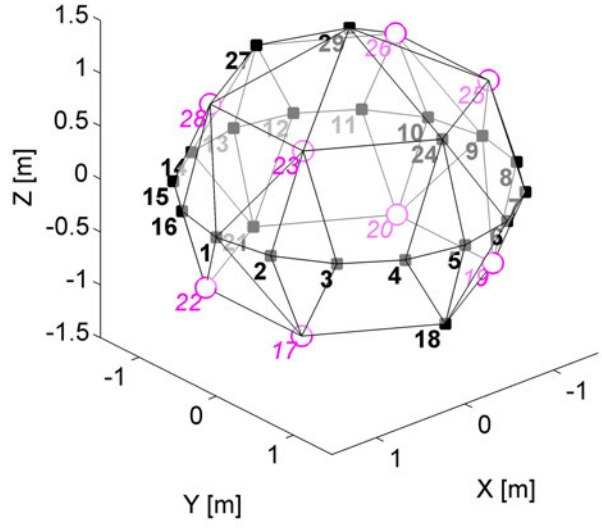

Figure 3.1: Loudspeaker positions of the array used in the experiments. The large open circles indicate the loudspeakers used for first-order Ambisonics auralization. In the single loudspeaker auralization condition, the direct sound was played with loudspeaker '1' only.

overall quality, Ambisonics represents a more versatile method and might be superior when only a limited number of loudspeakers are available.

An objective evaluation using different room acoustic measures showed that the spectral, temporal and spatial aspects of the room's response are preserved by the LoRA processing (Favrot and Buchholz, 2010). Since the LoRA system is primarily designed for auditory perception research, a subjective evaluation of the system is needed. This study presents a speech intelligibility measure-based subjective evaluation of the auralization techniques used in the LoRA system. The specific aim was to assess the influence of single loudspeaker, first-order Ambisonics and HOA auralization of the direct sound and early reflections on speech intelligibility scores. The results allow conclusions on the applicability of (higher-order) Ambisonics in VAEs (such as the LoRA system) for speech perception research and moreover, have direct implications for simulating moving speech sources as well as presenting speech signals recorded in real environments.

Shirley *et al.* (2007) has shown that speech intelligibility measures can produce significantly different results when speech is presented by a single loudspeaker than when mixed via a stereo loudspeaker pair. Results might expected to be different here from those of Shirley *et al.* (2007) because (i) first-order Ambisonic and especially high-order Ambisonic are expected to provide a more accurate reproduction of the sound field of a single sound source than stereo mixing and (ii) listeners sensitivity to the direct sound image is limited here due to the presence of (simulated) room reflections.

3.2 Method

The influence of auralization technique on speech intelligibility in virtual environments was investigated by applying a method inspired from Bradley *et al.* (2003). In order to separately evaluate the effect of the auralization technique on the direct sound and early reflections, speech intelligibility scores were measured as a function of direct sound level and early reflection level.

The experiment took place in an acoustically damped room where a 3-D array of 29 loudspeakers was used (see Fig. 3.1). Nine normal hearing persons participated in the experiment.

3.2.1 Stimuli

Room simulation

A classroom was simulated with ODEON (see Fig. 3.2), a room acoustic modeling software, where a source ('talker') and a receiver ('listener') position were defined. A reflectogram (i.e., the direction, latency and attenuation of the direct sound (DS) and early reflections (ER), see Fig. 3.3) was obtained for this source-receiver configuration.

The direction of each individual component of the reflectogram was manipulated to match the direction of the closest loudspeaker present in the listening room (see Fig. 3.1). This was done to ensure that the single loudspeaker and Ambisonic auralization

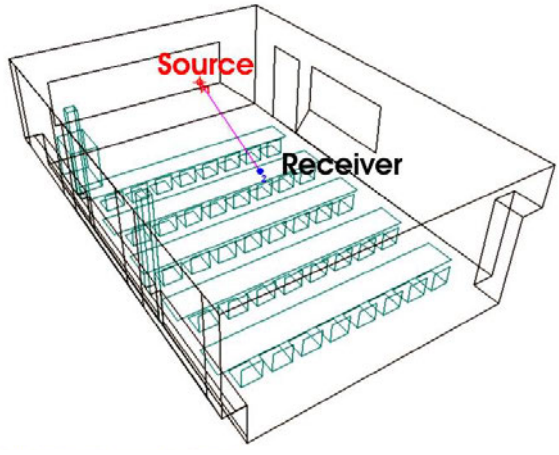

Figure 3.2: Position of the source and the receiver in the 3D model of a classroom (40 seats, 170 m^3). The listener was located at a distance of 3.5 m from the talker.

represent reflections from the same direction. The 'talker' was facing the 'source' such that the direct sound came from the frontal direction (0° azimuth, 0° elevation).

The reflectogram data were then processed separately for the direct sound and the early reflections with the LoRA toolbox. Each discrete component was treated as a source being reproduced either with a single loudspeaker (labeled as '0') , with fourth-order Ambisonics (labeled as '4') or with standard first-order Ambisonics (labeled as '1'). For the Ambisonic reproduction, a 'basic' decoding scheme was used for frequencies up to 2.8 kHz for fourth-order Ambisonics and up to 707 Hz for first-order Ambisonics. Above this frequency, the 'max r_E' decoding method was used in order to focus the energy in the expected direction (Daniel, 2000). Only 8 loudspeakers, indicated with large open circles in Fig. 3.1, were used for first-order Ambisonic auralization in order to limit coloration effects.

Since speech intelligibility in diffuse noise was measured here, the late reverberation part of the room's response was not reproduced. The presence of late reverberation is expected to deteriorate speech intelligibility, however the level of the late reverberation in the classroom was low relatively to the level of the diffuse noise.

For the source-receiver configuration in the considered room, the obtained mul-

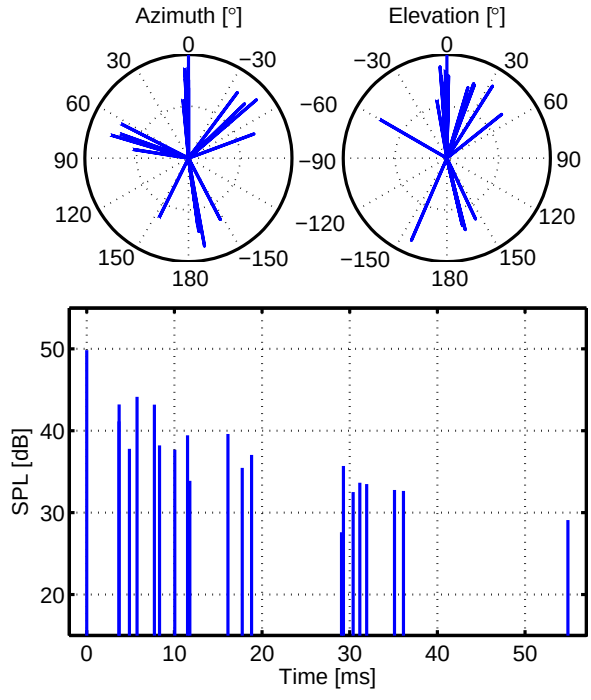

Figure 3.3: Reflectogram for the source and receiver configuration in the classroom showed in Fig. 3.2). The directional characteristic (azimuth and elevation) is indicated in the polar plots and the temporal behavior is shown in the lower panel.

tichannel room impulse response (mRIR) was about 55 ms long and the level of the total mRIR (DS+ER) was 5.7 dB higher than the level of the direct sound (DS) alone.

The level of the direct sound and/or of the early reflections in the mRIR was varied to obtained different signal-to-noise ratios (SNRs).

Speech corpus and speech-shaped noise

The speech corpus consisted of the Dantale II Danish Hagerman sentences (Wagener *et al.*, 2003); each sentence containing five words following the structure: 'Name' + 'Verb' + 'Number' + 'Adjective' + 'Noun'. The sentences were auralized by convolving them with the classroom mRIR reproduced with different techniques.

Diffuse speech-shaped noise was obtained from cutting the monaural speech-shaped noise track from the Dantale II material in 29 noise signals. These uncorrelated noise signals were played simultaneously via the 29 loudspeakers. The diffuse noise started 1 s before the sentence with a cosine-shaped ramp of 0.6 s and stopped 0.5 s after the sentence with a decreasing cosine-shaped ramp of 0.3 s. Diffuse noise was played at a fixed level of 60 dB SPL.

Calibration

Loudspeaker equalization was performed on the loudspeaker array to flatten each of the loudspeaker frequency responses recorded at the center of the array (cf. Appendix B). In order to calibrate the system, speech-shaped noise convolved with the DS-only and the ER-only for the different conditions were recorded with an omnidirectional microphone at different positions within the sweet spot (7.5 cm around the center of the loudspeaker array).

The SPL of the recordings for the DS-only auralization with the three different techniques showed discrepancies with the simulated level ranging from -1 dB to +2.5 dB. A level adjustment was then performed to ensure identical SPL independent of the direct sound reproduction technique. A similar adjustment was realized for the ER-only responses which initially showed discrepancies from 2.2 dB to 3.6 dB. When the whole mRIRs (with the adjusted DS and ER levels) were used, the obtained SPLs varied only from 0 to +1.5 dB compared to the simulated SPLs. Different dummy-head recordings in the sweet spot as well as slightly off-center confirmed the applicability of the level-adjustment method (see Fig. 3.4 for the sweet spot recordings).

Discrimination task

A preliminary test was carried out to determine if listeners could discriminate short sentences presented with the different auralization techniques. DS-only and DS+ER mRIRs auralized with the three previously mentioned techniques were used for this test. Diffuse noise was not played simultaneously during the short sentences. Results showed that all techniques could be clearly discriminated. This was a prerequisite for

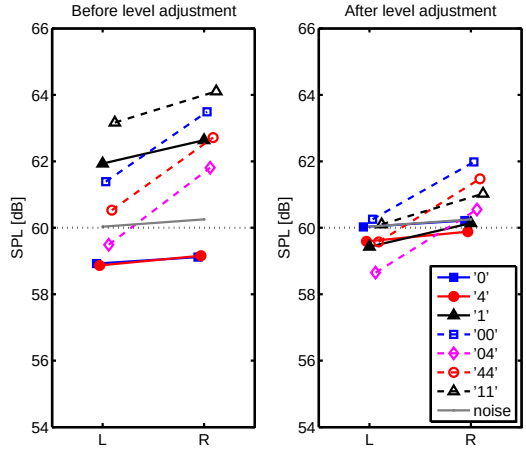

Figure 3.4: Binaural levels before and after level adjustment for the seven conditions. Each condition was auralized at 60 dB SPL. The measured SPL of the diffuse speech-shaped noise is also plotted.

the speech intelligibility experiment as non discriminable stimuli would have provided similar intelligibility scores.

3.2.2 Procedure

First, the speech reception threshold (SRT) was derived with an adaptive method where the whole mRIR (DS+ER) level was varied relative to the fixed level of the diffuse speech-shaped noise (60 dB SPL). The SRT corresponded to the signal-to-noise ratio (SNR) at 50 % intelligibility. The DS reference level was defined as the level of the direct sound in the whole mRIR at the SRT level. The mRIR used in this part of the experiment was obtained by the single loudspeaker technique for both the DS and the ER (condition '00'). The SRT was measured for each subject with two lists of 10 sentences.

Second, intelligibility scores were measured for fixed SNRs at the DS level at SRT +2, +4, +6, and +8 dB by using only the direct sound and varying its level. Measures were taken for the direct sound auralized by (i) a single loudspeaker, (ii) forth-order HOA and (iii) first-order Ambisonics (conditions '0', '4' and '1' respectively).

Third, the direct sound level was kept at the previously derived DS reference level

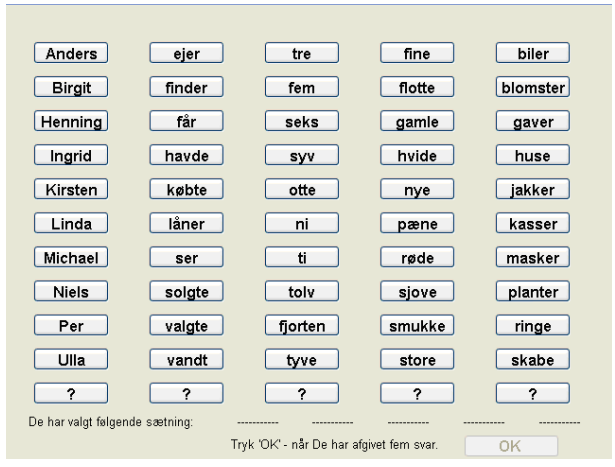

Figure 3.5: Touch screen caption for the sentence selection.

and early reflections were added to the direct sound at level +2, +4, +6, and +8 dB. Three conditions were measured for this part of this experiment: the same technique was used for the DS and ER (conditions '00', '44' and '11'). For the second and third part of this experiment, a list of 10 sentences was used to determined the intelligibility scores.

In each part of the experiment, after each sentence was played with diffuse noise, subjects were asked to select the five words s/he had heard on a touch screen (see Fig. 3.5). There were ten possible choices for each word plus a "I don't know" (labeled as "?") one which, when selected, randomly choose a word from the list of 10 words. The "I don't know" button was present to not force the subject to randomly pick a word if it was not heard. The selection screen was shown only after the sentence was played. After the selection, the subject pressed the "OK" button to play the next sentence.

Each test-subject participated in a training phase which consisted of 4 repetitions of the first part of the experiment (80 sentences). After the first part of the experiment, intelligibility scores for the different SNRs and conditions were measured in a random order. The entire experiment was carried out for each subject in two sessions of one hour and a half, including breaks.

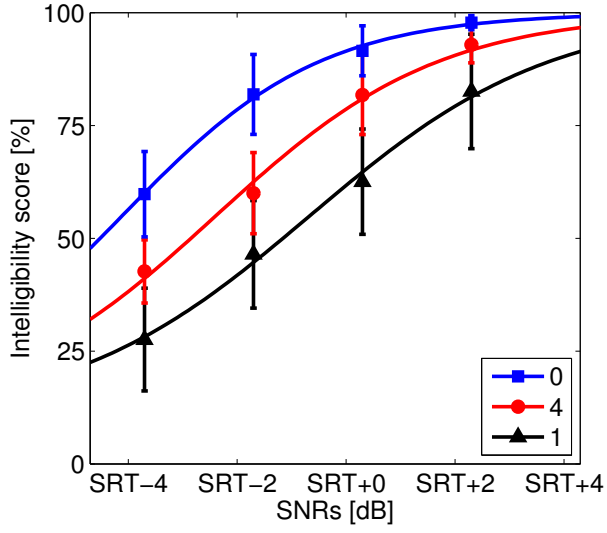

Figure 3.6: Direct sound only mean speech intelligibility scores (markers) with $\pm$ 1 standard deviation. Solid lines represent the fitted sigmoid function.

3.3 Results

In the first part of the experiment, speech reception thresholds were measured between -13 and -10.3 dB for the nine test-subjects. These values are substantially lower than the -8.4 dB normally observed for normal hearing subjects with the classic Dantale II test. This difference might be explained by the playback method, i.e., using loudspeakers instead of headphones, and the experimental procedures, i.e., applying a user interface with a restricted choice of words rather than the subject telling the operator what s/he had heard.

Fig. 3.6 and Fig. 3.7 show the inter-subject mean intelligibility scores for the DS-only conditions and the DS+ER conditions respectively. The corresponding standard deviations are indicated by error-bars. Results from the nine test-subjects were analyzed for these two groups of conditions. To assess the statistical significance of the comparison of any two conditions, a paired *t*-test was performed.

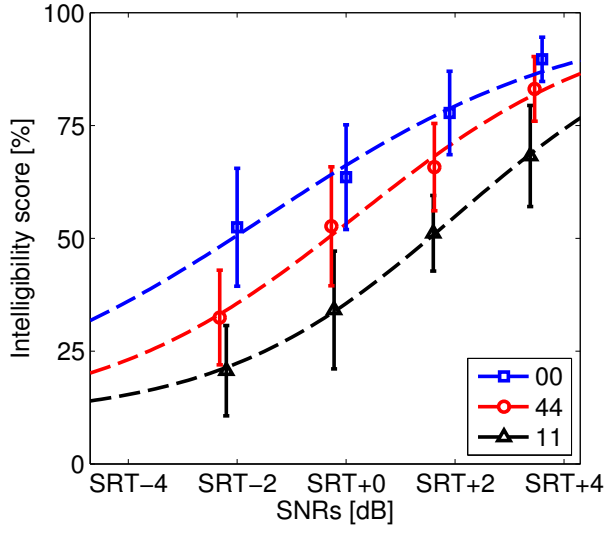

Figure 3.7: Direct sound and early reflections mean speech intelligibility scores (markers) with ± 1 standard deviation. Dashed lines represent the fitted sigmoid function.

3.3.1 Modeling the data

As expected, mean intelligibility scores plotted over SNR exhibited a psychometric function. In order to quantify the measured psychometric functions, a sigmoid function $P(L)$ was fitted to the data with the parameter L_{55} (threshold, SNR L at 55 % percent correct) and s_{55} (slope at the inflection point, $L = L_{55}$) for each condition according to the following formula (Brand and Kollmeier, 2002):

$$P(L) = \frac{1-\alpha}{1+exp(4 \cdot s_{55} \cdot (L_{55}+L))} + \alpha. \tag{3.1}$$

The chance level α was here 10 % since there were 10 possible answers.

The fitted parameters can be found in Table 3.1 and the corresponding sigmoid functions were plotted in Fig. 3.6 (solid lines) and Fig. 3.7 (dotted lines). The sigmoid functions fitted the measured data with an RMS error smaller than 2 %. Theses parameters characterize the significant increase of intelligibility score with increasing SNR and were analyzed for all the considered conditions.

Conditions	0	4	1	00	44	11
L_{55} [dB]	-4.1	-2.4	-0.7	-1.5	0.2	2.0
s_{55} [%/dB]	13.7	12.2	11.3	8.8	10.6	11.4
error [%]	0.7	1.7	1.6	2.2	1.7	0.5

Table 3.1: Fitted parameters of the sigmoid function for each group. L_{55} are relative to the SRT.

3.3.2 Effect of the reproduction technique for the direct sound only conditions

For the direct sound only conditions (see Fig. 3.6 and Table 3.1), the intelligibility scores are significantly dependent on the auralization method. Highest scores are observed for the single loudspeaker technique and lowest scores for first order Ambisonics. The difference is mainly due to a shift in the L_{55} threshold, indicating a shift in effective SNR: a threshold shift of 1.7 dB is observed from single loudspeaker ('0') to forth-order HOA ('4') and a shift of 3.4 dB from single loudspeaker to first-order Ambisonics ('1').

