Master Thesis

Auditory object recognition of normal hearing and hearing impaired listeners in virtual acoustic environments

Author:	*Sascha L. Reidt*
Supervisors:	*Prof. Norbert Dillier*
	Dr. Wai Kong Lai
Track advisor:	*Prof. Janos Vörös*

UZH, Laboratory of Experimental Audiology

ETHZ, Institute for Biomedical Engineering

April 2014

Abstract

Common evaluation tests of hearing impairments and their corresponding solutions with hearing instruments in a clinical setting often ignore many important real-life aspects. These real-life challenges can be reproduced in a controlled laboratory setting using virtual acoustic environments. In this Master thesis a system to create realistic acoustic scenes in free field out of a surround sound recording has been implemented and validated. The multi-channel recordings have been reproduced by using a white noise calibration method and inverse filters. Objective evaluation showed good spatial resolution and a flat frequency response. Subjective listening tests aimed at quantifying the binaural localization ability of static and dynamic auditory objects in potentially dangerous traffic situations. It has been demonstrated that the results of the listening tests are reproducible and can be used to differentiate between normal hearing listeners and listeners with a simulated conductive hearing loss.

Acknowledgements

I would like to thank my supervisors Prof. Norbert Dillier and Dr. Wai Kong Lai and my track advisor Prof. Janos Vörös for their great support and for making it possible to work at the Laboratory of Experimental Audiology of the University Hospital Zurich. I am also grateful to the whole LEA team for many recreational coffee breaks in an enjoyable atmosphere. Special thanks go to Dr. Eleftheria Georganti for helping with many technical questions and to Andrea Kegel for assisting the psycho-acoustical experiments. Last but not least I would like to thank all the volunteers that participated in one of the pilot experiments.

Preface

This Master Thesis is part of my graduate study at the Institute for Biomedical Engineering at the Swiss Federal Institute of Technology (ETH).

I certify that this Master Thesis, and the research to which it refers, is the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referring practices of the discipline.

Sascha L. Reidt

Author:	Sascha L. Reidt	slreidt@gmail.com
Supervisors:	Prof. Norbert Dillier	norbert.dillier@usz.ch
	Dr. Wai Kong Lai	waikong.lai@usz.ch
Track advisor:	Prof. Janos Vörös	janos.voros@biomed.ee.ethz.ch

Table of Contents

List of Figures

List of Tables

Abbreviations

ANOVA	Analysis of variance
AP	Action potential
CI	Cochlear implant
dB	Decibel
DFT	Discrete Fourier transform
DSP	Digital signal processing
fb	front-back
FFT	Fast Fourier transform
HRIR	Head related impulse response
HRTF	Head related transfer function
ID	Interaural difference
IHC	Inner hair cells
ILD	Interaural level difference
ITD	Interaural time difference
MAA	Minimum audible angle
MAMA	Minimum auditory movement angle
MLS	Maximum length sequence
MSE	Mean squared error
OHC	Outer hair cells
RIR	Room impulse response
RMS	Angular root mean squared error
RT	Reverberation time
SIL	Sound intensity level
SNR	Signal to noise ratio
SPL	Sound pressure level
STFT	Short-time Fourier transform
STI	Speech transmission index
SVD	Singular value decomposition
VAE	Virtual acoustic environment
VBAP	Vector based amplitude panning
XTC	Cross talk cancellation

1 Introduction

1.1 Motivation

The auditory system of a human listener performs multiple relevant tasks in daily life. It enables communication through spoken language, helps orienting in certain environments and is an important tool in identifying dangerous situations, for example warning signals or vehicle noise in traffic. An important mechanism of this ability is binaural auditory object detection and tracking. It makes use of binaural cues to identify, localize and track an auditory object. Different hearing impairments influence the use of binaural cues to varying degrees and so do hearing aids and cochlear implants. The ability to use these binaural cues is often tested in speech and localization tests in a controlled laboratory environment. However, these tests often ignore important aspects of real acoustic environments. Virtual acoustics in free field allows recreating realistic acoustic scenarios in a clinical setting and enables the development of a new test battery that quantifies performance of a listener in real life environments.

1.2 Aim

Realization of a free-field virtual acoustic environment (VAE) system for real recordings is the final goal of this project. The different acoustical properties of the original setting and the potential reproduction cabin have to be taken into account and relevant acoustical challenges for human listeners in daily life are identified, recorded and reproduced using said tool. This results in a new listening test that is evaluated and validated in a last step.

1.3 Methods

A multi-channel surround sound microphone is used to record realistic everyday scenarios. These scenarios include cocktail party, street or office environments.

Ideally the sound signal is divided into background and one or more targets that have to be identified by a test subject. Identification includes direction as well as intelligibility of speech. Performance of a subject can be classified by root mean azimuthal error and number of front-back misclassifications in direction and by number of errors in repetition of a spoken sentence.

Reproducing these scenarios in a multi-loudspeaker system requires a frequency dependent calibration of the microphone to obtain the effect of each azimuthal sound direction on each channel. This allows convolving an arbitrary number of microphone channels back to an arbitrary number of loudspeakers. Frequency response of each loudspeaker, the room and of the test subject within the room has to be taken into account and can be implemented using filters. Calibration has to be validated using test recordings by re-recording reconstructed signals and compare it to the original signal. Static recorded acoustic scenes are reconstructed using the calibration for further validation. Having a valid recording-reconstruction-system, different acoustic environments are recorded and carefully protocolled with a main focus on traffic scenarios.

1.4 Outline

This Master's thesis is divided into following chapters:

Theory: Gives an introduction into acoustics, perception of sound, source localization and speech intelligibility.

Realization: Explains the methods for implementing the free-field virtual acoustics system.

Evaluation: Describes the validation of the virtual acoustics system and shows results of performed listening tests.

Conclusion: Summarizes the key points and gives an outlook for future research.

2 Theory

2.1 Fundamental acoustics

This section gives an introduction to fundamental linear acoustics by presenting the most important equations, expressions and parameters.

2.1.1 Sound as a wave

Oscillations of pressure, density and particle velocity in matter are called sound. Pressure defines harmonic sound fields uniquely and can be linearly approximated by a second-order linear partial differential equation in a source free area [1]:

$$\Delta p = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} \tag{2.1}$$

Here, *p(x, t)* is the varying pressure, *x* is location, *t* is time and *c* the propagation velocity. Using a separation approach Equation 2.1 can be simplified.

$$\mathrm{p(x,t)} = \mathrm{p(t)p(x)}$$

$$\Rightarrow \quad \frac{\Delta p(x)}{p(x)} = \frac{1}{c^2 p(t)}\frac{d^2 p(t)}{dt^2}$$

$$\Rightarrow \quad \begin{cases} -k^2 := \dfrac{\Delta p(x)}{p(x)} \\ -k^2 := \dfrac{1}{c^2 p(t)}\dfrac{d^2 p(t)}{dt^2} \end{cases}$$

$$\Rightarrow \quad \begin{cases} (\Delta + k^2)\, p(x) = 0 \\ \left(\dfrac{d^2}{dt^2} + (kc)^2\right) p(t) = 0 \end{cases} \tag{2.2}$$

The first expression of 2.2 is the famous Helmholtz equation that can be solved by another variable separation in spherical coordinates (r, θ, φ):

$$p(x) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} (a_{lm} j_l(kr) + b_{lm} y_l(kr)) Y_l^m(\theta, \varphi) \qquad 2.3$$

Equation 2.3 describes a sound pressure wave as a superposition of spherical Bessel functions j_l and y_l with constants a_{lm} and b_{lm} multiplied by spherical harmonics Y_l^m. It can be shown that the latter span the Hilbert space of the square-integrable functions as an orthonormal basis and thus the angular components of an arbitrary wave can be written as a series of spherical harmonics. For a detailed analysis of the possible solution methods of the Helmholtz equation and their consequences the reader is referred to the text book by Jackson [2]. The second expression can be solved by a superposition of a normal and a back propagating plane wave in the time domain.

2.1.2 Sound as a particle

Particle wave dualism states that every particle can be described as a wave and vice versa. This means for certain applications that sound can be treated as a particle travelling in certain direction with certain energy. This approximation is usually valid for wavelengths much smaller than the geometry of the objects it interacts [3]. However, in audiology the important frequency bands are mostly in a range where wave effects dominate and should not be neglected.

2.1.3 Sound measures

To quantify a sound field different measures exist. It is common to define sound pressure level (SPL) of multiple sound waves p_i relative to a reference value p_{ref} in decibel (dB) [4]:

$$\mathrm{SPL} := 10\, log_{10}\left(\frac{\sum {p_i}^2}{{p_{ref}}^2}\right) dB \qquad 2.4$$

Usually particle velocity *v* is related to sound pressure *p* by the acoustic impedance

$$\mathrm{Z} := \frac{p}{vS} \tag{2.5}$$

where *S* is the surface area. The speed of sound, not to be confused with the particle velocity, is a function of temperature *T* and given by $c = \left(331.4 + \frac{0.6T}{^{\circ}C}\right) m/s$ in dry air [4]. The acoustic intensity $I := pv$ can be normalized also by a reference value by using the sound intensity level (SIL) [1]:

$$\mathrm{SIL} := 10\ log_{10}\left(\frac{\sum {I_i}^2}{{I_{ref}}^2}\right) dB \tag{2.6}$$

2.1.4 Sound representation

An audio signal can be illustrated by an oscillogram that is usually sound pressure as a function of time. However, in many applications it is favorable to have a better representation of the frequency components. Plotting the intensity of the short-time Fourier transform (STFT) is called a spectrogram. A STFT is a sequence of the discrete Fourier transform (DFT) of windowed time bins that overlap, usually computed using the fast Fourier transform (FFT) algorithm in digital signal processing (DSP) [5]. An illustration of a STFT is shown in Figure 2.1.

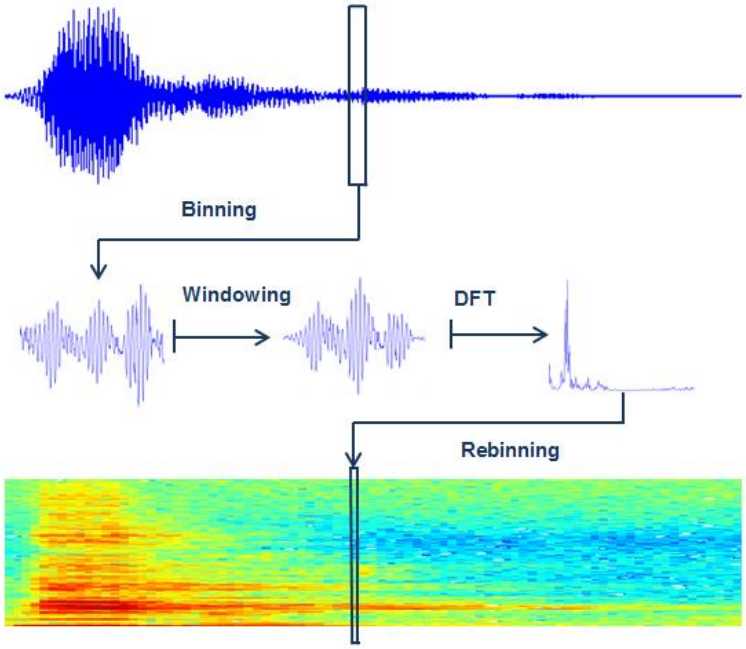

Figure 2.1: STFT of an audio signal to get spectrogram representation.

2.1.5 Room acoustics

Describing sound in free field is rather simple, but introducing reflecting components of a real room makes the problem more complex. The distribution of sound in a room depends mainly on the following:

- Geometry of room and obstacles
- Absorption coefficient of different materials
- Reflection coefficient of different materials
- Scattering on obstacles and non-uniformities
- Diffraction on obstacles

For simple boundary condition it is possible to solve the Helmholtz equation analytically by using Green's function [4] or directly by applying conditions to the general solution in Equation 2.3 However, in practical applications it is more common to use statistical methods for simulation or measurements to determine characteristic parameters of the room as described in [3] or [6]. The reaction of a room to a given acoustical stimulus is called room impulse response (RIR). The

room affects the amplitude of the stimulus in time as well as in frequency. When a sound source in a room is turned off, the sound wave gets reflected many times until it gets completely absorbed by the room. This effect is called reverberation. The time it takes for the SPL to decrease by 60 dB is called reverberation time RT_{60} and widely used to classify rooms quantitatively. An empirical formula for this measure given the volume V, surface S_i, and absorption coefficient α_i has been introduced by Sabin [4]:

$$RT_{60} \approx 0.161 \frac{V}{\sum \alpha_i S_i} \tag{2.7}$$

This formula only holds for uniformly distributed sound fields [7]. An actual measurement of RT_{60} can be performed by linear fitting of decay curves using a maximum length sequence (MLS) based method [4]. Since the influence of the room is highly frequency dependent, these decay curves vary in frequency.

2.2 Psychoacoustics

An important aspect that has to be considered in VAE's is the perception of sound by a target listener. In the following sections the hearing organ is described biologically (2.2.1 and 2.2.2) before introducing the concepts of source localization (2.2.3) and speech understanding (2.2.4) as presented in [4], [8], and [9].

2.2.1 Peripheral hearing organ

The peripheral hearing organ of humans is divided into three parts: The outer, middle and inner ear (Figure 2.2).

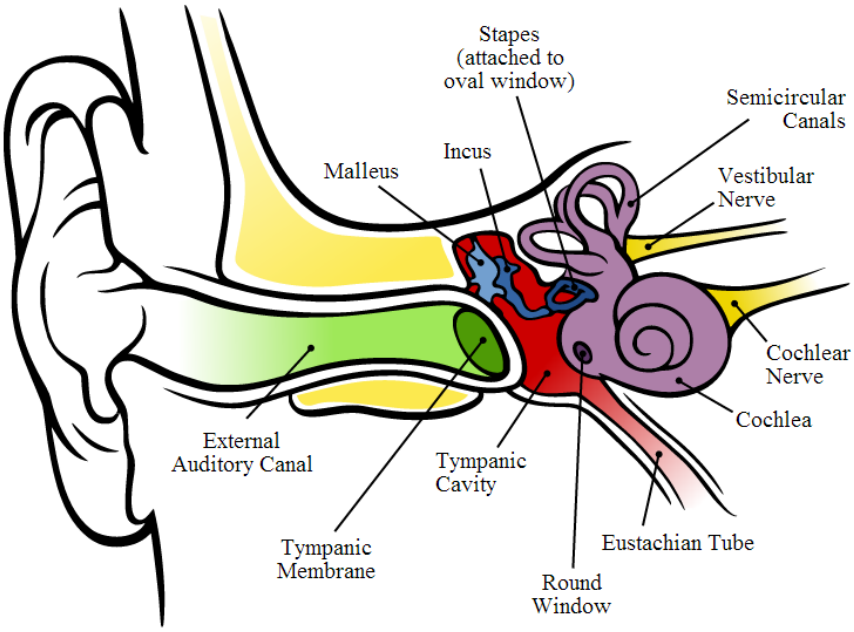

Figure 2.2: Biology of human peripheral hearing organ [10].

