Binaural Audio Signal Processing Using Interaural Coherence Matching

THÈSE N^{O} 4643 (2010)

PRÉSENTÉE LE 27 AVRIL 2010

À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS

LABORATOIRE DE COMMUNICATIONS AUDIOVISUELLES 1

PROGRAMME DOCTORAL EN INFORMATIQUE, COMMUNICATIONS ET INFORMATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Fritz MENZER

acceptée sur proposition du jury:

Prof. M. Hasler, président du jury
Prof. M. Vetterli, Dr C. Faller, directeurs de thèse
Dr J. Breebaart, rapporteur
Dr J.-M. Jot, rapporteur
Dr H. Lissek, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2010

Contents

Abstract

Binaural room impulse responses (BRIRs) characterize the transfer of sound from a source in a room to the left and right ear entrances of a listener. Applying BRIRs to sound source signals enables headphone listening with the perception of a three dimensional auditory image. BRIRs are usually linear filters of several hundred milliseconds to several seconds length. The waveforms of the BRIRs contain therefore a vast amount of information. This thesis studies the modeling of BRIRs with a reduced set of parameters. It is shown that late BRIR tails can be modeled perceptually accurately by considering only the time-frequency energy decay relief and frequency dependent interaural coherence (IC). This insight on BRIR modeling enables a number of algorithms with advantages over the previous state of the art. Three such algorithms are proposed:

The first algorithm makes it possible to obtain BRIRs by measuring room properties and listener properties separately, vastly reducing the number of measurements necessary to measure listener-specific BRIRs for a number of listeners and rooms. The listener properties are measured as a head related transfer function (HRTF) set and the room properties are measured as a B-format[1] room impulse response (RIR). It is shown how to combine the HRTF set of the listener with a B-format RIR to obtain BRIRs for that room individualized for the listener. This technique uses the insight on BRIR perception by computing the BRIR tail as a frequency dependent, linear combination of B-format channels, designed to obtain the desired energy decay relief and interaural coherence.

A serious problem related to convolving sound source signals with BRIRs is the computational complexity of implementing long BRIRs as finite impulse response (FIR) filters. Inspired by the perceptual experiments on BRIR tails, a modified Jot reverberator is proposed, simulating BRIR tails with the desired frequency dependent interaural coherence, requiring significantly less computational power than direct application of BRIRs. Also inspired by the perception of BRIRs, an extension of this reverberator is proposed, modeling efficiently the reverberation tail with the correct coherence and also distinct early reflections using two parallel feedback delay networks.

If stereo signals are played back using headphones, unnatural binaural cues are given to the listener, e.g. interaural level difference (ILD) changes not accompanied by corresponding interaural time difference (ITD) changes or diffuse sound with unnat-

[1] B-format refers to a 4-channel signal recorded with four coincident microphones: one omni and three dipole microphones pointing in orthogonal directions.

ural IC. In order to simulate stereo listening in a room and to avoid these unnatural cues, BRIRs can be applied to the left and right stereo channels. Besides the computational complexity associated with applying the BRIR filters, this technique has a number of disadvantages. The room associated with the used BRIRs is imposed on the stereo signal, which usually already contains reverberation and applying BRIRs leads to a change in reverberation time and to coloration. A technique is proposed in which the direct sound is rendered using data extracted from HRTFs and the ambient sound contained in the stereo signal is modified such that its coherence is matched to the coherence of a binaural recording of diffuse sound, without modifying its spectrum.

Implementations of reverberators based on general feedback-delay networks (e.g. Jot reverberators) can require a high number of operations for implementing the so-called feedback matrix. For certain applications where the number of channels needs to be high, such as decorrelators, this can pose a real problem. Special types of matrices are known which can be implemented efficiently due to matrix elements having the same magnitude. However, the complexity can also be reduced by introducing many zero elements. Different types of such sparse feedback matrices are proposed and tested for their suitability in Jot reverberators. A highly efficient feedback matrix is obtained by combining both approaches, choosing the nonzero elements of a sparse matrix from efficiently implementable Hadamard matrices.

Keywords: signal processing, binaural audio, binaural reverberation, binaural rendering, 3D audio, interaural coherence, auditory perception, diffuse sound, early reflections, late reverberation, stereo playback, B-format.

Résumé

Des binaural room impulse responses (BRIRs) caractérisent le transfert de son d'une source dans une salle aux deux oreilles d'un auditeur. Appliquer des BRIRs à des signaux de sources sonores permet la perception d'une image auditoire en trois dimensions lors de l'écoute avec un écouteur. Les BRIRs sont normalement des filtres linéaires d'une longueur de plusieurs centaines de millisecondes. La forme d'onde d'une BRIR contient donc une quantité d'information considérable. Cette thèse étudie la modélisation de BRIRs avec un nombre de paramètres réduit. Il est montré que la réverbération tardive contenue dans une BRIR peut être modélisée sans dégradation perceptible en considérant seulement le "energy decay relief" (EDR) et la cohérence interaurale en fonction du temps et de la fréquence. Ce résultat sur la modélisation de BRIRs permet l'implémentation d'algorithmes avancés. Trois algorithmes sont proposés:

Le premier algorithme permet d'obtenir des BRIRs en mesurant des propriétés de salles et des propriétés d'auditeurs séparément, réduisant ainsi considérablement le nombre de mesures nécessaire pour obtenir des BRIRs individuelles pour plusieurs auditeurs et plusieurs salles. Les propriétés d'un auditeur sont mesurés comme un ensemble de head related transfer functions (HRTFs) et les propriétés d'une salle sont mesurés comme une réponse impulsionnelle B-format[2]. La technique proposée utilise les connaissances sur les BRIRs pour modéliser la partie tardive de la BRIR comme une combinaison linéaire et dépendante de la fréquence des canaux B-format, conçue pour obtenir la cohérence et l'EDR désiré.

Une problème majeur qui empêche l'utilisation de BRIRs dans des systèmes à temps réel est la complexité de calcul d'une implémentation d'une BRIR comme filtre FIR. Inspiré par les résultats des expériences sur la perception de BRIRs, un réverbérateur de Jot modifié est proposé, permettant de contrôler précisément la cohérence interaurale des signaux de sortie. Egalement inspiré par les résultats perceptuels, une extension de ce réverbérateur est proposée, modélisant non seulement la cohérence mais aussi les réflexions précoces de façon efficace avec deux feedback delay networks parallèles.

Quand des signaux stéréo sont reproduits avec des écouteurs, ce ne correspondent pas à une situation d'écoute naturelle. Par exemple, en champ libre le rapport d'intensité entre l'oreille gauche et l'oreille droite est lié à une différence de temps d'arrivée, ce qui n'est pas le cas pour un signal stéréo reproduit avec un casque et

[2]B-Format désigne un signal à 4 canaux enregistré avec quatre microphones coïncidents: un omni et trois dipôles orientés perpendiculairement

également la cohérence d'un champ sonore diffus enregistré en stéréo ne correspond pas à la cohérence interaurale d'un champ sonore diffus. Une méthode connue pour résoudre ce problème est d'appliquer des BRIRs aux canaux d'un signal stéréo. A part la complexité de calcul, cette méthode a également le désavantage que l'effet d'une salle est appliqué au signal stéréo, changeant les temps de réverbération et le spectre du signal. Dans cette thèse une méthode est proposée où le son direct est rendu en utilisant des données HRTF, et le son diffus est rendu en contrôlant la cohérence interaurale, sans modification du spectre.

L'implémentation d'un réverbérateur de Jot peut nécessiter un grand nombre d'opérations de calcul pour la multiplication avec la matrice de feedback. Certaines matrices sont connues pour avoir une implémentation efficace, grace à un nombre d'éléments ayant la même valeur absolue. Cependant, il est possible de réduire la complexité en introduisant un grand nombre d'éléments égaux à zéro. Différents types de matrices creuses sont proposés et testés dans des réverbérateurs. Une matrice très efficace est obtenue en combinant les deux approches, remplissant les coefficients non nuls avec des coefficients de matrices de Hadamard.

Mots-clés: traitement du signal, audio binaural, réverbération binaurale, rendu binaural, audio 3D, cohérence interaurale, perception auditive, son diffus, réflexions précoces, son stéréo, B-format.

Acknowledgments

I would like to thank my two thesis supervisors Martin Vetterli and Christof Faller for giving me the opportunity to study for a Ph.D. degree at LCAV. In particular, I would like to thank Martin for the freedom he gave me in choosing research topics and Christof for generously sharing his knowledge on spatial audio with me.

A special thank you is reserved for the experts of my thesis committee, Jean-Marc Jot, Jeroen Breebaart, and Hervé Lissek, whose suggestions and insightful comments were very useful for me.

Many thanks go also to everybody at EPFL's Electromagnetics and Acoustics Laboratory (LEMA) for their kind support with HRTF and BRIR measurements and I want to thank Olaf Blanke, Anna Brooks and Pär Halje for the collaboration on the footsteps experiment which was an exciting and enriching experience for me.

Furthermore, many thanks to all the students working on the semester projects and the internship that I supervised and to all the subjects of the listening tests and the footsteps experiment for spending their time for my research.

And last but not least I want to say "kheily mamnoon" to my wife for her patience and her support while I was working on the thesis and "vielen Dank" to my parents for their support and their encouragement.

This research project was partially funded by Swiss National Science Foundation (SNSF) grant 200021-109406.

Chapter 1

Introduction

1.1 Binaural audio

Binaural audio refers to techniques that produce two-channel signals to be played back with headphones and that contain so-called binaural cues, i.e. signal properties that enable the listener to determine the position of a sound source, the type of the listening environment and even the material with which the walls of the listening environment were covered (e.g. carpets or tiles).

While one can only marvel at the capabilities of the human auditory system to extract so much information from just two audio channels, these capabilities in general are not matched by the current state of the art in audio signal processing. Even a seemingly simple task such as determining the number and the positions of sound sources present in a signal can bring state-of-the-art localization models to their limits [Faller, 2004, see Figure 7.4].

Even though there is no complete understanding yet of how the perception of binaural audio signals functions, a number of techniques has been developed to generate audio signals that contain realistic binaural cues. The most straightforward such technique is binaural recording, i.e. recording a two-channel signal by either using two microphones placed near the ear canal entrances of a listener or by using an artificial head with built-in microphones. High-quality binaural recordings never fail to surprise the unsuspecting listener who, disbelievingly, needs to take off the headphones to make sure that the sound is really emanating from them. Good examples of binaural recordings are the Holophonic sound demos [Holophonic SA, 2006] and the "virtual barbershop" demo [Starkey Laboratories Inc., 2007]. While both examples are essentially binaural recordings, it is very likely that they have been enhanced using signal processing techniques. Unfortunately the exact nature of the applied algorithms remains the secret of the respective companies. While there are two patents related to the Holophonic recording method [Zuccarelli, 1982, 1987] describing a device for binaural recordings, similar to an artificial head, as well as phase and amplitude equalization, hardly anything is known to the general public about how the "virtual barbershop" demo was produced. Starkey Laboratories have however confirmed to be the author of this recording [Starkey Laboratories Inc., 2008].

In cases where binaural recording is not applicable, good results can be obtained

by applying so-called binaural room impulse responses to dry (i.e. reverberation-free) audio signals. A binaural room impulse response (BRIR) is an impulse response recorded using an artificial head and a sound source in a room. It generally depends on the room, the position of the source, the position and the orientation of the head, and, if the source is not omnidirectional, the orientation of the source. In principle, making a binaural recording and applying the correct BRIRs to the dry source signals is equivalent. In the first case, the convolution of the dry signals with the BRIRs is performed naturally by the room and in the second case the convolution is performed artificially by a signal processor, but in both cases simply a convolution with an impulse response is applied to a signal. Unfortunately, BRIRs are a real alternative to binaural recordings only in simple cases where neither the sources nor the listener are moving. In cases with moving sources or a moving listener, a very high number of BRIRs would have to be recorded, which is likely to be more complicated than making a binaural recording.

In complex cases that involve moving sources, often other binaural rendering algorithms are applied. These algorithms, known as binaural reverberation algorithms, simulate the convolution with binaural room impulse responses. Normally this is done by simulating the reflections introduced by the environment using delays, filters, and head-related transfer functions (HRTFs). HRTFs, also known as head-related impulse responses (HRIRs), can be considered as BRIRs recorded in an anechoic environment, i.e. containing only the direct sound and no reflections.

Besides the above-mentioned techniques for generating binaural audio content "from scratch", there has also been a recent interest in converting existing non-binaural recordings into simulated binaural recordings. One application is the creation of binaural sound tracks on movie DVDs based on the original stereo or surround sound track. There are companies specialized in this application, e.g. [Mo'Vision]. In the same field there is also an interest in enhancing stereo signals such that they sound more realistic when played back using headphones [Breebaart and Schuijers, 2008].

1.2 Thesis motivation

The number of devices with a potential for binaural audio applications has dramatically increased in the past three decades, since the introduction of the Walkman by Sony in 1979. While 30 years ago most people were not regularly listening to audio content using headphones, nowadays headphone playback has become ubiquitous, for example to listen to music using a portable MP3 player, a mobile phone, or a laptop. Furthermore, since the introduction of the Game Boy by Nintendo in 1989 until spring 2010, approximately 400 million handheld game consoles with stereo headphone connectors have been sold [Wikipedia, 2010b,c,g,e,f,d], 255 million out of which belong to the latest generation [Wikipedia, 2010a], brought to the market starting from 2004 (numbers including Apple's iPhone and iPod Touch series).

While a "binaural audio application" for a mobile cassette player from the 1980's would essentially be a binaural recording stored on the magnetic tape of a stereo cassette, current audio devices contain a significant processing capability enabling them to perform binaural processing of audio signals in real time. Therefore it can be expected that binaural processing will appear in more and more devices in the future, possibly even by means of a simple software upgrade to be installed by the user.

Besides the mentioned mobile devices, there are also other applications of binaural audio, such as computer games, where binaural audio is already used and advanced algorithms may enable a more widespread and consequent use in the future, or the so-called "headphone parties" where people listen to music using wireless headphones and where interesting application for binaural audio exist.