Moreover, the psychometric functions for the '0' condition are steeper at the inflection point than for the Ambisonic ones: the s_{55} slope decreases by 1.5 %/dB for HOA and by 2.4 %/dB for the first-order Ambisonics one. These small variations indicate that, with Ambisonics (especially at first-order) slightly larger SNRs than with the single loudspeaker technique are required to increase intelligibility scores.

3.3.3 Effect of the auralization technique on the reproduction of the whole response

For the DS+ER conditions (Fig. 3.7 and Table 3.1), the intelligibility scores show a similar significant dependency on auralization technique as observed for the direct sound only conditions (section 3.3.2). For HOA auralization a L_{55} threshold shift of 1.6 dB is needed to obtain similar intelligibility scores than with the single loudspeaker technique. For first-order Ambisonic a 3.5 dB threshold shift is required to match the scores obtained with the single loudspeaker auralization.

The slope of the psychometric functions are slightly steeper with Ambisonics than

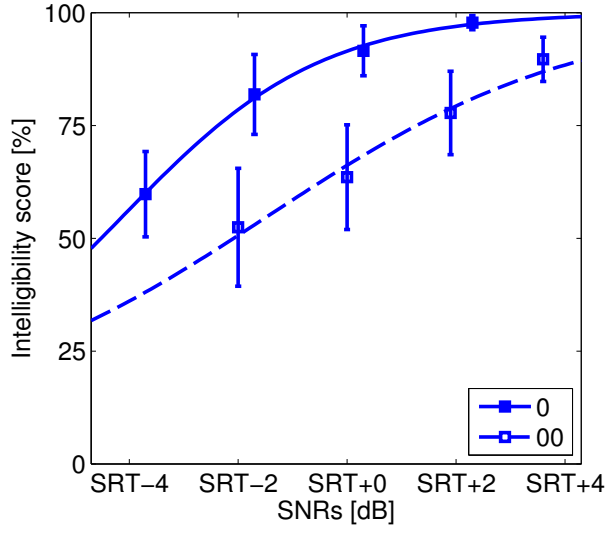

Figure 3.8: Mean speech intelligibility scores (markers) and psychometric curves for DS-only and DS+ER conditions auralized with single loudspeaker.

with the single loudspeaker technique. These slope variations with the reproduction technique are the opposite of the ones observed for the DS-only conditions. However, these variations are rather small and might thus be disregarded.

3.3.4 Effect of the addition of early reflections

Although overall intelligibility is modified by the playback technique, the addition of reflections has a similar effect for all playback techniques.

Comparing the intelligibility scores measured in the DS-only condition (Fig. 3.6) and DS+ER conditions (Fig. 3.7), it can be observed that the increase in intelligibility is larger when adding direct sound energy than when adding reflection energy. Accordingly, the $s55$ slopes of the psychometric functions (Table 3.1) are shallower for the DS+ER conditions than for the DS-only conditions. The effect is highlighted in Fig. 3.8 for the single loudspeaker technique where the experimental data is replotted for condition '0' and condition '00'. The decrease in s_{55} slope is slightly more pro-

nounced for the single loudspeaker technique (-4.9 %/dB) than for HOA (-1.6 %/dB) and first-order Ambisonics (+0.2 %/dB).

The L_{55} thresholds increase for DS+ER conditions compared to DS-only conditions with slightly variation for different reproduction technique. A threshold increase of 2.7 dB was observed for the single loudspeaker technique, 2.6 dB for HOA and 2.7 dB for first-order Ambisonics (see Table 3.1).

3.4 Discussion

The fitted psychometric functions showed that in the present experiment the absolute intelligibility scores were dependent on the auralization technique for both the DS-only and DS+ER conditions. The dummy-head recording levels for the DS-only conditions were 0.3 dB smaller for the two Ambisonics techniques compared to the single loudspeaker auralization (see Fig. 3.4). Since thresholds shifts of the psychometric functions for the different auralization techniques were in the order of 1.7-3.4 dB (Fig. 3.6 and Table 3.1), they can not be solely explained by the DS level calibration. For the DS+ER conditions, the dummy-head recording levels showed a decrease of about 0.5 dB also failing to account for the shifts of the psychometric function of 1.6-3.5 dB (see Fig. 3.7).

The decrease in intelligibility was probably due to the imperfect sound field reconstruction with Ambisonics. This typically leads to degraded spatial cues and coloration due to comb-filtering effects. These effects are even more pronounced when the Ambisonic order is low, which is here reflected by the lower scores for first-order Ambisonics than for forth-order Ambisonics. These detrimental effects for an anechoic source (DS-only) and for a source in a reverberant environment (DS+ER) did also lower the intelligibility of forth-order Ambisonics compared to the single loudspeaker technique. This observation is in line with Shirley *et al.* (2007) where a degraded speech intelligibility was measured when using a stereo phantom image compared to a single loudspeaker due to the cross-talk caused by the phantom image. With Ambisonics, the cross-talk can be described with the energy ratio or spatial spread r_E (Daniel, 2000) which is increasing with Ambisonics order but does not reach the value of 1 which would be obtained for the reference (single loudspeaker).

The SNR increase by raising direct sound level versus addition of early reflections led to a decrease in the slope of the psychometric function as well as a threshold increase for all the considered reproduction techniques. This can be interpreted as the consequence of a temporal and spatial spread of energy when early reflections are added. The benefit of early reflections in speech intelligibility has been pointed out in numerous studies (Lochner and Burger, 1964; Nábělek and Robinette, 1978). However, in the study by Bradley *et al.* (2003), the benefit of the early reflections was of the same order as for an increase of the direct sound level only. This finding was not observed here and might be due to a difference in the employed speech test as well as the considered stimuli.

3.5 Conclusion

This study investigated the impact of the auralization method on speech intelligibility. It was found that the overall intelligibility threshold (SNR at 55 % word correct) was lower when using a single loudspeaker technique than when using Ambisonics. This threshold also increased when the Ambisonic order decreased. Moreover, the addition of early reflections increased the intelligibility to a lesser extent than when the direct sound alone was raised, resulting in psychometric functions with shallower slopes. However, the addition of early reflections induced a similar threshold shift for all the considered reproduction techniques.

It can be concluded that speech intelligibility experiments can be run with the LoRA system with either the single loudspeaker or HOA technique as the reproduced early reflections provide the same benefit on intelligibility. However, intelligibility scores need to be equalized for the individual auralization method by a simple SNR shift. This encourages the use of HOA microphone arrays to record complex auditory scenes for speech intelligibility experiments. However, further evaluation is required to investigate the effect of the physical properties of such microphone arrays on the captured auditory scene.

4

Distance perception in a loudspeaker-based room auralization system [‡]

Abstract

A loudspeaker-based room auralization (LoRA) system has been recently proposed which efficiently combines modern room acoustic modeling techniques with high-order Ambisonics (HOA) auralization to generate virtual auditory environments (VAEs). The reproduction of the distance of sound events in such VAE is very important for its fidelity. A direct-scaling distance perception experiment was conducted to evaluate the LoRA system including the use of near-field control (NFC) for HOA. Experimental results showed that (i) loudspeaker-based auralization in the LoRA system provides similar distance perception to that of the corresponding real environment and that (ii) NFC-HOA provides a significant increase in the range of perceived distances for near sound sources as compared to standard HOA.

4.1 Introduction

Virtual auditory environments (VAEs) have received increased attention in the past decades and are used in a variety of domains from entertainment to psychophysical research. VAEs ideally create the percept of acoustic events that are generated by sound sources in a given space, whereas neither the sound sources nor the space are physically present in the actual listening environment (Blauert, 1997). In order to dis-

[‡] This chapter was published as Favrot and Buchholz (2009a).

play sound sources at specific positions in the virtual space, understanding of human auditory localization is required. Whereas the perception of the direction of the source has been widely studied (Blauert, 1997), much less is known about auditory distance perception. Besides the fact that reproduction of the distance of sound sources in a VAE is very important for its fidelity, VAEs provide a tool to systematically study the influence of individual distance cues (e.g., Zahorik, 2002a) and thus to better understand the mechanisms underlying distance perception.

A change of the physical distance between a listener and a sound source induces the change of four main characteristics of the acoustical waveforms at the listener's ears: intensity, direct-to-reverberant energy ratio, binaural differences (for sources in the acoustic near-field) and spectrum (Zahorik, 2002a). The intensity and the spectrum are ambiguous cues since their variation can also result from a change of the sound source characteristics (e.g., the input signal to a loudspeaker source), whereas the direct-to-reverberant ratio and the binaural differences are independent of the source characteristics. This implies that a realistic reproduction of the direct-to-reverberant ratio and the binaural differences (for near-field sources) potentially allows for an independent control of the distance and intensity of a source in a VAE. This is of particular importance in applications such as auditory displays where the distance can code the importance of an event.

Strategies for accurately reproducing these cues in a VAE are closely linked to the playback technique: either binaural or loudspeaker-based. This study focuses on the ability of loudspeaker-based systems to provide realistic distance perception for both near- and far-field sources in various acoustic environments. The loudspeaker-based room auralization (LoRA) system (Favrot and Buchholz, 2010) is here taken as an example to demonstrate that realistic distance perception can be achieved in loudspeaker-based VAEs. The LoRA system efficiently combines modern room acoustic modeling techniques with high-order Ambisonics (HOA) auralization.

The reproduction of the different distance cues to listeners in VAEs is of various degree of complexity. While the intensity cue and the spectrum cues are merely controlled by the overall playback gain in different frequency channels, the reproduction of natural (i) direct-to-reverberant ratio cue and (ii) binaural differences cue for near-field sources with loudspeakers requires more technical effort.

Accuracy of the provided direct-to-reverberant ratio cues in the VAE are determined by the precision of the different components in the overall chain of the LoRA system, i.e. the acoustic room model providing the room impulse response (RIR), the HOA auralization of the early part of the (RIR) and the reverberation algorithm for the late part of the RIR. The goal of the first part of this study is to assess the quality or realism of this cue provided by VAEs such as the LoRA system. For this purpose, subjective distance estimations in VAEs need to be compared to estimations in corresponding real environments. Since distance perception is strongly influenced by visual cues, an appropriate auditory-only reference condition needs to be defined. Individualized binaural room impulse responses (BRIRs) have been previously used for studying distance perception in real environments (Zahorik, 2002a) and seem to provide a suitable auditory-only reference condition. However, this technique requires a substantial amount of time and it has been shown that auditory distance perception is not significantly modified by the use of non-individualized BRIRs in reverberant rooms for distances between 0.3 and 13.8 m (Zahorik, 2002b). Thus, dummy-head recorded BRIRs provided the reference condition for the present distance perception experiments.

Binaural differences are an important cue for distance perception of nearby sources which, unlike binaural techniques, are challenging to render with a loudspeaker array as signals at the listener ears are not precisely controlled. Ambisonics is a commonly used spatialization technique which primarily reproduces far-field sources (i.e. producing plane waves). The near-field compensation (NFC) method (Daniel, 2003) has been proposed to reproduce spherical wave sound fields produced by near-field sources, typically closer than the loudspeaker array radius, and has been evaluated physically. The second part of this study provides a perceptual evaluation of the distance of near-field sources rendered with the NFC-HOA technique in both anechoic and reverberant environments in the framework of the LoRA system. Whereas the reverberant condition reflects the most common case in VAEs, the anechoic condition is expected to provide the most sensitive case for studying the influence of the NFC technique.

4.2 Methods

In this study, the perception of distance was assessed by using a direct+scaling method in environments restricted to the auditory modality. Both the fidelity of distance perception in loudspeaker-generated VAEs and the effect of the NFC-HOA method on the reproduction of near-field sources were investigated.

The experiments took place in an acoustically damped room where a spherical 3D array of 29 loudspeakers with a radius of 1.8 m was used. Seven normal-hearing persons participated in the experiment.

4.2.1 Stimuli

Acoustic environments

Three rooms were considered in this experiment: a classroom, a large auditorium and an anechoic room (Fig. 4.1). In each room, a set of receivers positioned at various distances (see Tab. 4.1) from an omni-directional source was selected. The tested conditions are listed in Tab. 4.2. Conditions C1 to C8 refer to the investigation of distance perception in LoRA generated VAEs and conditions C9 to C15 to the NFC investigation for which the range of distances was restricted from 0.6 to 4.3 m.

Rooms	Distances [m]					
Classroom	0.6	0.8	1.3	2.3	4.3	6.3
Auditorium	0.5	1.0	2.2	4.0	7.5	11.2
Anechoic	0.6	0.8	1.3	2.3	4.3	6.3

Table 4.1: Table of the source-receiver distances used in the experiment.

Binaural recordings for the reference conditions

For conditions C5 to C8 and C15, binaural room impulse responses (BRIRs) were recorded with a head and torso simulator (HATS) from Brüel & Kjær at the receiver positions defined in Fig. 4.1 in the actual physical rooms. BRIRs have also been recorded in an anechoic room (C15) for comparison to the anechoic NFC-HOA con-

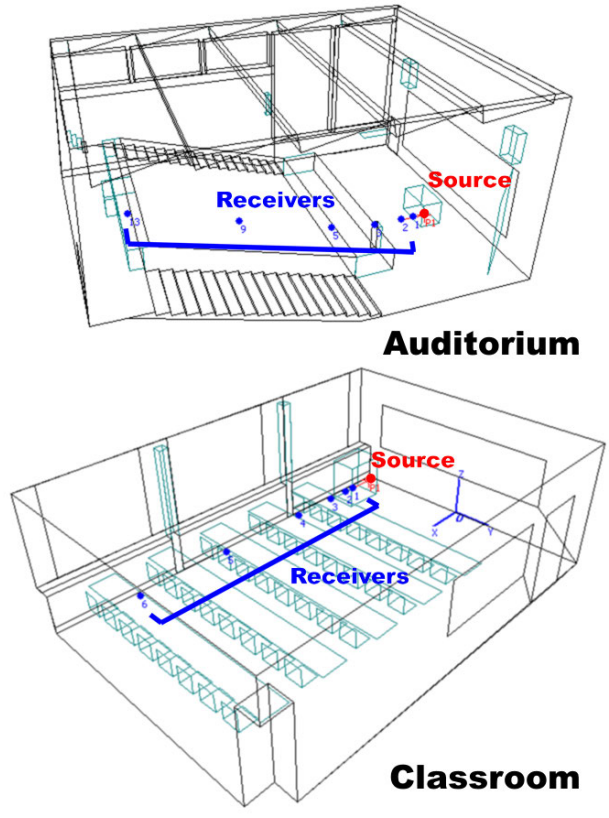

Figure 4.1: 3D models of the rooms and positions of the sources and receivers used in the experiment.

ditions. The non-individualized BRIRs measured in the anechoic room are known to greatly degrade distance perception and typically result in in-the-head-localization (Blauert, 1997). However, this anechoic condition was included to allow for a comparison between non-individual tailored binaural and loudspeaker-based systems. The source was placed in the $90°$ azimuth direction of the receiver to provide significant binaural differences. The same condition at $0°$ azimuth was not tested because a preliminary test showed that the stimuli at different distances were not discriminable.

	Auralization technique	Room	Az.	Process
C1	LoRA	Class.	0°	
C2	LoRA	Class.	0°	IN
C3	LoRA	Aud.	0°	
C4	LoRA	Aud.	0°	IN
C5	Binaural	Class.	0°	
C6	Binaural	Class.	0°	IN
C7	Binaural	Aud.	0°	
C8	Binaural	Aud.	0°	IN
C9	LoRA	Class.	0°	NFC+IN
C10	LoRA	Class.	90°	NFC+IN
C11	HOA 4^{th} 3D	Anec.	0°	NFC+IN
C12	HOA 4^{th} 3D	Anec.	90°	NFC+IN
C13	HOA 7^{th} 2D	Anec.	0°	NFC+IN
C14	HOA 7^{th} 2D	Anec.	90°	NFC+IN
C15	Binaural	Anec.	90°	IN

Table 4.2: Test conditions are listed with the auralization technique, the room (classroom, auditorium or anechoic), the azimuth of the direction of the source (Az.) and with either no processing, intensity normalization (IN) or NFC and intensity normalization (NFC+IN).

Room simulation

For condition C1 to C4, the classroom and the auditorium were simulated with ODEON (Fig. 4.1), a room acoustic modeling software, where the source and receiver positions where set according to Fig. 4.1. For each source-receiver configuration, room impulse responses (RIRs) were derived from the ODEON software. These RIRs were processed by the LoRA system (Favrot and Buchholz, 2010). They were first decomposed into an early discrete part (i.e., the direct sound and the early reflections) and a late diffuse part. For these two rooms, the discrete part of the room's response was auralized with fourth-order 3D HOA and the diffuse reverberation part was auralized by multiplying intensity envelopes of the room's response with noise

that is uncorrelated across loudspeakers. A set of multi-channel RIRs (mRIRs) was thus obtained for the classroom and the large auditorium.

Near-field control with high-order Ambisonics

For conditions C9 to C12, another set of mRIRs was derived by integrating near-field control (NFC) filters to HOA (Daniel, 2003). This technique consists of filtering the Ambisonic encoded components according to the distance of the source (ρ) and the radius of the loudspeaker array (R) to represent finite distance sources with HOA. The formula describing these filters, $H_m^{NFC(\rho,R)}(\omega)$, is shown in equation (10) in Daniel (2003) and their frequency responses are shown in Fig. 4.2 for each Ambisonic component m equals 0 to 7 for a source of 0.5 m. These filters result in a very large bass-boost when reproducing very close sources. A Thikonov regularization, inspired by Moreau *et al.* (2006), was used to limit the amplification at 30dB. Regularized filters were described by:

$$H_{reg} = \frac{H_m^{NFC(\rho,R)} \cdot \lambda^2}{{H_m^{NFC(\rho,R)}}^2 + \lambda^2} \tag{4.1}$$

with $\lambda = 10^{30/20}$ and are plotted (dashed curved) in Fig. 4.2.

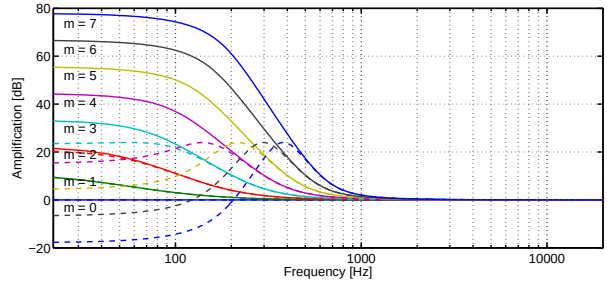

Figure 4.2: Frequency response of the near-field coding (NFC) and the regularized NFC (dashed curves) filters.

An anechoic room was simulated by selecting the direct sound alone from the classroom RIRs for each source-receiver configurations. The anechoic room and the

classroom RIRs were auralized with both fourth-order 3D and seventh-order 2D Ambisonics with NFC filters for an incoming direction of 0 and 90 degree azimuth.