Outer ear: The outer part of the ear that surrounds the ear canal is called pinna. It gathers sound energy and directs it to the auditory canal. The folds of the pinna have a different impulse response for different sound directions. Additionally, sound frequencies between 3-12 kHz are amplified due to resonance in the auditory canal that ends at the tympanic membrane.

Middle ear: The acoustic impedance of air and liquid differ considerably. This means most of the sound energy would be reflected at an air-liquid interface. Since the outer and middle parts are filled with air but the inner ear with perilymph, humans and other mammals developed a mechanism to translate the energy via three vibrating bones to the oval window.

Inner ear: Vibrations of the oval window create pressure waves in one of the compartments of the inner ear, the scala vestibuli. The wave propagates along the coils of the cochlea to the apex, before it travels back in the scala tympani. The pressure is relieved at the round window. This process produces a force on the third compartment, the scala media, which bends the inner hair cells (IHC) of the organ of Corti. The position of the maximal force depends on the frequency of the pressure wave. The IHC's are the mechanosensors of the Cochlea,

translating the sound pressure into electrical nerve signals. The outer hair cells (OHC) on the other hand generate a voltage controlled force to amplify the signal.

2.2.2 Central processing

Hyperpolarization or depolarization of the IHC's affects the amount of neurotransmitter that is released. Neurotransmitters trigger an action potential (AP) at the spiral ganglion of the post-synaptic cell that is further transmitted by the auditory nerve. The average spike rate increases roughly logarithmically with sound pressure level for single tones. For low frequency tones the firing rate is phase locked.

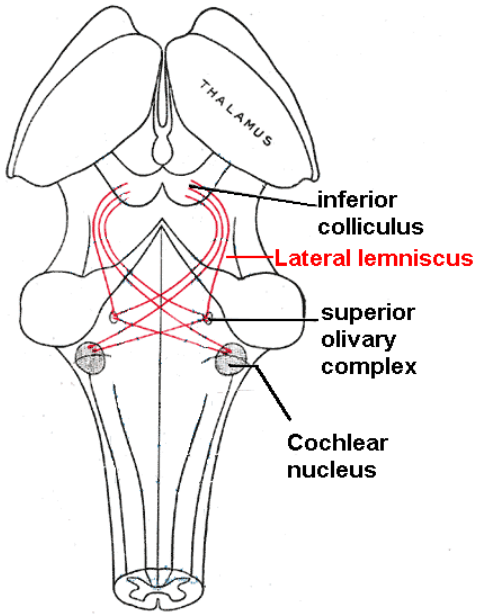

Figure 2.3: Auditory pathway in brainstem.

The signal arrives at the cochlear nucleus in the brainstem that projects to the superior olivary complex. Here, the signals of the two ears are combined and the brain can make use of binaural information. The cochlear nucleus also projects to the inferior colliculus via the lateral lemniscus (Figure 2.3) where the auditory information proceeds to the primary auditory cortex in the thalamus. In the primary auditory cortex and higher auditory areas the sound is processed and perceived consciously.

2.2.3 Sound source localization

An essential capability of our auditory system is localization of a sound source in all three dimensions. The brain is able to resolve direction as well as distance a sound comes from by making use of different cues coming from the head related impulse response (HRIR). Interaural differences (ID's) are cues that rely on the difference of the perceived signal between the two ears. The time of arrival of a sound wave coming from a certain horizontal direction varies between the two ears, resulting in an interaural time difference (ITD). Furthermore the head shadow will attenuate the signal at the ear further away from the source leading to an interaural level difference (ILD). Because sound waves with frequencies below 1000 Hz are diffracted strongly at the head, ITD's are mainly used for localization in this region [11]. Above these 1000 Hz, ILD's become more important as illustrated in Figure 2.4.

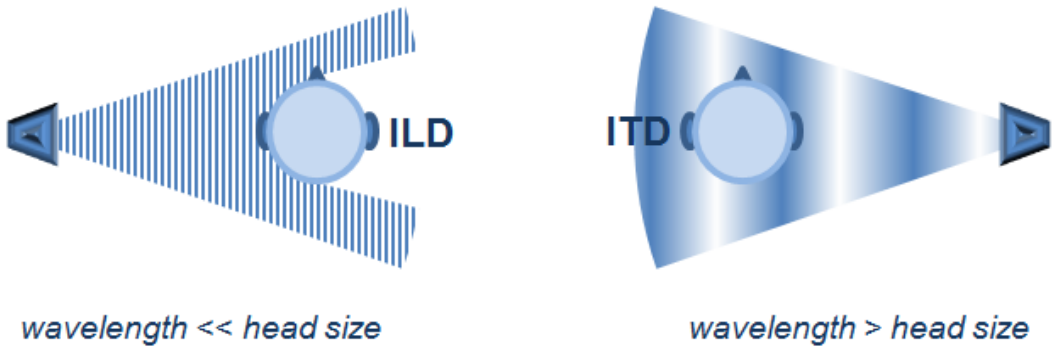

Figure 2.4: Influence of head shadow for two different frequencies of an incoming sound wave.

Obviously these one dimensional measures are not able to completely describe the three dimensional space, resulting in a cone of confusion that is defined as all potential sound source positions which would lead to identical ILD's and ITD's at the ears. In horizontal localization this cone of confusion has the consequence that a listener might confuse sound waves that come from the front with ones that come from the back and vice versa. Theoretically, these so called front-back confusions can be resolved by rotating and tilting the head for triangulation. Other cues include the impulse response of the head and the pinna. The shape of

these body parts adds spectral information to the signal depending on the angle of incidence. However, the most important cues for source localization are ID's, which are binaural effects. Asynchronous hearing impairments and audio processing in hearing aids may not preserve the ID's and negatively affect localization ability of a patient [12].

2.2.4 Speech intelligibility

Speech is a sequence of phonemes produced by the vocal chords and the oral cavity of humans for communication and speech understanding is considered the most important feature of the auditory system. Conversations are taking place in the frequency band 125 Hz to 4000 Hz [8]. The ability of the normal hearing human to resolve and understand speech in noise is called cocktail party effect originally proposed by Cherry [13] and still unmatched by any algorithm. Therefore one of the fundamental challenges in hearing aid and cochlear implant (CI) signal processing is to improve speech intelligibility [14].

2.3 Evaluation methods

Important aspects of audiology are methods to quantify hearing ability. Many different test batteries have been developed not only for diagnosis of hearing impairment, but also to evaluate the performance of hearing instruments. Examples of the latter are discussed in the following and can be divided into objective (2.3.1) and subjective measures (2.3.2).

2.3.1 Objective measurements

As the name suggests, these experiments deal with an object rather than an individual and are usually carried out prior to doing a subjective test in hearing aid or CI development. An example for an objective test is given by Keidser et al. in [12] where they measured ID's on a dummy head in free-field for different microphone placements and DSP strategies of hearing aids. Another approach would be to measure signal to noise ratio (SNR) difference between pre- and post-processed signals or the speech transmission index (STI).

2.3.2 Subjective measurements

In the end, hearing instruments have to be worn by a human listener which means the overall performance has to be optimized with respect to that. Subjective evaluation often aims at quantifying localization ability, speech intelligibility or quality of sound. Horizontal localization experiments often measure the angular root mean squared error (RMS) and the percentage of front-back confusions of a participant according to Equation 2.8 [15].

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{N} \sum (\arcsin(\sin(\theta_i)) - \arcsin(\sin(\hat{\theta}_i)))^2}{N}} \qquad 2.8$$

In this equation θ_i are the *N* azimuthal positions of a target signal and $\hat{\theta}_i$ the corresponding responses of a subject. The minimum angle that can be resolved by a human listener was introduced by Mills [16] and called minimum audible angle (MAA). In localization of dynamic targets subjects are often ask to determine whether a source has been moved which results in a minimum auditory movement angle (MAMA) [17]. Other researcher performed velocity estimation tests [18] or asked to indicate perceived location with a pointer [19]. The measurement of speech intelligibility is a more complex topic with countless different test batteries. A listener is often asked to identify phonemes, words or complete sentences in a certain noise level. The performance can be rated in percentage correct or speech reception threshold (SRT) that is defined as the SNR of 50 % correct repetitions. Listening effort is a term defining the cognitive work of a subject to perform a certain task. Possibilities to access this measure are dual tasks, pupil dilation, EEG activity or reaction times as reviewed in [20] and [21]. The disadvantage of the tests mentioned above is always the laboratory setting that is never equal to real-life. A way of evaluating certain hearing aids or CI's real-life performance are self-reports with standardized questionnaires, for example *Speech, Spatial, Quality* (SSQ) [22]. Another way to do so would be to create realistic acoustic scenes in the lab and evaluate the performance of a subject in different tasks.

2.4 Recording and Reproduction of auditory scenes

There exist different methods to generate an auditory scene for a human listener. Binaural reproduction aims at creating a desired audio signal at the ears of a listener, whereas the idea of surround sound is to create a desired sound field in space. The first one is introduced quickly in 2.4.1 and the second one in section 2.4.2.

2.4.1 Binaural approach

Binaural techniques are widely used and described in great detail in literature, for example in [23], [24] or [25]. They make use of the fact that theoretically any auditory scene can be reproduced for a human listener by controlling a processed acoustic signal at the eardrums. These acoustic signals can either be synthesized or recorded. In both cases the HRIR has to be taken into account. One way to do so is by recording binaurally using a dummy head like *TORONTO*, *VALDEMAR* or *KEMAR* with microphones mounted inside the ear canals. For non-binaurally recorded sounds or synthesized signals HRIR has to be measured and results in a head related transfer function (HRTF) that is used to reproduce said audio signal with headphones. Binaural reproduction via loudspeaker systems works in a similar fashion by cancelling cross-talk (XTC) as presented in [26] or [27].

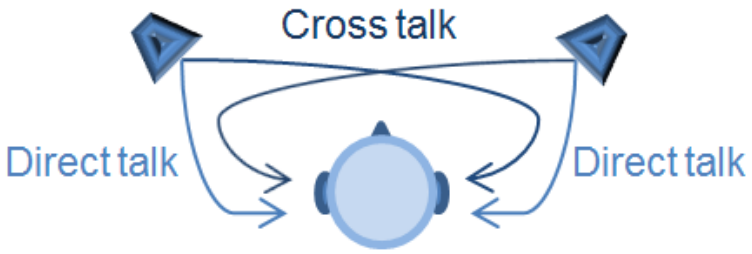

Figure 2.5: Idea of cross talk and direct talk of a binaural reproduction setup via two loudspeakers.

Binaural systems have two main disadvantages: Firstly, no HRIR of two different subjects are the same. A non-individualized binaural recording or HRTF resulting from a dummy head will never lead to an exact reproduction at the eardrums of a real human listener. Secondly, movements of the target subject have to be tracked and the audio signals processed in real time accordingly which is a complex DSP task.

2.4.2 Surround sound

Stereophony is the most popular sound reproduction format nowadays. Since in two-channel stereophony sound is only coming from the front, it is not suitable to create an auditory scene. The first attempt to extend a reproduced sound field into two dimensions has been made by four-channel quadrophony, the ancestor of 5.1 surround sound systems widely used in entertainment [24]. The setups of these systems are shown in Figure 2.6.

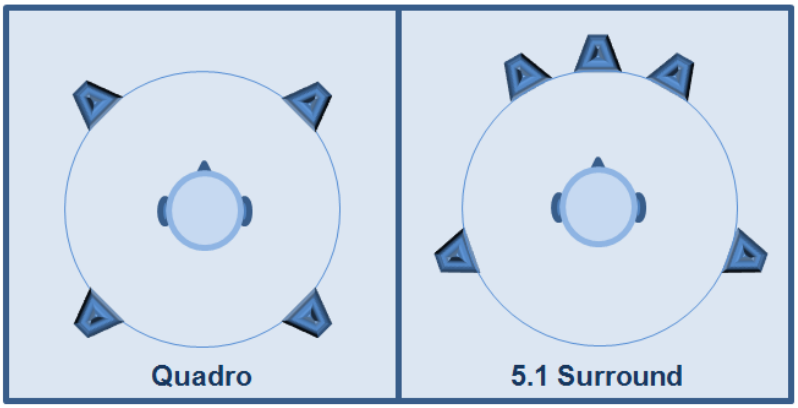

Figure 2.6: Setup for quadrophony and 5.1 surround sound.

However, these systems mainly focus on home applications. More advanced methods to produce multidimensional VAE's have been developed and three of them important for this project are introduced in sections 2.4.2.1 to 2.4.2.3. For an overview of the various techniques the reader is referred to [28].

2.4.2.1 Vector based amplitude panning

Many surround sound reproduction techniques rely on the amplitude panning principle. A virtual source is perceived between loudspeakers if one applies different gains to the same signal and plays it on the speakers simultaneously [29]. Vector based amplitude panning (VBAP) as proposed by Pulkki [30] defines these gains ***g*** as shown in Equation 2.9:

$$\boxed{\begin{aligned} \boldsymbol{g} &= \boldsymbol{p}^T L^{-1} \\ \boldsymbol{g}^T \boldsymbol{g} &= c \end{aligned}} \qquad 2.9$$

Vector ***p*** is the position of the virtual source, *L* the position of the closest loudspeakers arranged in a matrix and *c* is a normalization constant that defines the overall loudness of the signal and therefore determines the perceived distance of a source to some extent. Physical sound pressure decays with 1/r, but perceived sound distance is often a function of $1/r^{\alpha}$, where $\alpha < 1$ is a distance factor. Many other factors, especially reflection and reverberation, play also an important role in sound source distance perception [24].

2.4.2.2 Ambisonics

Ambisonics is a surround sound system initially proposed by Gerzon [31] in 1973. After specifying recording format and coding principles it became a flexible alternative to 5.1 surround working with an arbitrary number of loudspeakers even in three dimensions [32]. Ambisonics is built on the principle that an arbitrary sound field can be described as a superposition of spherical harmonics in angular direction, as shown in 2.1.1. First-order Ambisonics approximates these fields by the set of zero- and first-order spherical harmonics as a basis. Figure 2.7 represents this basis by spherical plots of the real values of the harmonics as a function of the angles, where blue parts illustrate positive values and red parts illustrate negative values.

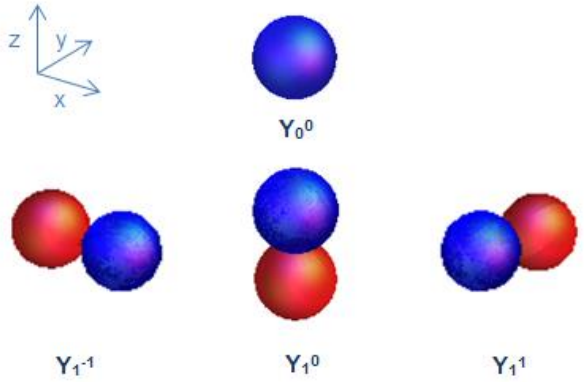

Figure 2.7: Zero- and first-order spherical harmonics.