While there is a great potential for binaural audio due to the enormous number of devices with a headphone connector and adequate processing capabilities, only limited knowledge exists on how binaural cues are actually perceived. It is known that human beings locate sound sources using time delay and level difference cues as well as spectral cues [Blauert, 1997] and recent investigations [Faller and Merimaa, 2004] give explanations how processing in the human auditory system based on interaural coherence may make sound source localization robust. Yet, many open questions related to binaural audio remain. For example, there is no conclusive answer to the question what causes some audio signals to sound clearly externalized (i.e. "out of the head") while other signals do not have this property. Furthermore, not many convincing binaural audio rendering algorithms exist, and most commercially available tools sound clearly less good than binaural rendering based on BRIRs.

1.3 Contributions and overview

In the field of binaural audio, many methods have been proposed for adding binaural cues to audio signals, most of them based on HRTFs or BRIRs, the latter being motivated by the insight that reverberation and early reflections increase the perceived spaciousness and externalization of binaural audio signals [Begault, 1992; Begault et al., 2001]. Binaural reverberators have been developed and optimized for many years, and countless methods were developed for calculating, measuring, or improving HRTFs. However, in most of these efforts, the main attention was paid to binaural cues related to direct sound and early reflections while only basic processing was used to generate interaural cues related to diffuse sound. The main goal of this thesis was to make improvements to binaural audio signal processing methods by considering binaural cues related to diffuse sound, in particular the interaural coherence as a function of time and frequency.

In **Chapter 2** different methods for modeling reverberation tails of BRIRs are investigated. The interaural coherence as a function of time and frequency is identified as an important measure describing the interaural properties of BRIR tails, and the impact of the head orientation of the listener on the interaural coherence in the early part of the BRIR is studied. Methods for analyzing the interaural coherence of audio signals and for synthesizing signals with a given time- and frequency-dependent interaural coherence are proposed. Two subjective tests were carried out to investigate the perceptual importance of the time- and frequency-dependent interaural coherence of BRIR tails.

Chapter 3 describes a method for generating BRIRs from B-format RIRs and a set of HRTFs. Measuring binaural room impulse responses for different rooms and different persons is a complex and time consuming task. The proposed method allows to measure the room related and the head related properties of BRIRs separately, reducing the amount of measurements necessary for obtaining BRIRs for different rooms and different persons to a number of B-format RIR measurements per room

and one HRTF set per person. The BRIRs are modeled by applying an HRTF to the direct sound part of the B-format RIR and using a linear combination of the reflections part of the B-format RIR. The linear combination is determined such that the spectral cues as well as the frequency-dependent interaural coherence cues match those of corresponding directly measured BRIRs. A subjective test indicates that the computed BRIRs are perceptually similar to corresponding directly measured BRIRs.

Room impulse responses are in general considered to be separable temporally in a direct sound part, an early reflections part and a late reverberation part, and little attention is paid to the fact that in many impulse responses the early reflections and diffuse reverberation are highly overlapping in time. In **Chapter 4**, a novel method for efficiently implementing a binaural reverberator using two parallel feedback delay networks is presented, modeling the overlap of reflections and diffuse reverberation. Furthermore, a simple method of adapting a Jot reverberator in order to model BRIRs is described, taking into account the insights on the perception of BRIRs.

A method for adding realistic binaural cues to a stereo signal while maintaining all the other cues such as direct to reverberant sound ratio, reverberation time, early reflections and – to some extent – the overall spectrum of the signal is presented in **Chapter 5**. The stereo signal is separated into coherent and diffuse sound based on basic assumptions that hold for signals coming from a symmetric coincident microphone setup as well as for signals produced using amplitude panning and stereo reverberators. The coherent part is rendered using HRTFs and the diffuse part is rendered by matching its interaural coherence to the coherence of a binaural recording of diffuse sound.

Appendix A presents different methods for designing unitary mixing matrices for Jot reverberators with a particular emphasis on cases where no early reflections are to be modeled. Possible applications include diffuse sound reverberators and decorrelators. The trade-off between effective mixing among channels and the number of multiply operations per channel and output sample is investigated as well as the relationship between the sparseness of powers of the mixing matrix and the sparseness of the impulse response.

Appendix B describes the measurement of an extensive set of room impulse responses in a lecture hall. This set contains BRIRs, HRTFs, and B-format RIRs measured with a single setup. Using this set, it was possible to compare the BRIRs generated from HRTFs and B-format RIRs using the method presented in Chapter 3 with actually measured BRIRs. Furthermore, BRIRs from this set covering 72 azimuth angles and 7 elevation angles were also used in Chapters 2 and 4.

Chapter 2

Perceptual Aspects of Binaural Room Impulse Responses

2.1 Introduction

In the design of binaural reverberators [Jot et al., 1995; Borss and Martin, 2009] much attention is dedicated to precisely modeling reflections. While early reflections are known to influence the perception of spaciousness and externalization [Begault, 1992; Begault et al., 2001], literature on the precedence effect [Blauert, 1997; Litovsky et al., 1999] as well as the cue selection model [Faller and Merimaa, 2004] stipulate that reflections have only little or no influence on sound source localization. One of the goals of this chapter is to investigate different simplifications of BRIRs allowing to replace precise modeling of reflections by matching only statistical aspects of a BRIR. The statistical aspects considered are the power spectrum as a function of time for the left and the right channel and the interaural coherence between the channels as a function of time and frequency.

Two subjective tests were conducted in order to study the perceptual impact of replacing the reverb tail by two channels of filtered Gaussian noise whose interaural coherence was matched to a measured BRIR. Previously, both frequency-dependent and wideband coherence matching had been proposed for modeling the late BRIR tail [Menzer and Faller, 2008; Borss and Martin, 2009; Jot et al., 1995]. The first test investigated the differences between the two methods while the second test compared different ways of modeling also the time dependence of the interaural coherence.

This chapter is organized as follows: Section 2.2 introduces the concept of interaural coherence in the context of BRIRs and presents different ways of calculating the interaural coherence. Section 2.3 describes how the synthetic BRIRs for the listening test are generated and Section 2.4 provides a signal-level evaluation of the synthetic BRIRs. Sections 2.5 explains the stimuli set and the procedure of the listening tests. The results are shown in Section 2.6 and discussed in Section 2.7. Conclusions are drawn in Section 2.8.

2.2 Interaural coherence of BRIRs as a function of time and frequency

2.2.1 Definitions

The similarity between two audio signals has been studied extensively using statistical methods, and two measures are commonly used in the case of binaural signals. The first method produces a single value for a pair of signals, called interaural cross-correlation coefficient (IACC) [Damaske and Ando, 1972]:

$$\mathrm{IACC}_{s_L,s_R} = \max_{\tau} \left[\frac{\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} s_L(t) s_R(t+\tau) dt}{\sqrt{\left(\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} s_L^2(t) dt\right) \left(\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} s_R^2(t) dt\right)}} \right] , \qquad (2.1)$$

where $s_L(t)$ and $s_R(t)$ denote the left and right channels of the binaural signal in time domain.

The second method applies the frequency-dependent measure of coherence [Bendat and Piersol, 1971] to a pair of signals, resulting in the so-called interaural coherence (IC):

$$\mathrm{IC}_{S_L,S_R}(f) = \frac{|\langle S_L(f) S_R^\star(f) \rangle|}{\sqrt{\langle S_L(f) S_L^\star(f) \rangle \langle S_R(f) S_R^\star(f) \rangle}} , \qquad (2.2)$$

where $S_L(f)$ is the Fourier transform of $s_L(t)$ and $S_R(f)$ is the Fourier transform of $s_R(t)$, * denotes the complex conjugate and $\langle x \rangle$ is the expected value of x.

Both IACC and IC measure the similarity between two signals and have the property that independent signals will result in the measure being 0 and identical signals (up to a positive amplitude scaling factor) will result in the measure being 1. These measures are considered to be an important cue for binaural hearing related to perceptual aspects such as source width, envelopment and spaciousness [Blauert, 1997], and to the detection of direct sound in reverberant environments [Faller and Merimaa, 2004].

In this chapter, the signals of interest are two channels of a binaural room impulse response, supposed to be two sampled signals of length N, and denoted $h_L(n)$ and $h_R(n)$. Different measures of similarity between the two channels are used. A single, frequency independent value Φ_{FI} for the two signals is defined based on the IACC:

$$\Phi_{FI} = \max_{l} \left[\frac{\sum_{m=\max(0,-l)}^{\min(N-1,N-1-l)} h_L(m) h_R(l+m)}{\sqrt{\left(\sum_{m=0}^{N-1} h_L(m)^2\right) \left(\sum_{m=0}^{N-1} h_R(m)^2\right)}} \right] , \qquad (2.3)$$

where l is the temporal offset between the left and the right channel (assuming $|l| \ll N$) and m is the summation index used for summation over the whole length of the impulse response. This is a full-band interpretation of the interaural coherence.

A measure for the frequency-dependent interaural coherence is calculated based on the short-time Fourier transform (STFT) of $h_L(n)$ and $h_R(n)$, called hereafter

$H_L(i,k)$ and $H_R(i,k)$, where i denotes the frequency bin and k is the time frame index:

$$\Phi_{FD}(i) = \frac{\Re\left(\sum_{k=0}^{K} H_L(i,k)H_R(i,k)^\star\right)}{\sqrt{\sum_{k=0}^{K}|H_L(i,k)|^2 \sum_{k=0}^{K}|H_R(i,k)|^2}}, \tag{2.4}$$

where $\Re(x)$ stands for the real value of x and $y^\star$ denotes the complex conjugate of y. K stands for the number of timeframes that each signal has in the chosen STFT domain (K depends on frame size and overlap).

A difference between the proposed measure and the widely used measure (2.2) is that instead of the absolute value, the real value is used in the numerator of (2.4). This is necessary in order to detect negative correlation. For example in the case $s_l(n) = -s_r(n)$, $\Phi_{FD}(i)$ should be -1, not 1. Experimental data comparing (2.4) with its corresponding imaginary part [Borss and Martin, 2009] shows that for a binaural recording of diffuse sound the imaginary part is small compared to the real part. This means that the absolute value of (2.4) is close to the coherence as defined by [Bendat and Piersol, 1971], or in other words, for diffuse BRIRs (2.4) can be interpreted as a signed approximation of (2.2).

One reason why the interaural coherence of a BRIR should be considered as in (2.4) and not as in (2.3) is that due to the spacing of the ears of a listener, there is an inherent frequency-dependent bias of the interaural coherence [Cook et al., 1955]. For diffuse sound, the interaural coherence at low frequencies is generally higher than for high frequencies, and can even be negative at certain frequencies.

There are reasons to consider the interaural coherence of a binaural room impulse response (BRIR) also as a function of time. A BRIR normally consists of the head related impulse response (HRIR) for the direct sound, some early reflections, and a diffuse reverberation tail [Gardner, 1998]. Generally, it may be expected that the interaural coherence decreases with time, as the reflection density becomes higher.

To obtain a measure for the interaural coherence as a function of time and frequency, we replace the summation over all time frames in (2.4) by a smoothing operator $\mathcal{S}\{\}$ of the type of a weighted moving average operating along the time dimension:

$$\mathcal{S}\{H(i,k)\} = \sum_{m=-l}^{l} w(m)H(i,k+m), \tag{2.5}$$

where $w(m)$ is a set of $2l+1$ weights for the moving average, e.g. a raised cosine window with 5 samples length. Using this operator instead of a summation leads to the expression of the time-frequency dependent interaural coherence $\Phi_{TF}(i,k)$ as

$$\Phi_{TF}(i,k) = \frac{\Re\left(\mathcal{S}\{H_L(i,k)H_R(i,k)^\star\}\right)}{\sqrt{\mathcal{S}\{|H_L(i,k)|^2\}\mathcal{S}\{|H_R(i,k)|^2\}}}. \tag{2.6}$$

2.2.2 Coherence changes due to head orientation

The significance of interaural coherence as a function of time and frequency can be illustrated by examining a set of BRIRs measured for rotational positions every 5°

using a KEMAR head and torso simulator mounted on a turntable. The room used for the recordings was a small lecture hall at our university (approximately 10 m wide and 14 m long) and the fixed sound source was in front of the listener for the head orientation 0°.

It has previously been shown that in closed rooms with one or more sound sources, the frequency-dependent interaural coherence depends on the orientation of the head and that this coherence variation is perceptually relevant [Mason et al., 2009]. In this section we investigate how the interaural coherence as a function of *time and frequency* changes when the head is turned.

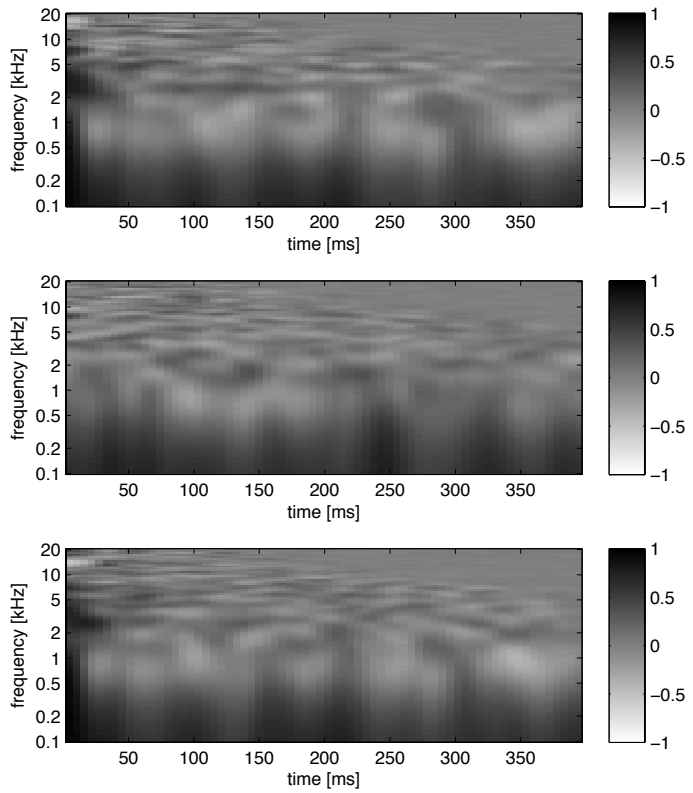

Figure 2.1: Interaural coherence as a function of time and frequency for BRIRs with different orientations of the head. **Top panel:** head orientation 0°. **Middle panel:** head orientation 90°. **Bottom panel:** head orientation 180°. The direct sound parts of the BRIRs (i.e. the first 5 ms) were not taken into account for these plots. An increased interaural coherence in the first 30 ms can be seen for the head orientations 0° and 180°. Coherence curves for this interval are shown in Figure 2.2.