Calibration

The mRIRs were obtained for all rooms with a source sound power set at 72 dB re. 1W in ODEON (corresponding to 61 dB SPL at 1 m in free field). For each room, the recorded BRIRs where adjusted such that for a source-receiver distance of 2.3 m, the computed SPL in the 500 Hz octave band derived in ODEON matches the SPL in the 500 Hz octave band of the BRIR.

Intensity normalization was performed for the conditions specified in Tab. 4.2 and relied on the SPL level in the 500 Hz octave band where the maximum energy of the speech signals used in this experiment was found. For each room, the normalization gains for the mRIRs were obtained by subtracting the 500 Hz octave band SPL of the RIRs derived by ODEON for the 2.3 m configuration to the one for the considered distance configuration. Normalization gains for the BRIRs where similarly obtained by deriving the 500 Hz octave band SPLs for the 0° azimuth configuration.

Playback

Obtained binaural and multichannel responses were convolved with the same anechoic Danish speech sentence (approximately 2 s long) for all distances of the same condition. Convolved sentences for multichannel conditions were played on the 29 3D loudspeaker array or the restricted 16-loudspeaker horizontal array for the HOA seventh 2D conditions (C13 and C14). A loudspeaker equalization was performed to flatten each of the loudspeaker frequency responses recorded at the center of the array (cf. Appendix B). The binaural convolved sentences were played back on headphones that were equalized to flatten the average frequency response between the headphone and the ears of the HATS for different headphone positions.

4.2.2 Procedure

Each condition was divided into three blocks consisting of three repetitions of each distance configuration presented in a random order. To avoid any interference by visual cues provided by the playback room, the listeners were unfamiliar to the laboratory area and were blindfolded before entering the playback room. Subjects were seated on a chair at the center of the loudspeaker array and asked to remain still during each block. In the beginning of each block, the auralized sentences corresponding to the closest and furthest distance for this condition were played. It provided the subject with the distance range of the sound events together with the distance variation cue in the present block. The value of the actual physical distance range was not revealed to the subject. Then, after each auralized sentence, subjects were asked to answer aurally the question: "How far away does the sound appear?" with a distance in meter. Each subject performed a short training session beforehand in order to be familiarized with the task.

4.3 Results

4.3.1 Overall data and power-law fitting

Means and standard deviations of the logarithmic transform of the apparent distance across repetitions are plotted in Fig. 4.3 for each subject and all conditions. In each panel, a small variation was introduced in the physical distance in order to avoid occlusion of symbols for like responses across subjects.

Standard deviation across repetitions were ranging from 0 to 0.35 for all subjects, conditions and distances except for subject 'S' who showed a range between 0.12 and 0.66 for the binaural conditions. For all conditions, these within-subjects variations were small in comparison to the large inter-subject variability of the data shown in Fig. 4.3. In order to model the data, the apparent distances were fitted with a simple linear function in the double logarithmic domain, which in the linear domain corresponds to a power-law fitting $r' = kr^a$, where r and r' represent the physical and apparent distances in meters respectively, a the power-law exponent and k a constant. The fitted curves for an example subject ('T') were indicated in Fig. 4.3 by a blue solid

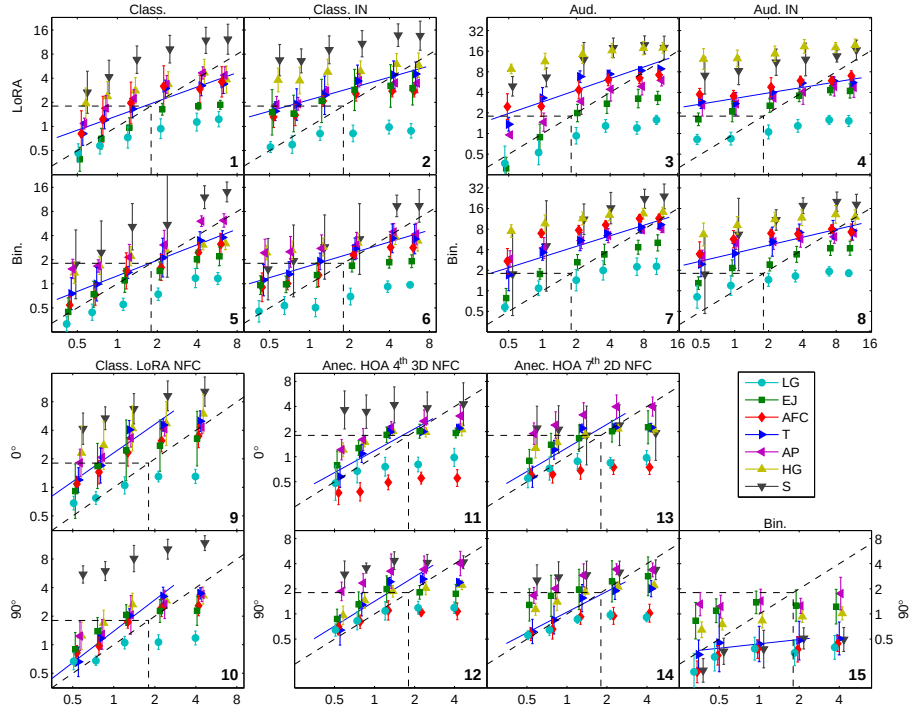

Figure 4.3: Mean values and standard deviations of the logarithmic apparent distance plotted against the logarithmic physical distance for all conditions. Axis labels units are in meters. The dashed vertical and horizontal lines represent the loudspeaker array radius and the diagonal dashed line, the physical-estimated identity. The condition number is written on the bottom right corner of each plot.

line. For conditions when all natural cues were presented (i.e., conditions without intensity normalization: C1, C3, C5 and C7), the average value of a across subjects was 0.52 (s.d. 0.16) and always lower than its veridical value of 1. This highlights the compressive behavior of the distance perception as seen in various auditory studies reviewed in Zahorik *et al.* (2005). The constant k represents the offset in the absolute estimation of distances and was ranging, for these conditions, from 0.5 to 4.7 m for the classroom and from 0.6 to 11.1 m in the auditorium, reflecting the above mentioned large inter-subject variability.

4.3.2 Distance perception with the LoRA system

Power-law exponents, i.e., the degree of compression of the fitted power function, for all subjects for the first part of the experiment (i.e., conditions C1 to C8) showed significant positive values. The mean values and confidence intervals of these exponents across subjects are shown in Fig. 4.4 and the lowest value of the confidence intervals was 0.16. All subjects were thus able to perceive a change of distance even when the intensity of the stimuli was normalized with distance (IN), indicating the usefulness of the direct-to-reverberation cue provided by both auralization techniques. A further analyze showed that the exponent a of the natural intensity conditions was significantly larger than the one of the corresponding intensity normalized conditions for the classroom and the auditorium for both auralization techniques (approximatively by a value of 0.2). As expected, when both intensity and the direct-to-reverberant ratio cue are provided, subjects had a greater distance discrimination compared to when only the later cue was provided.

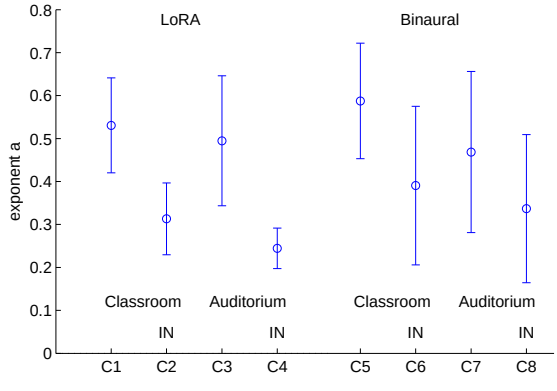

Figure 4.4: Mean values and confidence intervals of the fitted power-law exponent a for conditions C1 to C8.

When comparing distance perception between both auralization techniques, mean values of the difference of exponent a and constant $log(k)$ between the LoRA and the binaural conditions were not significantly different from 0 for both rooms, with or without intensity normalization. Hence, the experiment show no significant difference

in quality or realism in distance perception between the room simulation based LoRA system and the dummy-head recording-based binaural system.

4.3.3 Distance perception of near-field sources

In order to evaluate the effect of the NFC method on HOA for near sound sources, i.e., for sources within the loudspeaker array of 1.8 m radius, only the data corresponding to physical distances smaller than 2.3 m were considered in the power-law data fitting. In the anechoic environment, the means of the power-law exponents across subjects for conditions when the distance was simulated with NFC filters only (C11 to C14) were significantly positive (see Fig. 4.5) and showed an average value across these four conditions of 0.45 which is comparable to the average exponent a of 0.52 for the natural intensity condition in reverberant environments (i.e., conditions C1, C3, C5 and C7). Sources were thus perceived from within the loudspeaker array when reproduced with the near-field control method for HOA. However, absolute distance estimations varied greatly across subjects (average k of 1.6 m, s. d. 1 m) as shown in Fig. 4.3 indicating that sources closer than the loudspeaker array radius (1.8 m) were not always estimated below 1.8 m.

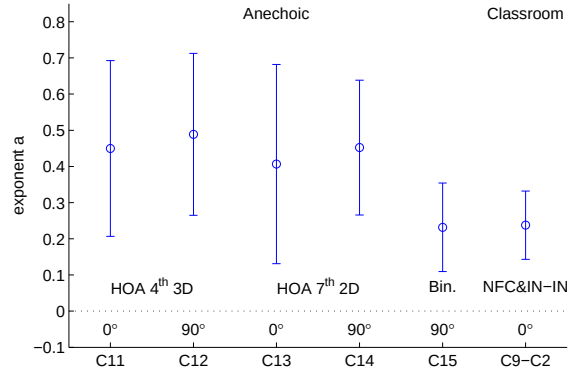

Figure 4.5: Mean values and confidence intervals of the fitted power-law exponent a for conditions C11 to C15 and C9-C2.

The mean values of the difference of exponent a between conditions auralized

with HOA fourth order 3D and HOA seventh order 2D were not significantly different from zero. The direction of the source did also not have any influence on the exponent.

The mean value of the exponent a for the binaurally auralized condition (C15) in the anechoic environment was of 0.23 (see Fig. 4.5), which was lower than the mean for the NFC-HOA auralized conditions (C12 and C14, average a value of 0.47). However, the mean of the differences (C12-C15 and C14-C15) was not significantly different from zero. The value of the average constant k was 0.6 m, significantly lower than the two loudspeaker-based conditions (C12 and C14, average k value of 1.5 m). Hence, distances were globally underestimated with the binaural technique auralization compared to the loudspeaker-based one. This might be due to the limitation of the use of non-individual BRIRs in anechoic environments that often leads to in-the-head-localization (Blauert, 1997).

In the classroom with intensity normalized stimuli, apparent distances showed a larger power-law exponent when the NFC method was used: the mean of the difference of a was of 0.24 (See Fig. 4.5, C9-C2).

4.4 Discussion

The inter-subject variability in distance estimations was very large in all experimental conditions, an aspect that has previously been pointed out by other studies (e.g., Nielsen, 1993; Zahorik *et al.*, 2005), highlighting the rather poor auditory sensitivity to distance. The large variability might have been further influenced by the fact that evaluation of distances based solely on auditory cues was not a familiar task for the test-subjects and that the direct-scaling method was used without prior reference in meter to avoid bias of the responses.

The study showed significant different degrees of compression of the fitted power function for different conditions. First, in reverberant rooms, the power-law exponent was decreased when the stimuli intensity was normalized with distance for both binaural and loudspeaker-based auralization techniques (by approximatively 0.2). Second, the power-law exponent was increased when near-field compensation filters were applied to HOA auralization (by approximatively 0.24). This showed that the removing of the intensity cue decreased the range of estimated distances and that NFC signif-

icantly contributes to distance perception in HOA-based room auralization respectively.

In anechoic environment, NFC-HOA (at least for fourth- and seventh-order) auralized sources positioned as close as 0.6 m and within the loudspeaker array radius were perceived closer (average a of 0.45) compared to the same source auralized without NFC. This demonstrates the ability of rendering sources within the loudspeaker array with NFC-HOA.

The experimental results showed no significant differences between the binaural recording-based and the loudspeaker-based (LoRA system) auralization technique both in terms of the power-law exponent a and the absolute values k. This was the case when all natural cues were presented as well as when the intensity of the stimuli was normalized with distance. Loudspeaker-based auralization in the LoRA system seems thus able to reproduce auditory environments in which distance perception is similar to that of the corresponding real environment. However, one could argue that the experiment was not able to show a significant effect between the two auralization techniques due to limitations in the statistical relevance of the study. In this regards, the sensitivity of the experiment could be improved by increasing the number of subjects and repetitions. Nevertheless, even if an effect on distance perception between the two auralization techniques exists, it is expected to be smaller than the effect of removing the intensity cue or applying NFC, which the experiment provided evidence for.

4.5 Conclusion

The present auditory distance perception experiment provided evidence for the following points:

1. Loudspeaker-based auralization in the LoRA system provides similar distance perception to that of the corresponding real environment,

2. NFC-HOA provides a significant increase in the range of perceived distances for near sound sources as compared to standard HOA,

3. NFC-HOA auralization allows for rendering sources within the loudspeaker array both in anechoic and reverberant environments,

4. The use of non-individual BRIRs led to underestimated distances compared to the use of NFC-HOA.

5

Reproduction of nearby sound sources using higher-order Ambisonics with practical loudspeaker arrays[§]

Abstract

In order to reproduce nearby sound sources with distant loudspeakers to a single listener, the near field compensated (NFC) method for higher-order Ambisonics (HOA) has been previously proposed. In practical realization, this method requires the use of regularization functions. This study analyzes the impact of two existing and a new proposed regularization function on the reproduced sound fields and on the main auditory cue for nearby sound sources outside the median plane, i.e., low-frequencies interaural level differences (ILDs). The proposed regularization function led to a better reproduction of point source sound fields compared to existing regularization functions for NFC-HOA. Measurements in realistic playback environments showed that, for very close sources, significant ILDs for frequencies above about 250 Hz can be reproduced with NFC-HOA and the proposed regularization function whereas the existing regularization functions failed to provide ILDs below 500 Hz. A listening test showed that these lower-frequency ILDs provided by the proposed regularization function lead to a significantly improved distance perception performance. This test also showed that the distance of virtual sources are perceived less accurately than corresponding physical sources when amplitude cues are not available.

[§] This chapter was published as Favrot and Buchholz (2012).

5.1 Introduction

Virtual auditory environments (VAEs) ideally reproduce recorded or simulated auditory environments to one or several listeners (Blauert, 2005) depending on their applications (e.g. auditory display, vehicle simulators, psychophysical research, Chap. 2). In some of these applications, nearby sound sources, also known as focused sound sources (typically located within 1 m from the listener), need to be reproduced. Simulating VAEs including nearby sources is especially challenging when they are generated by loudspeaker-based systems, i.e., sound field reproduction techniques such as wave field synthesis (WFS, Berkhout *et al.*, 1993) or higher-order Ambisonics (HOA, Gerzon, 1973; Fellgett, 1974; Poletti, 2005). The main difference between these techniques is that WFS provides a reproduced sound field to large listening areas whereas HOA reproduces the sound field around a defined point inside the loudspeaker array (sweet spot). In this manuscript, the HOA method is considered for a single listener located in the center of a loudspeaker array.

The human auditory system makes use of different cues to evaluate the distance of sound events. For far away sources (more than 1 m away from the listener), distance perception mainly relies on monaural cues such as amplitude cues, direct-to-reverberant energy ratio cues and, to some degree, spectral cues (Zahorik *et al.*, 2005). For nearby point sources (closer than 1 m), the interaction of the listener's body with the sound field changes with distance because of the curvature of the sound field. This can lead to very large interaural level differences (ILDs) at low-frequencies for sources outside the median plane (Brungart and Rabinowitz, 1999) and provides a strong cue for nearby distance perception (Brungart, 1999). Interaural time differences (ITDs) have however only little impact on distance perception for these sources. In the median plane, the ILD cue is not available. Finally, distance perception of nearby sound sources is in better agreement with the physical distances than for distant sources.

For higher-order Ambisonics, Daniel (2003) developed the near field compensated (NFC) HOA method for the reproduction of monopole point sources. This method introduces near field coding filters applied to the Ambisonic components and, in theory, results in correct sound field reproduction in the sweet spot. However, these filters exhibit large gains at low frequencies for short distances and high-order

components, which can lead to extremely large loudspeaker signals at low frequencies. Although the increased low-frequency energy is compensated inside the sweet spot by the interference of the different loudspeaker signals, a strong low-frequency boost can be observed in the sweet spot with the presence of the listener's head or small (realistic) sound path differences from the different loudspeakers to the receiver. Therefore, in practical implementations, NFC filter gains need to be limited, which can be achieved by regularization functions (RFs) such as high-pass filters (Daniel and Moreau, 2004) or cosine-shaped weighting functions (Ahrens and Spors, 2009). Since a preliminary study showed that the performance of these RFs was found to be insufficient (Favrot and Buchholz, 2009a), an alternative RF is considered in this study, which consists in a modified version of the Tikhonov regularization function. The choice of the RF thereby impacts the reproduced sound field depending on the listening room type and loudspeaker array characteristics.

For practical reasons, loudspeaker arrays are often used in acoustically damped rooms rather than in anechoic rooms. In these conditions, small wall reflections deteriorate the distance cues depending on the applied RF. In HOA, the sound field reproduction is often restricted to the horizontal plane (2D), which allows for a better directivity and requires less loudspeakers for a given order compared to full 3D representations. The Fourier-Bessel expansion theory underlying the 2D restriction results in a $1/\sqrt{r}$ amplitude decay with distance r (Morse and Feshbach, 1953) and thus is different from the decay of a point source ($1/r$). This is an important limitation for realizing nearby sound sources and therefore 2D and 3D NFC-HOA performance are separately discussed in this paper.

The main goal of this study is to assess if a NFC-HOA reproduced nearby sound source can be perceived at the same distance as the corresponding physical one. In order to realize this, the impact of (1) two existing and a proposed RF, (2) the loudspeaker setup (2D or 3D) and (3) the listening room (anechoic or acoustically-damped) on distance perception of these sources is evaluated. This evaluation is realized both objectively by studying resulting sound field characteristics and distance cues as well as subjectively by performing listening tests. The distance perception performance of nearby sound sources is discussed in regards of the physical distance cues (i.e., low

frequency ILDs) constrained by the reproduction system, hence providing insight into basic perception of nearby sound sources.