In practice, a sound wave can be measured in this basis directly by three figure-to-eight in each Cartesian direction and one omni-directional microphone, because the pattern of this microphone configuration corresponds to the spherical harmonics shown in Figure 2.7. Such a recording *(X, Y, Z, W)* is called Ambisonics B-format and can be decoded for an arbitrary set of loudspeakers in three dimensions. Many different decoding strategies have been developed taking psychoacoustics and listening area into account, but all are based on a linear combination of the B-format. A simple decoder for a set of concentrically arranged loudspeakers *l*(θ, φ) on a sphere would be Equation 2.10:

$$\mathrm{l}(\theta,\varphi) = \begin{pmatrix} X \\ Y \\ Z \\ W \end{pmatrix}^T \begin{pmatrix} Re({Y_1}^{-1}(\theta,\varphi)) \\ Im({Y_1}^{1}(\theta,\varphi)) \\ {Y_1}^{0}(\theta,\varphi) \\ {Y_0}^{0}(\theta,\varphi) \end{pmatrix} = \begin{pmatrix} X \\ Y \\ Z \\ W \end{pmatrix}^T \begin{pmatrix} \sin(\theta)\cos(\varphi) \\ \sin(\theta)\sin(\varphi) \\ \cos(\theta) \\ 1 \end{pmatrix} \qquad 2.10$$

Equation 2.10 uses the spherical harmonics in the real basis, where the first two orders are identical to the spherical coordinates. Higher-order Ambisonics improves spatial resolution by including more directional components of the higher-order spherical harmonics. These do not longer correspond to real microphone patterns which make decoding and encoding more complex.

2.4.2.3 Microphone inversion

Another general approach to reproduce a recorded sound field is by inverting a multi-channel microphone. For example, Kahana [33] derives linear inverse filter sets from mean squared error (MSE) minimization based on earlier work by Kirkeby [34] and others. In this section the mathematical background of linear microphone inversion as described by Xie in [24] for general binaural reproduction is introduced: Let ***l*** be the signal of q loudspeakers, ***p*** the vector of sound pressures at n points and ***e*** an m-dimensional input signal. The relation between the input signal ***e*** and the loudspeaker signal ***l*** is called XTC matrix $A \in Mat(q,m)$ and the acoustic transmission matrix $H \in Mat(n,q)$ maps the loudspeaker signal ***l*** to the sound pressures ***p*** as showed below:

$$\begin{cases} \boldsymbol{l} = A\boldsymbol{e} \\ \boldsymbol{p} = H\boldsymbol{l} \end{cases}$$

$$\Rightarrow \qquad \boxed{\boldsymbol{p} = C\boldsymbol{e}} \qquad 2.11$$

This means $C = HA$ is a $n \times m$-matrix mapping an input signal ***e*** onto desired pressures ***p*** in space. In many applications it is desired to calculate matrix A to know how to mix a known input signal ***e*** for a set of loudspeakers to perform XTC. Depending on the dimensionality of this inversion problem different methods can be applied as listed in Table 2-1.

Dimensionality	Solution
$rank(H) = n = q$	Inversion
$rank(H) = min(n,q)$	Pseudo inversion using MSE minimization
$rank(H) < min(n,q)$	Singular value decomposition (SVD)

Table 2-1: Dimensionality of inversion problem and proposed solution method.

For detailed information about the individual solution method the reader is referred to a linear algebra text book.

3 Realization

3.1 Material

The technical devices and computer programs that have been used for the final version of the free-field virtual acoustics system are described in the following.

3.1.1 Recording devices

The acoustic scenes are recorded with a multi-channel microphone *Zoom H2N* after comparing it to a *SoundField ST250*. The *H2N* features Mid-Side (MS) microphones and 90° stereo (XY) channels allowing 4-channel horizontal surround recordings in the format (M, S, X, Y). All the recordings are taken with a standard sample rate of 44100 kHz and 16 Bit. According to the manufacturer the frequency response varies about ± 10 dB and it can be equipped with a tripod and a windshield. For exact measurements of SPL's a *B+K 2218* sound pressure meter is available. SPL over a certain time is monitored by a *Samsung® Galaxy S3* equipped with the *Noise Meter* application.

Figure 3.1: Zoom H2N four channel surround sound microphone mounted on a tripod.

3.1.2 Playback system

The sound signals are reproduced in *Room 7* of *ORL-LEA* at the *University Hospital Zurich* that is a shoebox type room with measured reverberation times RT_{60} as in Table 3-1.

Frequency in Hz	125	250	500	1000	2000	4000	8000
RT_{60} in ms	230	270	270	210	230	300	300

Table 3-1: Reverberation times RT60 as a function of frequency in Room 7 of ORL-LEA as measured in [35].

The loudspeaker system consists of twelve concentric *Genelec® Active Monitor 1029A* loudspeakers arranged on a circle with radius of 1.5 meter and equal angular spacing of 30°. These speakers are controlled by two *RME: Hammerfall DSP Multiface II* soundcards connected to a *Windows® XP SP3* workstation with *Intel® Core 2Duo* CPU at 2.66 GHz and 3.5 GB RAM. The two soundcards are synchronized via a world clock I/O in a master-slave mode. A *Windows® 7 SP1* computer with an *Intel® Core i7* CPU at 2.2 GHz and 8 GB RAM has been chosen for audio processing.

3.1.3 Feedback capturing

In subjective measurements feedback of a participant can be captured by an *ELO Touchscreen Control Panel* connected as a secondary monitor via VGA to the workstation. Head movements of the subject are monitored by *an Xsens MTx 3DOF Orientation Tracker* equipped with turn sensors, accelerometers and magnetometers via a serial port.

3.1.4 Software

The software for carrying out the experiments is written in *MATLAB® 8.1.0.604* 32-bit and 64-bit versions. *HoerTech SoundMexPro 1.5.4.13* is used as an interface between *MATLAB®* and the multi-channel *ASIO* driver for playback. Some computations and plots are made in *Wolfram® Mathematica 8.0.4.0*. Recording samples are cut and edited in *Audacity 2.0.5* and *Adobe® Audition*.

3.2 Virtual acoustic environment designer

In this section the calibration procedure of a given multi-channel recording to create a VAE using a microphone calibration method (Figure 3.2) is described in detail. A general derivation of the algorithm is given in 3.2.1 and the resulting program is described in 3.2.2.

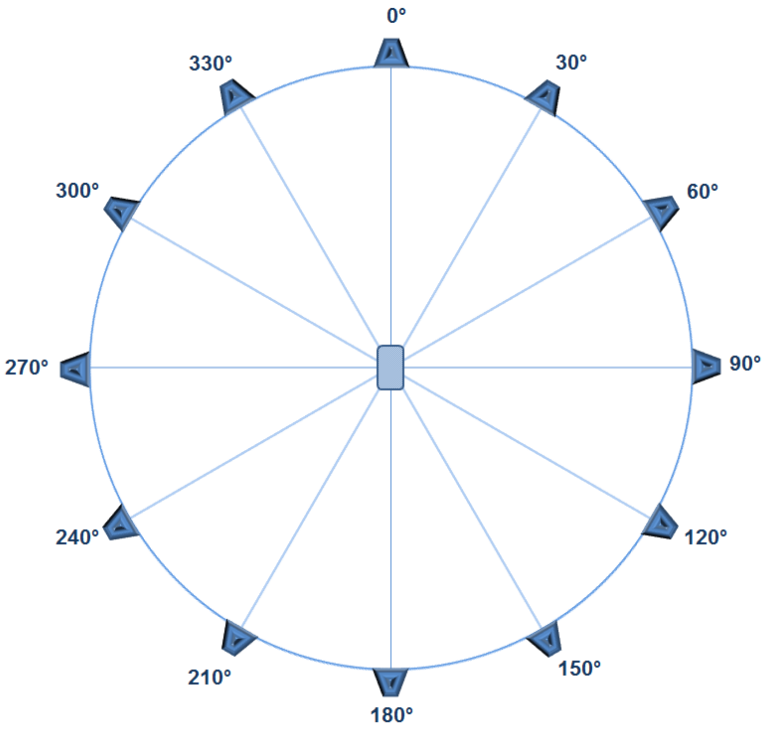

Figure 3.2: Calibration setup of ORL-LEA Room 7 with loudspeakers surrounding a multi-channel microphone in the center.

3.2.1 Mathematical background

The algorithm derived here is an adaption of the microphone inversion technique presented in section 2.4.2.3. Because of the linearity of the model below the following calculations are shown for a specific frequency only without loss of generality. Let ***p*** be the vector of sound pressures at *n* points and ***e*** an *m*-dimensional input signal of a certain frequency. Assuming a linear relation $C \in Mat(n,m)$ between these two variables and a set of input vectors in rows of matrix *E* with corresponding sound pressures *P* allows writing Equation 3.1 in matrix form:

$$P = CE \tag{3.1}$$

A way to determine matrix *C* is a set of *n* calibration measurements for an input signal $\hat{E}$ that results in pressures $\hat{P}$. In case of $n < m$, the matrix *C* can be pseudo-inverted with an MSE minimization approach:

$$C^+ := argmin_{C'}(\underbrace{(C'\hat{P} - \hat{E})^T(C'\hat{P} - \hat{E})}_{MSE})$$

$$\Rightarrow \quad \frac{\partial MSE}{\partial C'}(C^+) = -2\hat{P}^T\hat{E} + 2\left(\hat{P}^T\hat{P}\right)C^+ = 0$$

$$\Rightarrow \quad C^+ = \left(\hat{P}^T\hat{P}\right)^{-1}\hat{P}^T\hat{E} \tag{3.2}$$

For many applications the inversion of $\hat{P}^T\hat{P}$ in Equation 3.2 is a badly conditioned problem that can be solved by introducing a distortion matrix Γ, a so called regularization, which results in the final form 3.3:

$$\boxed{C_\Gamma^+ = \left(\hat{P}^T\hat{P} + \Gamma^T\Gamma\right)^{-1}\hat{P}^T\hat{E}} \tag{3.3}$$

Theoretically, this means that an arbitrary multi-channel microphone recording ***p*** can now be reproduced in a calibrated laboratory setting using virtual input signals $\boldsymbol{e} = C_\Gamma^+\boldsymbol{p}$ computed by convolution in the frequency domain.

3.2.2 Implementation

This microphone calibration method has been practically implemented in the *MATLAB*® program '*VAE_Designer*' that is described here. A multi-channel microphone is placed at the location where a potential sound field has to be created, the so called sweet spot, in the middle of a loudspeaker ring (see for example Figure 3.2). To calibrate the microphone for all frequencies simultaneously a white noise signal coming from one speaker at a time is recorded, resulting in a response spectrogram of each microphone channel to each loudspeaker. The spectrograms are obtained by an STFT with Hamming windows of 512 samples length and a 50 % overlap for an FFT length of 512. The mean recorded amplitude can be calculated in the frequency domain resulting in a matrix $\hat{P}_k$ as in Equation 3.2 for each frequency band *k* and is smoothened by a zero-phase filter over a length of 20 frequencies. This is basically an impulse response, written as a three dimensional tensor with elements for each loudspeaker-microphone channel pair in each frequency band (Figure 3.2).

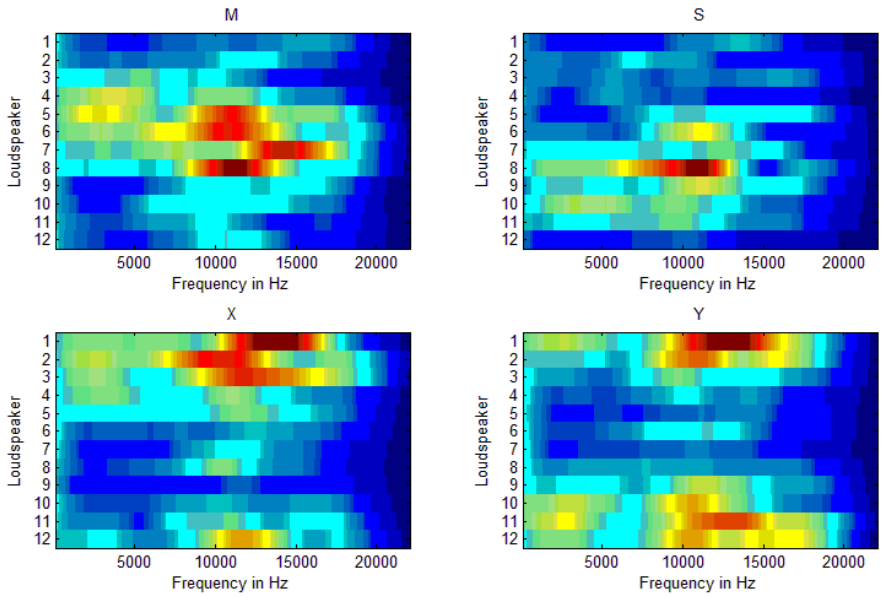

Figure 3.3: Impulse response for channels (M, S, X, Y) of a *H2N* microphone in ORL-LEA Room 7 to a white noise stimulus on twelve loudspeakers.

Different methods to compute a set of input signals for reproduction of a recorded sound field with an *H2N* are implemented in '*VAE_Designer'* and presented in the following:

Mean amplitude weighting: Similarly to a first-order Ambisonics system, the mean amplitude over all frequencies is taken as a weight connecting each loudspeaker to each microphone channel. The input signals $\boldsymbol{e}$ for each loudspeaker are computed by Equation 3.4.

$$\boldsymbol{e} = \left(\sum_{k=1}^{\#k} \frac{\hat{P}_k}{\#k}\right)^T \boldsymbol{p} \tag{3.4}$$

This reconstruction method assumes a flat frequency response of the calibration procedure that is clearly not the case as shown in Figure 3.3. Therefore frequency dependent reconstruction methods have to be taken into consideration.

Impulse response weighting: This reconstruction method works in the same way as mean amplitude weighting, but for each frequency band separately. The recorded signals $\boldsymbol{p_k}$ are convolved with the element-wise reciprocal of the impulse response, written as $1./\hat{P}_k$, to get input signals $\boldsymbol{e_k}$:

$$\boldsymbol{e_k} = (1./\hat{P}_k)^T \, \boldsymbol{p_k} \tag{3.5}$$

Although this method corrects for the impulse response, directionality in the final sound field is still poor. Angular resolution can be improved by inverting the impulse response matrices instead of a simple weighting approach.