Figure 2.1 shows $\Phi_{TF}(i,k)$ as defined in (2.6) for different orientations of the head relative to the sound source. On a qualitative level, it can be noticed that the

plots for 0° and 180° show a strong interaural coherence in the time interval up to approximately 35 ms. This is not observed in the BRIR for the head orientation 90°. A possible explanation is that for 0° and 180° the floor and ceiling reflections from the source at 0° add up coherently while this is not the case for the head orientation 90°. After approximately 35 ms all three graphs are very similar.

That the variation of the interaural coherence due to head orientation takes place mainly in the beginning of the BRIR is illustrated on a more quantitative level in Figure 2.2: the frequency dependent interaural coherence as defined in (2.4) is plotted for integration intervals 5 ms to 35 ms and 35 ms to 400 ms. For the first interval, in the frequency range from 2 kHz to 5 kHz, strong interaural coherence variations between -0.5 and 0.7 occur. For the second interval, the coherence varies only little and is qualitatively similar to the theoretical coherence for perfectly diffuse sound as derived by [Cook et al., 1955] for two omni microphones spaced at a distance of 25 cm. A larger distance than the actual ear-to-ear distance of the KEMAR head had to be used to compensate a bias in the theoretical model due to considering only two spaced microphones in free field and not modeling the diffraction of sound around the head. A more detailed model has been described in [Avni and Rafaely, 2009], showing that a rigid sphere model of the head makes the first minimum in the coherence curve appear at a lower frequency than for the spaced omni microphone model.

Figure 2.3 shows the interaural coherence as a function of frequency and head orientation, again separately for the early reflections and the late reverb tail. Strong and systematic variations depending on the head rotations can be observed in the frequency dependent interaural coherence integrated between 5 ms and 35 ms after the direct sound, in particular a high IC across all frequency bands up to 8 kHz at the head orientations 0° and 180°. Given that from the listener's point of view the sound source as well as the floor and ceiling reflections are at azimuth direction 0°, it can be expected that the coherence should be high for head orientations 0° and 180°, because for these orientations the direct sound and the floor and ceiling reflections add up coherently at the two ears. For the integration window 35 ms to 400 ms the coherence seems to be largely independent from the head rotation. This is not surprising because under the hypothesis that late reflections arrive with equal energy from all directions at the listener's head, the late BRIR should not depend on the head orientation. However, it may surprise that the BRIR seems to behave like perfectly diffuse sound already 35 ms after the direct sound. This may be due to the fact that the room in question was relatively small. For bigger rooms we would expect that the interaural coherence depends on the head rotation longer than only 35 ms, due to the presence of more distinct and later arriving early reflections.

2.3 BRIR synthesis

This section describes the different ways used to model BRIR tails based only on statistical aspects of the original BRIR. In all cases, the tail of the synthetic BRIR was generated from white Gaussian noise processed in a short-time discrete Fourier transform (STFT) domain in order to have a given interaural coherence and the same energy decay relief (EDR) as the measured BRIR. The EDR has been defined as the frequency distribution of the remaining energy in an impulse response as a function of time [Jot, 1992]. For the analysis and synthesis of BRIRs, STFTs with a frame

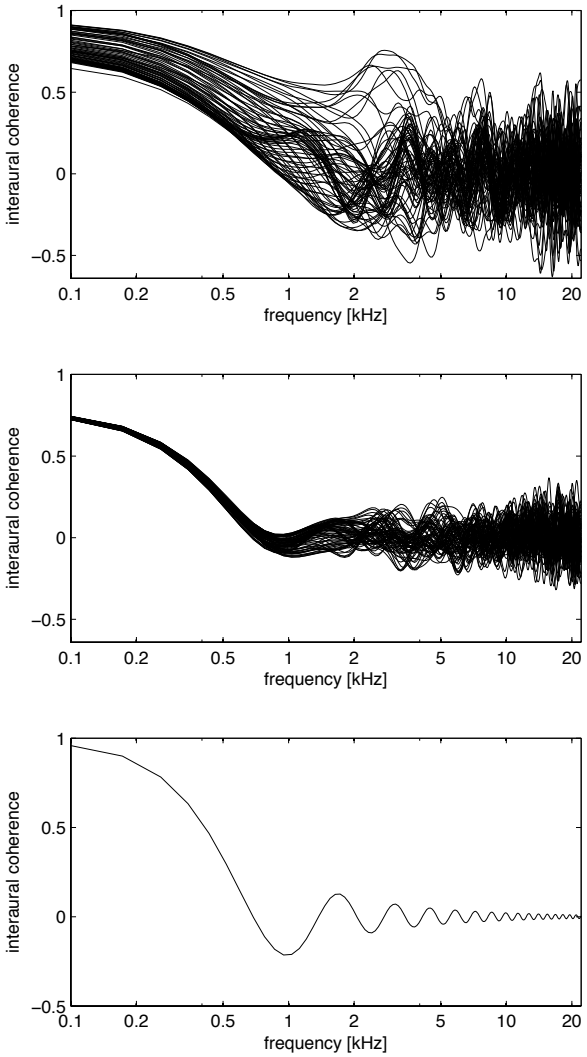

Figure 2.2: Frequency-dependent interaural coherence of a measured BRIR for different integration intervals and theoretical curve for diffuse sound. **Top panel:** Measured BRIR, time interval for integration 5 ms to 35 ms (dominated by early reflections). **Middle panel:** Measured BRIR, time interval for integration 35 ms to 400 ms (mostly diffuse reverberation). **Bottom panel:** Theoretical interaural coherence for perfectly diffuse sound as derived in [Cook et al., 1955] for two omni microphones spaced at a distance of 25 cm.

size of 1024 samples and a frame overlap of 50% were used and the sampling rate was 44.1 kHz. The first part of the synthetic BRIR was always taken from the measured BRIR. The only differences between the synthetic BRIRs are the split point between

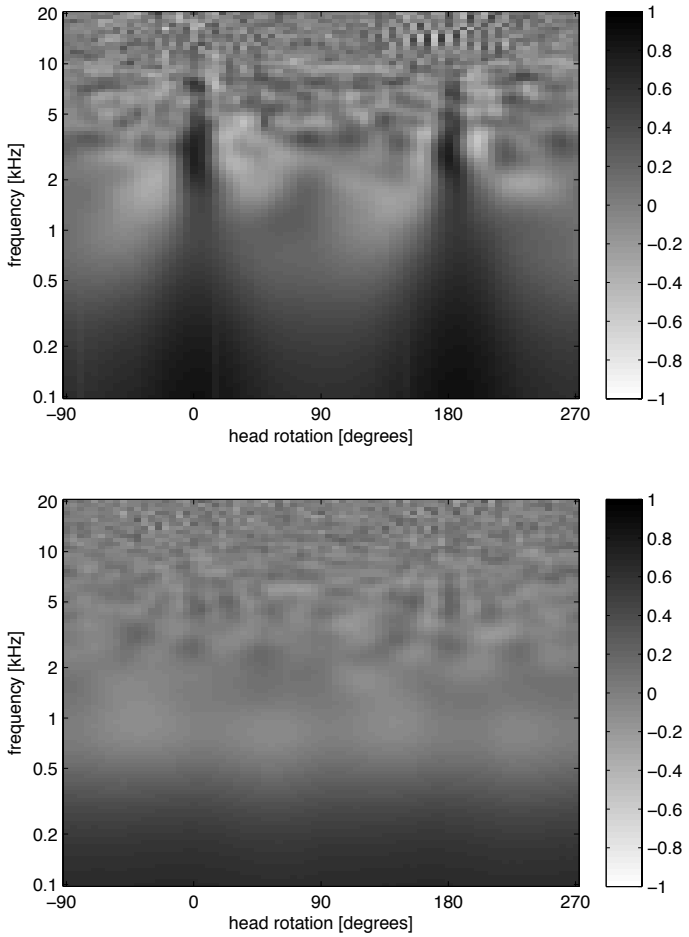

Figure 2.3: Frequency-dependent interaural coherence as a function of head rotation. **Top panel:** Time interval for integration 5 ms to 35 ms (dominated by early reflections). **Bottom panel:** Time interval for integration 35 ms to 400 ms (dominated by diffuse reverberation).

the measured and synthesized parts and the interaural coherence as a function of frequency and time in the synthesized part.

In the following, the synthetic BRIRs will be referenced by short names containing two letters and one of two numbers (see Table 2.1). The two letters stand for the type of coherence matching: **FI** stands for “frequency independent”, **FD** for "frequency dependent”, and **TF** for “time-frequency”. The number following the letters indicates how many milliseconds at the beginning of the BRIR are copied from the reference BRIR. For example, FI80 stands for a BRIR where the first 80 milliseconds are taken from the reference and the rest is synthesized using frequency independent coherence

matching. If the number in the short name is of the form **FD**x/y, where x and y are two numbers, this indicates that the coherence matching was done separately in two parts: the first part starts at x ms and ends at y ms and the second part goes from y ms to the end of the BRIR. This takes into account the strong temporal dependence of coherence in the beginning of the BRIR. FD5/35 means therefore that the first 5 ms were taken from the reference BRIR, the next 30 ms were synthesized with frequency dependent coherence matching using parameters estimated in the interval between 5 ms and 35 ms after the direct sound, and everything after 35 ms was synthesized using frequency dependent coherence matching based on the parameters estimated from the reference BRIR in the interval between 35 ms and the end of the BRIR. One exception of the rule described above is the BRIR names IC1. It refers to a BRIR generated using a mono noise source for the synthesis of the BRIR tail, leading to an interaural coherence close to 1.

Coherence matching	**Measured/synthetic split point**			
	5 ms	20 ms	80 ms	200 ms
Frequency-independent	FI5		FI80	
Frequency-dependent	FD5	FD20	FD80	FD200
FD, 2nd split point 25 ms	FD5/25			
FD, 2nd split point 35 ms	FD5/35			
Time-frequency	TF5			
IC=1	IC1			

Table 2.1: Synthetic BRIR types

In the following, the left and right measured BRIRs are denoted $h_L(n)$ and $h_R(n)$ and the left and right synthetic BRIRs will be denoted $\widetilde{h}_L(n)$ and $\widetilde{h}_R(n)$.

For the generation of the synthetic BRIRs, the following procedure is used: first, the interaural coherence of the measured BRIR tail for the desired time interval(s) of the synthetic BRIR tail is computed using the appropriate coherence measure. Then the synthetic BRIR tail is calculated in the STFT domain from two independent white Gaussian noise signals, denoted $N_1(i,k)$ and $N_2(i,k)$ in STFT domain (i is the frequency bin and k is the time frame index of the STFT coefficients):

$$\begin{aligned} \widetilde{H}_L(i,k) &= c(i,k)\,(a(i,k)N_1(i,k) + b(i,k)N_2(i,k)) \\ \widetilde{H}_R(i,k) &= d(i,k)\,(a(i,k)N_1(i,k) - b(i,k)N_2(i,k))\ , \end{aligned} \tag{2.7}$$

where $c(i,k)$ and $d(i,k)$ are factors that adapt the energy decay relief (EDR) of $\widetilde{H}_L(i,k)$ and $\widetilde{H}_R(i,k)$ to the EDR of the measured left and right BRIRs, respectively, while $a(i,k)$ and $b(i,k)$ control the interaural coherence of the synthetic BRIR. In order not to impose the fine structure of the measured BRIR's amplitude envelope on $\widetilde{H}_L(i,k)$ and $\widetilde{H}_R(i,k)$, $c(i,k)$ and $d(i,k)$ are smoothed in time using the smoothing operator defined in (2.5). The coefficients $a(i,k)$ and $b(i,k)$ are calculated as a function of the (potentially time- and frequency-dependent) interaural coherence $\Phi(i,k)$ and the time-smoothed short-time power spectrum estimates $P_1(i,k)$ and $P_2(i,k)$ of

the noise signals:

$$\begin{aligned} a(i,k) &= \sqrt{\frac{P_2(i,k)^2(1+\Phi(i,k))}{P_1(i,k)^2(1-\Phi(i,k))+P_2(i,k)^2(1+\Phi(i,k))}} \\ b(i,k) &= \sqrt{1-a(i,k)^2} \\ &= \sqrt{\frac{P_1(i,k)^2(1-\Phi(i,k))}{P_1(i,k)^2(1-\Phi(i,k))+P_2(i,k)^2(1+\Phi(i,k))}}\,, \end{aligned} \tag{2.8}$$

where $P_1(i,k) = \mathcal{S}\{|N_1(i,k)|^2\}$ and $P_2(i,k) = \mathcal{S}\{|N_2(i,k)|^2\}$ and $\mathcal{S}$ is the smoothing operator defined in (2.5).

In the case where the time-frequency interaural coherence is applied (TF5), the value of $\Phi(i,k)$ is calculated using (2.6), while in the cases where only the frequency-dependent interaural coherence is applied (e.g. FD5), the value calculated in equation (2.4) is used (i.e. $\Phi(i,k) = \Phi_{\mathrm{FD}}(i)$).

For FI5 and FI80, $a(i,k)$ and $b(i,k)$ used in (2.8) are calculated as follows:

$$\begin{aligned} a(i,k) &= a_{FI} = \sqrt{\frac{1+\Phi_{FI}}{2}} \\ b(i,k) &= b_{FI} = \sqrt{\frac{1-\Phi_{FI}}{2}}\,. \end{aligned} \tag{2.9}$$

Using single time-invariant parameters a_{FI} and b_{FI} corresponds to matching the coherence on average for the considered BRIR tail. This is in contrast to (2.8), where the parameters are adjusted as a function of time and frequency.