In the first section, the implementation of NFC filters is presented together with two existing and one proposed RFs. In the second section, physical properties of resulting sound fields are analyzed with the help of numerical simulations for both 2D and 3D representations. Then, the ILD cues available to the listener are analyzed both in an anechoic (ideal) and in an acoustically-damped (practical) listening environment . Finally, the listening test to assess the distance perception of NFC-HOA virtual nearby sources is presented and the results are discussed.

5.2 Methods

5.2.1 Near field compensated higher-order Ambisonics

Near field compensated higher-order Ambisonics (NFC-HOA, Daniel, 2003)) aims at recreating a sound field produced by a virtual point source with a loudspeaker array. For the 3D representation, this technique is based on a Fourier-Bessel expansion of a point source (Williams, 1999), and relates to the HOA plane wave description (Malham and Myatt, 1995; Daniel, 2003; Poletti, 2005) by applying near field coding (NFC) filters to Ambisonic components. In the following, the nomenclature follows the numerical approach and notation of Daniel (2000, 2003) and the spherical coordinate system with azimuth $\theta \in [0; 2\pi[$, elevation $\delta \in [-\pi/2; \pi/2]$ and distance ρ from the coordinate origin O (center of the spherical loudspeaker array) is considered.

For a 3D NFC-HOA representation, a point source situated at point O_s $(\theta_s, \delta_s, \rho_s)$ carrying a signal $S(\omega)$ (captured at the center O) is encoded up to the Ambisonic order M with Ambisonic components ${B_{mn}^{\sigma}}^{NFC(R)}(\omega)$ (for $0 \leq n \leq m \leq M, \sigma = \pm 1$) expressed in the frequency domain ($k = \omega/c = 2\pi f/c$) described by:

$$ {B_{mn}^{\sigma}}^{NFC(R)}(\omega) = S(\omega)\ H_m^{NFC(\rho_s,R)}(k)\ Y_{mn}^{\sigma}(\theta_s, \delta_s) \tag{5.1} $$

where R is the radius of the loudspeaker array and $Y_{mn}^{\sigma}(\theta, \delta)$ the "real-valued" spherical harmonics function (Morse and Feshbach, 1953) described as N3D normalized by

Daniel (2003),

$$Y_{mn}^{\sigma}(\theta,\delta) = \sqrt{\epsilon_n(2m+1)\frac{(m-n)!}{(m+n)!}}P_{mn}(\sin\delta)\begin{cases}\cos n\theta \text{ if } \sigma = +1\\ \sin n\theta \text{ if } \sigma = -1\end{cases} \tag{5.2}$$

where $\epsilon_n = 2, n > 0$ and $\epsilon_0 = 1$, and $P_{mn}(\sin\delta)$ are the "Schmidt semi-normalized" associate Legendre functions of degree m and order n. $H_m^{NFC(\rho_s,R)}(k)$ represents the NFC filters and are expressed for $0 \leq m \leq M$ by:

$$H_m^{NFC(\rho_s,R)}(k) = \frac{F_m(k\rho_s)}{F_m(kR)} \tag{5.3}$$

where the transfer function $F_m(kr)$ is described by:

$$F_m(kr) = \frac{h_m^-(kr)}{h_0^-(kr)} \tag{5.4}$$

with h_m^- the spherical Hankel function of the second kind. The normalization by $h_0^-(kr)$ accounts for the attenuation and the delay of the pressure signal S which is measured at the origin O (Daniel, 2003; Morse and Feshbach, 1953). The frequency response of NFC filters is plotted in Fig. 5.1 for a source at distance $\rho_s = 0.5$ m for $m = 0$ to 7. The decoding of NFC-HOA signals is realized in the same manner as for HOA (Daniel, 2003) and the same limitations concerning the number of loudspeakers and the Ambisonic order holds. It should be noted that the Fourier-Bessel expansion of a point source used to derive Eq. 5.1, is only valid for a center region free of source (Williams, 1999) , i.e., the region $\rho < \rho_s$.

For a 2D NFC-HOA representation, a monopole line source expansion leads to similar NFC filters with Hankel functions (H_m^-) instead of spherical Hankel functions (h_m^-) in Eq. 5.4 (Ahrens and Spors, 2009). This 2D representation relies on monopole infinite line sources for the virtual source and the loudspeakers. Such a source produces a sound field with amplitude decaying by $1/\sqrt{r}$ with distance r to the source. However, sound propagation of cabinet loudspeakers is more accurately modeled by a point source with an amplitude decay of $1/r$. In order to approximate a virtual point source with a 2D representation, Daniel and Moreau (2004) utilized NFC filters based

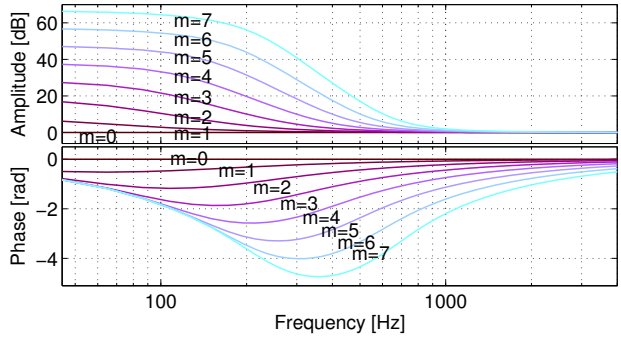

Figure 5.1: Amplitude and phase responses of NFC filters for $m = 0$ to 7, for a virtual source at $\rho_s = 0.5$ m and a loudspeaker array radius of $R = 1.5$ m and $c = 340$ m/s.

on *spherical* Hankel functions. This approximation provides an amplitude decay that is closer to the one of a point source and is also applied in this study. The amplitude decay issue for 2D HOA representation is also found in other sound field reproduction techniques, such as wave field synthesis (Nicol and Emerit, 1999).

5.2.2 Regularization functions for NFC-HOA

Even though the NFC filters are stable filters, they exhibit large gains at low frequencies and high order m (tends to m 20 $log_{10}(R/\rho_s)$ for $f \rightarrow 0$) when the virtual point source is situated inside the loudspeaker array ($\rho_s < R$). The large resulting loudspeaker signals are destructively interfering for locations closer than the virtual source $\rho < \rho_s$ to realize the desired sound field, while at locations further than the virtual source $\rho > \rho_s$, they lead to extremely large sound pressure (e.g., Ahrens and Spors, 2009). The latter is shown in the top left panel of Fig. 5.3 and 5.4 where sound pressure values are clipped in the region $\rho > \rho_s$, i.e., the sound pressure level (SPL) is 6 and 10 dB larger than the SPL at the origin O respectively. In practice, for high orders and short distances, the required energy at low frequencies can not be produced by monitor-size loudspeakers. Moreover, the large low frequency energy from the different loudspeaker does not cancel each other in positions between the reference center and the virtual source when small (realistic) sound path differences occur (i.e., from the different loudspeakers to the receiver and/or small wall reflections from a

non-anechoic playback room). This eventually leads to a low-frequency boost and arises from the fact that the underlying theory of NFC-HOA is only valid in the region $\rho < \rho_s$ (Eq. 5.1).

The high-order components are responsible for introducing very large energy in the loudspeaker signals at low frequencies, but they do not significantly contribute to the sound field reproduction in the sweet spot below a certain frequency (Daniel and Moreau, 2004). Therefore, practical implementations of NFC-HOA introduce regularization functions (RFs) to the NFC filters in order to restrict the contribution of high-order components at high frequencies. These weighting functions can be interpreted as high-pass filters with a certain cut-off frequency depending on the order m and the ratio ρ_s/R. The use of RFs results in low energy loudspeaker signals adding up in the listening area to create the required sound field instead of large energy signals leading to destructive interference with the non-regularized NFC filters. However, to our knowledge, no perceptual study exists that evaluates the potential influence of the provided Ambisonic order at low frequencies on perceived distance.

Different RF implementations have been proposed by Daniel and Moreau (2004) and by Ahrens and Spors (2009). The overall goal of these functions is to avoid positive gains in the NFC filters and to maintain their phase response which is mainly responsible for the curvature of the reproduced sound field. In this paper, the effect of three RFs w_m (applied to $H_m^{NFC(\rho_s,R)}$) on the reproduced sound field is investigated: (i) the cosine-shaped RF (noted 'coswin') as described by Ahrens and Spors (2009):

$$w_m^{coswin} = \begin{cases} (cos(\frac{m\pi}{k\ \rho_s}) + 1)/2 & \text{if } m \leq k\ \rho_s \\ 0 & \text{if } m > k\ \rho_s \end{cases} \tag{5.5}$$

(ii) the high-pass (HP) filter (noted 'hp') proposed by Daniel and Moreau (2004)

$$w_m^{hp} = \frac{\left(f/f_{lim}^{(m)}\right)^{4m}}{1 + \left(f/f_{lim}^{(m)}\right)^{4m}} \tag{5.6}$$

where the cut-off frequency $f_{lim}^{(m)}$ is derived from generic values listed in Daniel and Moreau (2004, Table 1) which can be approximated by $f_{lim}^{(m)} = mc/(2\pi\rho_s)$, and (iii)

a novel RF (noted 'MTreg') which is a modified version of Tikhonov regularization function (Vogel, 2002):

$$w_m^{MTreg} = \frac{2}{\left|H_m^{NFC(\rho,R)}\right|^2 + 1} \tag{5.7}$$

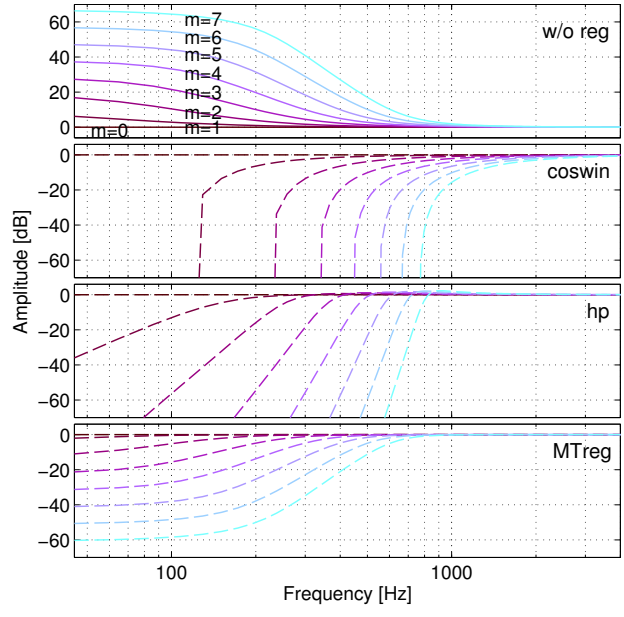

Figure 5.2: Amplitude spectrum of the regularized near field coding (NFC) filters (dashed lines) for the cosine RF (coswin), the HP filter (hp) and the MTreg RF for $m = 0$ to 7 and the same condition as Fig. 5.1. The non-regularized NFC filters for the same conditions as Fig. 5.1 are replotted on the top panel (w/o reg).

The resulting amplitude spectra of the regularized NFC filters are shown in Fig. 5.2 for the three RFs. The phase spectra shown in Fig. 5.1 are not affected by these RFs. The coswin RF realizes a HP filter that strongly attenuates the high-order components over a large frequency range. With the hp RF, the high-order components are activated at lower frequencies and the regularized NFC filters exhibit small positive overshoot (2.5 dB for $m = 7$). The MTreg RF leads to activation of high-order components at even lower frequencies but without providing filters with positive

gains. This function is a compromise between (i) activating high-order components at sufficiently low frequencies where the phase deviation is the largest (see bottom panel in Fig. 5.1) and (ii) restricting their amplitude in order to minimize the introduction of low frequencies in the loudspeaker signals.

The different RFs show different robustness in realistic conditions, i.e., when loudspeaker differences or loudspeaker-receiver path differences due to a non-anechoic playback room are present. Moreover, they show different low-frequency characteristics which potentially influences human distance perception (Brungart, 1999). These aspects will be investigated in the following sections.

5.3 Physical properties of NFC-HOA reproduced sound fields

In this section, the impact of the regularization functions on reproduced sound fields is investigated. Loudspeakers are considered as perfect monopole point sources in free field for both 3D and 2D NFC-HOA representations (see justification in section 5.2.1). The pressure sound field generated by such a monopole point source located at $\vec{R}$ is modeled at a point $\vec{r}$ as

$$p(\vec{r},k) = S\ \frac{e^{-j\ k\ |\vec{r}-\vec{R}|}}{|\vec{r}-\vec{R}|}\ \frac{|\vec{R}|}{e^{-j\ k\ |\vec{R}|}} \tag{5.8}$$

where k is the wave number and S the pressure measured at the origin O. For simplicity, in this section, the virtual sources are only considered in the horizontal plane, i.e., $\delta_s = 0°$, but similar considerations apply to elevated sources.

5.3.1 Sound field simulations

Instantaneous pressure sound fields in the horizontal plane resulting from a virtual point source located at point O_s ($\theta_s = 0°$ and $\rho_s = 0.5$ m) with 2D NFC-HOA for an order of $M = 15$ are shown in Fig. 5.3 for the non-regularized NFC filters and the NFC filters with the three considered RFs. A circular 32-loudspeaker array of radius $R = 1.5$ m with equally spaced angles was considered. In each panel, the contour

encloses the reconstructed area within 25 % absolute error of the complex pressure. The arrows indicate the local propagation direction, which is opposite of the gradient of the pressure phase. The SPL of these reproduced sound fields are shown in Fig. 5.4.

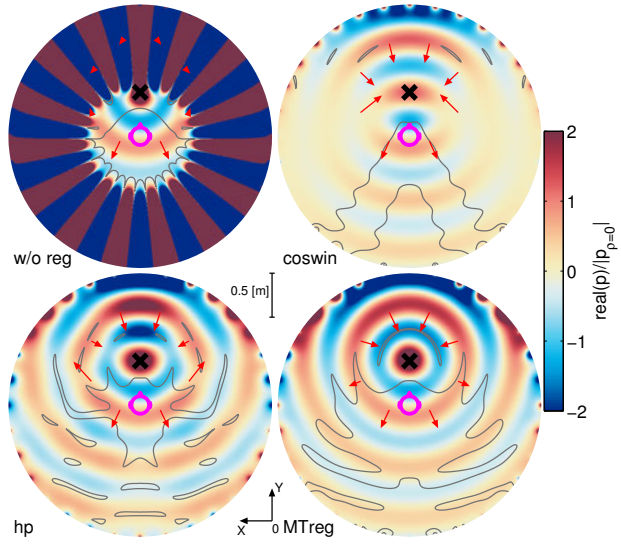

Figure 5.3: Real value of the instantaneous pressure of the reproduced sound field with 2D NFC-HOA (order $M = 15$) for a monopole point source placed at location O_s ($\theta_s = 0°, \rho_s = 0.5$ m, cross marker) without RF (w/o reg), with the coswin, the hp and the MTreg RF for a frequency $f = 600$ Hz. The arrows indicate the local propagation direction and the thin lines enclose the reconstructed area within 25 % absolute error.

For NFC, the sound wave propagation converges to the point source between the array and the source and diverges from the point source (as desired) between the source and the listener position. This property is the reason for the term "focused virtual sound source". Due to the causality principle, it is impossible to realize a converging sound field simultaneously in these two areas ($\rho < \rho_s$ and $\rho_s < \rho$). This limitation is also found in the reproduction of focused sound sources with WFS (Spors *et al.*, 2009) and is problematic for arbitrarily placed listeners. This phenomenon does not constitute a limitation for NFC-HOA when a single listener placed in the sweet spot is considered.

Without RF, extremely large sound pressure values occur in the region $\rho > \rho_s$

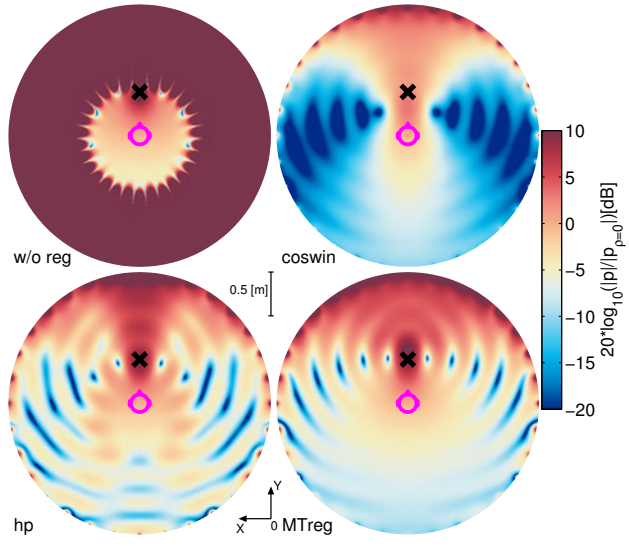

Figure 5.4: SPL of the reproduced sound fields from Fig. 5.3

(values are clipped at 10 dB relative to the SPL at the origin O in the top left panel in Fig. 5.4) as expected. The use of the three RFs limits these excessive sound pressure values for $\rho > \rho_s$ to acceptable values (bottom left and right panels). When the coswin RF is used, the correct reconstructed area defined by this contour is very tight around the head size of the listener. This area is larger with the hp RF and even larger with the MTreg RF. This effect is also seen at other frequencies and indicates that the curvature of the sound field is well reproduced over the largest area around the listener when the MTreg RF is used.

Similarly, the SPL distribution in the region closer than the source ($\rho < \rho_s$) is the closest to the non-regularized case when the modified Tikhonov RF (MTreg) is used (bottom right panel, Fig. 5.4). Because of the strong attenuation of high-order components by the coswin RF, a drop of level is observed laterally with this function. With the hp RF, some irregularities can be observed on the edge of this area.

In order to compare NFC-HOA sound field reconstruction in 2D and 3D, instantaneous pressure sound fields simulated for a lower Ambisonic order of $M = 7$ are shown in Fig. 5.5 (left panel). For the 2D case, a horizontal array consisting of 16

equally spaced loudspeakers was used. For the 3D case, a quasi-regular 3D array of 92 loudspeakers (obtained after 10 subdivisions of an octahedron, Hollerweger (2005)) was used. When comparing Fig. 5.3 (lower right panel) and 5.5 (left panel), it seems like the 7$^{\text{th}}$- and 15$^{\text{th}}$-order decomposition provide similar sound fields for the region $\rho < \rho_s$. This is not surprising since the RFs are designed to effectively reduce the contribution of high order Ambisonic components at low frequencies. In this simulation, at 600 Hz, for a virtual source at 0.5 m and a loudspeaker radius of $R = 1.5$ m, the contribution of components of orders higher than 7 are attenuated by more than 10 dB (Fig. 5.2, lower panel). Since the closer the source is to the origin the higher the cut-off frequency of the regularized NFC filters is, the RFs inhibit the HOA algorithm to improve the sound field reproduction of very close sound sources at low frequencies when increasing the Ambisonic order.