Impulse response inversion: Instead of just taking the reciprocal of each element in the impulse response, all the matrices $\hat{P}_k$ are inverted with Equation

3.2 to get a new filter set. A distortion matrix $\Gamma = \alpha I$, where *I* is the identity matrix and α a regularization parameter, is introduced. The rows of the matrix that has to be inverted are permutated in a way that the first microphone channel faces the first few loudspeakers; the second channel the next loudspeakers and so on. This motivates an additional weight on the diagonal introduced by a diagonal distortion matrix. The regularization parameter is iteratively increased to reduce the condition number below a threshold before the matrices are inverted to get $C^{+}_{\Gamma\,\mathrm{k}}$ as in Equation 3.2 for each frequency band *k*. Finally this results in a filter bank specific to the loudspeaker setup and microphone used for the calibration.

$$\boldsymbol{e_k} = C^{+}_{\Gamma\,k}\boldsymbol{p_k} \qquad 3.6$$

Instead of convolving a recording with this filter bank to get the virtual input signals directly, a set of finite impulse response filters (FIR) is designed with a frequency response corresponding to the measured filters. The recordings are now filtered in the frequency domain after performing FFT and reconstructed in the time domain with an overlap-add algorithm (Appendix B) to reduce computation time. This approach has been proven to result in spatial blur and distorted signals because in this method a XTC is included theoretically by means of negative values in the filter bank, but accurate performance is not possible due to varying phase between the recorded signals.

Amplitude inversion: The idea of amplitude inversion is motivated by the assumption that relative sound energy is distributed identically for all frequencies. This means all measured $\hat{P}$ are averaged resulting in a single matrix that is inverted in the same way as above with Equation 3.2. The problem of negative weights in this matrix can now be eliminated by using prior information about the microphone channels. Only the microphones facing a certain loudspeaker are taken into account for computing its virtual signal, while the others are set to zero. The room impulse response as shown in Figure 3.2 is considered by filtering the signals with the reciprocal values of this response, again by designing FIR filters and the overlap-add algorithm.

$$W(i,j) := \begin{cases} \sum_{k=1}^{\#k} \frac{C^{+}_{\Gamma_k}(i,j)}{\#k} & if \sum_{k=1}^{\#k} \frac{C^{+}_{\Gamma_k}(i,j)}{\#k} > 0 \\ 0 & if \sum_{k=1}^{\#k} \frac{C^{+}_{\Gamma_k}(i,j)}{\#k} \leq 0 \end{cases}$$

$$\boxed{\boldsymbol{e_k} = (1./\hat{P}_k)^T W \boldsymbol{p_k}} \qquad 3.7$$

The pathway defined by Equation 3.7 of the recorded microphone signal to derive a virtual signal is illustrated for one loudspeaker in Figure 3.4.

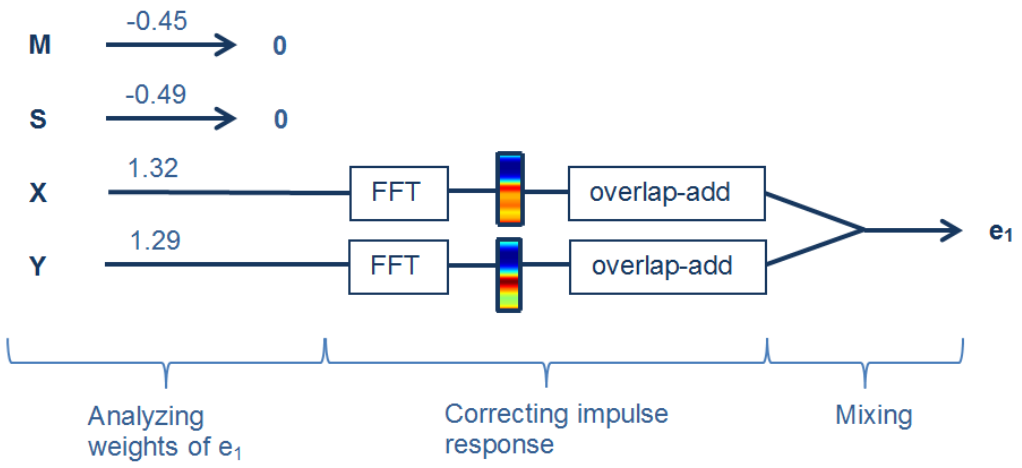

Figure 3.4: Example of signal processing of *H2N* channels (M, S, X, Y) to get a virtual signal e_1 for a first loudspeaker.

The computed virtual signals for each loudspeaker are then normalized and saved in a library where they can be accessed by other programs to carry out experiments.

3.3 Virtual acoustic environment experimenter

Having a program to design arbitrary virtual acoustic environments from real microphone recordings as described in the previous seciton 3.2 enables a whole set of possible experiments. The *MATLAB*® program '*VAE_Experimenter'* is a tool to design experiments based on the sound files computed by '*VAE_Designer'*. The basic concept is to mix a target signal with background noise and ask for identification. This target signal can either be static for a simple localization experiment, dynamic for a more advanced task or a spoken sentence and can be chosen from a library by the user together with other settings. After starting an experiment the program generates an instruction screen on the feedback monitor that enables a participant to begin with a target presentation round, training or an actual task. The signals are generated offline according to a user's input using *SoundmexPro*. The coefficient for a master volume is set according to the SPL, and a target track volume according to the SNR. These coefficients are linked to the actual values in dB by an uncorrelated white noise calibration. Each loudspeaker channel is mapped to two tracks, one to load the background and one for the targets. The background track is loaded with looped reconstructed background recordings and the targets are loaded to the target track at the determined times while the gaps are filled with zeroes. The two tracks are mixed to a channel and sent to the ASIO driver for playback. The code enters the main loop that captures events from the feedback touchscreen and reads out the head tracker and saves the data in a struct.

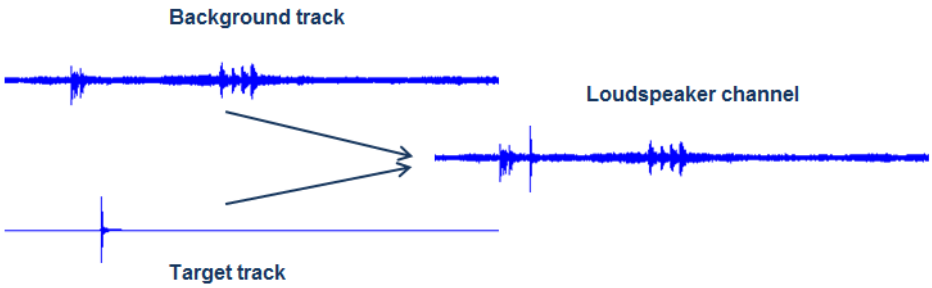

Figure 3.5: Offline mixing of a target track with a background track to final loudspeaker signal.

4 Evaluation

4.1 Instrumental measures

A first set of evaluation experiments is carried out objectively with the aim to investigate the spatial resolution 4.1.1, volume distribution 4.1.2 and the frequency response 4.1.3 of the system together with the reconstruction methods presented in section 3.2. All the experiments take place in Room 7 of ORL-LEA with an *H2N* microphone as described in section 3.1.

4.1.1 Spatial resolution

An important aspect of reproducing acoustic environments in the laboratory is preservation of spatial information. A back projection of a multi-channel recording to even more channels of a reproduction system is an ill-posed mathematical problem and results in spatial blur.

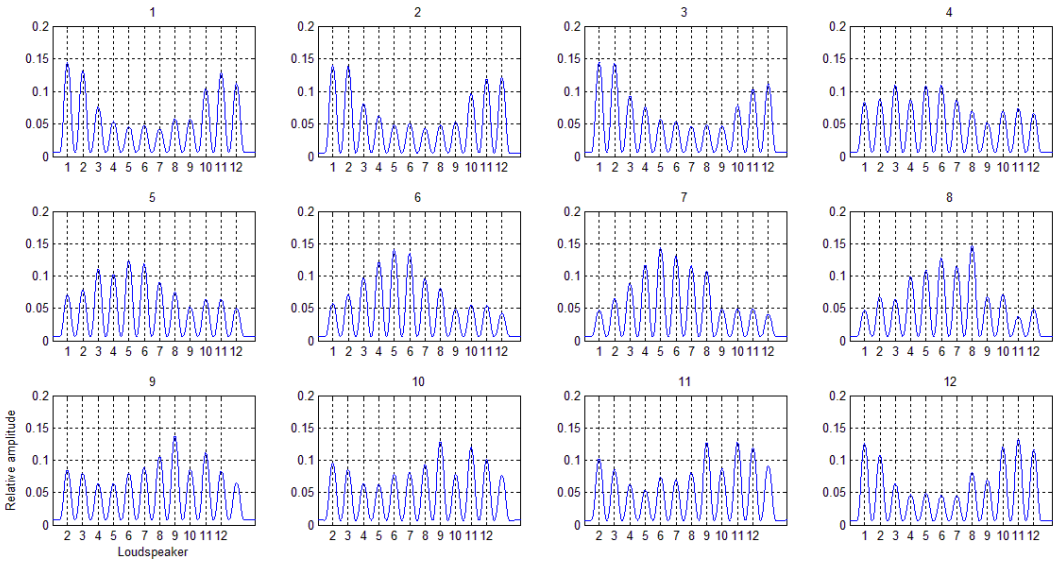

Figure 4.1: Spectral amplitude for each reconstructed virtual signal as a function of the loudspeaker where the white noise stimulus has been played. Shown is an example of the amplitude inversion method using two microphone *H2N* channels per loudspeaker. Ideally each of these plots would feature exactly one peak at the point where the stimulus loudspeaker is the same as the virtual channel.

Spatial blur for different reproduction methods is analyzed by a white noise stimulus with a sample rate of 44.1 kHz for one second. The signal is played sequentially on one loudspeaker after another and recorded by the microphone. Virtual input signals are computed and the blur of the sound energy to other loudspeakers is analyzed in the frequency domain by taking the averaged spectral amplitude over a certain time (see Figure 4.1) and assuming normal distribution.

Reconstruction method	**Microphone channels per loudspeaker**	**Mean standard deviation in degrees (spatial blur)**
Impulse response inversion	4 (all)	93.1 ± 3.3
	2	90.8 ± 3.1
	1	91.1 ± 3.1
Amplitude inversion	4 (all)	103.3 ± 4.9
	2	92.4 ± 4.0
	1	91.0 ± 5.4

Table 4-1: Mean standard deviation in degrees of sound energy distribution across neighboring loudspeakers for different reconstruction methods indicating spatial blur.

The mean standard deviation of the fitted normal distributions is listed in Table 4-1 for the two different reconstruction methods using a varying number of microphone channels per loudspeaker and setting other channels to zero. Theoretically, the XTC included in both reconstruction methods would result in higher spatial resolution if more microphone channels are involved. Table 4-1 shows that the opposite is the case indicating that spectral subtraction of microphone channels is prone to errors due to phase mismatch.

4.1.2 Sound pressure level

The volume between the different constructed virtual channels should not vary for a recording of the same sound intensity coming from all directions. The equal

distribution of the sound energy to the virtual channels is tested in the following experiment: A white noise stimulus is played on each loudspeaker simultaneously three times for one second and recorded with the microphone placed in the center of the ring. Averaged spectral amplitude is computed for each virtual channel reproduced from the recording with different methods. The relative standard deviation between the volumes of the virtual channels is listed in Table 4-2 for each reproduction method.

Reconstruction method	**Microphone channels per loudspeaker**	**Mean standard deviation in percent**
Impulse response inversion	4 (all)	20.1 ± 0.1
	2	28.3 ± 0.0
	1	33.1 ± 0.1
Amplitude inversion	4 (all)	5.7± 0.0
	2	9.3 ± 0.1
	1	21.8 ± 0.2

Table 4-2: Relative standard deviation of spectral amplitude between virtual channels for recording of an equally distributed sound field. Relative standard deviation is averaged over three white noise test stimuli coming from all speakers simultaneously for the different reconstruction methods.

According to Table 4-2 there are several over- and under-expressed virtual channels for the impulse response inversion method, while the sound energy is almost equally distributed for amplitude inversion with all or two microphones.

4.1.3 Frequency analysis

The impulse response of the experimental setup together with the reconstruction method is analyzed by designing a white noise stimulus again with a sample rate of 44.1 kHz for one second. The signal is played on one speaker, recorded on an *H2N* placed in the center and a virtual input signal is generated with different reconstruction methods and analyzed in the spectral domain. This analysis is

carried out for two and four H2N channels per loudspeaker for impulse response inversion and amplitude inversion. For comparison the result is also shown for frequency independent amplitude inversion. The relative standard deviation of the spectral amplitude serves as a measure for variation in the frequency domain, is averaged over the different virtual channels and listed in Table 4-3.

Reconstruction method	**Microphone channels per loudspeaker**	**Mean standard deviation in percent**
Impulse response inversion	4 (all)	41.0 ± 8.8
	2	39.8 ± 9.2
Amplitude inversion	4 (all)	32.0 ± 13.6
	2	30.7 ± 12.2
	2 without impulse response correction	43.0 ± 10.2

Table 4-3: Relative standard deviation of spectral amplitude averaged over the virtual channels for the different reconstruction methods. Lower standard deviations mean flatter frequency responses and the virtual signals are closer to the initial white noise stimulus.

4.1.4 Comparison of methods

The previous analysis of the different reconstruction methods regarding spatial resolution, impulse response and volume distribution should help making a decision on which method to use for the behavioral experiments. Amplitude inversion with all microphones is disregarded due to the significantly increased spatial blur. The distribution of the sound energy to the different virtual signal channels is sufficiently equal for amplitude inversion with four and two microphone channels. Also the frequency response of the methods supports these methods while showing clearly the benefit of impulse response correction. Furthermore an advantage of the methods with less microphone channels per loudspeaker is reduced computation time. Therefore, the reconstruction method of choice for section 4.2 is amplitude inversion with correction of the room impulse response and with two microphone channels per loudspeaker.

4.2 Behavioral measures

For further evaluation of the system different localization experiments are designed where the ability of different human listeners to identify and localize a certain auditory object in background noise is investigated. Here, the experiments focus on realistic traffic situations with a potentially dangerous target that can be identified acoustically. The recording samples, procedure and participants of this pilot study are described in 4.2.1- 4.2.3 and the results of the experiments are presented in 4.2.4 and discussed in 4.2.5. The original settings and instructions can be found in Appendix C.

4.2.1 Recording samples

Three different acoustic environments are recorded with an *H2N*, reconstructed and mixed with a target to create a test scenario as listed in Table 4-4.

Background	**Target**	**SNR in dB**	**SPL in dB**
Street	Bicycle bell	+5	70
Train station	Horn	+3	72
Square	Tram	+10	65

Table 4-4: Different recorded acoustic scenarios with certain sound pressure level and corresponding target at a certain signal to noise ration.

Street has been recorded in Zurich Niederdorf on a crossing. The recording is dominated by passing cars and walking pedestrians. *Train station* recording has been taken place in Zurich HB with announcements, walking and talking pedestrians, background noise and the horn of a luggage transporter. *Square* has been recorded at Zurich Bellevue with cars, pedestrians, background noise and a tram approaching. The background recordings are cut to a length of roughly 60 seconds at zero-crossings of all channels to enable looping.