To obtain the raw synthetic BRIR tails $\widetilde{h}_L(n)$ and $\widetilde{h}_R(n)$, the inverse STFTs of $\widetilde{H}_L(i,k)$ and $\widetilde{H}_R(i,k)$ are calculated. These tails are further refined by matching their magnitude spectra to the measured BRIRs' magnitude spectra in the Fourier domain. Finally, the reverberation tail is joined with the first part of the original BRIR using a 0.2 ms cross-fade.

For the cases IC1, FI5, FI80, and FD5 to FD200, two different instances of Gaussian white noise were used for the synthesis in order to be able to estimate the influence of the particular noise instance on the perception of the synthesized BRIR. For the other cases (used only in the second listening test), only one noise instance was used because preliminary results showed that the noise instance had only little influence on the perception.

2.4 Objective evaluation

2.4.1 Measured reference BRIRs

For the study presented here, two BRIRs were used. BRIR number 1 was measured in a lecture hall at our university which is 10 m wide, 14 m long and whose height varies between 4 m in the front of the room and 2 m in the back of the room. A loudspeaker and a KEMAR head and torso simulator were placed in the front of the room, roughly corresponding to the position of a lecturer and a student in the first row,

respectively. The measured BRIR has a reverberation time RT60 of about 1 s. This BRIR represents a common listening situation, suitable for 3D audio reproduction. For the objective evaluation, this BRIR was used.

BRIR number 2 was measured with a Neumann KU 80 dummy head in an empty lecture hall with a reverberation time of about 2 s. The azimuth of the loudspeaker relative to the dummy head was 30°.

2.4.2 Analysis of synthetic BRIRs

In order to validate the BRIR synthesis method, the synthesized BRIRs were analyzed on the waveform level, in the spectral domain, and also in terms of their interaural coherence.

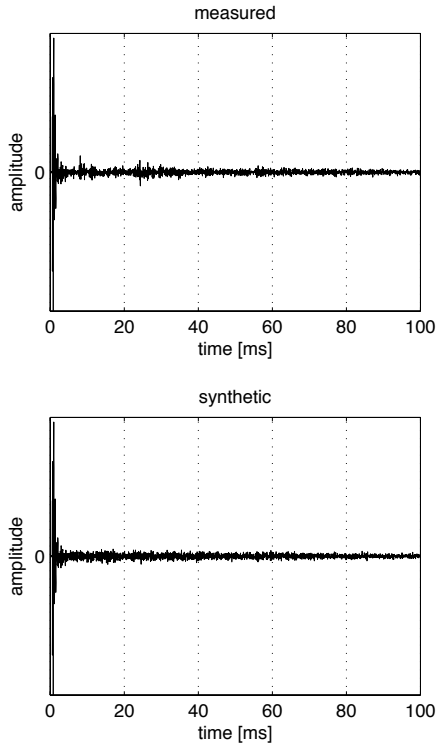

Figure 2.4: Waveforms of the first 100 ms of the measured BRIR and one synthetic BRIR. Only the left BRIR is shown. **Top panel**: measured BRIR. **Bottom panel**: synthetic BRIR (FD5). The BRIRs are matched to have the same EDR and frequency dependent interaural coherence, but the synthetic BRIR does not model any early reflections, which leads to visible differences in the waveform plots.

On the waveform level, it may be observed that only the exponential decay of the measured BRIR's amplitude envelope is imposed on the synthetic reverb tails,

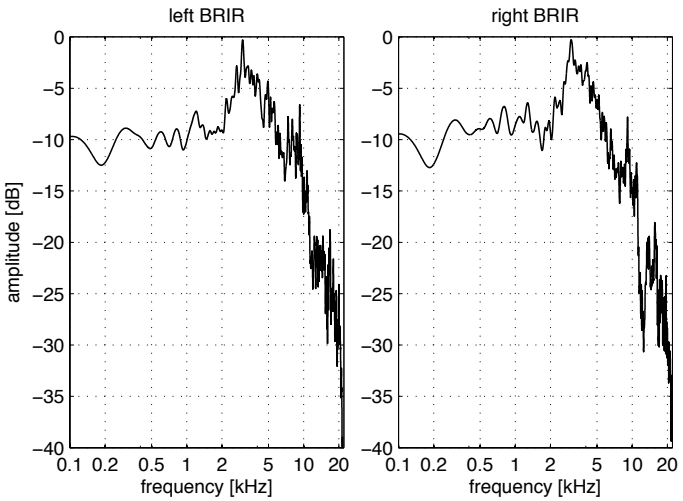

Figure 2.5: Spectrum of the left and right measured BRIR tail (starting 5 ms after the beginning of the BRIR).

but not the details of the amplitude envelope corresponding to the individual early reflections. This is illustrated in Figure 2.4 where the measured BRIR and a synthetic BRIR (FD5) are shown. In the measured BRIR, several early reflections can be seen, approximately between 5 ms and 40 ms. In the same time interval, no distinct early reflections can be observed in the synthetic BRIR. It has therefore been shown that the synthetic reverb tail does not model the amplitude envelope changes due to the early reflections. The only way in which the early reflections of the measured BRIR influence the synthetic BRIR tail is through their impact on the interaural coherence and on the spectrum.

Figure 2.5 shows the spectra of left and right measured BRIR tails. The spectral deviations of a synthetic BRIR (FD5) from the measured BRIR are shown in Figure 2.6. It can be seen that the spectra of the synthetic BRIR tail match the spectra of the measured BRIR closely, with a maximum deviation below 1 dB for frequencies up to 15kHz.

Besides the waveform and the spectrum, also the frequency-dependent interaural coherence of the synthetic BRIR tails was examined. Figure 2.7 shows the coherence of the frequency-independent cases FI5, FI80, and IC1, while Figure 2.8 shows the coherence of the frequency dependent cases FD5, FD20, FD80, and FD200. Figure 2.9 shows the coherence of the time-frequency matched cases FD5/25, FD5/35, and TF5. In all cases, regardless of the split point, the coherence was calculated on the time interval from 5 ms after the direct sound to the end of the BRIR.

For the cases FI5 and FI80 the frequency dependent interaural coherence is very different from the frequency dependent interaural coherence of the measured BRIR, as illustrated in Figure 2.7. In particular the high interaural coherence at low frequencies (up to 1 kHz) is not reproduced correctly, which can be perceived as a “roughness”

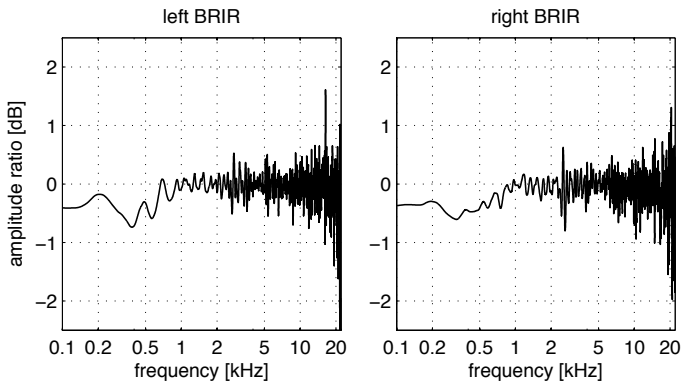

Figure 2.6: Deviation of the spectra (in dB) of the reverb tail of a synthesized BRIR (FD5) from the spectra of the measured BRIR tail.

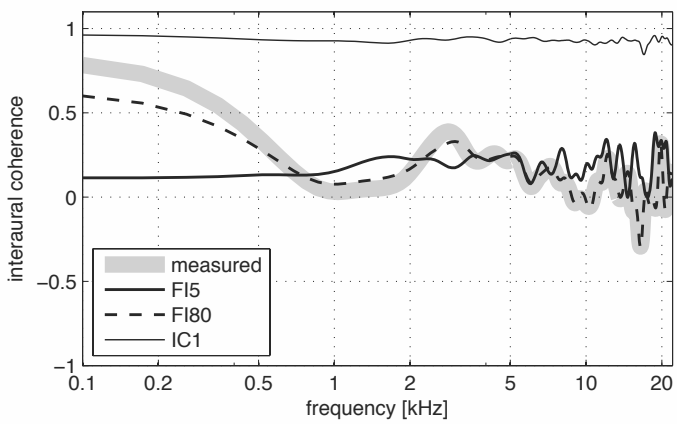

Figure 2.7: Interaural coherence of FI5, FI80, and IC1. The frequency-independent interaural coherence matching used in FI5 and FI80 produces a frequency-dependent coherence that is very different from the measured BRIR. This effect is stronger for FI5 than FI80. Even though for the synthesis of IC1 only a single noise channel was used, the modifications introduced by the EDR matching slightly modify the interaural coherence.

or "tickling" when an audio signal convolved with such a BRIR is played back using headphones. It has also been shown previously that variations in interaural coherence are particularly noticeable if the interaural coherence is close to 1 [Robinson and Jeffress, 1963; Culling et al., 2001; Breebaart and Kohlrausch, 2001b], which explains the particular sensitivity to the IC in the low frequency range.

It may also be noted that the frequency-independent coherence matching (case FI5) does not produce a perfectly flat coherence curve. One reason is that the spectral

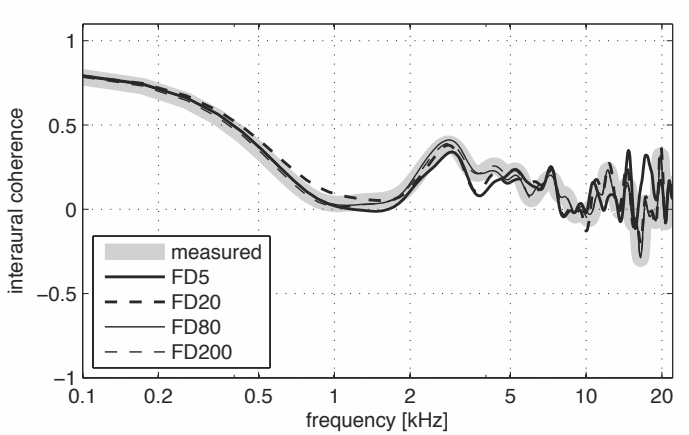

Figure 2.8: Interaural coherence of four synthetic BRIRs (FD5, FD20, FD80, FD200) and the coherence of the measured BRIR. Notice that the coherence matching works nearly perfectly in the low frequency range between 100 Hz and 500 Hz.

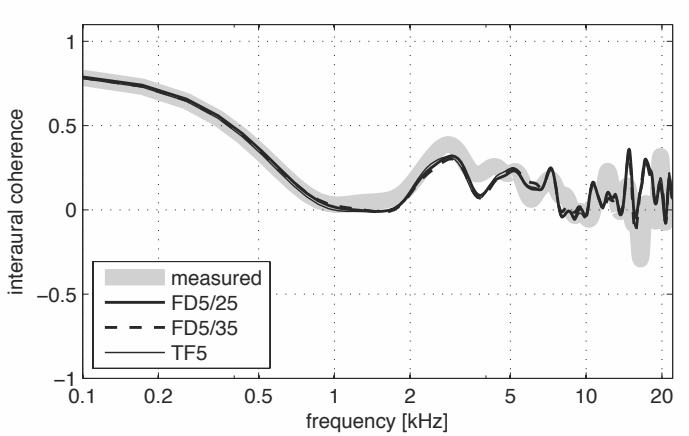

Figure 2.9: Interaural coherence of FD5/25, FD5/35, TF5. The 3 curves are almost identical to the frequency dependent interaural coherence of FD5 show in Figure 2.8.

power fluctuations of the noise are not taken into account, contrary to what is done in the cases FD5 to FD200. Figure 2.7 also shows that even in the case IC1 the interaural coherence is not perfectly 1 despite the fact that the whole reverb tail is synthesized from a single channel of noise. This shows that the EDR matching introduces slight variations in the interaural coherence.

For the other cases, where the interaural coherence was matched as a function of frequency, the coherence matching works very well in the low frequencies. The deviations are mostly below 0.1 for frequencies up to 10 kHz and below 0.02 for frequencies up to 500 Hz. These small deviations are assumed not to be perceptible [Culling et al., 2001]. Relatively large deviations occur above 10 kHz, reaching up to 0.3. However, since there is only little energy in the reverb tail above 10 kHz, we do not expect this to be a problem.

It may be noted that the frequency dependent interaural coherence for the time-frequency matched cases shown in Figure 2.9 are all very similar and resemble the coherence curve for the case FD5 shown in Figure 2.8. This indicates that modeling the time dependence of the interaural coherence does not significantly change the frequency-dependent coherence as defined in (2.4).

Finally, also the time-frequency interaural coherence as defined in (2.6) was evaluated for the cases FI5, FD5, FD5/35, TF5 and for the measured BRIR and the resulting "coherence reliefs" are shown in Figure 2.10. It can be seen that the cases FI5 and FD5 don't model the time dependence of the interaural coherence, while the cases FD5/35 and TF5 model this time dependence reasonably well.

2.5 Subjective evaluation

Two listening tests were conducted in order to answer the following questions:

- Does frequency-dependent interaural coherence matching perform significantly better than frequency-independent interaural coherence matching?
- Can coherence-matched noise transparently replace the reverb tail of a measured BRIR after a certain split point?
- Can the modeling of the temporal structure of the interaural coherence improve the perceived quality of a synthetic BRIR with a split point of 5 ms? In other words, can we improve the quality of a synthetic BRIR where everything except for the direct sound is synthesized by applying the correct time-frequency interaural coherence?

For both listening tests different synthetic BRIRs plus the reference BRIR were convolved with 4 different audio signals. The audio signals were samples of male speech, female speech, drums, and of a hand-clap cut to 8.5 ms length followed by 1 s of silence.