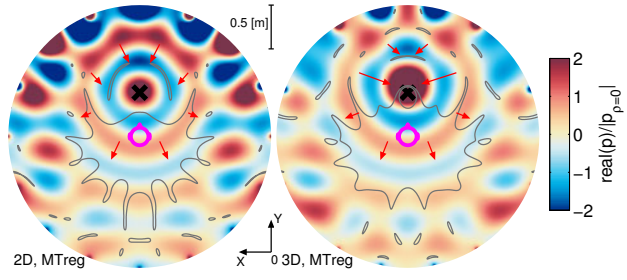

Figure 5.5: Real value of the instantaneous pressure of the reproduced sound field in the horizontal plane with 2D (left panel) and 3D (right panel) NFC-HOA (order $M = 7$) for a monopole source placed at location O_s ($\theta_s = 0°, \rho_s = 0.5$ m) using the MTreg RF for a frequency $f = 600$ Hz.

The 2D and 3D representations used with the MTreg RF leads to similarly well reconstructed areas (absolute error < 25 %) except for positions between the source and the origin where the pressure level appears to be larger in the 3D case.

5.3.2 Pressure field amplitude decay

In order to analyze the pressure sound field's amplitude decay for the different RFs, the SPL along the straight line connecting the source and origin O (y-axis) is plotted in Fig. 5.6 (panel A and B) for a virtual source at 0.5 m. The dashed lines represent the amplitude decay of an ideal monopole point source ($\propto 1/|\rho - \rho_s|$). The level of

all the amplitude decay curves is normalized by the SPL at origin O ($p_{\rho=0}$). For the 2D case (panel A), the amplitude decay for $|\rho| < \rho_s$ has a slightly more shallow slope than that for a point source for the three RFs. As mentioned in section 5.2.1, the use of spherical Hankel functions in the 2D case helps to improve the amplitude decay, but can not fully reproduce the decay of a point source. For the 3D representation (panel B), the best match of the amplitude decay to the ideal monopole source is observed when the hp and the MTreg RFs are used. It should also be noted that when the coswin RF is used, the maximum level occurs between the origin and the source and is substantially lower than for the two other RFs for both 2D and 3D representations.

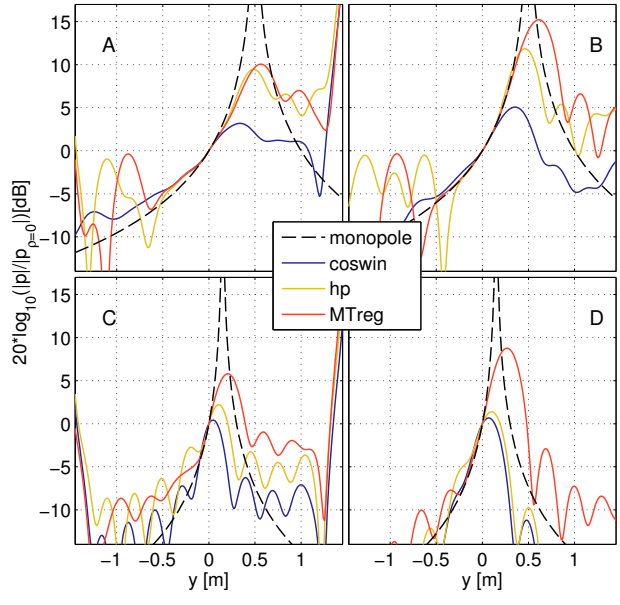

Figure 5.6: Amplitude decay along the y-axis for a virtual source at distance $\rho_s = 0.5$ m (A, B) and at $\rho_s = 0.15$ m (C, D) for a 2D (A, C) and 3D (B, D) NFC-HOA reproduction.

Amplitude decays for a very close source ($\rho_s = 0.15$ m) are shown in panels C and D in Fig. 5.6. The coswin and hp RFs lead to similar amplitude decay curves for both 2D and 3D representations with a maximum very close to the origin (i.e., $y = 0$ m). The MTreg RF shows an SPL maximum that is larger by about 5 dB (2D)

and 8 dB (3D) and shifted beyond the virtual source location, which results in a better fit around the origin O.

5.3.3 Frequency response outside the origin

The preceding analyzes focused on one single frequency (600 Hz). In order to characterize the reproduced sound field over the whole audible frequency range, amplitude spectra are shown in Fig. 5.7 for positions O, A, B and C (see inset) at distances 0, 0.08, 0.15 and 0.21 m from the origin and at a constant distance of 0.15 m from the virtual source at point O_s ($\rho_s = 0.15$ m, $\theta_s = 0°$). Here, only 2D NFC-HOA reproduction (order $M = 7$) is considered since a 3D representation shows a very similar behavior.

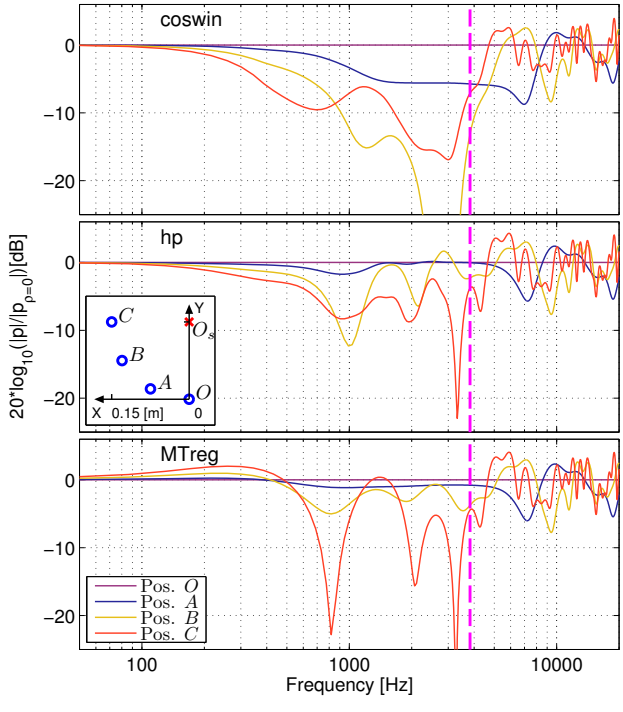

Figure 5.7: Amplitude spectra at positions O, A, B and C for a virtual source at point O_s ($\rho_s = 0.15$ m, $\theta_s = 0°$) reproduced with 2D NFC-HOA at order $M = 7$ for the three RFs.

For the tested RFs, the sound pressure at mid-frequencies (approx. from 300 Hz to 4 kHz) decreases when distance increases from the origin O. This decrease is larger with the coswin RF (in line with Fig. 5.4) and smaller with the MTreg RF for positions where $\rho \leq \rho_s$ (positions O, A and B). For higher frequencies (above approx. 4 kHz), amplitude fluctuations outside the origin O can be observed for all three RFs. These fluctuations arise from the different loudspeaker-receiver path lengths and are typical of HOA for which a flat amplitude response across frequencies can only be ideally obtained at the origin O. The frequency limit f_{lim} for accurate reproduction (error < 4 %) is defined in HOA (Moreau *et al.*, 2006; Ward and Abhayapala, 2001; Poletti, 2005) as $f_{lim} = Mc/2\pi r$ where M is the order and r the radius of the sweet spot. Using this definition, with $r = 0.1$ m as an approximation of the head radius and $M = 7$, $f_{lim} = 3.8$ kHz (vertical dashed line in Fig. 5.7). The MTreg RF also shows the best performance in the area $\rho \leq \rho_s$ in simulations with 3D NFC-HOA.

5.4 Auditory distance cues

The important physical properties of NFC-HOA reproduced sound fields with nearby sound sources have been analyzed in the previous sections. This section investigates the sound pressure at the listener's ears that is inferred by the reproduced sound fields. In particular, the ILDs are considered, which provide the predominant distance cue for nearby virtual sources outside the median plane (Brungart *et al.*, 1999). ILDs are in general influenced by the Ambisonics order and potentially further modified at low frequencies due to the RFs. The simulation of the NFC-HOA reproduction was realized here with either a 2D or 3D loudspeaker array in either an anechoic (ideal) or acoustically-damped (practical) listening environment.

5.4.1 Simulated anechoic listening environment

In ideal conditions, defined here as a listener perfectly aligned at the center of the loudspeaker array placed in free field, signals at the listener's ears were derived using the KEMAR head related transfer functions (HRTFs) database (Gardner and Martin, 1995), and an array of radius $R = 1.4$ m (loudspeaker distance used in KEMAR's

HRTFs recording) was considered. The equalized version of KEMAR's HRTFs were used in order to remove the influence of the loudspeaker's frequency response. Two loudspeaker array layouts were simulated to realize 7^{th}-order 2D and 4^{th}-order 3D Ambisonics: a 2D array consisting of 18 loudspeakers equally spaced on a circle and a 3D array consisting of the horizontal 2D ring plus two rings of 7 loudspeakers at elevation $-40°$ and $+40°$ plus one loudspeaker at the zenith (i.e., 33 loudspeakers in total). These layouts were selected from the HRTFs directions available in the KEMAR database to correspond best to the layout used in the practical listening environment described in the following section. The amplitude spectra of the simulated binaural signals are shown in Fig. 5.8 for a NFC-HOA reproduced source at $\rho_s = 0.15$ m, $\theta_s = 90°$ and the MTreg RF. The figure also shows the amplitude spectra of the binaural signals for a source at $\rho_s = R = 1.4$ m simulated using "plane-wave" HOA (i.e., without the near-field compensation) which are considered to correspond to far-field HRTFs. A "basic" HOA decoding was used for all frequencies (see Daniel, 2000 for details).

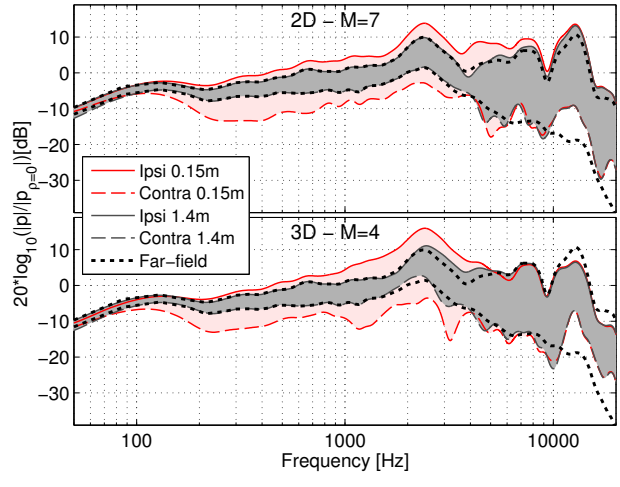

Figure 5.8: Binaural amplitude spectra for 2D (top, M=7) and 3D (bottom, M=4) NFC-HOA reproduction of a source at azimuth $\theta_s = 90°$ and distance $\rho_s = 0.15$ m (red curves) and 1.4 m (gray curves). Thick dashed curves represent the corresponding far field HRTF. Shaded areas represent the ILD.

Amplitude spectra of binaural signals for a source simulated at 1.4 m (gray

curves) match the corresponding far field HRTFs (dashed thick curves) from about 200 Hz to about 4 kHz for the 2D representation and to about 2 kHz for the 3D representation. The high-frequency limits are directly linked to the Ambisonic order used for each representation. The frequency limit f_{lim} (see section 5.3.3) for a sweet spot of radius $r = 0.076$ m (corresponding to the KEMAR's head radius) is 5 kHz for 7^{th}-order 2D and 2.9 kHz for 4^{th}-order 3D.

For a close source at $\rho_s = 0.15$ m (red curves), in both 2D and 3D, the amplitudes at the ipsi-lateral ear are larger and at the contra-lateral ear smaller compared to the corresponding far-field HRTF. This reflects the pressure field amplitude decay shown in Fig. 5.6 and corresponds to an increase of ILDs (shaded area in Fig. 5.8) when source distance decreases. For frequencies higher than f_{lim}, the amplitude spectra for the 0.15 m distant source is similar to the 1.4 m-distant source for both ears.

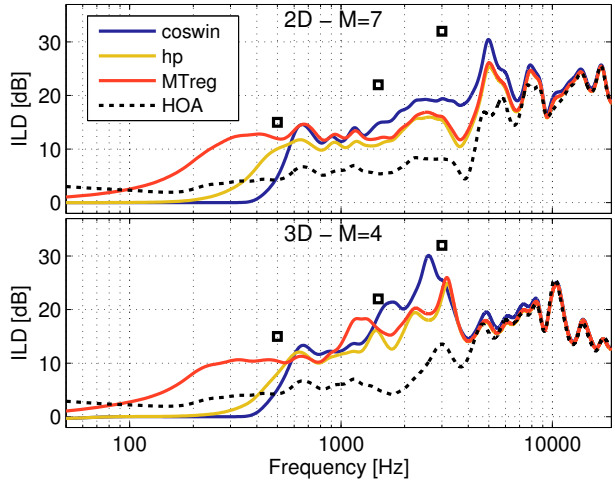

Figure 5.9: ILDs for a virtual source at O_s ($\rho_s = 0.15$ m, $\theta_s = 90°$) simulated with three RFs (solid curves) and without any near-field compensation (dashed curves). Black squares indicate ILDs measured for real sources by Brungart and Rabinowitz (1999).

In order to evaluate the effect of the three RFs on ILDs, ILDs are shown in Fig. 5.9 for the 2D and the 3D array. The MTreg RF provides a substantial ILD increase between 250 Hz to f_{lim} compared to for the non-compensated HOA (dashed curve), whereas the other RFs do not show any significant ILD for frequencies lower than

500 Hz. Between about 1.5 kHz and f_{lim}, 3D NFC-HOA provides larger ILDs than 2D NFC-HOA and ILDs obtained with the coswin RF were consistently larger than with the other two RFs.

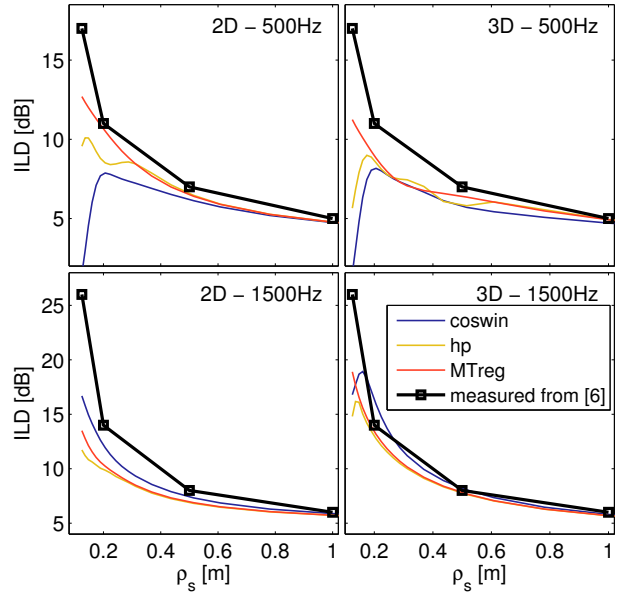

Figure 5.10: ILDs versus distance for a source at azimuth $\theta_s = 90°$ at 500 Hz (top) and 1.5 kHz (bottom) for simulated 2D (left) and 3D (right) arrays for NFC-HOA with the three considered RFs. Black squares represent measured data from Brungart and Rabinowitz (1999).

For both 2D and 3D cases, these simulated ILDs are lower than real ILDs measured by Brungart and Rabinowitz (1999) with the KEMAR manikin for a physical point source (square markers). ILDs naturally result from the amplitude decay of the sound field (proximity effect) and the diffraction of the sound by the head (Brungart and Rabinowitz, 1999). In Fig. 5.8, the interaction of the head with reconstructed standard HOA sound fields (here $\rho_s = 1.4$ m) compared to the real sound fields (far-field) leads to similar ILDs (within 1 dB at $\theta_s = 90°$) for frequencies below 1.5 kHz for $M \geq 4$. Even though the interaction with NFC-HOA sound fields potentially differs from this value, the lower simulated ILDs by about 3 dB compared to real ILDs at 600 Hz for the 2D case with MTreg for a close source ($\rho_s = 0.15$ m) can thus mainly

be explained by the lower simulated amplitude decay shown in Fig. 5.6, which was, between the two ear positions, 2.4 dB lower than the ILDs for the ideal point source (with a $1/r$ decay).

Simulated ILDs of a lateral virtual source ($\theta_s = 90°$) are plotted against distance in Fig. 5.10 and compared to ILDs measured by Brungart and Rabinowitz (1999) . At 500 Hz, simulated ILDs with the MTreg RF show similar behavior as real ILDs, but with lower values (especially in 3D). For very close sources, the two other RFs fail to provide monotonously increasing ILDs with decreasing distance, as seen in Fig. 5.9 at low frequencies. At 1.5 kHz, simulated ILDs better match real ILDs in the 3D representation and the coswin RF provides the largest ILDs for distances below 0.5 m.

In order to analyze the obtained ILDs from a reproduced NFC-HOA source at different azimuth angles, ILDs are plotted for all directions in Fig. 5.11 for distances of 0.12, 0.2 and 0.5 m for both a 2D (left side) and a 3D (right side) representation. For clarity, the hp RF is not considered here. At 500 Hz, the trend described above for a lateral source, i.e., largest ILDs for 2D and for the MTreg RF, also applies to other azimuth angles. At 1.5 kHz, measured ILDs change non-uniformly with the direction of the virtual source with peaks around 45° and 120° for the closer sources. The ILD pattern follows best the one for real sources (black markers) when the coswin RF is used with the 2D representation (from 0.2 m and above) whereas, the MTreg RF provides the best performance for the 3D representation where very large ILDs of up to 30 dB can be observed.

In synopsis, NFC-HOA with the modified Tikhonov RF leads to an increase of ILDs over a large frequency range for decreasing source distance. The use of the coswin RF leads to noticeably larger ILDs for very close sources at mid-frequencies (from around 1.5 kHz to f_{lim}) but is unable to provide low-frequency ILDs.