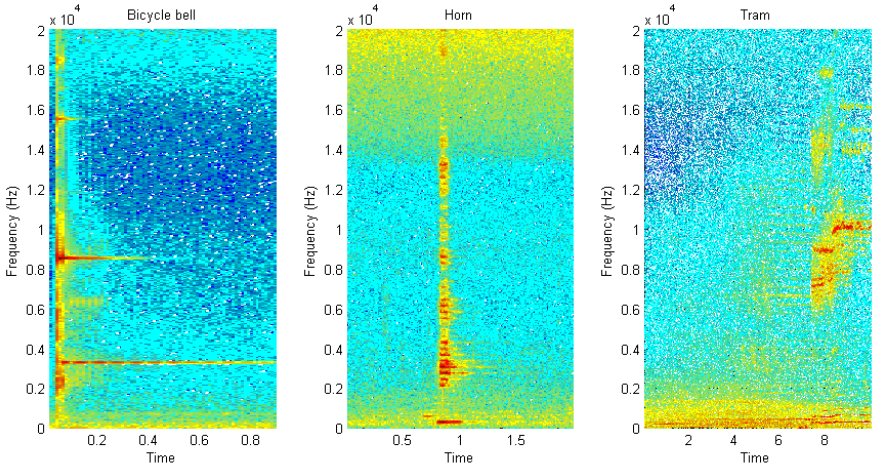

Figure 4.2: Spectrogram of different targets that have to be localized in the experiments by a participant.

The targets are cut from the original recording; noise is reduced in *Audacity* and they are filtered by the RIR function creating an individual signal for each loudspeaker (Figure 4.2). The tram target is simulated to cover a braking distance tangential to the loudspeaker ring with filters derived from VBAP with a distance factor of 0.8 (see Figure 4.3).

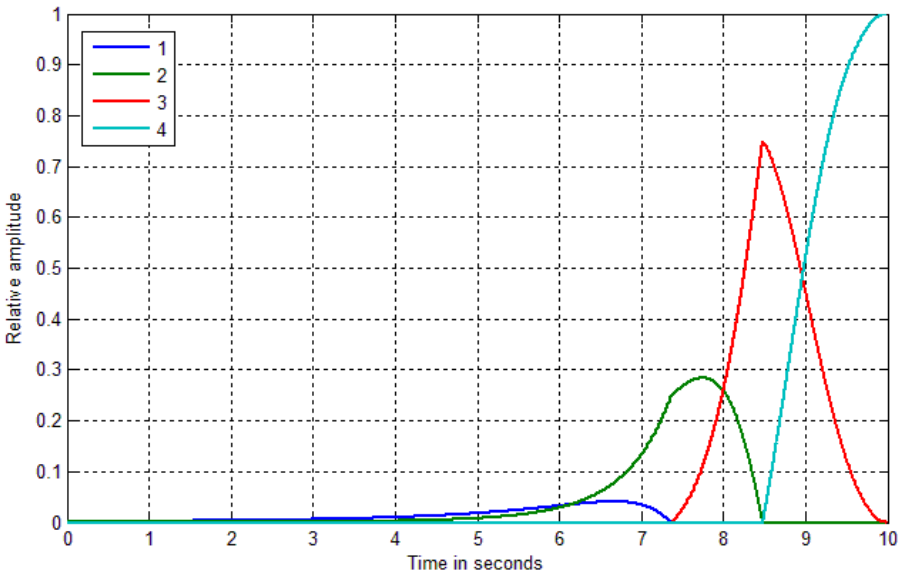

Figure 4.3: Filters to simulate a dynamic target driving tangentially to the loudspeaker ring and finally stopping derived from VBAP and a distance factor of 0.8.

The sound pressure levels of the backgrounds vary in time meaning that the targets have to be carefully placed in certain time windows to provide comparable

conditions. The criterion for these time windows is that the total amplitude level has to be within 5 dB of the mean amplitude for *Street*, 1 dB for *Train station* and 3 dB for *Square*. Additionally the minimal time between two targets is set to be 5 seconds to give participants enough time for feedback. In the static case, a total of six targets per loudspeaker is randomized and mixed with the looped backgrounds. The dynamic target is randomized and mixed such that it stops twice on each loudspeaker, once from each side. Because volume calibration is performed with uncorrelated white noise, the peak SPL is additionally logged with a *Samsung® Galaxy S3* and *Noise meter* and checked with a sound pressure meter for each background (Figure 4.4).

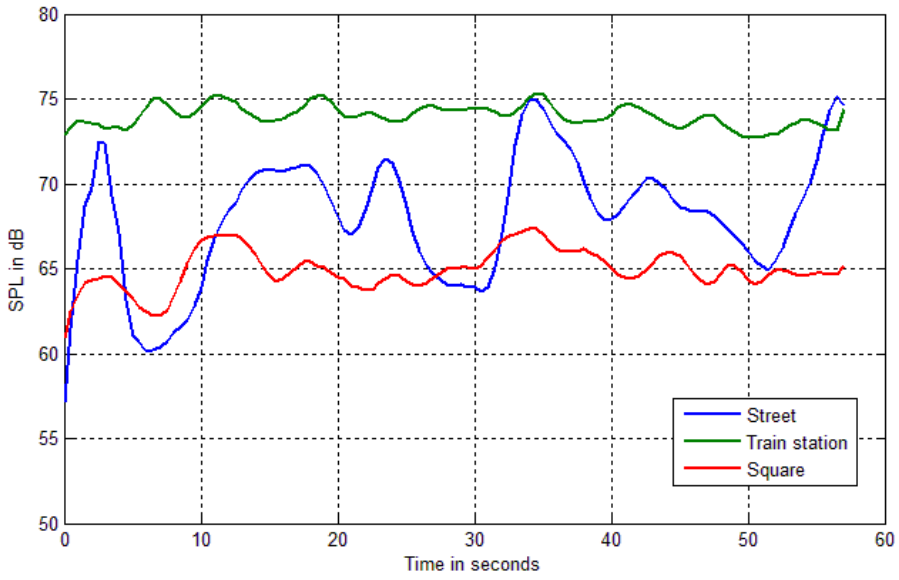

Figure 4.4: SPL of different backgrounds in dB as a function of time monitored by *Noise Meter* application and controlled by a sound pressure meter.

4.2.2 Experimental procedure

At the beginning of each new experiment the participant is placed on a chair in front of the feedback monitor in the center of the ring (Figure 4.5), instructed with general information about the tasks and equipped with a motion tracker on top of the head. The participants are especially encouraged to move their heads as it

might help localizing a sound source, and to ask for breaks if necessary. The task for static localization is to push the labelled button indicating perceived position of the presented target. For dynamic localization the participants are asked to follow the perceived position of the tram with their heads until it stops. After that they should push the button of the current position before they are asked to indicate the direction of the tram. The procedure to test each scenario is divided into four steps:

1. The scenarios are introduced with a symbolic picture and additional information on the monitor.
2. Targets of static scenarios *Street* and *Train station* are presented clockwise on each loudspeaker once. The dynamic target is presented on the front, on the back and once on each side.
3. A training session is performed to get familiar with the task. The targets are presented randomized on the side, front and back and a feedback is provided to the participant by showing the correct answer. If necessary the training session is repeated until the participant agrees to proceed.
4. The actual test is performed by presenting the targets and capturing feedback on the feedback monitor and the head tracker. Static targets are presented six times on each loudspeaker, while the dynamic target is presented twice on each loudspeaker, once from each direction. Additionally a progress bar is shown on the feedback monitor.

Figure 4.5: Experimental setup in ORL-LEA Room 7 with loudspeakers, feedback monitor and head tracker.

4.2.3 Participants

A total of seven normal hearing listeners performed the test in a first round serving as a control group. Five of them were randomly chosen to repeat the experiment several days later for analysis of reproducibility and training effect. Four participants agreed to repeat the experiment wearing *Ohropax® Soft* earplugs to simulate a conductive hearing loss. The severity of the hearing loss is quantified with a warble tone audiogram shown in Figure 4.6.

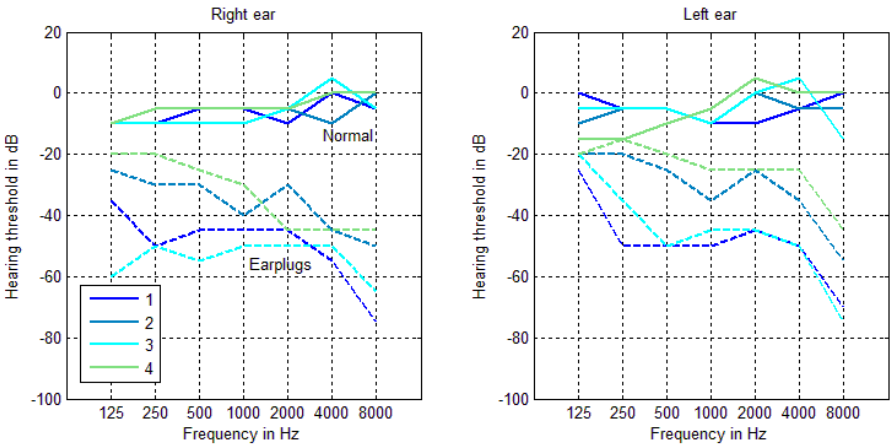

Figure 4.6: Warble tone audiogram of four participants with simulated conductive hearing loss compared to their normal hearing ability.

Additionally, a participant with a sensorineural hearing loss took part in the experiment wearing a behind-the-ear *Audeo S Smart III.* The test is performed in different conditions, namely without hearing aid, in an omni-directional mode and a directional mode.

4.2.4 Results

The feedback given by the participants is analyzed separately for the static and the dynamic localization task. In sections 4.2.4.1 and 4.2.4.2, the results are presented quantitatively using RMS, front-back confusions and the head movements. The dynamic results are analyzed in relation with the data from the head tracker, in section 4.2.4.3.

4.2.4.1 Root mean squared error

The RMS for test and retest of the normal hearing group for the localization tasks is shown with standard deviations in Figure 4.7.

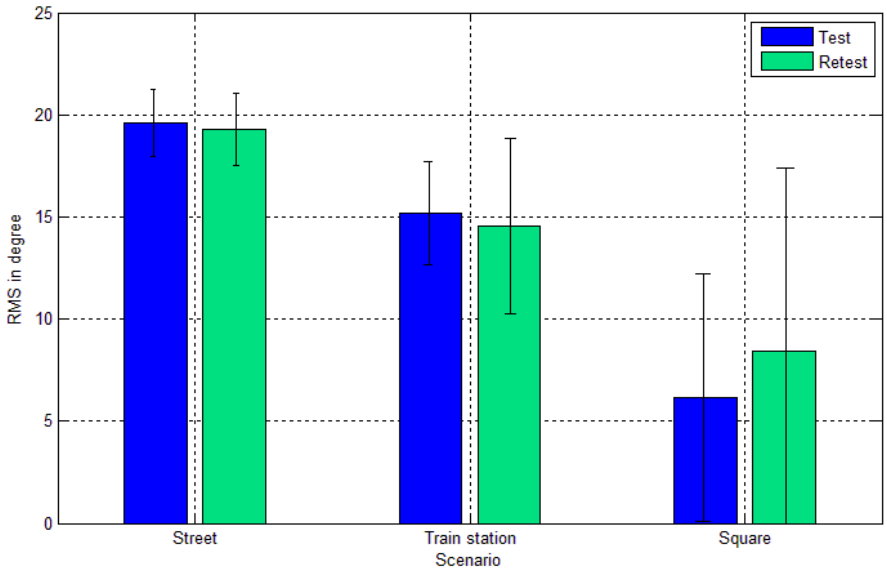

Figure 4.7: RMS in degree for the two static scenarios *Street* and *Train station* and the stopping position of the dynamic target in test and retest with standard deviation.

The correlation coefficient between test and retest of the group that performed both experiments was $\rho=0.93$, meaning that the results are reproducible. An analysis of variance (ANOVA) of the normal hearing results reveals a significant difference in difficulty of the tasks with a p-value of $p=0.0003$ for the hypothesis that all results were drawn from the same scenario. The p-value for the hypothesis that all RMS values came from the same group was $p=0.81$. In Figure 4.8, the results of test and retest are combined and serve as a control group to compare the RMS for different hearing impairments.

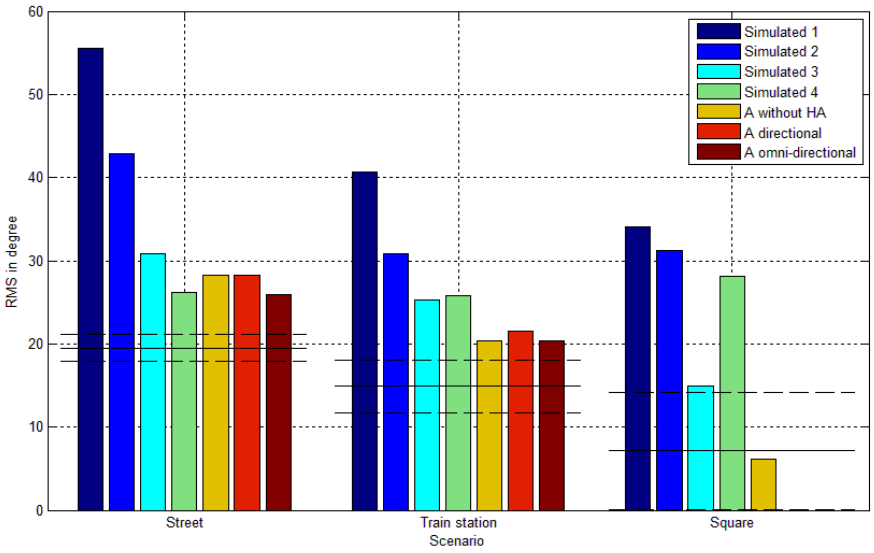

Figure 4.8: RMS for the three scenarios of listeners with simulated or sensorineural hearing loss compared to the normal hearing group indicated by the black lines with standard deviation.

RMS value in this experiment is increased in the simulated case and correlates with the severity of the hearing loss according to the audiogram. Here, the hypothesis that the simulated hearing loss and normal hearing results were drawn from the same group is rejected with a p-value of *p=0.000001* in an ANOVA, indicating a significant difference in performance between these two conditions. The participant with sensorineural hearing loss performed similarly in all conditions, meaning that the hearing aid seems not to improve RMS in these tasks significantly. In the last scenario, candidate A scored zero RMS while wearing a hearing aid.

4.2.4.2 Front-back confusions

Another measure to quantify localization performance is the number of front-back confusions in percent, which are illustrated for test and retest in Figure 4.9. The movement of the dynamic target together with head rotations allowed resolving front-back confusions completely for most presentations in the *Square* test. Because of this ceiling effect it is not considered in this analysis.