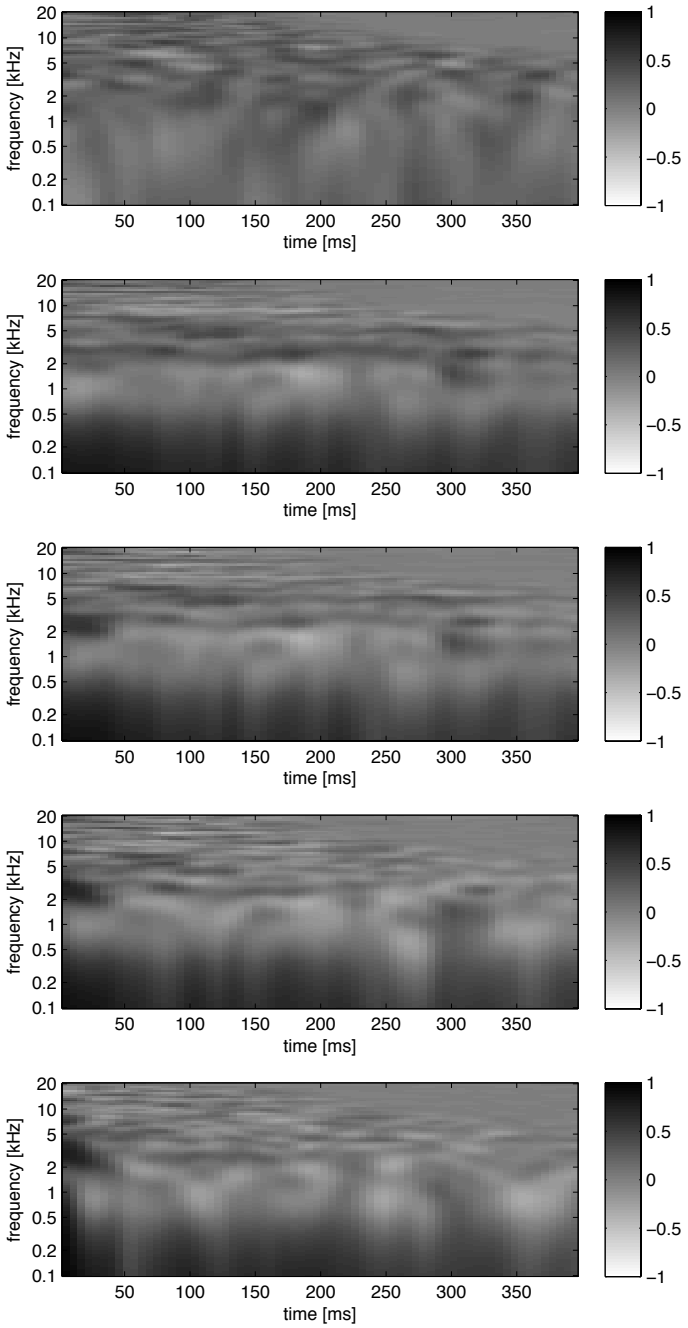

Figure 2.10: Interaural coherence as a function of time and frequency for measured BRIR 1 and different synthetic BRIRs. **From top to bottom:** FI5, FD5, FD5/35, TF5, measured BRIR.

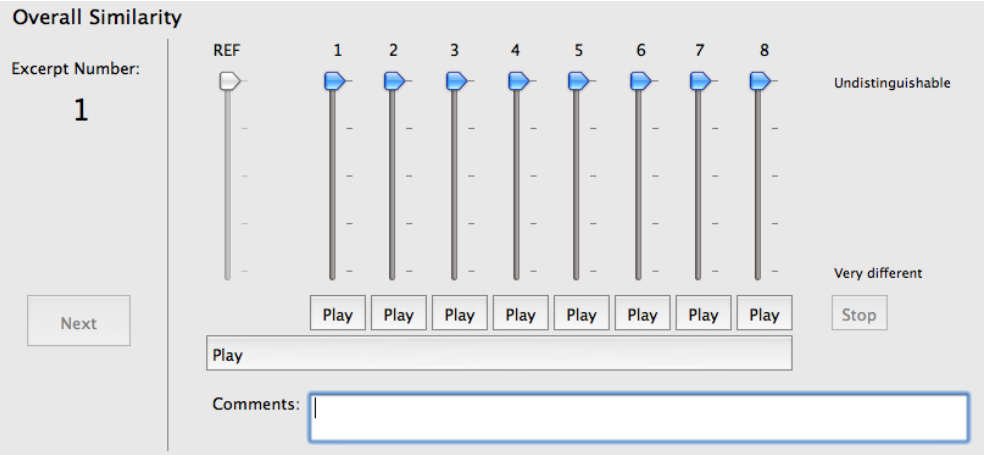

Figure 2.11: Graphical user interface of the subjective test software. The frozen slider to the left corresponds to the reference while the eight sliders to the right correspond to the other BRIRs (including the hidden reference).

2.5.1 Test setup and subjects

Both listening tests were carried out using an automated subjective test software. The audio signals presented to the subjects were D/A converted at a sampling rate of 44.1 kHz with a MOTU Traveler sound interface connected via firewire to the personal computer running the test software. High-quality headphones (Sennheiser HD 650) were used.

We asked 8 subjects to participate in the first listening test. All of them reported normal hearing and 6 out them participated also in the second listening test.

2.5.2 Test procedure

A MUSHRA [Rec. ITU-R BS.1116.1, 1997] type subjective test using an absolute grading scale was conducted. The subjects were asked to grade the similarity between a reference (the measured BRIR) and other BRIRs on a continuous scale from "undistinguishable" to "very different". A hidden reference was used to test the reliability of the subjects. In the first listening test also an "anchor" (IC1) was introduced in order to obtain results on a similar scale from all subjects. In informal listening tests the anchor was found to be perceptually so different from the reference that is was expected to obtain marks close to "very different". In the second test the anchor was not used but two cases from the first test (FD5 and FI80) were introduced again to be able to precisely compare these cases with the new cases tested in the listening test, which may have had a "normalizing" effect similar to using an anchor.

Figure 2.11 shows the graphical user interface of the subjective test software. The subjects were presented with 8 respectively 6 play buttons and sliders to judge the

stimuli. Furthermore there was a play button and a slider frozen at “undistinguishable" for the reference. The subjects could switch between the stimuli at any time while the sound was instantly crossfaded from one stimulus to the other. Informal listening indicated that such instant switching facilitates the discrimination of the BRIRs.

Written instructions were given to the subjects before the test started. The test contained 8 panels, each panel corresponding to one audio signal convolved with all the BRIRs, plus two additional panels which were taken from the 8 other panels and which were used as training items. The training items used a male speech sample and a drum sample. The order of the audio signals presented was randomized differently for each subject and the order of the BRIRs (i.e. the correspondence between sliders and BRIRs) was randomized for each panel.

The duration of the test session varied between the listeners due to the freedom to repeat the stimuli as often as they requested. Typically the time needed to complete one of the listening tests was between thirty minutes and one hour.

2.5.3 Stimuli

For the first listening test, synthetic BRIRs of the types FI5, FI80, FD5, FD20, FD80, FD200 and IC1 were generated based on BRIR 1 (described in Section 2.4.1). For each case, two different instances of Gaussian white noise were used to generate the synthetic BRIRs. In total, each synthetic BRIR type appeared 8 times in the listening test (two noise instances, each applied to four different audio signals). The goal of the first listening test was to compare frequency-dependent coherence matching to frequency-independent coherence matching, the influence of the noise instance used for the BRIR synthesis, and to determine if there is a threshold after which the diffuse reverb tail of a measured BRIR can be perceptually transparently replaced by filtered and coherence-matched noise.

For the second listening test, synthetic BRIRs of the types FI80, FD5, FD5/25, FD5/35 and TF5 were generated both based on BRIR 1 and BRIR 2. Only one of the two noise instances used in the first listening test was used because preliminary results of the first listening test indicated that the influence of the noise instance is negligible. The goal of the second listening test was to examine the improvement of the perceptual quality due to modeling not only the frequency dependence of the interaural coherence but also the time dependence.

2.6 Results

2.6.1 Listening test 1

The results of the first listening test for the drum, male speech, female speech, and clap samples are shown in Figures 2.12, 2.13, 2.14 and 2.15, respectively. The average results for all samples can be seen in Figure 2.16. Since the clap sample is a special and – compared to the other samples – less natural case, the average results only for the drum, male speech, and female speech samples were analyzed separately and are shown in Figure 2.17.

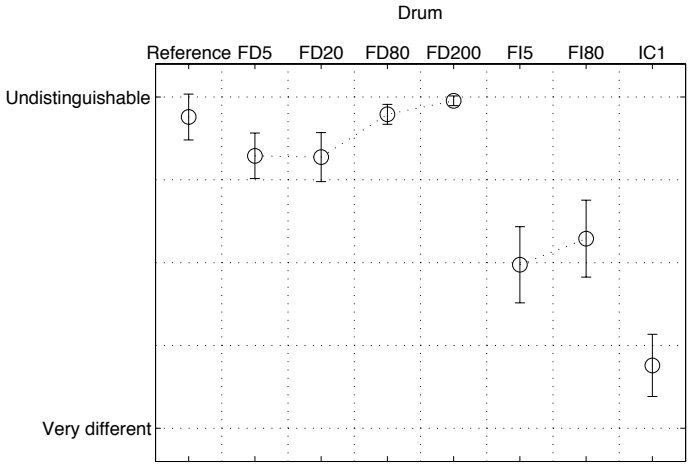

Figure 2.12: Results for the first listening test for the drum sample, average over all subjects, with 95 % confidence intervals.

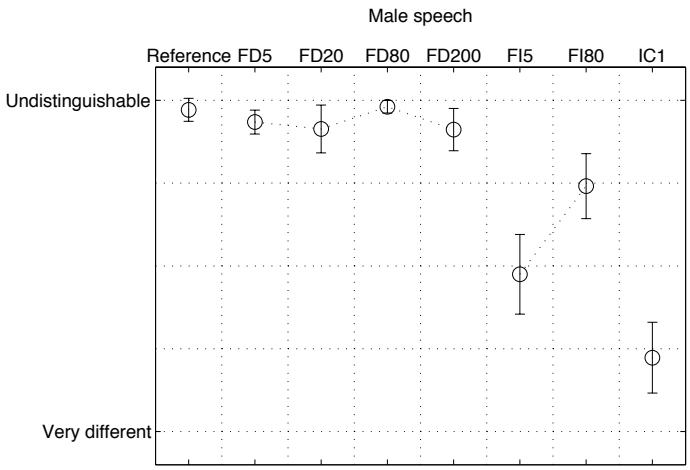

Figure 2.13: Results for the first listening test for the male speech sample, average over all subjects, with 95 % confidence intervals.

For all cases, the frequency dependent interaural coherence matching performs significantly better than the frequency independent interaural coherence matching. The BRIR with frequency-dependent coherence with the split point at 80 ms (FD80) is judged to be significantly better than its frequency-independent equivalent FI80 in all cases, as can be seen from the non-overlapping confidence intervals in Figures 2.12 to 2.15. Furthermore, a one-sided pairwise T-test shows significance on a 99 %

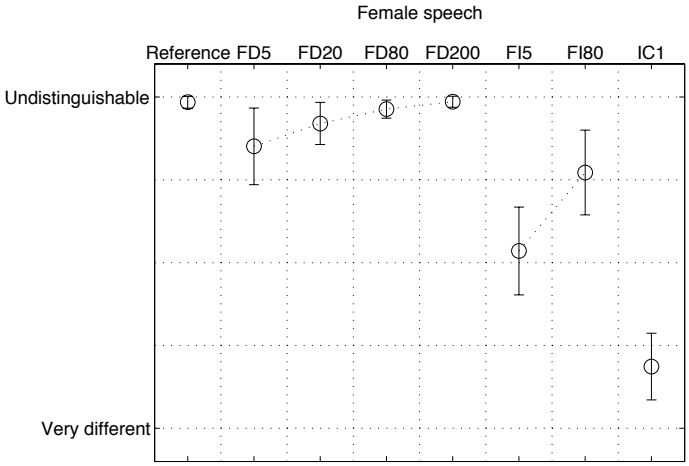

Figure 2.14: Results for the first listening test for the female speech sample, average over all subjects, with 95 % confidence intervals.

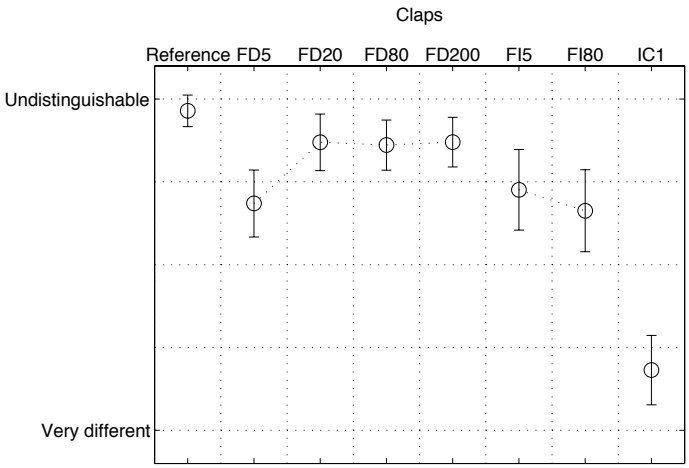

Figure 2.15: Results for the first listening test for the clap sample, average over all subjects, with 95 % confidence intervals.

confidence level for all test cases. The same result holds also for the split point at 5 ms, except for the clap sample.

For all samples except for the clap sample, the BRIRs FD80, FD200, and the reference are very close. For the average over all sounds except claps (Figure 2.17), even with a 90 % confidence level, there is no significant difference between the reference and FD80 or FD200. However, at lower confidence levels, the difference becomes

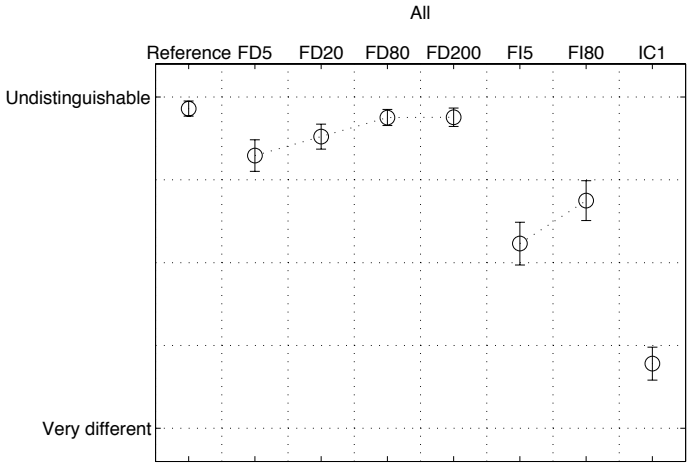

Figure 2.16: Results for the first listening test for all samples, average over all subjects, with 95 % confidence intervals.