5.4.2 Acoustically-damped listening environment

In practical situations, loudspeaker arrays are placed in acoustically-damped rooms rather than in anechoic rooms. In order to investigate how room reverberation disturbs the provided ILD cues, binaural room impulse responses were recorded for each loudspeaker of the “SpaceLab” facility using a head and torso simulator (HATS, from Brüel and Kjær) placed at the center of the array. The SpaceLab facility at the Cen-

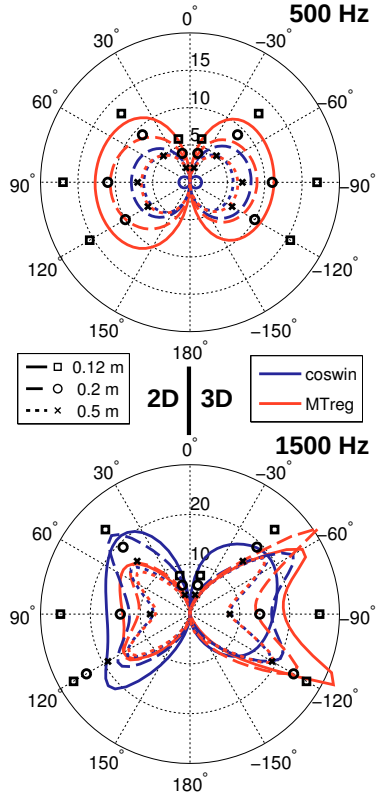

Figure 5.11: ILD versus azimuth angle for three distances at 500 Hz (top) and 1.5 kHz (bottom) for 2D (left side) and 3D (right side) arrays. Markers represent ILDs measured with real sources from Brungart and Rabinowitz (1999).

tre for Applied Hearing Research (CAHR) consists of an acoustically-damped room ($4.5 \times 4.4 \times 2.5$ m) with a reverberation time of $T_{30} = 0.16$ s at 125 Hz and below 0.1 s for higher frequencies. In this room, 29 loudspeakers were placed at a distance of $R = 1.8$ m from the center point. The loudspeaker array consisted of a 16-loudspeaker horizontal ring (at elevation 0°), two 6-loudspeaker rings at elevation of -34° and +37° and one loudspeaker at the zenith (Favrot and Buchholz, 2010). The ring at elevation 0° was used for the 2D HOA with $M = 7$ and all loudspeakers were used for the 3D HOA with $M = 4$.

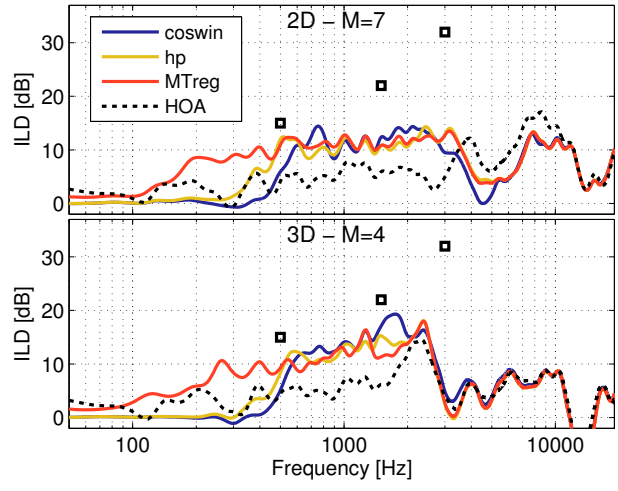

Figure 5.12: ILDs measured with a virtual lateral source ($\rho_s = 0.15$ m, $\theta_s = 90°$) reproduced with the three RFs (solid curves) and without any near-field compensation (dashed curves) in an acoustically damped room.

The ILDs for a lateral virtual source placed in the distance of 0.15 m are plotted in Fig. 5.12. Obtained ILDs show spectral ripples, which resulted from the soft reflections in the reverberant playback room. ILD values are generally lower than in anechoic environments and drastically drop above f_{lim}. The impact of the different RFs on ILDs is similar as in anechoic environments. Lower ILDs than in anechoic conditions are also observed for other azimuth angles (Fig. 5.13). At 500 Hz, the same impact of the two RFs on ILDs is found whereas, at 1.5 kHz, the coswin RF provided the largest ILDs in the 3D representations with most variation with distance.

To conclude, nearby sources reproduced with 2D or 3D NFC-HOA with the MTreg RF provided increasing ILDs with decreasing distance to a centered listener for a large frequency range (between about 250 Hz and f_{lim} for a source at 0.15 m), although to a lesser extent than real physical sources. At mid-frequencies (between 750 Hz and 3 kHz), ILDs were larger in 3D than in 2D representations. The coswin RF showed much smaller ILDs at lower frequencies than the MTreg RF, but it showed larger ILDs at frequencies around 1.5 kHz. Generally, the ILDs were noticeably lower

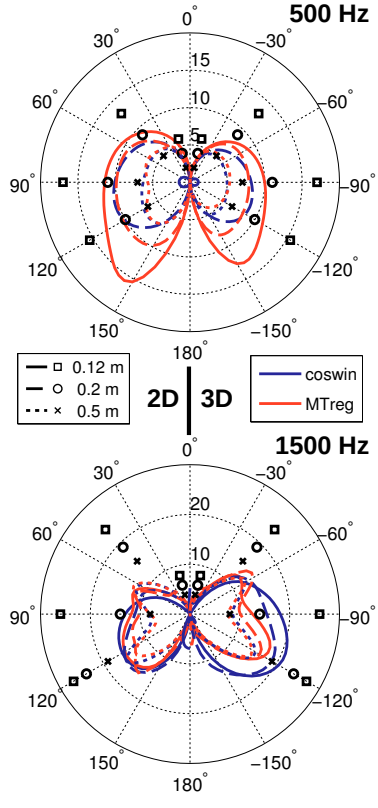

Figure 5.13: ILD versus azimuth angle for three distances at 500 Hz (top) and 1.5 kHz (bottom) for 2D (left) and 3D (right) arrays in an acoustically damped room. Same markers as in Fig. 5.11.

when measured in an acoustically-damped room compared to that derived from situations for free field.

5.5 Subjective evaluation

The increase of low-frequency ILDs (up to 3 kHz) with decreasing distance has been shown to be the most important cue for distance perception of nearby lateral sound sources in anechoic environments (Brungart, 1999). In contrast, interaural time differ-

ences have only a minor impact. In the previous section, it was shown that NFC-HOA techniques can provide substantially increased ILDs with decreasing distances for different frequency ranges depending on the applied RF.

A listening test was conducted in order to (i) assess if the distance of sources reproduced with NFC-HOA with RFs can be similarly discriminated as for real sources and to (ii) evaluate the impact of the different RFs on distance perception of nearby sound sources. The test consisted in comparing the distance perception of frontal nearby sources, where no ILD cues are available, to that of lateral sources, which naturally provide strong ILD cues.

5.5.1 Methods

The experiment was performed by using the same set-up as described to measure ILDs in section 5.4.2. Six normal hearing test-subjects participated in the experiment.

Nearby sound sources were reproduced at different distances either in front (0°) or on the left side (90°) of the listener with NFC-HOA, using either the coswin or the MTreg RF. In a 2D condition, 16 loudspeakers were used to realize 7^{th}-order NFC-HOA, and in a 3D condition, 29 loudspeakers were used to realize 4^{th}-order NFC-HOA. In total, the experiment consisted of 8 conditions, each corresponding to one azimuth (0° or 90°), one RF (coswin or MTreg) and one loudspeaker array (2D or 3D). Each of these conditions were realized by 6 blocks of 24 stimuli at random distances selected on a logarithmic axis between 0.125 and 1.7 m. In addition, 4 stimuli at a distance of the loudspeaker array (R =1.8 m) were considered, corresponding to standard far-field HOA (where the distance encoding was not used). The 48 blocks were presented in a random order.

During the experiment, subjects were seated in the center of the loudspeaker array. The height of the non-rotatable seat was set such that his/her ears were at the height of the horizontal 2D loudspeaker ring (at elevation 0°). In order to avoid any interference by visual cues and to focus the subject's attention on the auditory modality, subjects were blindfolded during the experiment. After each stimulus presentation, test-subjects were asked to orally give an estimate of the distance of the auditory event in centimeters relative to the center of their head. Subjects were asked to answer 10 cm when they perceived the stimulus inside their head. No feedback was provided to the

subjects. The operator was present in the playback room (outside the loudspeaker array) during the whole experiment to note down the estimated distances on a touch screen and to monitor the position of the subject's head. Before each session, subjects participated in a training session consisting of 4 blocks of 28 stimuli. Each block of 28 stimuli lasted approximately 5 min and subjects participated in 3 sessions of 1.5 h including training and breaks.

The virtual sound sources played a train of five 150-ms long bursts of pink noise, separated by pauses of 30 ms. The SPL of the stimuli at the center of the loudspeaker array was roved over a 12-dB range (between 46 and 58 dB SPL) to avoid the influence of natural amplitude cues which could otherwise affect the provided ILD cues by NFC-HOA. Loudspeaker equalization was performed on the loudspeaker array to flatten each of the loudspeaker frequency responses recorded at the center of the array.

5.5.2 Results

The logarithmic distance estimates ρ' of the subject 'TL' are plotted against the logarithmic source distance ρ in Fig. 5.14 for the eight stimulus conditions.

In order to assess the distance perception performance of the subjects, a power-law function ($\rho' = k\rho^a$) was fitted to the raw data and plotted as solid lines. Correlation coefficients were computed between log distance estimates and log stimulus distances and represent to what extent distance estimates can be determined by a power-law function. More generally, they represent the distance perception performance of the subject, i.e., the "reliability" of the distance estimate. The closer the correlation coefficient is to one, the better the distance perception performance is (see Brungart *et al.*, 1999 for details). Typically, data from this experiment as well as data from the experiment in Brungart *et al.* (1999) show correlation coefficients following the same variations as fitted power-law exponents. The correlation coefficient approach was used in this study to allow for a direct comparison to the results from Brungart (1999).

Results presented in Fig. 5.14 show positive correlation coefficients (0.28 to 0.56) for lateral sources and negative correlation coefficients (-0.15 to -0.38) for frontal sources. An overview of the correlation coefficients of all subjects is presented in Fig. 5.15.

It should be noted that inside-the-head perception (distance estimates of 10 cm)

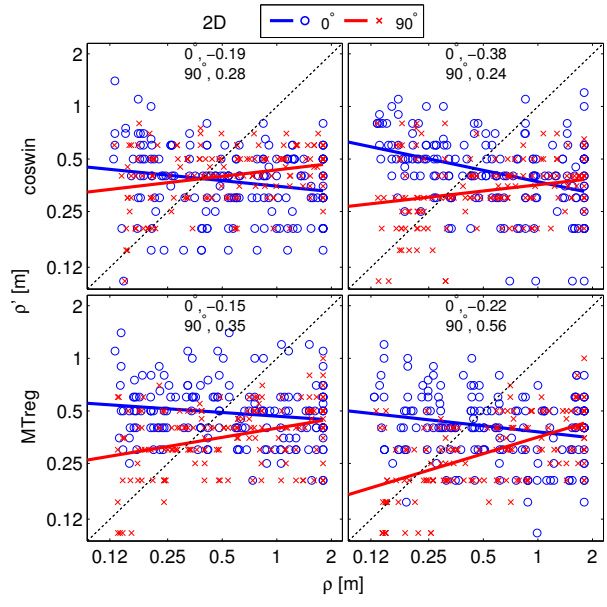

Figure 5.14: Distance estimations for the subject 'TL'. The solid lines represent the fitted power-law function. The number represents the correlation coefficient between the log estimated distances ρ' and the log stimulus distances ρ for the 0° and the 90° condition.

occurred mainly for one subject ('MM') who perceived around 10 % of the frontal stimuli inside his head. While subjects indicated that the task was rather simple to perform, some subjects reported a change of the perceived stimulus direction within a block, especially for the blocks with frontal sources. This change of perceived direction mainly appeared as front-back confusions which made the subjects task more difficult.

Although correlation coefficients vary across subjects and conditions, they were always negative for frontal sources (0°). For lateral sources, correlation coefficients were generally higher than for the frontal conditions, but always lower than 0.6, indicating a rather poor distance perception performance for all subjects.

Correlation coefficients were averaged across subjects for the eight stimulus conditions using the Fisher transform (Devore, 1991) and plotted in Fig. 5.15 (Mean). For frontal sources, mean correlation coefficients are very similar for both RF and

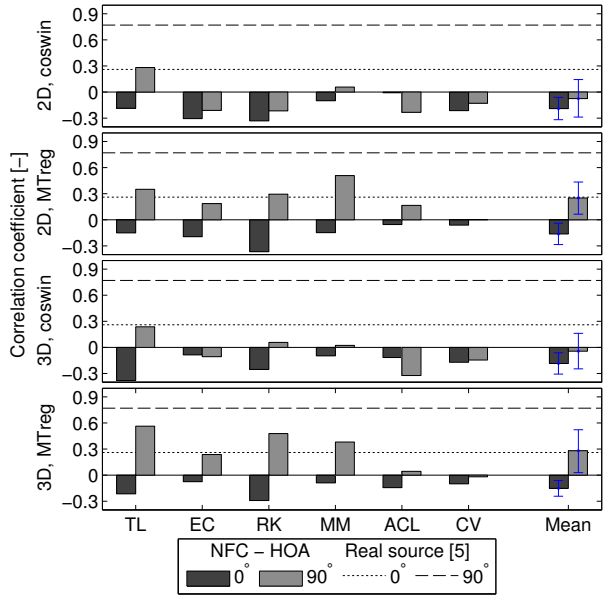

Figure 5.15: Correlation coefficients between the estimated distance ρ' and the log stimulus distance ρ for all subjects and the mean across subjects (x-axis). Error bars indicate confidence intervals ($\alpha < 0.05$) for the mean across subjects. Horizontal lines indicates data for real sources are read from (Brungart, 1999, Fig. 8)

for 2D and 3D conditions and are significantly negative (overall mean -0.18). The low negative mean correlation coefficients indicate that distances were consistently poorly estimated and that further sources appeared slightly closer than nearby sources. Therefore, the use of the two RFs seems to lead to artifacts which were interpreted as "opposite" distance cues.

For lateral sources, with the coswin RF, correlation coefficients were not significantly different from zero (overall mean -0.06) showing that NFC-HOA with this RF was unable to provide any noticeable auditory distance cues. Although, as shown in section 5.4.2, this method provided some low-frequency ILDs from 500 Hz to 3 kHz (but to a lesser extent than real sources), this cue was apparently not salient enough to provide any distance discrimination with random levels. A further analysis showed that distance estimates were correlated with the stimulus level both in the 2D (-0.54)

and 3D (-0.60) condition indicating that subjects were judging distances mostly relying on the random amplitude cues. Considering the MTreg RF, correlation coefficients for lateral sources were significantly positive with a mean value across subjects of 0.26 for the 2D condition and of 0.30 for the 3D condition. In addition, for these lateral sources, the mean of the difference between the correlation coefficients for the MTreg and the coswin RFs was significantly positive both in the 2D (0.33) and the 3D (0.32) conditions. Hence, although the amplitude cues were conflicting with the stimulus distance (applied to the NFC filters), subjects used extra auditory cues, i.e., which appear to be low-frequency ILDs, to estimate the distance of the stimuli. The mean of the difference of the correlation coefficients between the 2D and 3D conditions was not significantly different from zero for both the coswin and the MTreg RFs.

5.5.3 Discussion

In order to compare the results described in section 5.5.2 to those obtained with real physical nearby sources, correlation coefficients measured by Brungart (1999, Fig. 8) for similar source positions are plotted in Fig. 5.15 (horizontal lines). It should be noted that many parameters in Brungart's experiment differed from the present study, such as the azimuth and elevation range of the stimulus and the distance reporting method. Therefore, the results of this study should only be regarded as an indication of distance perception performance for real sources.

Correlation coefficients measured for frontal NFC-HOA sources were significantly negative (overall mean -0.18) and lower than for real sources (0.26, see Fig. 5.15). NFC-HOA, as well as plane-wave HOA, leads to spectral artifacts, particularly at high frequencies above f_{lim} (see section 5.3.3), which might have been interpreted, at least for some subjects, as opposite distance cues. The interaction of the arriving sound with the listener's pinna at these high frequencies (above 4 or 5 kHz) is crucial for the perception of direction and externalization (Blauert, 1997). This means that the incorrect sound field reproduction of NFC-HOA in the sweet spot at high frequencies has likely disturbed the perceived stimuli direction and might have led to inside-the-head perception. This was sometimes reported afterwards by subjects and has potentially disturbed their distance perception judgments. Further investigations are needed to assess the impact of these spectral artifacts as well as the perceived di-

rection localization. For lateral sources, mean correlation coefficients with the RF for 2D (0.26) and 3D (0.30) were considerably lower than for real sources (0.77 as shown in Fig. 5.15). This observation is somehow expected from the lower measured ILDs with these RFs compared to real ILDs, which is mainly due to the combination of the imperfect sound field amplitude decay and the reverberation of the listening room as shown in section 5.4.2.

The NFC-HOA technique was able to provide some relevant distance information for lateral sources when the MTreg RF was used, although to a lesser extent than for real sources. Moreover, NFC-HOA associated with the MTreg RF led to significantly higher correlation between the estimated and reproduced distances than with the coswin RF. Since the main difference between the two considered RFs (according to section 5.4.2) is that the MTreg RF provides significant ILD cues down to about 250 Hz (and the coswin RF only down to about 500 Hz), it can be concluded that ILD cues below about 500 Hz appear to be very important for auditory perception of very close sound sources (at least for virtual sources). Hence, in NFC-HOA, such low-frequency ILD cues need to be provided, which requires a significant contribution of higher-order Ambisonics components at low frequencies. Although section 5.4 showed larger ILDs in the 3D condition than in the 2D condition for frequencies between 750 Hz and 3 kHz, which was expected to enhanced distance perception performance Brungart (1999), this did not lead to a significant improvement in listener's distance discrimination performance in 3D over 2D NFC-HOA.

For real-time implementation, Daniel (2003) developed infinite impulse response (IIR) filters for NFC filters (without RF). This parametric implementation provides a lower computing cost compared to the use of finite impulse response (FIR) filters that would need to be recomputed for each new distance parameter. However, straightforward IIR implementation of the MTreg RF does not seem realizable. Instead, an FIR solution should be developed with similar characteristics as the RF, i.e., with low cut-off frequencies and steep slopes. The NFC-HOA approach with RF does not seem to be able to provide substantially larger ILDs (to match real ILDs) as lowering the RF cut-off frequencies would eventually lead to an unnatural low-frequency boost. However, the loudness increase induced by this low-frequency shaping may provide a distance cue, which might be useful in some auditory display applications that do not

strictly require a realistic reproduction of nearby sound sources (Favrot and Buchholz, 2009a).

Instead of the RF-based NFC-HOA realization for nearby sound sources, alternative approaches should be considered which inherently do not have the low-frequency amplification problem. For instance, explicit modeling of focused sources (Ahrens and Spors, 2009) and removing the evanescent contribution of the reproduced sound field that was shown to be responsible for the excessive low frequency energy (Spors and Ahrens, 2010) have been proposed. The performance of these methods in terms of ILDs and distance perception needs to be evaluated in a similar fashion as described in this paper.

Moreover, it should be noted that only distance perception for simulated anechoic sound sources has been considered. In many applications, room reflections might be additionally simulated, which will also have an influence on distance perception (Shinn-Cunningham *et al.*, 2005).