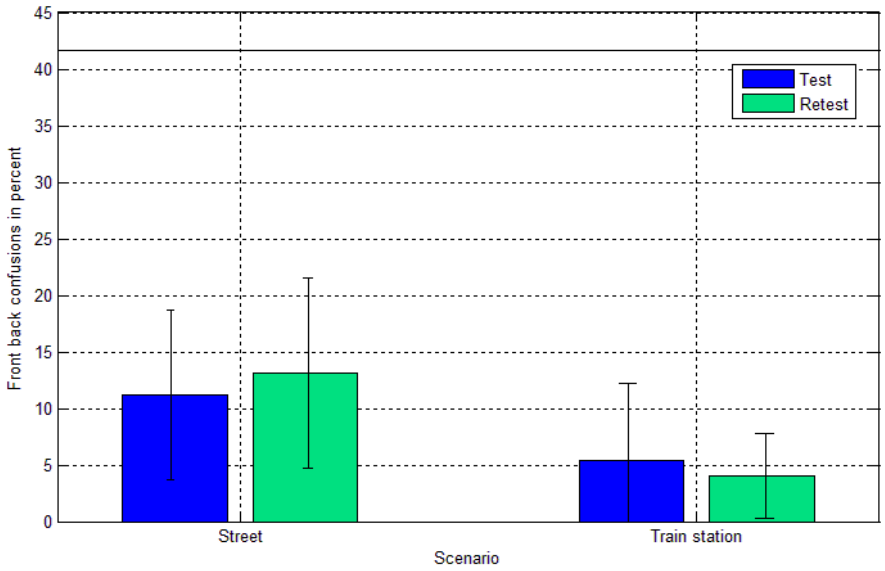

Figure 4.9: Front-back confusions in percent of normal hearing listeners in the two static scenarios with chance level indicated by the black line.

Although the standard deviation is large indicating high inter-subject variability, the front-back confusions are reproducible with correlation coefficient of $\rho=0.70$. The number of front-back confusions of the two different scenarios is significantly different with an ANOVA p-value of $p=0.004$, and no training effect can be observed. Similarly to the RMS, the results of the hearing impaired listeners are compared to the control group in Figure 4.10.

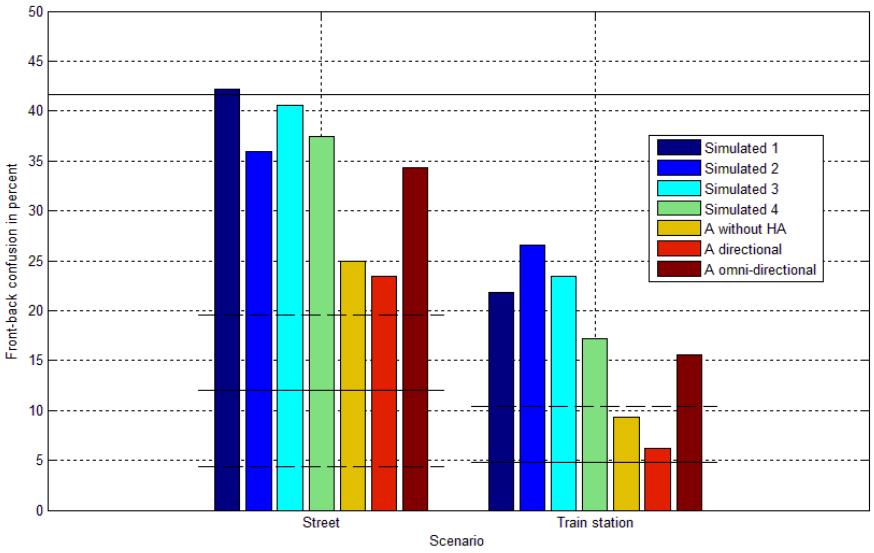

Figure 4.10: Front-back confusions in percent for both static scenarios and the different hearing impaired participants compared to the normal hearing group indicated by lower black line with standard deviation. The upper black line shows chance level of a potential randomly answering listener.

The participants with simulated conductive hearing loss show almost chance level performance in the first scenario while the results of the sensorineural hearing impaired is slightly increased compared to the normal hearing group. In the second scenario the group with the earplugs is still significantly worse than the normal hearing people, but the hearing impaired performs at a similar level with no hearing instruments and in the directional mode. There is a trend for participant A that the omni-directional mode of the hearing aid increases the number of front-back confusions.

4.2.4.3 Dynamic target

The feedback given by the participants by pushing buttons for direction of the dynamic target is analyzed similarly with wrong directions in percent. The normal hearing group scored a mean of 99.3 % correct with a standard deviation of 2.3 % suggesting a ceiling effect. The performance on direction perception of the hearing impaired group is summarized in Figure 4.11.

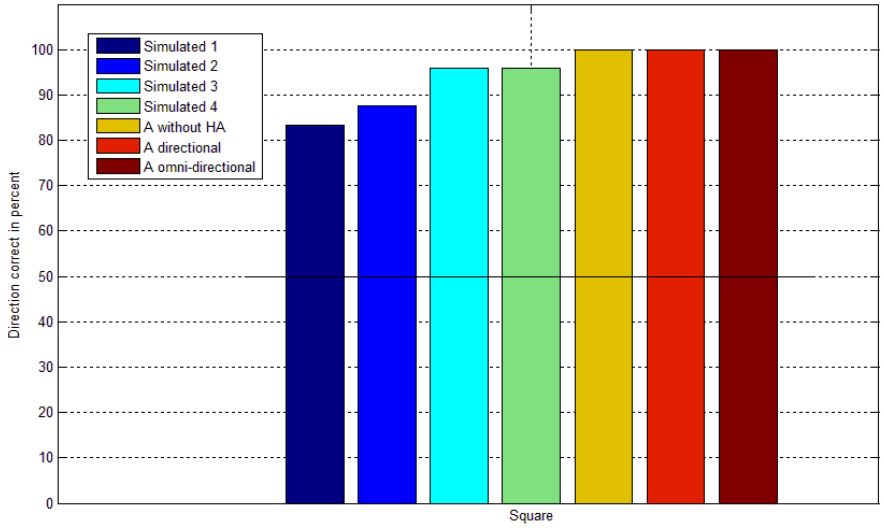

Figure 4.11: Percent correct of perceived direction of moving tram target for hearing impaired participants. Normal hearing score is almost perfect and chance level is indicated by the black line.

Additionally to the direction responses the dynamic target is analyzed with the data from the head tracker that indicates time until the participants were able to localize the tram, accuracy as well as scanning strategy. Examples of representative head trajectories are shown in Figure 4.12, where they are compared to the actual position of the tram.

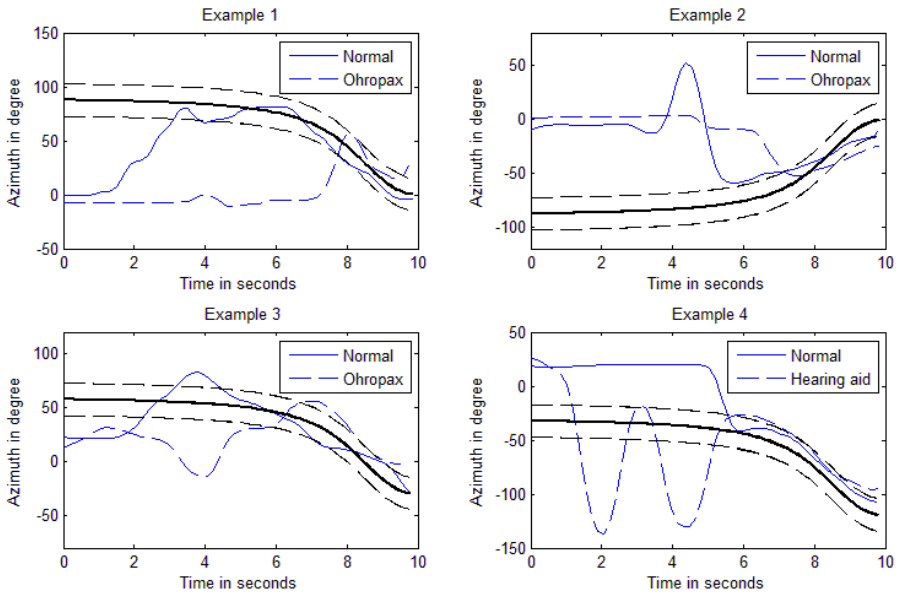

Figure 4.12: Examples of head trajectories while following a virtual approaching tram in two cases for four subjects. Black line indicates actual position of tram with tolerance level of ± 15°.

Participants usually wait until they hear the dynamic target and then start an active scanning process by rotating their heads to localize the target. The following analysis is restricted to the targets presented on the front loudspeakers. Performance is quantified by a dynamic RMS that is the mean RMS of all data points of a presentation. Test-retest correlation for this measures was found to be $\rho=0.99$ but varied between subjects because every individual listeners had a different extent of head rotations during the experiments. Therefore, the results of the hearing impaired group are normalized with respect to the performance of the normal hearing group for each target direction and illustrated in Figure 4.13.

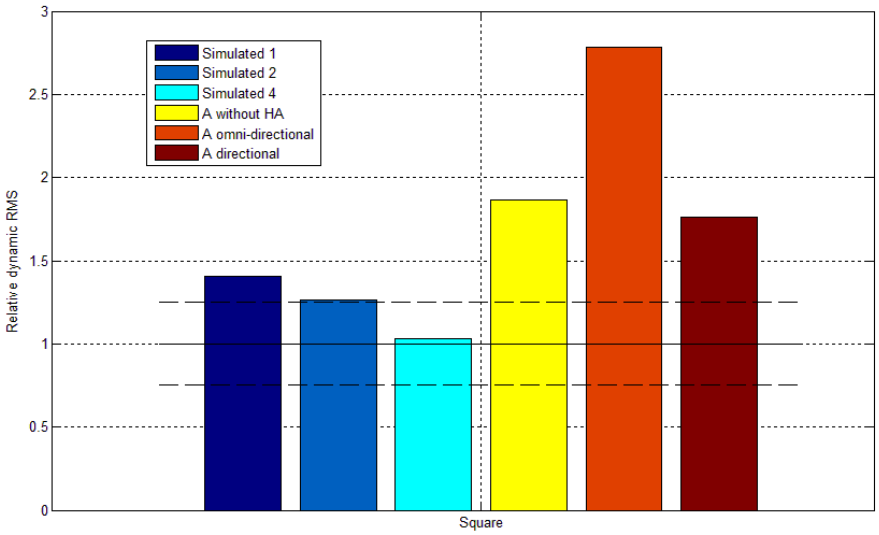

Figure 4.13: Relative score for dynamic RMS error of hearing impaired listeners compared to normal hearing control group indicated by black line with standard deviation.

While the participants with a simulated conductive hearing loss wearing earplugs performed worse in the static test, the opposite is the case here. Furthermore, wearing a hearing aid in this task seems not to be beneficial for subject A. Another measure to analyze dynamic localization in this experiment is given by the time a listener needs until the head is turned in the correct direction and stays there within a tolerance level of ±15°. This reaction time is not sufficiently reproducible with a correlation coefficient of $\rho=0.53$ and no conclusions are drawn from comparing normal hearing with hearing impaired performance. A histogram of the overall performance of all participants is shown in Figure 4.14.

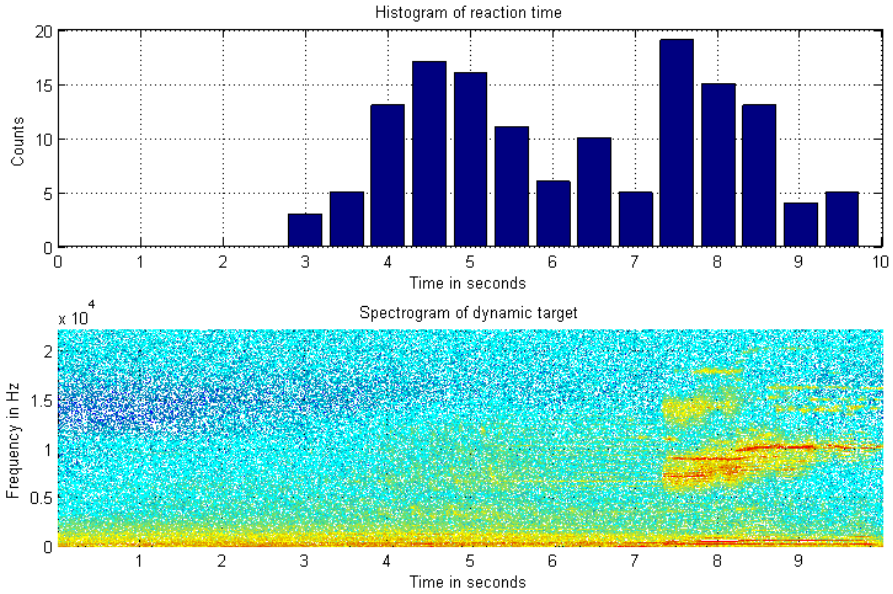

Figure 4.14: Histogram of reaction times for all participants compared to spectrogram of dynamic target.

The braking sound of the dynamic target visible in the high frequency domain of the spectrogram after seven seconds corresponds to a second peak in the histogram. It can be assumed that this high frequency domain provides additional information that is eventually sufficient for localization.

4.2.5 Discussion

The results presented in the previous chapter and the experiment in general are summarized and discussed in the following. Static localization tasks *Street* and *Train station* revealed expected results for RMS and front-back confusions. Although orally reported differently by many participants the target signal in *Street* causes a higher RMS and more front-back confusions than the other target in *Train station.* This is explained by the length of the signals and the broader spectral information of the horn signal in *Train station.* Furthermore the rather short signals are the reason for a rather high number of front-back confusions occurring in this experiment. Longer targets that offer enough time for localization with head movements would reduce this number but are prone to ceiling effects with full score for every presentation.

The normal hearing group performs in both static scenarios reproducibly better than the simulated conductive hearing impaired and sensorineural hearing impaired listeners. Remarkably, the simulated hearing loss has a much more negative effect on the performance, suggesting that adaption plays an important role. For the simulated hearing loss the performance correlates with the loss measured by the warble tone audiogram. However, this may be not the case for hearing aid wearers, because the hearing aid of participant A does not improve the test performance. And although the stopping position RMS for the dynamic target in scenario *Square* is improved with a hearing aid, the dynamic RMS is increased in the omni-directional mode. This means that at the beginning of a new dynamic target presentation the participant seems to be confused and scans for the unknown source position more extensively in the omni-directional setting. The reaction time for finding the dynamic target is not a characteristic value in this test. A more sophisticated experimental setup especially designed for dynamic target localization might lead to another conclusion. The results from all the different tasks are compared and their pairwise correlation is shown in Table 4-5.

		Street		Train station		Square		
		RMS	front-back	RMS	front-back	RMS	Dynamic RMS	Reaction time
Street	RMS		0.76	0.95	0.73	0.72	0.50	0.07
	front-back			0.76	0.77	0.70	0.55	0.43
Train station	RMS				0.77	0.70	0.45	0.14
	font-back					0.73	0.35	0.20
Square	RMS						0.13	0.41
	Dynamic RMS							0.41
	Reaction time							

Table 4-5: Pairwise correlation of all measurement variables for all participants. The dark grey background indicates coefficients larger than the 1 % - significance level of 0.68.