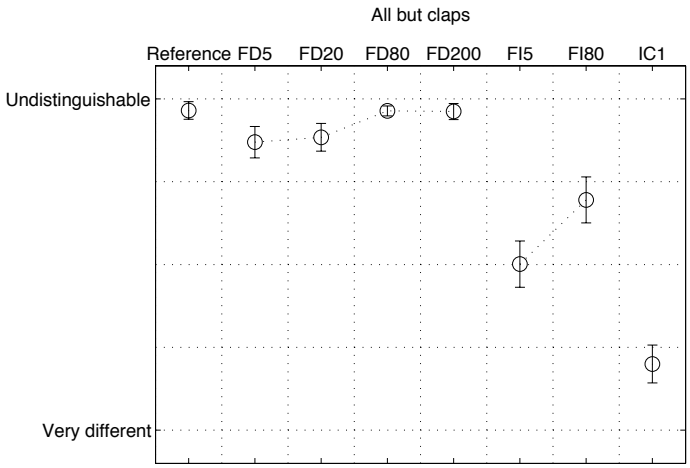

Figure 2.17: Results for the first listening test for the female speech, male speech, and drum samples, average over all subjects, with 95 % confidence intervals.

significant (with 53.3 % confidence we can say that FD80 is different). However, one has to keep in mind what these numbers mean: that with a probability of 46.7 %, the mean of eight listeners trying very hard to hear the most subtle differences does not give any useful information to distinguish FD200 from the reference. While this is by itself an interesting information, for practical purposes the perception of the

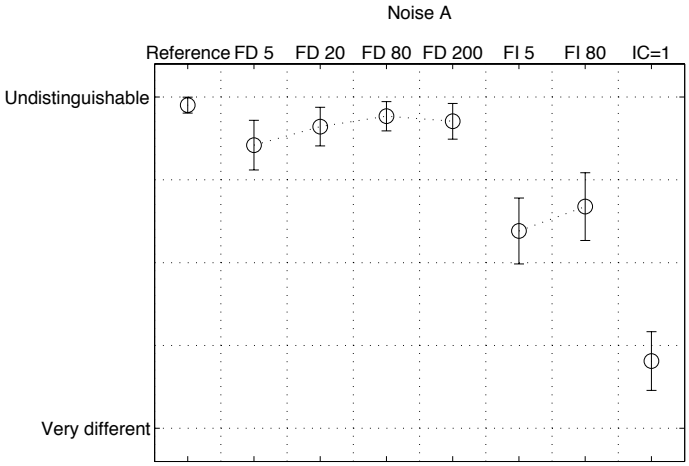

Figure 2.18: Results for the first listening test for all samples, average over all subjects, noise instance A, with 95 % confidence intervals.

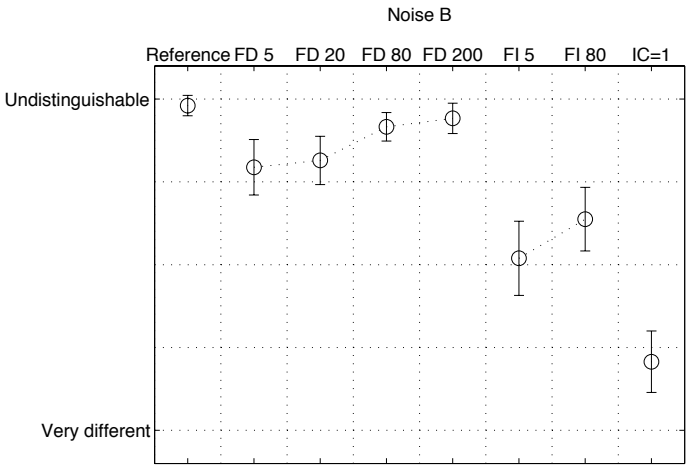

Figure 2.19: Results for the first listening test for all samples, average over all subjects, noise instance B, with 95 % confidence intervals.

BRIRs by individual listeners is more relevant. Therefore it is important to know which proportion of the listeners are capable of distinguishing the synthetic BRIRs from the reference, rather than considering the mean of all listeners.

With a confidence level of 95 %, considering all source files except for the claps, 2 out of 8 subjects were able to distinguish FD80 from the reference and none of the subjects was able to distinguish FD200 from the reference. Lowering the confidence

level to 75 %, 3 out of 8 subjects were able to distinguish FD80 from the reference and 2 out of 8 subjects were able to distinguish FD200 from the reference. These results show that the majority of listeners cannot distinguish FD80 and FD200 from the original BRIR.

Informally, all subjects reported that picking the reference from the presented audio signals was at the limit of their hearing capabilities and that they sometimes thought to hear differences which would vanish after more thorough listening.

The two noise instances produced similar results as can be seen by comparing Figures 2.19 and 2.18. Because noise instance A produced slightly better results in most cases, only this noise instance was used in the second listening test.

2.6.2 Listening test 2

The averaged results of the second listening test in Figure 2.20 show a slight but not significant trend towards better results with the time-frequency coherence matching (TF5) than with time-independent frequency dependent interaural coherence matching (FD5). Analyzing the results for the two BRIRs separately, it can be seen that for BRIR 1 the trend is stronger (see Figure 2.21) and for BRIR 2 no trend is present (see Figure 2.22). A one-sided pairwise T-test shows that for BRIR 1 TF5 performs significantly better than FD5 on a 95% confidence level.

Even though TF5 performs better than all of the other BRIR types in the task of modeling BRIR 1, the non-overlapping confidence intervals show that TF5 can be distinguished from the reference BRIR. Therefore it cannot be claimed that TF5 is a perceptually transparent model for the measured BRIR. The cases FD5/25 and FD5/35 (matching the frequency-dependent interaural coherence separately in two time intervals) show no significant gain compared to FD5.

Conclusively, the second listening test implies that there is only little gain when the interaural coherence is not only modeled to be frequency dependent, but also time dependent. Furthermore, the second listening test also confirmed the finding from the first listening test that frequency-dependent coherence matching performs significantly better than frequency-independent coherence matching: in all cases FD5 performs significantly better than FI80, even though this comparison is biased in favor of the frequency-independent coherence matching because FI80 contains a larger portion of the measured BRIR than FD5.

2.7 Discussion

The original aim of this study was to find a way of generating perceptually transparent synthetic BRIRs without modeling early reflections. The most advanced method to generate such BRIRs used in this study was to process two channels of white Gaussian noise to match the energy decay relief [Jot, 1992] and the time- and frequency-dependent interaural coherence of the measured BRIR. Even though the listening test showed that in most cases this method performed better than all the other methods and the absolute results showed that the method produces very good results, the goal of perceptual transparency was not reached. So it is interesting to discuss the reasons that may have led to the slight but audible differences between the measured

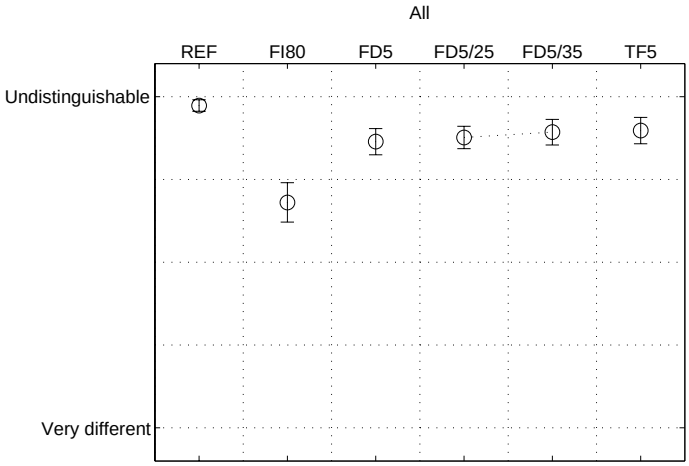

Figure 2.20: Results for the second listening test for all samples and both BRIRs, average for all subjects, with 95 % confidence intervals.

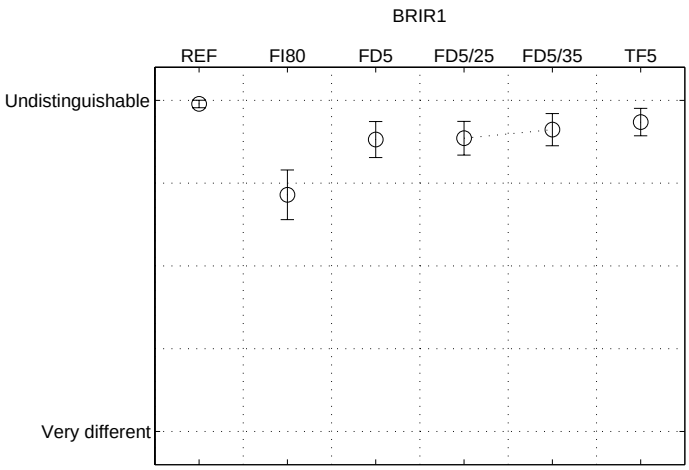

Figure 2.21: Results for the second listening test for BRIR 1 and all samples, average for all subjects, with 95 % confidence intervals.

BRIR and the best synthetic BRIR consisting of a measured direct sound part and a completely synthetic reverb tail.

From a purely perceptual point of view, the main difference that can be heard between the measured BRIR and the synthetic BRIR with time-frequency coherence matching is that there seems to be a low-frequency boost in the synthetic BRIR, even though the spectra of the reverb tails have been matched accurately in the frequency-

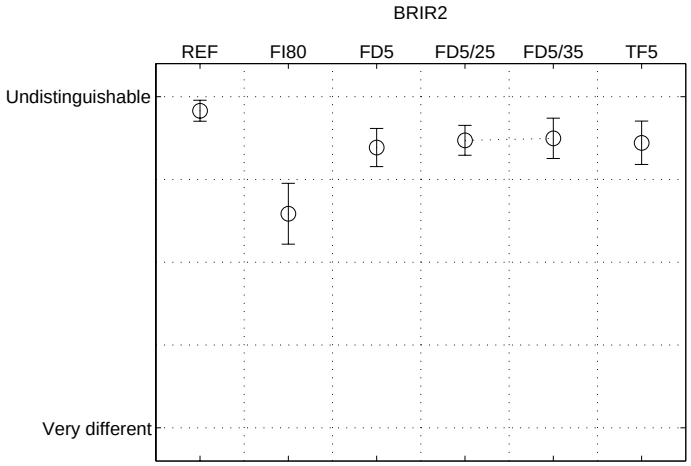

Figure 2.22: Results for the second listening test for BRIR 2 and all samples, average for all subjects, with 95 % confidence intervals.

domain.

One explanation may be that in the low frequency range the energy of the direct sound is not only contained in the first 5 ms that we considered as the direct sound part of the BRIR. This leads to a part of the direct sound being modeled by the noise tail, resulting in spreading in time of energy from the direct sound and a perceived low-frequency boost in the reverb tail. The rapid crossfade between the direct sound part and the reverb tail may also lead to a spreading of the energy between frequency bands, in particular at the low frequencies.

Furthermore, informal listening using BRIRs that were highpass-filtered at 100 Hz (removing the aforementioned low-frequency problems) showed that the synthetic BRIRs sound slightly more “cluttered" than the measured BRIRs. This may be due to not modeling the early reflections individually.

Early reflections have been shown to be important for the externalization of sound source [Begault et al., 2001] and in order to study the effect of early reflections in more detail, it is possible to consider not only the interaural coherence in the early part of the BRIR, but also the early lateral energy fraction [Barron and Marshall, 1981]. This measure is known to be related to the impression of spaciousness and often assumes values close to 0.2 in concert halls [Barron, 2000]. However, to measure it, two coincident microphones are needed, meaning that it cannot be calculated directly from a BRIR. Therefore, no “lateral energy fraction matching” could be introduced in this study.

Since a lot of research on binaural reverberation has focused on early reflections, it may be disputed if it is possible to generate a perceptually transparent BRIR without modeling individual early reflections. However, it is worth noting that without modeling any early reflections, a synthetic BRIR can be generated that is perceptually very similar to the measured BRIR.

It is therefore imaginable that for applications where a very accurate modeling of the acoustics is not necessary, but the impression of spaciousness needs to be created, a simple reverberator modeling the energy decay relief and the frequency- and time-dependent interaural coherence may be a good compromise between simplicity and accurate acoustical modeling. Such an approach was presented in [Menzer and Faller, 2009a], where a Jot reverberator was modified to model not only the energy decay relief but also the frequency-dependent interaural coherence.

2.8 Conclusions

Two listening tests were performed to examine to which extent BRIR tails can be replaced by filtered and coherence-matched Gaussian noise. It was shown that noise matching the interaural coherence of a measured BRIR individually in each frequency band performs significantly better than noise matching only the overall (wideband) interaural coherence of the measured BRIR.

It was also shown that for a specific measured BRIR the reverb tail 80 ms after the beginning of the BRIR can be replaced by filtered and coherence matched noise in a perceptually transparent way, implying that the frequency dependent interaural coherence is the only relevant binaural cue for diffuse sound.

Further improvements using time-dependent interaural coherence matching were studied and it was shown that for one out of the two BRIRs used in the test, time-dependent coherence matching leads to an improvement.

Furthermore, it was shown by analyzing a set of BRIRs with a 5° azimuth resolution that the interaural coherence changes due to changes in the orientation of the listener affect only the early reflections but not the diffuse reverberation.