5.6 Conclusion

The reproduction of nearby sound sources to a single listener placed in a center of the loudspeaker array using near field compensated higher-order Ambisonics requires the use of regularization functions in practical realizations. The impact of two existing and one novel RFs on (i) the reproduced sound field, (ii) the auditory distance cues (i.e., low-frequency ILDs), and (iii) the relative distance perception have been investigated in this study. The main outcomes of this study are summarized as follows:

1. The point source sound field curvature was appropriately reproduced in the sweet spot when either of the three RFs was used, both in 2D and 3D representations. NFC-HOA associated with the MTreg RF correctly reproduced the sound field over the largest area, and moreover matched the best the pressure field amplitude decay of a point source.

2. The proposed MTreg RF for NFC-HOA provided for very close sources significant low-frequency ILDs down to about 250 Hz, which, below about 750 Hz, were following similar variations with distances than for real nearby sources,

albeit with slightly lower values. In contrast, the other considered RFs failed to provide any significant ILDs below 500 Hz. The availability of this important low-frequency distance cue was confirmed by dummy-head recordings in a practical loudspeaker installation for different azimuth angles. In addition, it was found that 3D NFC-HOA representations provided larger ILDs than 2D representations for frequencies between 750 Hz and 3 kHz (even though a higher order was used in 2D).

3. A low correlation between the listener's perceived distances and the simulated distances of lateral sources at random level was observed when the MTreg RF was used, indicating that NFC-HOA with this RF provided salient ILD cues. The use of the MTreg RF led to significantly higher correlation between estimated and reproduced distances than when using the coswin RF, which did not provide any ILD cues below about 500 Hz for very close sources. This highlights the importance of low-frequency ILDs between 250 and 500 Hz for nearby distance perception. The listening test was not able to show any significant influence of using 3D representations compared to 2D representations. By comparing the distance perception of virtual nearby sources reproduced by NFC-HOA (using the MTreg RF) with the perception of real sources (Brungart, 1999), it was found that the distance of virtual sources were perceived less accurately and less precisely than corresponding physical sources when amplitude cues were not available.

Acknowledgment

The authors would like to thank Sandra Rodiño Palacios for organizing the test-subject appointments and carrying out the experiment. The authors would like to thank the reviewers for further improving the quality of the manuscript.

6 General discussion

The development and evaluation of a loudspeaker-based room auralization (LoRA) system[1] were presented in this thesis. The aim was to realize an auralization system that can be used to investigate auditory functions in reverberant environments as well as for testing of hearing instruments.

The LoRA system was described in chapter 2. Considerations about the auralization technique (higher-order Ambisonics) and auditory perception (precedence effect) led to a solution where the early and late part of the room impulse response (RIR) are processed separately. In particular, the late reverberation part is processed to ensure incoherent loudspeaker signals in order to reduce coloration artifacts. An objective evaluation of the LoRA system was carried out (section 2.3) by means of room acoustic parameters such as reverberation time, clarity, strength, speech transmission index and interaural cross correlation coefficient (IACC). Simulations using a 29-loudspeaker array showed that the characteristics of the input RIR provided by the room acoustic model were preserved by the involved signal processing of the system. The special processing of the late reverberation part was able to provide natural diffuse reverberation (low IACC) when a sufficient number of loudspeakers was used.

In the LoRA system, the auralization method of the early part of the RIR can either be higher-order Ambisonics or single loudspeaker. For many applications, the use of the single loudspeaker auralization method is recommended for far field sources and when a large number of loudspeakers is available. The source direction should be chosen to match the direction of one loudspeaker of the array. However, the single

[1] The LoRA toolbox written for Matlab is freely available at `http://www.dtu.dk/centre/cahr/English/Downloads.aspx` and includes functions to compute multichannel RIRs and to perform loudspeaker equalization (see appendix B).

loudspeaker method can not be used for the reproduction of very close sources or in a possible future implementation of moving sources in the system.

In order to assess the auditory perception in VAEs reproduced by the LoRA system, two subjective evaluations were carried out. The impact of the Ambisonic order used in the early part of the RIR on speech intelligibility was investigated in chapter 3. The direct sound and the early reflections were either auralized with 1^{st}- or 4^{th}-order HOA or with a single loudspeaker (the closest to the reflection direction). The latter auralization technique (also mentioned in Seeber *et al.*, 2010), is an alternative to HOA auralization and can more accurately reproduce the important characteristics of direct sound avoiding the high-frequency limitation of HOA (see section 2.2.2). The listening test showed that HOA-auralized speech provided less intelligibility than single loudspeaker presented speech and that a decrease in Ambisonic order further decreased speech intelligibility. Moreover, HOA-auralized early reflections provided a similar enhancement of speech intelligibility as early reflections presented by single loudspeakers. These two results indicated that the HOA-auralized speech level can be equalized to provide the same intelligibility as that obtained with single loudspeaker presented speech, even for first order Ambisonics. Therefore, speech intelligibility experiments can be conducted with the LoRA system either with HOA or single-loudspeaker auralization for the first part of the RIR. The required level equalization may suggest that the lower intelligibility of HOA-auralized speech is linked to its lower perceived loudness.

The second subjective evaluation of this thesis was presented in chapter 4. Distance perception in VAEs reproduced by the LoRA system (with Ambisonic 4^{th}-order for the first part of the RIR) was compared to that obtained in corresponding real environments. Since the use of non-individualized binaural room impulse response (BRIR) measurements in reverberant environments does not significantly modify auditory distance perception (Zahorik, 2002b), dummy-head recorded BRIRs provided in this experiment the reference condition representing the real environment. This test showed that, for far field sources, the LoRA system provides VAEs that create a similar distance perception as that in corresponding real environments. The important distance cue in these conditions, the direct-to-reverberant energy ratio, was found to influence distance perception similarly in both auralization techniques. This in-

dicated that this cue is not significantly degraded by the overall chain of processing from the acoustic room model to the computation of the mRIRs by the LoRA system. This result is consistent with the non degradation of the early-to-late energy ratio C_{80} parameter by the LoRA processing (section 2.3). This experiment also investigated the distance perception of nearby sources auralized by near field compensated (NFC) HOA. This preliminary study showed that NFC-HOA nearby sources were perceived within the loudspeaker array, both in anechoic and reverberant environments.

In practical realizations, NFC-HOA requires the use of angular weighting windows (AWWs) to avoid excessive energy in the loudspeaker signals. In the study presented in chapter 4, the weighting window consisted of a regularization function that limited NFC filter gains below +30 dB. Further analyses showed that these weighted NFC filters led to an increase of energy at low frequencies at both ears with decreasing source distance, of about 20 dB for frequencies below 300 Hz and for sources at 60 cm. This bass-boost did not contribute to an increase of low-frequency ILDs, i.e., the main auditory cue for nearby distance perception, but probably produced an unnatural distance cue in this experiment.

A deeper investigation of NFC-HOA and the role of the angular weighting windows on distance perception of nearby sources was presented in chapter 5. A novel regularization function was proposed, this time restricting NFC filter gains below 0 dB. The impact of this AWW as well as of two existing AWWs on reproduced sound fields and ILDs at the listener's ears was investigated. Sound field simulations showed that the proposed regularization function allowed for a better reproduction of the curvature and amplitude decay of the sound field than the other two windows. In both anechoic and acoustically-damped rooms, this AWW provided increased ILDs with decreasing distance for frequencies above 250 Hz for very close sources whereas the two other AWWs only showed an ILD increase for frequencies above 500-600 Hz. A listening experiment showed that, with the proposed AWW, virtual lateral sources reproduced by NFC-HOA were perceived nearly as accurately as corresponding real ones (measured in Brungart, 1999) when natural amplitude cues were provided. When these natural amplitude cues were not available, distances were still discriminated albeit to a lesser extent than for real sources. The use of another AWW did not allow for

the distance discrimination of lateral sources indicating the importance of large ILDs between 250 Hz and 600 Hz for nearby distance perception.

This NFC study focused on the reproduction of a single virtual source. When reverberant environments such as classrooms are considered, Shinn-Cunningham *et al.* (2005) showed that low-frequency ILDs are robust to the effect of reverberation. They suggested that distance perception of nearby sources may be improved in rooms compared to anechoic environments since reverberant energy may provide additional cues. Therefore, the use of NFC-HOA in the LoRA system may show a better distance perception performance compared to single NFC-HOA sources. Even though this result was already shown in section 4.3.3, it was probably due to the effect of the previously mentioned bass-boost which was induced by the use of an inappropriate AWW.

In order to validate the use of the LoRA system for a larger range of applications, additional subjective evaluations are required. The directional aspect of sound localization, in addition to the dimension of distance, is required to be accurately reproduced in many applications. The accuracy of the reproduction of the direct sound being crucial in direction localization (cf. precedence effect) suggests that the use of a single loudspeaker for the direct sound could be preferable to HOA, which degrades localization cues at high-frequencies (above f_{lim} cf. section 2.2.2). It should be noted that head rotations, if allowed in the experiment, will help the listener in localization tasks (direction and distance). Another evaluation is required for the use of hearing instruments such as binaural hearing aids in LoRA generated VAEs. These instruments often combine information from two or more microphones for spatial processing such as beamforming, which needs to provide similar performance in VAEs than in real environments.

Complex VAEs (with several static sources) can be obtained with the LoRA system by superposing multichannel RIRs. However, the creation of very complex scenes with many sources, such as in a train station or in a supermarket, represents a significant effort. Such environments often include moving sources which, so far, can not be reproduced with the LoRA system. The implementation of such feature would first require the use of a room acoustic model that is able to update the RIR (at least the early part) as the source is moving and this would rule out the use of the single loudspeaker auralization method. Another approach to reproduce very complex auditory

scenes is to use a spherical microphone array to record HOA signals. Recording of real scenes could then be played back on a loudspeaker array. While modifications of the acoustic scene would not be possible, this system could be used to reproduce very complex auditory scenes which, for example, could be used for testing or fitting hearing instruments. For the evaluation of such a system, the LoRA framework can be used for systematically studying the impact of the Ambisonics order on the perception of the scene. In particular, the subjective evaluation of the impact of 1^{st}- and 4^{th}-order HOA auralization on speech intelligibility presented in this thesis is relevant in this context.

A

Listening room characteristics

The SpaceLab facility at the Centre for Applied Hearing Research (CAHR) in DTU consists of an acoustically-damped room ($4.5 \times 4.4 \times 2.5$ m) and a 29-loudspeaker array. A picture of the room (after the acoustic treatment) can be found on the cover of this thesis.

A.1 Loudspeaker setup

The loudspeaker array consists of a 16-loudspeaker horizontal ring, two 6-loudspeaker rings (at -34° and +37° elevation) and one loudspeaker at the zenith at a radius of $R = 1.8$ m (see Fig. 2.6, p. 26). The loudspeakers are passive 2-way studio monitor BM6P from Dynaudio. The signal-flow diagram of the SpaceLab is shown in Fig. A.1.

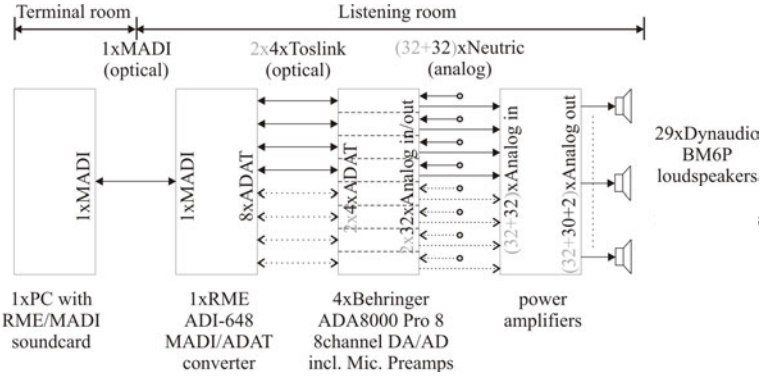

Figure A.1: Signal-flow diagram of the SpaceLab facility. The PC is located outside the actual listening room due to the produced noise.

A.2 Acoustic treatment

The room was acoustically treated to lower its reverberation time in order to minimize the surface reflections while reproducing VAEs. Acoustical damping of the room surfaces is described below.

A.2.1 Walls

All 4 walls of the room were covered with 125 mm thick Basotect material from BASF, which is an open cell foam made from melamine resin. The absorption coefficient α for perpendicular sound incident measured in an impedance tube according to ISO 10534-2 (1998) using the transfer function method is shown in Fig. A.2 (magenta, squares).

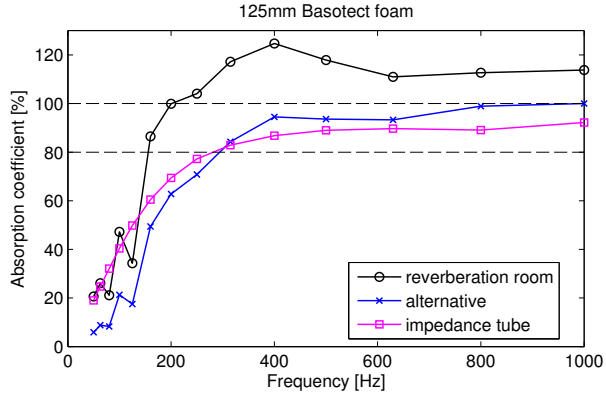

Figure A.2: Absorption coefficient α measured in the impedance tube for perpendicular sound incidence.

The absorption coefficient measured in the reverberation chamber for diffuse sound incident according to ISO 354 (1985) is shown in Fig. A.2 (black, circles). In addition, the alternative absorption coefficient is shown in Fig. A.2 (blue, crosses), which is derived from the "standard" absorption coefficient by applying the transformations proposed by Thomasson (1980, 1982). This alternative absorption coefficient takes into account the limited area of the absorber specimen, which results in much more realistic absorption coefficients, in particular at low frequencies. Based on the

impedance tube measurements and the alternative absorption coefficient, it can be concluded that the applied porous wall absorber achieves significant absorption (i.e., $\alpha \geq 80\%$) for frequencies down to about 300 Hz.

A.2.2 Ceiling

The ceiling is suspended by about 37 cm below the actual concrete ceiling using a thin metal grid. In this ceiling cavity directly on top of the metal grid, three layers of 50 mm thick glass wool (i.e., a total thickness of 150 mm) of type Isover diffuse ruller 320 were installed. This material is highly porous with a density of about 13 kg/m^3 and a specific flow resistance of about $5 \cdot 10^3$ kg/(s m)^3. This ceiling will be approximated here as a 150 mm thick porous layer with a 220 mm air gap towards the hard concrete ceiling. Knowing the characteristic impedance W_a and the propagation coefficient γ_a of the porous material, the absorption coefficient of the suspended ceiling can be estimated according to Mechel (1989, Band 1, p. 41). Considering porous mineral wool absorbers, Delany and Bazley (1970) provide approximations for W_a and γ_a that are purely based on the specific flow resistivity. The estimation of the absorption coefficient for the given ceiling is shown in Fig. A.3, which provides significant absorption down to about 100 Hz.

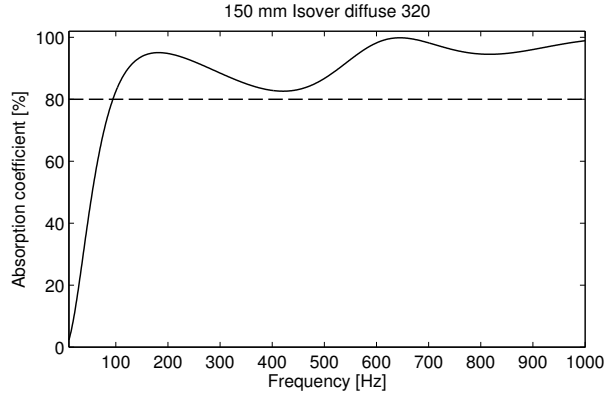

Figure A.3: Estimates absorption coefficient of the suspended ceiling with 150 mm of glass wool mounted with an air gap of 220 mm.

A.2.3 Floor

The floor consisted of PVC glued onto concrete and covered by a 1 cm thick woolen carpet with “cut threats” (i.e., not looped) stitched onto a thin layer of felt.

A.2.4 Room corners

In order to improve low frequency absorption, 1.2 m wide wooden frames filled with 125 mm thick Basotect foam was placed into three corners of the room, ranging from the floor to the suspended ceiling.

A.3 Reverberation time

The reverberation time of the listening room was measured for each loudspeaker of the array in the center of the room before and after the use of absorption material on the different surfaces of the listening room. The mean reverberation time measured from the first 30 ms (T_{30}) and the mean early decay time (EDT) across loudspeaker response are plotted in Fig. A.4. The use of the different absorption material led to a

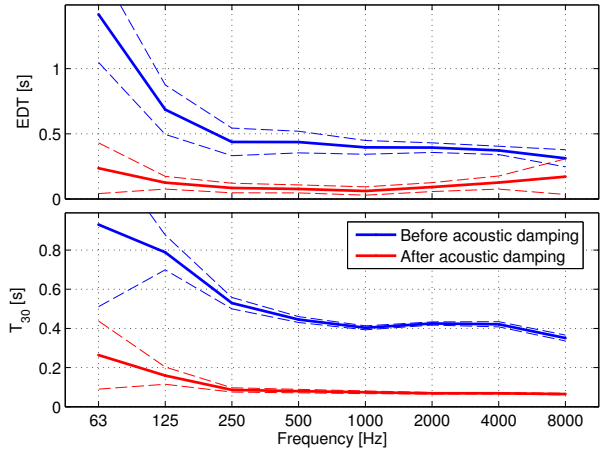

Figure A.4: Mean reverberation time measured before and after the acoustic treatment of the listening room across loudspeaker responses.

large decrease of the reverberation time (T_{30} and EDT). The absorption material led to a T_{30} of 0.16 s at 125 Hz and below 0.1 s for higher frequencies. This low reverberation time makes the listening room suitable for sound field reproduction techniques such as higher-order Ambisonics.

B

Loudspeaker equalization

A loudspeaker equalization was performed before using the SpaceLab facility in order to (i) flatten the amplitude frequency response of each loudspeaker, to (ii) ensure that each loudspeaker provide the same sound pressure level at the center and to (iii) time-align the impulse response recorded at the center.

First, impulse responses were measured at the center of the array for each loudspeaker. A gammatone filter smoothing (Kohlrausch and Breebaart, 2001) of the amplitude frequency response was performed and plotted in the top panel of Fig. B.1 for each loudspeaker n.