The rather low correlation coefficients of dynamic reaction time and dynamic RMS with the other results can be interpreted in various ways. Either the inter-subject variability dominates the results and is too high to get significant results with this number of subjects. Another explanation would be that these measures provide additional important information about the ability of a listener to orientate in an acoustic environment and to identify and localize a certain auditory object. More test results for other dynamic targets with different spectral characteristics might help to solve this problem.

5 Conclusion

The task of this Master's thesis was to implement and evaluate a tool that enables testing of the performance of normal hearing and hearing impaired listeners in a virtual acoustic environment emulating realistic and daily life scenarios. These scenarios are recorded in a first step and then reproduced in the lab by preserving spatial and spectral information of the original signal in a free field surround sound setup. The task is motivated by the fact that common schemes for evaluating performance of different hearing aid or CI audio signal processing algorithms and microphone placement take place in artificial laboratory settings that don't reflect real life performance adequately.

The system that has been developed within this thesis works with a multi-channel surround sound microphone. Calibration of this microphone in an arbitrary room with a loudspeaker ring setup playing white noise stimuli results in an impulse response of each microphone channel to each loudspeaker that combines RIR and XTC as well as the impulse response of the loudspeakers and the microphone itself. A recording made with this microphone is now reproduced for this room by filtering the signal with FIR filters designed from these impulse responses. The recording channels are weighted by an additional factor that is derived from inverting the matrix of the partial sound energies that come from one speaker and go to each microphone channel, leading to improved spatial resolution.

Instrumental evaluation of this reproduction method revealed a good azimuthal spatial resolution, an equal distribution of sound energy to all loudspeaker channels and a flat frequency response. Behavioral measures have been conducted in a pilot study. Many acoustic scenarios have been identified as potentially challenging for hearing impaired listeners, recorded and saved in a library. Three different traffic scenarios have been chosen for further testing, namely a street, a train station and a busy square. These scenes are mixed with auditory objects that have to be identified by a human listener and are static in the first two cases and dynamic in the last one. A participant is asked to indicate perceived location of these targets on a feedback screen while wearing a head tracker. RMS values and number of front-back confusions have been significantly

different for normal hearing listeners and listeners with a simulated conductive hearing loss. No significant benefit of wearing a hearing aid by a participant with sensorineural hearing loss was found in this experiment. Finally the dynamic scenario gave insight into orientation ability of a listener in a certain auditory scene, but more data is required for further analysis.

5.1 Advantage and disadvantage

One advantage of this free field reproduction system is that there is no need to render new signals to compensate for head movements, as common in binaural reproduction systems. Generally, this means that the audio signals can be preprocessed without computational load being an issue. The calibration method allows reproduction of any recording made by this microphone for an arbitrary room and an arbitrary number of loudspeakers, making the system very flexible. Together with the mixing of targets to these recordings a new test battery can be designed, including static and dynamic localization and speech in realistic noise. The great advantage of having realistic scenarios is also a weakness. SPL's of backgrounds and SNR's of targets in an experiment are never constant, making reproduction of the results in other laboratories difficult. Additionally, only a limited amount of scenes can be tested and those are never exactly identical to the challenges a hearing impaired listener faces in real life. Finally the torso and head of a potential participant in the center of the sound field induces distortions that are neglected in this setup.

5.2 Future prospects

As mentioned before, the reproduction system enables a large set of different tests. Daily challenges of hearing impaired listeners can be identified, recorded and tested in the lab. In an extended study it is possible to test speech in noise and other static as well as dynamic targets in traffic situations. These tests can be performed by real hearing impaired patients or using a hearing impairment simulator on a phantom head. A list of some of the possible scenarios is proposed below in Table 5-1:

Static and dynamic localization	Speech in noise
Warning signal as pedestrian	Concert hall
Warning signal as car driver	Church
Construction sites	Vehicles
Train station	Restaurant
Work places	Work place
	Class room

Table 5-1: List of proposed scenarios that might challenge a hearing impaired listener.

In addition to a more extended study the existing system can also be improved technically on a software and hardware level. Obviously a more sophisticated loudspeaker setup, for example with arrays, would improve spatial resolution of a reconstructed sound field. The handheld microphone *H2N* used so far is very easy to use, but has a rather low SNR and high background noise level. Better sound quality and better spatial resolution could be achieved with a professional surround sound microphone with more directive channels with a fixed phase relation. A microphone that records three dimensional sound fields with eight channels and also takes the HRTF into account would be an *H2-PRO* by *Holophone*®. Finally the setup of the head tracking can be improved by a better mounting system, a higher sampling rate and a more stable data capturing.

Appendix

A Task description

Auditory object recognition of normal hearing and hearing impaired listeners in virtual acoustic environments

1. INTRODUCTION

The acoustic environment in our society provides a number of challenges for human listeners. Communication through spoken language as well as detection and classification of relevant sounds such as warning signals, vehicle noise and music are examples of the multiple tasks which the human auditory system executes continuously. Binaural auditory object detection and tracking constitutes an important mechanism of the human auditory system. Hearing impairment affects the effective use of binaural cues to varying degrees depending on the type and amount of auditory deficit.

The evaluation of specific aspects of hearing impairment and its consequences for binaural localization and communication abilities in a clinical setting is the main goal of this master thesis. Hearing tests in a controlled laboratory environment often ignore many real life effects of challenging acoustic environments. Using virtual acoustics it is hoped that more realistic test environments can be successfully reproduced for clinical use.

Portable miniature multimicrophone systems allow recordings of various auditory scenes such as street noise scenarios, cafeteria or cocktail party environments and construction or manufacturing noise settings.

Reproduction of these recordings for hearing impaired persons equipped with hearing aids or cochlear implants precludes headphones and requires free field multi-loudspeaker setups. Rendering the M multichannel recording tracks to N loudspeakers can be done using convolution techniques and calibration measurements of the recording and reproduction system.

2. TASK LIST

- Write a detailed time-table of the work to be performed
- Review the relevant literature
- Evaluate potential solutions/methods to address the aims of the project
- Develop a calibration application in Matlab for the multi-microphone recording (Zoom H2next) and multi-loudspeaker reproduction (ORL-LEA) system
- Perform a series of multimicrophone recordings for relevant environments and reproduce these environments in the laboratory
- Propose a few test paradigms for normal hearing and hearing impaired subjects using these virtual environments
- Perform a validation or verification of the obtained results using theoretical and experimental methods
- Write a detailed report of the project

3. LITERATURE

- Blauert, J (2005) Communication Acoustics. Springer
- Blauert, J (2013) The Technology of Binaural Listening : Modern Acoustics and Signal Processing. Springer

- Elen, R (2001) Ambisonics: The Surround Alternative. Surround 2001
- Grämer, T. (2010) Efficient modeling of head movements and dynamic scenes in virtual acoustics. Master Thesis IBT-ETHZ
- May, T, van de Par. S, Kohlrausch, A (2012) A Binaural Scene Analyzer for Joint Localization and Recognition of Speakers in the Presence of Interfering Noise Sources and Reverberation. IEEE TRANS. AUDIO, SPEECH, AND LANGUAGE PROC. 20/7: 2016-2030
- Mueller, M F; Kegel, A; Schimmel, S M; Dillier, N; Hofbauer, M (2012). Localization of virtual sound sources with bilateral hearing aids in realistic acoustical scenes. The Journal of the Acoustical Society of America, 131(6):4732-42
- Schimmel, S; Mueller, M; Dillier, N (2011). *Binaural models and virtual acoustics to study spatial perception.* Journal of Hearing Science, 1(2):79-82.
- Vorländer, M (2008) Auralization - Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality. Springer
- Vorländer, M (2013) Computer simulations in room acoustics: Concepts and uncertainties. JASA 133(3): 1203–1213

B Overlap-add

The overlap-add algorithm is a fast and efficient way to compute discrete convolutions with FIR filters in signal processing and described in many DSP books, for example Smith [36]. Let *f(n)* be a signal and *g(n)* a FIR filter. Because *g(n)* is finite the discrete convolution is given by Equation B.1:

$$(\mathrm{f} * \mathrm{g})(\mathrm{n}) := \sum_{m=1}^{M} g(m) f(n-m) \tag{B.1}$$

By dividing the signal into multiple segments $f_k(n)$ with length *L* and using linearity of the convolution the expression can be written as Equation B.2.

$$(\mathrm{f} * \mathrm{g})(\mathrm{n}) = \sum_{k} (\mathrm{f}_k * \mathrm{g})(\mathrm{n}) \tag{B.2}$$

Because the summand is zero for $n > L + M$ this equation corresponds to a circular convolution that can be computed efficiently with the circular convolution theorem as shown in Equation B.3:

$$\boxed{(\mathrm{f} * \mathrm{g})(\mathrm{n}) = \sum_{k} \mathrm{IFFT}[\ \mathrm{FFT}[\mathrm{f}_k\,(\mathrm{n})]\ \mathrm{FFT}[\mathrm{g}(\mathrm{n})]\ ]} \tag{B.3}$$

C Behavioral experiment

C.1. Settings

The settings of the behavioral experiments carried out in this project with *"VAE_Experimenter 1.0 a"* are saved in .mat files and listed in tables in the following.

C.1.1. Street

Training session has a length of 30 s, SPL of 70 dB with instructions, active motion tracker and randomized locations and the task is static localization of the targets listed in Table C-1.

Target	Loudspeaker	SNR	Time in seconds
'Bikebell'	12	5	3
'Bikebell'	3	5	9.5
'Bikebell'	6	5	16
'Bikebell'	9	5	23.5

Table C-1: Target of static localization task *Street* in training session.

Targets 'Bikebell' for the actual task that lasts 485 s are divided into six presentation sets on all 12 loudspeakers with an SNR of 5 dB. The valid times for a target presentation are presented below in Table C-2.

Set 1 in s	Set 2 in s	Set 3 in s	Set 4 in s	Set 5 in s	Set 6 in s
3.00	83.72	163.13	243.74	324.57	404.05
9.50	88.62	169.15	250.32	329.48	410.02
16.00	96.92	174.15	256.93	337.71	414.99
23.50	102.77	183.59	264.29	343.74	424.60
28.50	108.84	190.08	269.34	349.87	431.05
36.73	113.95	196.62	277.50	354.74	437.54
42.73	123.42	204.14	283.62	364.36	445.12
48.73	129.96	209.15	289.50	370.73	449.98
53.73	136.39	217.27	294.56	377.39	458.32
63.23	143.75	223.37	304.10	384.76	464.32
69.54	148.93	229.36	310.65	389.72	470.23
76.23	157.19	234.25	317.16	397.97	475.28

Table C-2: Six presentation sets with target times in seconds for the actual test *Street*.

C.1.2. Train station

Task *Train station* is again a static localization experiment with length of 35 s, SPL of 72 dB, active motion tracker and instruction screen. Also the target loudspeakers are randomized. The training settings are shown in Table C-3.

Target	Loudspeaker	SNR	Time in seconds
'Horn_hb'	12	6	7.37
'Horn_hb'	3	6	15.87
'Horn_hb'	6	6	23.38
'Horn_hb'	9	6	30.38

Table C-3: Training settings for *Train station*.

The actual task has a length of 550 s and the targets are presented in six sets at times shown in Table C-4.

Set 1 in s	Set 2 in s	Set 3 in s	Set 4 in s	Set 5 in s	Set 6 in s
7.37	97.54	188.16	278.41	364.69	457.56
15.87	104.65	196.86	285.35	374.07	464.40
23.38	112.69	204.17	293.50	382.56	471.28
30.38	118.77	211.36	299.53	390.19	479.31
37.19	127.93	217.97	308.64	397.35	485.21
44.22	136.51	225.24	317.30	403.95	494.57
52.38	144.09	233.20	324.81	411.06	503.27
58.28	151.15	239.05	331.68	419.10	510.58
67.66	157.98	248.39	338.55	425.18	517.77
76.14	164.87	257.02	345.42	434.34	524.39
83.78	172.90	264.48	353.57	442.92	531.65
90.94	178.80	271.44	359.63	450.50	539.61

Table C-4: Valid target times for testing the *Train station* scenario.

C.1.3. Square

The dynamic localization task *Square* is trained in a 65 s session with an SPL of 65 dB with active motion tracker and instruction screen with settings as in Table C-5. Here, a negative loudspeaker number denotes a dynamic target in counterclockwise direction, while a positive number describes a dynamic target moving in clockwise direction.

Target	Loudspeaker	SNR	Time in seconds
'Tram'	12	10	0.43
'Tram'	-3	10	16.10
'Tram'	6	10	31.10
'Tram'	-9	10	47.62

Table C-5: Training settings of dynamic localization task *Square*.

The actual test consists of 24 presentations, twice on each loudspeaker and once for each direction with duration of 365 s. The list of the targets that are randomized for the actual test can be seen in Table C-6.

Set 1 in s	Set 2 in s
0.43	180.83
16.10	196.50
31.10	211.50
47.62	228.02
60.56	240.96
76.23	256.63
91.23	271.63
107.75	288.15
120.70	301.10
136.37	316.77
151.37	331.77
167.89	348.28

Table C-6: Target times of dynamic localization task *Square* for actual test.

C.2. Instructions

At the beginning of each new experiment the participants are instructed with general information:

General information:

This is an experiment about localization of a sound source. Localization is the ability to tell from which direction the sound comes. There are 3 different realistic scenes with 3 different tasks to perform. Feel free to rotate your head during the experiment, but please try to stay in the center of the loudspeaker ring. All in all, the testing will take 45 minutes. If you need a break between the tasks, please let the investigator know. Before each of the tests there is a presentation and a training session to become familiar with the scenario.

Each task is introduced separately by an illustration of the scenario and some information about it. The participants are asked to read the text and then proceed to presentation and training session by pushing a button.

Street:

In this scenery you are standing at a crossroad. From time to time a bicycle bell will ring. Your task will be to identify the loudspeaker from where the bell came from. Please press the button on the screen which represents the identified loudspeaker. Every sound will be presented only once within a short time frame. The duration of this whole task will be around 8 minutes.

Train station:

In this scenery you are standing in the middle of a large train station. From time to time you will hear a horn of a luggage transporter. Your task will be to identify the loudspeaker from where the horn came from. Please press the button on the screen which represents the identified loudspeaker. Every sound will be presented only once within a short time frame. The duration of this whole task will be around 10 minutes.

Square:

In this scenery you are standing at a tram station. From time to time a tram will approach you. Focus on that tram, because there will also be a tram further away from you. Your task will be to follow the tram with your head. After the tram braked and stopped, press the button on the screen which represents the location of the tram. After that you will be asked to identify the direction in which the tram was moving. The duration of this whole task will be around 6 minutes.

C.3. Feedback screen

The feedback screen for localization tasks is a ring of buttons labelled with the corresponding loudspeaker numbers. Below is a white bar that shows progress in the experiment. In training session the correct button is marked with green color after feedback, while in case of a wrongly pushed button it is marked with red. During an actual task the pushed button is marked with blue for a short time (Figure C.1).