Chapter 3

Obtaining BRIRs from B-Format Room Impulse Responses

3.1 Introduction

Binaural room impulse responses (BRIRs) are important tools for high-quality 3D audio rendering [Huopaniemi, 1999]. BRIRs take into account both the properties of the listener (or dummy head) as well as the properties of the room in which the BRIRs have been recorded and give the listener the impression of being in the room and hearing a sound source in the position where the sound source used for the BRIR recording was placed. Head-related transfer functions (HRTFs) on the other hand are recorded in an anechoic environment and can be used to simulate listening to a loudspeaker in an anechoic environment. HRTFs completely lack room-related properties.

In this chapter, a method is proposed that allows to compute BRIRs using room impulse responses measured with a B-format microphone (B-format RIRs) and HRTF sets. This means that recording the listener-specific properties (HRTFs) is independent from recording room-specific properties (B-format RIRs). In particular, this very much simplifies the task of providing individualized BRIRs for a large number of different acoustic environments for many different persons – something which is relevant for providing high quality 3D audio for a large user base.

Previously, [Merimaa and Pulkki, 2005; Pulkki and Merimaa, 2006] proposed a method which can generate RIRs for multi-channel loudspeaker setups with up to approximately 20 channels [Merimaa, 2006] from B-format RIRs. This method, called spatial impulse response rendering (SIRR), uses a decomposition into direct and diffuse parts. It distributes the direct part on the loudspeakers using vector base amplitude panning [Pulkki, 1997] and de-correlates the diffuse part to obtain several uncorrelated diffuse impulse responses.

The goal of the method proposed here is to generate BRIRs relative to any look direction of the head in a simple and robust way. Unlike SIRR, our technique cannot

produce impulse responses for multi-channel loudspeaker systems. By applying the correct HRTFs to the impulse responses generated by SIRR it is possible to simulate the target loudspeaker setup in anechoic conditions and therefore generate an approximation of a BRIR. Thus, SIRR can be used to perform the same task as the method proposed in this chapter. The proposed method is simpler than SIRR and eliminates the intermediate step of a multi-channel impulse response. This is very important because it also eliminates the need for a de-correlation method such as reverberators or phase randomization, which is necessary in a setup with more than 2 channels and which may introduce artifacts to the impulse response [Merimaa, 2006].

Given the B-format RIR of a specific room and a HRTF set, BRIRs individualized to the same listener as the HRTF set are generated as follows. The B-format RIR is separated into a direct sound part, and a reflections part, containing the early and late reflections of the RIR. The direct sound part of the BRIR is modeled by applying to the direct sound the HRTFs corresponding to the estimated direction of arrival. The reflections part of the BRIR is modeled as a linear combination of the late B-format signal channels such that the relevant spectral cues and perceptual spatial cues are the same as would be expected for a BRIR measured in the same room as the B-format RIR was measured. The considered spatial cues are the left and right power spectra and the interaural coherence (IC) [Blauert, 1997].

The chapter is organized as follows. Section 3.2 describes the proposed method to compute BRIRs in detail. In Section 3.3 the results produced by the proposed method are examined from a signal processing point of view, while a subjective test to evaluate the proposed method is described in Section 3.4. The conclusions are in Section 3.5. Furthermore, Appendix B describes the room impulse measurements performed for the evaluation of the proposed method.

3.2 Processing B-format RIRs

3.2.1 B-format room impulse responses

A B-format room impulse response (B-format RIR) is a room impulse response measured with a B-format microphone [Gerzon, 1973; Farrar, 1979a]. Ideally, it corresponds to a 4-channel room impulse response measured with four coincident microphones: one omnidirectional microphone ($w(n)$) and three dipole microphones ($x(n)$, $y(n)$, $z(n)$), pointing in the X, Y, and Z directions of a Cartesian coordinate system. An example of the directional responses in the horizontal plane is shown in Figure 3.1. Note that B-format is defined such that the dipoles have a gain which is $\sqrt{2}$ larger than the omnidirectional gain.

Inspired by current models of reverberation [Gardner, 1998], we consider room impulse responses to consist of a large peak corresponding to the direct sound as well as several delayed and filtered copies of this first peak, corresponding to the early reflections, and a diffuse reverberation tail, which may overlap with the early reflections.

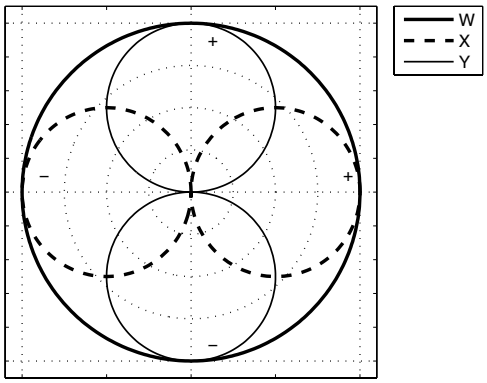

Figure 3.1: Directional responses of the W, X, and Y channels of B-format in the horizontal plane.

3.2.2 B-format RIR separation

Since the direct sound is processed in a different way than the reverberation, it is necessary to separate the B-format RIR into these two parts.

The split point between the direct sound and the late RIR is determined as the lowest local minimum of the energy envelope of $w(n)$ in the 10 ms after the absolute maximum of the energy envelope of $w(n)$. An example of such a separation can be seen in Figure 3.2. The 10 ms time interval after the direct sound was determined experimentally based on the RIRs at our disposal. For other rooms or other source and listener positions, there may be a need to slightly change the length of this interval in order to correctly separate the direct sound from the first reflection.

As opposed to an earlier implementation of the proposed method [Menzer and Faller, 2008] which, similar to SIRR, extracted the individual early reflections and convolved them with HRTFs corresponding to their directions of arrival, here both early and late reflections are processed by a single frequency dependent linear B-format decoding described in Section 3.2.4.

Two reasons led to this decision: When estimating the direction of arrival of the early reflections embedded in diffuse sound, errors are unavoidable. In practice, the linear decoding delivered better perceptual results than the directional modeling of the individual early reflections (i.e. the method presented here is both a simplification as well as an improvement compared to the method presented in [Menzer and Faller, 2008]). Furthermore, as will be shown in Section 3.3, the linear decoding method performs reasonably well even on the waveform level.

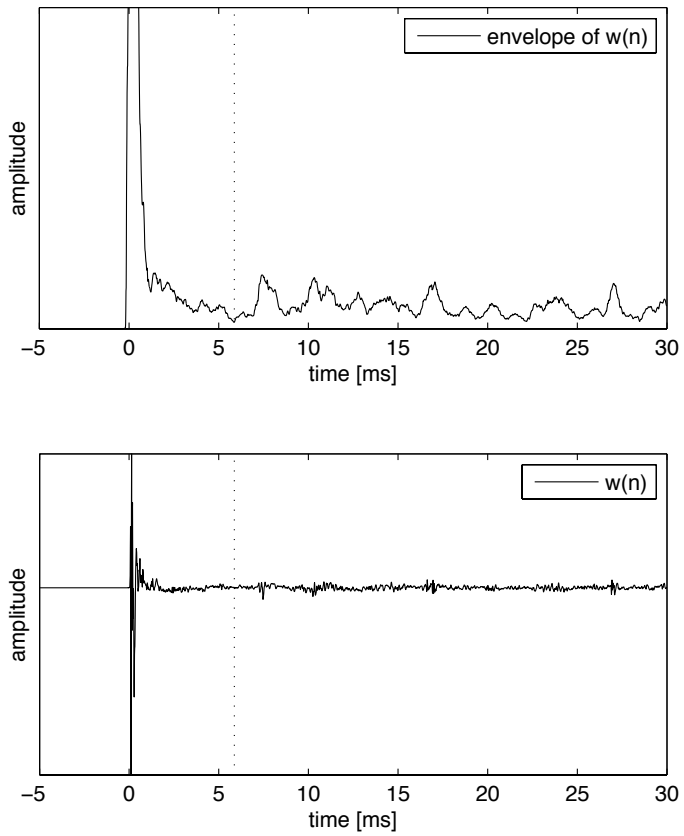

Figure 3.2: Separation of the B-format RIR in direct sound and reflections parts. **Top panel:** Envelope of $w(n)$ of B-format RIR. **Bottom panel:** $w(n)$ of B-format RIR. The separation is made at the lowest local minimum of the envelope of $w(n)$ in the first 10 ms after the direct sound.

3.2.3 Modeling the direct sound

The early BRIR corresponding to the direct sound is generated as follows. For the direct sound in the B-format RIR the direction of arrival is estimated by

$$\phi = \arg(I_x + iI_y) \tag{3.1}$$

$$\psi = \arg\left(\sqrt{I_x^2 + I_y^2} + iI_z\right), \tag{3.2}$$

where I_x, I_y and I_z are the components of the acoustic intensity vector $\vec{I}$ and are calculated as

$$I_x = \sum_{n \in D} x(n)w(n)$$

$$I_y = \sum_{n \in D} y(n)w(n)$$
$$I_z = \sum_{n \in D} z(n)w(n) \tag{3.3}$$

on the time interval D that corresponds to the direct sound.

Finally, the part of $w(n)$ corresponding to the direct sound is filtered with the HRTF closest to the estimated direction of arrival of the direct sound. Since the HRTF set used has resolution of 5° in the horizontal plane, for sources in the horizontal plane, the deviation from the estimated direction is 2.5° or less.

With respect to the direction of arrival estimate and the rendering of the direct sound, the presented method is very similar to [Merimaa and Pulkki, 2005].

3.2.4 Modeling the late BRIR

The late part of the BRIRs were obtained by linearly processing the late B-format RIR such that three conditions are fulfilled:

- The power spectra of the generated left and right late BRIR are the same as the power spectra of the true left and right BRIR.
- The coherence between the left and right generated late BRIRs is the same as the coherence between the true left and right late BRIRs at each frequency.
- At each frequency the temporal envelope of the generated late BRIR is the same as for the true late BRIR. In other words, the energy decay relief [Jot, 1992] is the same for generated late BRIR and the true late BRIR.

Note that method is designed such that the important perceptual spatial cues interaural level difference (ILD) and interaural coherence (IC) [Blauert, 1997] will be the same for the synthesized and the true late BRIRs at each frequency. The matching of the ILD follows from the matching of the power spectra. For a symmetric HRTF set (i.e. a HRTF set where one can obtain from the HRTF for azimuth $X°$ the HRTF for azimuth $-X°$ by swapping the channels) the ILD for perfectly diffuse sound is 0 dB. If the HRTF set was not symmetric, the resulting ILD for the diffuse sound would be correctly reproduced by the method proposed in this chapter. This does not hold for non-diffuse sound. However, we will show empirically in Section 3.3 that for non-diffuse sound the ILDs produced by the proposed method are in many cases approximately correct.

Note also that the third condition not necessarily implies the first. The temporal envelope measured in each frequency bin in a time-frequency representation has only a coarse frequency resolution, meaning that the power spectra calculated over the entire impulse response give additional information.

In the following we are computing the left and right true late BRIR power spectra and coherence as a function of frequency between the left and right late BRIR. Then, it is shown how to compute late BRIRs by linear B-format decoding from the B-format room impulse responses such that the power spectra and coherence are the same as in the true late BRIRs. The decay of the late BRIR is the same as the decay of the B-format RIR for each frequency. The linear B-format decoding is time-independent

and therefore has no impact on the decay which thus will be automatically correct, implying that also the frequency dependent reverberation time of the generated BRIR will be correct.

All of the linear B-format decoding described hereafter was implemented using an FFT, which is the natural choice since the B-format decoding is time-independent and frequency-dependent. However, alternative implementations, e.g. in STFT domain, are possible.

The proposed method for modeling the late BRIR is different from the diffuse sound rendering of SIRR because the late BRIR is calculated only by a linear decoding of the B-format RIR, with the aim of obtaining a BRIR with the correct interaural coherence directly, without using reverberators or other de-correlation techniques, which would be a possible source of artifacts.

Computation of the true BRIR parameters

In the following it is assumed that the late BRIR is ideally diffuse, i.e. sound arrives from all directions with the same power and sound arriving from each direction is independent of the sound arriving from all other directions. Further, diffuse sound is approximated by only considering directions for which HRTFs are available. The left and right HRTFs are denoted $L_i(\omega)$ and $R_i(\omega)$, where $i \in \{1, 2, \cdots, I\}$ is the direction index and I is the number of HRTFs in the set.

In the tests performed for this study, an HRTF set with an angular resolution of 5° in the horizontal plane was used. In previous tests, the proposed method was applied using the CIPIC HRTF set [Algazi et al., 2001] whose angular in the horizontal plane varies between 5° and 20°.

Given these assumptions, the late omnidirectional transfer function verifies:

$$W_{\text{late}}(\omega) = \sum_{i=1}^{I} D_i(\omega)\,, \tag{3.4}$$

where $D_i(\omega)$ is the diffuse sound arriving from the direction corresponding to index i. Note that the assumption about diffuse sound implies that

$$\mathrm{E}\{|D_i(\omega)|^2\} = \mathrm{E}\{|D_k(\omega)|^2\}$$

for all index pairs i and k, where $\mathrm{E}\{.\}$ is expectation and $|.|$ is the magnitude of a complex number. Also, the diffuse sound assumption implies that $\mathrm{E}\{D_i(\omega)D_k(\omega)\} = 0$ for $i \neq k$. Then with (3.4) it follows that the power spectrum of $D_i(\omega)$ is

$$\mathrm{E}\{|D_i(\omega)|^2\} = \frac{|W_{\text{late}}(\omega)|^2}{I}\,, \tag{3.5}$$

where $W_{\text{late}}(\omega)$ is the spectrum of $w_{\text{late}}(n)$.