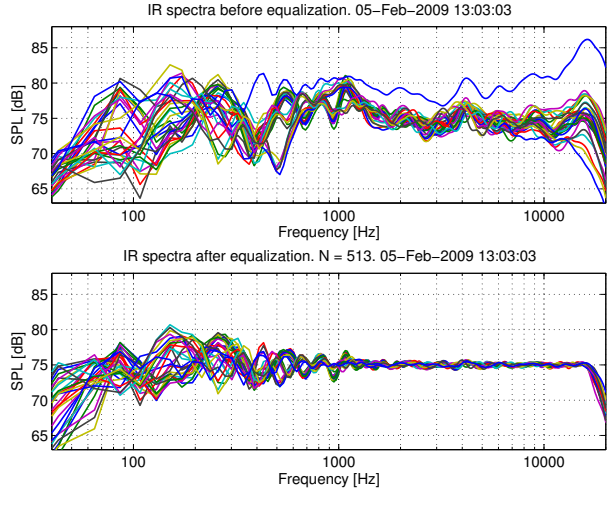

Figure B.1: Amplitude frequency response for all loudspeakers in the SpaceLab set-up before (top) and after (bottom) equalization.

In order to derive an equalization filter to flatten loudspeaker amplitude responses,

the mean m_H of the smoothed amplitude response $|H_n(f)|$ was first computed across loudspeaker responses and across frequencies between 100 Hz and 16000 Hz. This frequency range roughly corresponds to the bandwidth of the loudspeakers. Then, a finite impulse response (FIR) filter $H_{eq}(f)$ was derived from the inverse of the normalized smoothed amplitude responses according to:

$$H_{eq}(f) = \frac{m_H}{|H_n(f)|} \tag{B.1}$$

A 602-point minimum phase version of this filter was obtained with the real cepstrum method (Oppenheim and Schafer, 1975). Each loudspeaker filter was delayed such that each loudspeaker direct sound arrived at the exact same time at the center of the array. This was required to compensate for the small loudspeaker distance differences, especially for the zenith loudspeaker located 50 cm closer to the center than the other ones.

The smoothed equalized amplitude responses are plotted in the bottom panel of Fig. B.1. From about 1 kHz, the amplitude response are deviating ± 1 dB from the mean value. The presence of the listener typically results in larger variations in the amplitude response at the ears. The obtained equalization filter was convolved with the loudspeaker signals before being played.

Bibliography

Ahrens, J. and Spors, S. (**2009**). "Spatial encoding and decoding of focused virtual sound sources", in *Proc. of the 1st Int. Ambisonics Symposium*.

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

Azzali, A., Bilzi, P., Carpanoni, E., and Farina, A. (**2005**). "Comparison of different listening systems for speech intelligibility tests", in *Proc. of the 118th Audio Eng. Soc. Conv.*, volume preprint 6356.

Begault, D. R. (**1994**). *3-D sound for virtual reality and multimedia*, NASA technical memorandum edition (AP Professional, Cambridge, MA).

Begault, D. R., Wenzel, E. M., and Anderson, M. R. (**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.", J. Audio Eng. Soc. **49**, 904.

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

Bertet, S. (**2009**). "Formats audio 3d hierarchiques : caracterisation objective et perceptive des systemes ambisonics d'ordres superieurs", Ph.D. thesis, Insitut national des sciences appliquées de Lyon, France.

Bertet, S., Daniel, J., Gros, L., Parizet, E., and Warusfel, O. (**2007**). "Investigation of the perceived spatial resolution of higher order ambisonics sound fields: A subjective evaluation involving virtual and real 3d microphones", in *Proc. of the 30th Audio Eng. Soc. Int. Conf.*

Blauert, J. (**1997**). *Spatial hearing: the psychophysics of human sound localization*, third edition (MIT Press, Cambridge, MA).

Blauert, J. (**2005**). *Communication acoustics* (Springer Berlin Heidelberg New York).

Bork, I. (**2000**). “A comparison of room simulation software - the 2nd round robin on room acoustical computer simulation”, Acta Acustica **86**, 943–956.

Bork, I. (**2005**). “Report on the 3rd round robin on room acoustical computer simulation - part ii: Calculations”, Acta Acustica united with Acustica **91**, 753–763.

Bradley, J., Sato, H., and Picard, M. (**2003**). “On the importance of early reflections for speech in rooms”, J. Acoust. Soc. Am. **113**, 3233.

Bradley, J. S., Reich, R., and Norcross, S. (**1998**). “A just noticeable difference in C50 for speech”, Appl. Acoust. **58**, 99–108.

Bradley, J. S. and Soulodre, G. A. (**1995**). “The influence of late arriving energy on spatial impression”, J. Acoust. Soc. Am. **97**, 2263–2271.

Brand, T. and Kollmeier, B. (**2002**). “Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests”, J. Acoust. Soc. Am. **111**, 2801.

Brungart, D. (**1999**). “Auditory localization of nearby sources. III. Stimulus effects”, J. Acoust. Soc. Am. **106**, 3589.

Brungart, D., Durlach, N., and Rabinowitz, W. (**1999**). “Auditory localization of nearby sources. II. Localization of a broadband source”, J. Acoust. Soc. Am. **106**, 1956.

Brungart, D. and Rabinowitz, W. (**1999**). “Auditory localization of nearby sources. Head-related transfer functions”, J. Acoust. Soc. Am. **106**, 1465.

Buchholz, J. M., Mourjopoulos, J., and Blauert, J. (**2001**). “Room masking: understanding and modelling the masking of room reflections”, Proc. of the 110th Audio Eng. Soc. Conv. .

Christensen, C. L. (**2007**). *Odeon Room Acoustics Program, User Manual*, version 9.0 edition (Odeon A/S, Lyngby, Denmark).

Cox, T. J., Davies, W. J., and Lam, Y. W. (**1993**). “The sensitivity of listeners to early sound field change in auditoria”, Acustica **79**, 27–41.

Dalenbäck, B. I. (**1996**). “Room acoustic prediction based on a unified treatment of diffuse and specular reflection”, J. Acoust. Soc. Am. **100**, 899–909.

Daniel, J. (**2000**). “Representation de champs acoustiques, application a la transmission et a la reproduction de scenes sonores complexes dans un contexte multimedia (in french)”, Ph.D. thesis, 1996-2000 Universite Paris 6.

Daniel, J. (**2003**). “Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format”, in *Proc. of the 23rd Audio Eng. Soc. Int. Conf.*

Daniel, J. and Moreau, S. (**2004**). “Further study of sound field coding with higher order ambisonics”, in *Proc. of the 116th Audio Eng. Soc. Conv.*

Daniel, J., Nicol, R., and Moreau, S. (**2003**). “Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format”, in *Proc. of the 23rd Audio Eng. Soc. Int. Conf.*

Delany, M. and Bazley, E. (**1970**). “Acoustical properties of fibrous absorbent materials”, Appl. Acoust. **3**, 105–116.

Devore, J. (**1991**). *Probability and Statistics for Engineering and the Sciences* (Brooks-Cole, Belmont, MA, USA).

Djelani, T., Porschmann, C., Sahrhage, J., and Blauert, J. (**2000**). “An interactive virtual-environment generator for psychoacoustic research II: collection of head-related impulse responses and evaluation of auditory localization”, Acustica **86**, 1046–1053.

Farina, A. (**1993**). “Convolution of anechoic music with binaural impulse responses”, in *Proc. of the PARMA-CM Users Meeting*.

Farina, A. (**2000**). “Validation of the pyramid tracing algorithm for sound propagation outdoors: comparison with experimental measurements and with the ISO-DIS 9613 standards”, Advances in Eng. software **31**, 241–250.

Favrot, S. and Buchholz, J. (**2009**a). “Distance Perception in Loudspeaker-Based Room Auralization”, in *Proc. of the 127th Audio Eng. Soc. Conv.*, volume preprint 7854.

Favrot, S. and Buchholz, J. (**2009**b). “Validation of a loudspeaker-based room auralization system using speech intelligibility measures”, in *Proc. of the 126th Audio Eng. Soc. Conv.*, volume preprint 7763.

Favrot, S. and Buchholz, J. (**2010**). “Lora: A loudspeaker-based room auralization system”, Acta Acustica united with Acustica **96**, 364–375.

Favrot, S. and Buchholz, J. M. (**2012**). “Reproduction of nearby sound sources using higher-order ambisonics with practical loudspeaker arrays”, Acta Acustica united with Acustica .

Fellgett, P. B. (**1974**). “Ambisonic reproduction of directionality in surround-sound systems”, Nature **252**, 534–538.

Frank, M., Zotter, F., and Sontacchi, A. (**2008**). “Localization experiments using different 2d ambisonics decoders”, in *Proc. of the 25th Verband Deutscher Tonmeister Int. Conv.*

Gardner, B. and Martin, K. (**1994**). “HRTF measurements of a KEMAR dummy-head microphone”, MIT Media Lab Perceptual Computing - Technical Report .

Gardner, W. G. and Martin, K. D. (**1995**). “HRTF measurements of a Kemar”, J. Acoust. Soc. Am. **97**, 3907–3908.

Gerzon, M. A. (**1973**). “Periphony - with-height sound reproduction”, J. Audio Eng. Soc. **21**, 2–10.

Gerzon, M. A. (**1992**). “General metatheory of auditory localisation”, in *Proc. of the 92nd Audio Eng. Soc. Conv.*, preprint 3306.

Hirst, J. M., Davies, W. J., and Philipson, P. J. (**2006**). “Adaptation of concert hall measures of spatial impression to reproduced sound”, in *Proc. of the 120th Audio Eng. Soc. Conv.*, volume preprint 6732.

Hollerweger, F. (**2005**). “Periphonic sound spatialization in multi-user virtual environments”, Master’s thesis, Institute of Electronic Music and Acoustics.

ISO 10534-2 (**1998**). “Acoustics - determination of sound absorption coefficient and impedance in impedance tubes - part 2: Transfer-function method”, .

ISO 3382 (**1997**). “Acoustics - measurement of the reverberation time of rooms with reference to other acoustical parameters”, .

ISO 354 (**1985**). “Acoustics - measurement of sound absorption in a reverberation room considering amendment 1: Test specimen mountings for sound absorption tests.”, .

Kleiner, M., Dalenbäck, B. I., and Svensson, P. (**1993**). “Auralization - an overview”, J. Audio Eng. Soc. **41**, 861 – 875.

Kohlrausch, A. and Breebaart, J. (**2001**). “Perceptual (ir) relevance of HRTF magnitude and phase spectra”, in *Proc. of the 110th Audio Eng. Soc. Conv.*, preprint 5406.

Krokstad, A., Strom, S., and Sørsdal, S. (**1968**). “Calculating the acoustical room response by the use of a ray tracing technique”, J. Sound Vib. **8**, 118–125.

Kuster, M. (**2009**). “Combining Methods for Multichannel Room Impulse Response Generation and Objective and Subjective Performance Evaluation”, J. Audio Eng. Soc. **57**, 512–520.

Kuttruff, H. (**1991**). *Room Acoustics*, third edition (Taylor & Francis, London).

Lehnert, H. and Blauert, J. (**1992**). “Principles of binaural room simulation”, Appl. Acoust. **36**, 259–291.

Lindau, A., Hohn, T., and Weinzierl, S. (**2007**). “Binaural resynthesis for comparative studies of acoustical environments”, in *Proc. of the 122nd Audio Eng. Soc. Conv.*, volume preprint 7032.

Litovsky, R. Y., Colburn, H. S., Yost, W. A., and Guzman, S. J. (**1999**). “The precedence effect”, J. Acoust. Soc. Am. **106**, 1633–54.

Lochner, J. P. A. and Burger, J. F. (**1964**). “The influence of reflections on auditorium acoustics”, J. Sound Vib. **1**, 426–448.

Lokki, T. (**2002**). “Physically-based auralization Ű design, implementation, and evaluation”, Ph.D. thesis, PhD thesis, Helsinki University of Technology.

Malham, D. G. and Myatt, A. (**1995**). “3-D sound spatialization using ambisonic techniques”, Computer Music Journal **19**, 58 – 70.

Mechel, F. (**1989**). *Schallabsorber, Band I-III* (Hirtzel Verlag, Stuttgart, Germany).

Merimaa, J. and Pulkki, V. (**2005**). “Spatial impulse response rendering i: analysis and synthesis”, J. Audio Eng. Soc. **53**, 1115 – 1127.

Miller, J. D. (**2001**). “Slab: A software-based real-time virtual acoustic environment rendering system”, in *Proc. of the Int. Conference on Auditory Display*.

Minnaar, P., Olesen, S. K., Christensen, F., and Møller, H. (**2001**). “The importance of head movements for binaural room synthesis”, in *Proc. of the 2001 Int. Community for Auditory Display*, volume 21-25.

Møller, H. (**1992**). “Fundamentals of binaural technology”, Appl. Acoust. **36**, 171–218.

Moreau, S., Daniel, S., and Bertet, S. (**2006**). “3d sound field recording with higher order ambisonics”, in *Proc. of the 120th Audio Eng. Soc. Conv.*, volume preprint 6857.

Morse, P. and Feshbach, H. (**1953**). *Methods of theoretical physics.* (McGraw-Hill).

Nábělek, A. and Robinette, L. (**1978**). "Influence of the precedence effect on word identification by normally hearing and hearing-impaired subjects", J. Acoust. Soc. Am. **63**, 187.

Naylor, G. M. (**1993**). "Odeon - another hybrid room acoustical model", Appl. Acoust. **38**, 131 – 143.

Nicol, R. and Emerit, M. (**1999**). "3D-Sound Reproduction Over An Extensive Listening Area, A Hybrid Method Derived From Holophony And Ambisonic", in *Proc. of the 23rd Audio Eng. Soc. Int. Conf.*, 436–453.

Nielsen, S. (**1993**). "Auditory distance perception in different rooms", J. Audio Eng. Soc. **41**, 755–755.

Oppenheim, A. and Schafer, R. (**1975**). *Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, USA).

Orfanidis, S. J. (**1995**). *Introduction to signal processing* (Prentice Hall, Upper Saddle River, NJ).

Poletti, M. A. (**2005**). "Three-dimensional surround sound systems based on spherical harmonics", J. Audio Eng. Soc. **53**, 1004–1025.

Pulkki, V. (**1997**). "Virtual sound source positioning using vector base amplitude panning", J. Audio Eng. Soc. **45**, 456–466.

Pulkki, V. (**2007**). "Spatial sound reproduction with directional audio coding", J. Audio Eng. Soc. **55**, 503–516.

Rindel, J. H. (**2000**). "The use of computer modeling in room acoustics", J. Vibroengineering **3**, 41–72.

Rindel, J. H., Nielsen, G. B., and Christensen, C. L. (**2009**). "ODEON website", www.odeon.dk .

Seeber, B., Kerber, S., and Hafter, E. (**2010**). "A system to simulate and reproduce audio-visual environments for spatial hearing research", Hearing Res. **260**, 1–10.

Shinn-Cunningham, B., Kopco, N., and Martin, T. (**2005**). "Localizing nearby sound sources in a classroom: Binaural room impulse responses", J. Acoust. Soc. Am. **117**, 3100.

Shirley, B., Kendrick, P., and Churchill, C. (**2007**). "The Effect of Stereo Crosstalk on Intelligibility: Comparison of a Phantom Stereo Image and a Central Loudspeaker Source", J. Audio Eng. Soc. **55**, 852.

Solvang, A. (**2008**). "Spectral impairment for two-dimensional higher order Ambisonics", J. Audio Eng. Soc. **56**, 267–279.

Spors, S. and Ahrens, J. (**2010**). "Reproduction of focused sources by the spectral division method", in *Proc. of the 4th IEEE ISCCSP*.

Spors, S., Ahrens, J., Wierstof, H., and Geier, M. (**2009**). "Physical and perceptual properties of focused sources in wave field synthesis", Proc. of the 127th Audio Eng. Soc. Conv. .

Thomasson, S. (**1980**). "On the absorption coefficient", Acustica **44**, 265–273.

Thomasson, S. (**1982**). "Theory and experiments on the sound absorption as function of the area", Report No. TRITA-TAK, Department of Technical Acoustics, Royal Institute of Technology, Stockholm, Sweden **8201**.

Vogel, C. R. (**2002**). *Computational methods for inverse problems* (Society for Industrial and Applied Mathematics).

Vorlander, M. (**1995**). "International round robin on room acoustical computer simulations", in *Proc. of the 15th Int. Congress Acoust.*

Vorländer, M. (**2008**). *Auralization: fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*, volume 1 of *RWTHedition* (Springer).

Wagener, K., Josvassen, J., and Ardenkjær, R. (**2003**). "Design, optimization and evaluation of a Danish sentence test in noise", Int. J. Audiol. **42**, 10–17.

Ward, D. B. and Abhayapala, T. D. (**2001**). "Reproduction of a plane-wave sound field using an array of loudspeakers", IEEE T audio speech **9**, 697–707.

Wenzel, E., Arruda, M., Kistler, D., and Wightman, F. (**1993**). "Localization using nonindividualized head-related transfer functions", J. Acoust. Soc. Am. **94**, 111.

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

Zahorik, P. (**2002**a). "Assessing auditory distance perception using virtual acoustics", J. Acoust. Soc. Am. **111**, 1832.

Zahorik, P. (**2002**b). "Auditory display of sound source distance", in *Proc. of the Int. Conf. on Auditory Display*, 326–332.

Zahorik, P., Brungart, D., and Bronkhorst, A. (**2005**). "Auditory distance perception in humans: A summary of past and present research", Acta Acustica united with Acustica **91**, 409–420.

Zurek, P. M. (**1987**). “The precedence effect”, in *Directional Hearing*, edited by W. A. Yost and G. Gourevitch, 85–105 (Springer-Verlag, New York).

Contributions to Hearing Research

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

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

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

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

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

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

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

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

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

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

In complex acoustic environments, such as a train station or a café, hearing-impaired people often experience difficulties to communicate even when wearing hearing instruments, whereas normal-hearing people are typically able to communicate without effort in such conditions. In order to systematically study the signal processing of realistic sounds by normal-hearing and hearing-impaired listeners, a flexible, reproducible and fully controllable auditory environment is needed. A loudspeaker-based room auralization (LoRA) system was developed in this thesis to provide virtual auditory environments (VAEs) with an array of loudspeakers. The LoRA system combines state-of-the-art acoustic room models with sound-field reproduction techniques. Limitations of these two techniques were taken into consideration together with the limitations of the human auditory system to localize sounds in reverberant environments. In order to assess the usability of the LoRA system, one objective and two subjective evaluations were carried out. Beside investigating the auditory system, such virtual auditory environments (VAEs) are also relevant for evaluating and optimizing hearing instruments and communication devices.

DTU Electrical Engineering
Department of Electrical Engineering

Ørsteds Plads
Building 348
DK-2800 Kgs. Lyngby
Denmark
Tel: (+45) 45 25 38 00
Fax: (+45) 45 93 16 34
www.elektro.dtu.dk

ISBN 978-87-92465-23-8