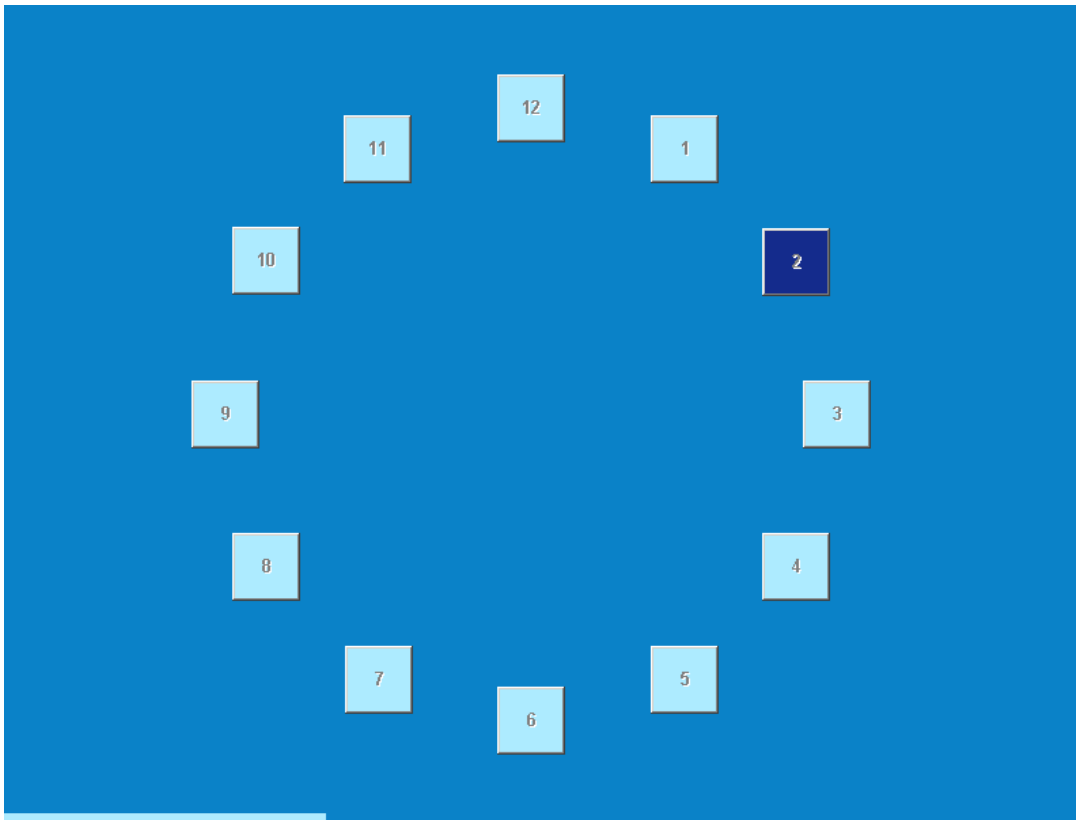

Figure C.1: Feedback screen on secondary monitor during localization experiment with given feedback on loudspeaker 2 and progress bar at the bottom.

For dynamic target localization arrows appear after indicating perceived position of the tram. These arrows indicate the direction the tram has been moving as shown in Figure C.2.

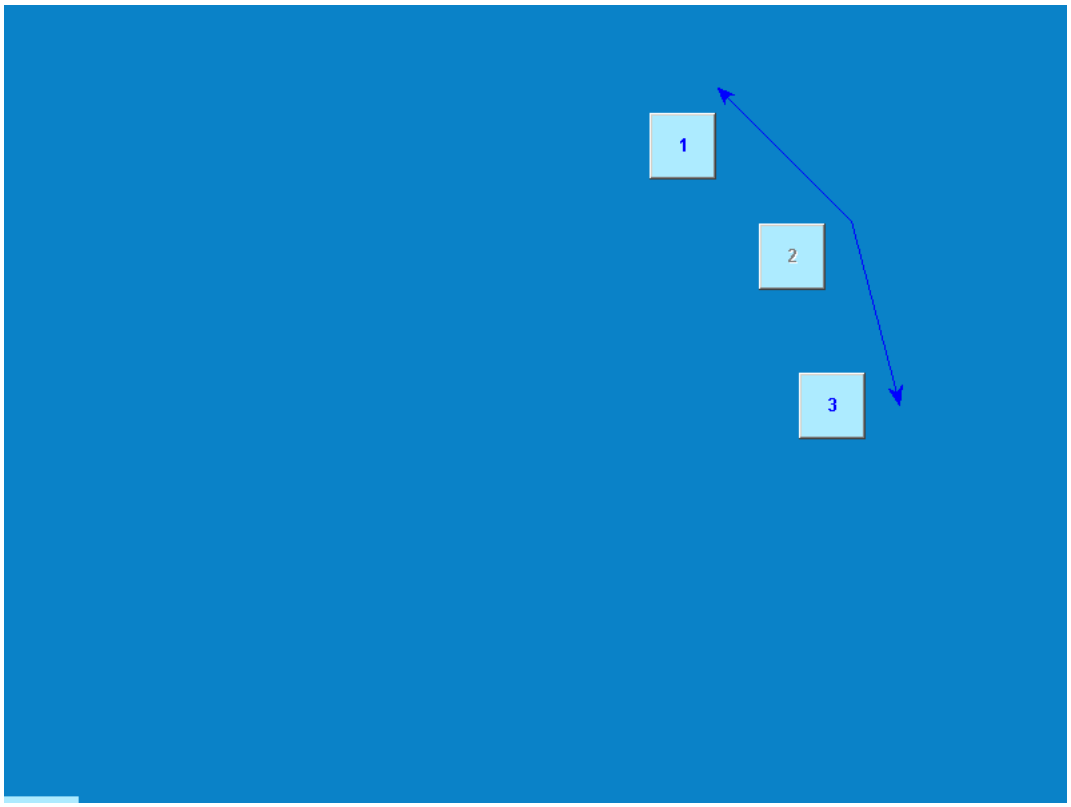

Figure C.2: Feedback screen to capture perceived direction of a dynamic target source.

C.4. Scenarios

A whole set of different scenarios and targets are currently saved in the library and can be accessed by the program as listed in Table C-7.

Background	Target
'street'	'bikebell'
'train_station'	'horn_hb'
'square'	'tram' (dynamic)
'construction_site'	'horn'
'car'	'BT1' (Speech: Basler sentences)
'train'	
'cafeteria'	

Table C-7: List of scenarios and targets in the library.

Literature

[1] A. D. Pierce, Acoustics: An Introduction to Its Physical Principles and Applications, Melville, NY: Acoustical Society of America, 1989.

[2] J. Jackson, Klassische Elektrodynamik, de Gruyter, 2006.

[3] M. Vorlaender, "Computer simulations in room acoustics: Concepts and uncertainties," *The Journal of the Acoustical Society of America 133(3),* pp. 1203-1213, 2013.

[4] S. Bharitkar and C. Kyriakakis, Immersive Audio Signal Processing, New York, NY: Springer, 2006.

[5] J. O. Smith, Mathematics of the Discrete Fourier Transform with Audio Applications, W3K Publishing, 2007.

[6] T. Funkhouser, I. Carlbom, G. Elko, G. Pingali, M. Sondhi and J. West, "A beam tracing approach to acoustic modeling for interactive virtual environments," *Proceedings of the 25th annual conference on Computer graphics and interactive techniques,* pp. 21-32, 1998.

[7] P. M. Morse and K. U. Ingard, Theroetical Acoustics, New Jersey: Princeton University Press, 1986.

[8] T. Haslwanter, Sensory Systems: Biological Organisms, an Engineer's Point of View., Wikibooks, 2013.

[9] J. Breebaart and C. Faller, Spatial Audio Processing, Blackwell Publishing, 2007.

[10] L. Chittka and A. Brockmann, "Perception Space—The Final Frontier," *PLoS Biol, 3(4),* p. 137, 2005.

[11] J. C. Middlebrooks and D. M. Green, "Sound localization by human listeners," *Annual Review Psychology, 42,* pp. 135-159, 1991.

[12] G. Keidser, K. Rohrseitz, H. Dillon, V. Hamacher, L. Carter, U. Rass and E. Convery, "The effect of multi-channel wide dynamic range compression, noise reduction, and the directional microphone on horizontal localization performance in hearing aid wearers," *International Journal of Audiology, 45,* pp. 563-579, 2006.

[13] E. Cherry, "Some experiments on the recognition of speech, with one and with two Ears," *The Journal of the Acoustical Society of America 25(5),* pp. 975-979, 1953.

[14] A. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acustica 86,* pp. 117-128, 2000.

[15] M. F. Mueller, A. Kegel, S. M. Schimmel, N. Dillier and M. Hofbauer, "Localization of virtual sound sources with bilateral hearing aids in realistic acoustical scenes," *Journal of the Acoustic Society of America, 131 (6),* pp. 4732-4742, 2012.

[16] A. W. Mills, "On the Minimum Audible Angle," *Journal of the Acoustic Society of America, 30 (4),* pp. 237-246, 1958.

[17] D. R. Perrott and A. D. Musicant, "Minimum auditory movement angle: Binaural localization of moving sound sources," *The Journal of the Acoustic Society of America, 62 (6),* pp. 1463-1466, 1977.

[18] S. Carlile and V. Best, "Discrimination of sound source velocity in human listeners," *The Journal of the Acoustic Society of America, 111 (2),* pp. 1026-1035, 2002.

[19] J. Lewald and W. H. Ehrenstein, "Spatial coordinates of human auditory working memory," *Cognitive Brain Research, 12,* pp. 153-159, 2001.

[20] K. B. Klink, M. Schulte and M. Meis, "Measuring listening effort in the field of

audiology: a literature review of methods, part 1," *Zeitschrift für Audiologie, 51 (2),* pp. 60-67, 2012.

[21] K. B. Klink, M. Schulte and M. Meis, "Measuring listening effort in the field of audiology: a literature review of methods, part 2," *Zeitschrift für Audiologie, 51 (3),* pp. 96-105, 2012.

[22] S. Gatehouse and W. Noble, "The Speech, Spatial and Qualities of Hearing Scale (SSQ)," *International Journal of Audiology, 43,* pp. 85-99, 2004.

[23] J. Blauert, Communication Acoustics, Springer, 2005.

[24] B. Xie, Head-Related Transfer Function and Virtual Auditory Display, Plantation, FL: J. Ross Publishing, 2013.

[25] J. Blauert, The Technology of Binaural Listening: Modern Acoustics and Signal Processing, Springer, 2013.

[26] M. Akeroyd, J. Chambers, D. Bullock, A. Palmer, A. Summerfield, N. P. and S. Gatehouse, "The binaural performance of a cross-talk cancellation system with matched or mismatched setup and playback acoustics," *The Journal of the Acoustical Society of America 121(2),* pp. 1056-1069, 2007.

[27] T. Lentz, D. Schroeder, M. Vorlaender and I. Assenmacher, "Virtual reality system with integrated sound field simulation and reproduction," *EURASIP Journal on Advances in Signal Processing,* 2007.

[28] G. Dickins, "Soundfield Representation, Reconstruction and Perception," Master Thesis, ANU Canberra, 2003.

[29] V. Pulkki, "Localization of Amplitude-Panned Virtual Sources II: Two- and Three-Dimensional Panning," *Journal of the Acoustic Society of America, 49:9,* pp. 753-767, 2001.

[30] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *Journal of the Acoustic Society of America, 45:6,* pp. 456-466, 1997.

[31] M. Gerzon, "Periphony: With-Height Sound Reproduction," *Journal of the Audio Engineering Society, 21,* pp. 2-10, 1973.

[32] R. Elen, "Ambisonics: The Surround Alternative," *Surround,* 2001.

[33] Y. Kahana, P. A. Nelson, O. Kirkeby and H. Hamada, "A multiple microphone recording technique for the generation of virtual acoustic images," *Journal of the Acoustic Society of America, 105(3),* pp. 1503-1516, 1998.

[34] O. Kirkeby, P. A. Nelson, F. Orduna-Bustamante and H. Hamada, "Local sound field reproduction using digital signal processing," *Journal of the Acoustic Society of America, 100(3),* pp. 1584-1593, 1996.

[35] M. Mueller, A. Kegel, S. Schimmel, N. Dillier and M. Hofbauer, "Localization of virtual sound sources in realistic and complex acoustical scenes," Technical Report, Zuerich, 2010.

[36] S. Smith, The Scientist & Engineer's Guide to Digital SIgnal Processing, California Technical Pub, 1997.

[37] M. Vorlaender, Auralization - Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality, Springer, 2008.

[38] E. Wenzel and S. Foster, "Realtime digital synthesis of virtual acoustic environment," *Proceedings of the 1990 symposium on interactive 3D graphics,* pp. 139-140.

[39] M. Hawley, R. Litovsky and H. Colburn, "Speech intelligibility and localization in a multi-source environment," *The Journal of the Acoustical Society of America 105(6),* pp. 3436-3448.

[40] T. May, S. van de Par and A. Kohlrausch, "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," *IEEE Transactions on Audio, Speech, and Language Processing 20/7,* pp. 2016-2030, 2012.

[41] W. Nelson, R. Bolia, M. Ericson and R. McKinley, "Spatial audio displays for

speech communications: Comparison of free field and virtual acoustic environments," *Proceedings of the Human Factors and Ergonomics Society 43,* pp. 1202-1205, 1999.

[42] M. Rychtarikova, T. van den Bogaert, G. Vermeir and J. Wouters, "Binaural sound source localization in real and virtual rooms," *Journal of the Audio Engineering Society 57,* pp. 205-220, 2009.

[43] S. Schimmel, M. Mueller and N. Dillier, "Binaural models and virtual acoustics to study spatial perception," *Journal of Hearing Science 1(2),* pp. 79-82, 2011.

[44] T. Takeuchi and P. Nelson, "Optimal source distribution for binaural synthesis over loudspeakers," *The Journal of the Acoustical Society of America 112(6),* pp. 2786-2797, 2002.

[45] D. Ward and G. Elko, "A robustness analysis of 3D audio using loudspeakers," *Proceeding 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,* pp. 191-194, 1999.

[46] J. Meyer, Acoustics and the Performance of Music, New York, NY: Springer, 2009.

[47] T. Graemer, "Efficient modeling of head movements and dynamic scenes in virtual acoustics," Master Thesis, D-ITET: ETH Zuerich, 2010.

[48] C. B. Lang and N. Pucker, Mathematische Methoden in der Physik, Heidelberg: Spektrum Akademischer Verlag, 2010.

[49] O. Kirkeby and P. A. Nelson, "Reproduction of plane wave sound fields," *Journal of the Acoustic Society of America, 94(5),* pp. 2992-3000, 1993.

[50] V. Pulkki, "Uniform Spreading of Amplitude Panned Virtual Sources," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, 1999.

[51] T. Lossius, P. Baltazar and T. de la Hogue, "DBAP- Distance-Based

Amplitude Panning," *Proceedings of 2009 International Computer Music Conference,* 2009.