The late left and right BRIRs are

$$\begin{aligned} B_{\text{L,late}}(\omega) &= \sum_{i=1}^{I} L_i(\omega)D_i(\omega) \\ B_{\text{R,late}}(\omega) &= \sum_{i=1}^{I} R_i(\omega)D_i(\omega)\,. \end{aligned} \tag{3.6}$$

From (3.5) and (3.6) it follows that the BRIR power spectrum is:

$$\begin{aligned} P_L(\omega) &= \frac{|W_{\text{late}}(\omega)|^2}{I} \sum_{i=1}^{I} |L_i(\omega)|^2 \\ P_R(\omega) &= \frac{|W_{\text{late}}(\omega)|^2}{I} \sum_{i=1}^{I} |R_i(\omega)|^2 \ . \end{aligned} \tag{3.7}$$

The magnitude of the coherence between the left and right BRIRs is

$$\Phi(\omega) = \frac{\left| \left\langle B_{\text{L,late}}(\omega) B^{\star}_{\text{R,late}}(\omega) \right\rangle \right|}{\sqrt{\left\langle |B_{\text{L,late}}(\omega)|^2 \right\rangle \left\langle |B_{\text{R,late}}(\omega)|^2 \right\rangle}} , \tag{3.8}$$

where * denotes the complex conjugate of a complex number and $\langle x \rangle$ denotes the expected value of x. This is equivalent to

$$\Phi(\omega) = \frac{|\sum_{i=1}^{I} L_i(\omega) R_i^{\star}(\omega)|}{\sqrt{\sum_{i=1}^{I} |L_i(\omega)|^2 \sum_{i=1}^{I} |R_i(\omega)|^2}} . \tag{3.9}$$

In the following, late BRIRs are generated in a way that their left and right power spectrum is equal to (3.7) and their coherence is equal to (3.9).

Note that equations (3.7) and (3.9) imply a set of HRTFs for directions evenly spaced on a sphere around the head of the listener. If such a set is not available, it is necessary to weight each HRTF by the area on a unit sphere that represents all directions which would be quantized to the HRTF in question [Jot et al., 1995; Larcher et al., 1998]. The HRTF set used for this study has HRTFs for equally spaced azimuth angles and 7 different elevation angles (see Appendix B) and before the coherence calculation the HRTFs for the different elevation angles were weighted proportionally to the surface on the unit sphere they represent.

Computation of the modeled BRIR

From the B-format late room impulse response signals, denoted $W_{\text{late}}(\omega)$, $X_{\text{late}}(\omega)$, $Y_{\text{late}}(\omega)$, and $Z_{\text{late}}(\omega)$, the left and right channels of the late BRIR, $B_{\text{L,late}}$ and $B_{\text{R,late}}$ are generated (assuming that the look direction of the head is along the X axis – other directions can be obtained by rotating the B-Format signal before the processing [Farrar, 1979a,b]):

$$\begin{aligned} \hat{B}_{\text{L,late}}(\omega) &= H_L(\omega) \left(v(\omega) W_{\text{late}}(\omega) + \frac{1 - v(\omega)}{\sqrt{2}} Y_{\text{late}}(\omega) \right) \\ \hat{B}_{\text{R,late}}(\omega) &= H_R(\omega) \left(v(\omega) W_{\text{late}}(\omega) - \frac{1 - v(\omega)}{\sqrt{2}} Y_{\text{late}}(\omega) \right) , \end{aligned} \tag{3.10}$$

where $v(\omega)$ is a frequency dependent constant and $H_L(\omega)$ and $H_R(\omega)$ are real-valued filters that model the modification of the power spectrum imposed by the HRTF

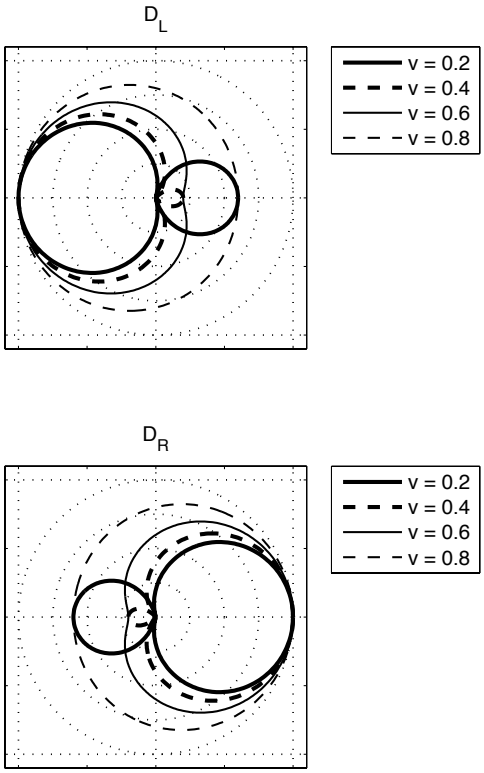

Figure 3.3: Directional responses D_L and D_R for various B-format decoding constants v (normalized, on a linear scale).

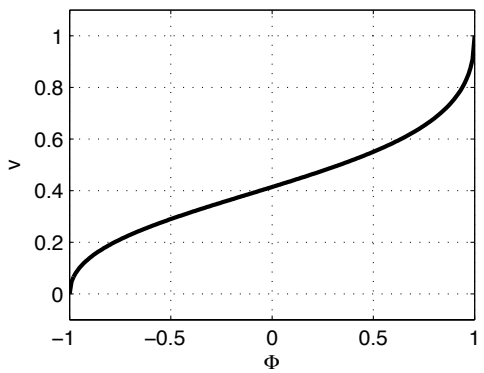

Figure 3.4: B-format decoding constant v as a function of the coherence Φ.

set. Note that the factor $1/\sqrt{2}$ is there to compensate the additional $\sqrt{2}$ gain in the B-format dipole gains.

First the constant $v(\omega)$ is determined. The directional responses of the two signals (3.10) are

$$\begin{aligned} D_L(\omega,\phi) &= H_L(\omega)\left(v(\omega)+(1-v(\omega))\cos\phi\right) \\ D_R(\omega,\phi) &= H_R(\omega)\left(v(\omega)-(1-v(\omega))\cos\phi\right)\,. \end{aligned} \tag{3.11}$$

Figure 3.3 shows a few example normalized directional responses for different B-format decoding constants v. As can be seen from Figure 3.3, the directional response of the linear B-format decoding has its global maximum on the left side, i.e. corresponds to a microphone pointing to the left. The decoding for the right channel is the same as for the left channel, but mirrored with respect to the median plane.

From these directional responses the magnitude of the coherence for the generated BRIRs (3.10) can be determined, assuming diffuse sound[1]:

$$\Phi(\omega) = \frac{|\int_{-\pi}^{\pi} D_L(\omega,\phi) D_R^{\star}(\omega,\phi)\mathrm{d}\phi|}{\sqrt{\int_{-\pi}^{\pi}|D_L(\omega,\phi)|^2\mathrm{d}\phi\int_{-\pi}^{\pi}|D_R(\omega,\phi)|^2\mathrm{d}\phi}}\,. \tag{3.12}$$

By substituting (3.11) into (3.12) it can be shown that

$$\Phi(\omega) = \frac{v^2(\omega)+2v(\omega)-1}{3v^2(\omega)-2v(\omega)+1}\,. \tag{3.13}$$

Equation (3.13) is equivalent to the quadratic equation

$$(3\Phi(\omega)-1)v^2(\omega) - 2(\Phi(\omega)+1)v(\omega) + \Phi(\omega) + 1 = 0\,. \tag{3.14}$$

The solution of (3.14) which fulfills $v(\omega)\in[0,\ 1]$ is

$$v(\omega) = \frac{\Phi(\omega)+1}{3\Phi(\omega)-1} - \frac{\sqrt{4(\Phi(\omega)+1)^2 - 4(3\Phi(\omega)-1)(\Phi(\omega)+1)}}{6\Phi(\omega)-2}\,.$$

Figure 3.4 shows the B-format decoding constant $v(\omega)$ as a function of the coherence $\Phi(\omega)$.

In addition to determining $v(\omega)$ in (3.10), the filters $H_L(\omega)$ and $H_R(\omega)$ need to be determined. From the condition that the power spectra of (3.10) need to be equal to the desired power spectra (3.7), it follows that

$$\begin{aligned} H_L(\omega) &= \frac{\sqrt{P_L(\omega)}}{\left|v(\omega)W_{\mathrm{late}}(\omega) + \frac{1}{\sqrt{2}}(1-v(\omega))Y_{\mathrm{late}}(\omega)\right|} \\ H_R(\omega) &= \frac{\sqrt{P_R(\omega)}}{\left|v(\omega)W_{\mathrm{late}}(\omega) - \frac{1}{\sqrt{2}}(1-v(\omega))Y_{\mathrm{late}}(\omega)\right|}\,. \end{aligned}$$

[1]For simplicity a horizontal diffuse sound model is considered here. A three dimensional diffuse sound model can be considered by integrating three dimensional directional responses.

3.3 Signal-level evaluation

The proposed method was implemented in Matlab and was applied to a B-format RIR measured in a lecture hall at our university. We also measured in the same room and with the same loudspeaker setup a set of BRIRs (see Appendix B), from which we could also obtain a set of HRTFs for the same source directions by isolating the direct sound. Therefore we could compare a measured BRIR with a BRIR generated from a B-format RIR and an HRTF set measured in the same room with the same loudspeaker setup and with the same microphone position. In the following, all data shown is for BRIRs with azimuth 0° and elevation 0° (i.e. the sound source is directly in front of the listener).

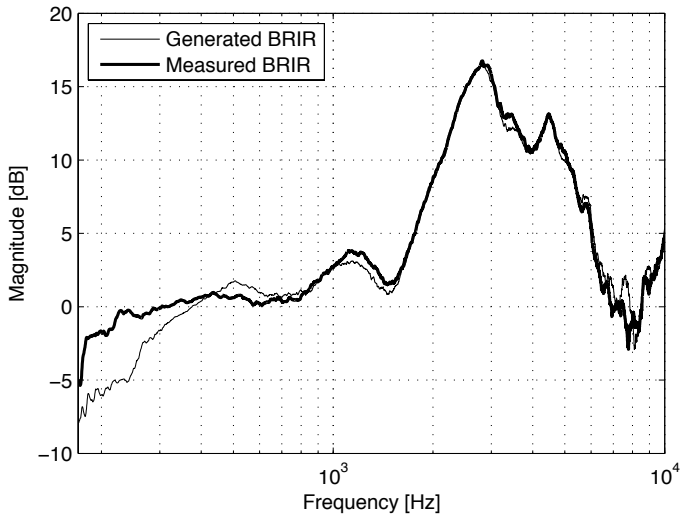

Figure 3.5: Spectra of a measured and a generated left BRIR.

The power spectra and coherence of the measured BRIR and the generated BRIR are shown in Figures 3.5, 3.6, and 3.7 respectively. Figure 3.5 compares the spectra of the entire BRIRs. The good match between the two BRIRs above 300 Hz is due to the fact that the direct sound, which contains most of the energy, is similar for the measured and for the generated BRIR. For simplicity, only the left channel is shown. However, one can observe a deviation of about 5 dB around 200 Hz. It may be that the separation of the direct sound from the rest of the BRIR is not well adapted to low frequencies, where artifacts may occur because of the abrupt transition from the HRTF-based direct sound processing to linear decoding of the late tail.

However, to evaluate the performance of the linear decoding of the reflections part of the B-format RIR, the spectra of the reflections parts of the BRIRs must be compared, as in Figure 3.6. The spectrum of the reflections part generated with the linear B-format decoding matches the measured BRIR up to 3 kHz, but above this frequency deviations of 5 dB and more occur. One possible source of these errors is that at high frequencies the directional responses of the Soundfield microphone used

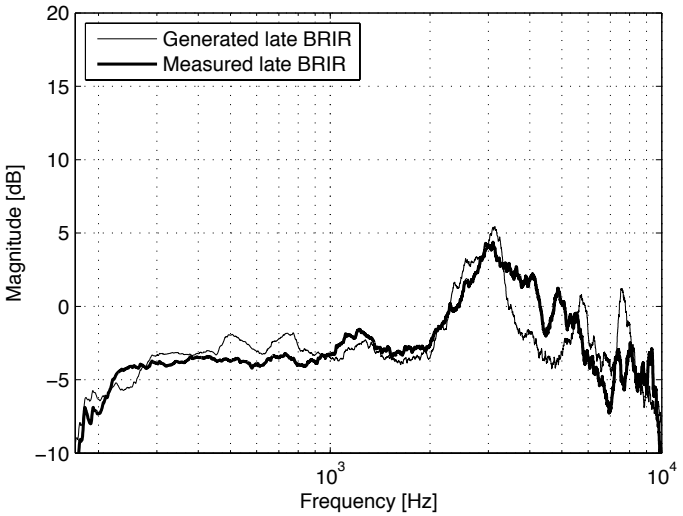

Figure 3.6: Spectra of the reflections part of the same measured and generated left BRIR as in Figure 3.5.

for the B-format RIR measurements start to deviate from the ideal responses [Faller and Kolundzija, 2009].

The coherence of the measured and the generated BRIR are shown in Figure 3.7. The top panel shows the interaural coherence for the late reverb tail, from 150 ms after the direct sound, as well as the HRTF-based prediction of the interaural coherence for diffuse sound. In this case the assumption of a perfectly diffuse sound in the late BRIR is approximately verified and all three curves match well up to 4 kHz, giving evidence that the proposed method for interaural coherence matching works as intended.

The bottom panel of Figure 3.7 shows the interaural coherence for the entire reflections part of the measured BRIR and the generated BRIR. Even though the assumption of a perfectly diffuse sound is not verified for single early reflections, the linear decoding technique based on this assumption produces a reverberation with a qualitatively similar interaural coherence.

It can be noticed that above 4 kHz and especially above 6 kHz, the coherence of the generated BRIR is generally too high. Again, imperfections of the Soundfield microphone may be the source of these errors.

In order to compare the proposed method with a more conventional way of generating BRIRs from a B-format RIR, a simple B-format RIR decoding with multiple directional RIRs obtained by simulating cardioid directional microphones was implemented. The directional RIRs were convolved with HRTFs for the corresponding directions in order to obtain a simulated BRIR. In particular, simulated BRIRs with three and four cardioid responses with elevation 0° and azimuths 0°, 120°, and 240° and 0°, 90°, 180°, and 270°, respectively, were calculated, where 0° corresponds to the azimuth direction of the direct sound. Informal listening showed that higher numbers of cardioids lead to less natural sounding BRIRs, therefore only the aforementioned

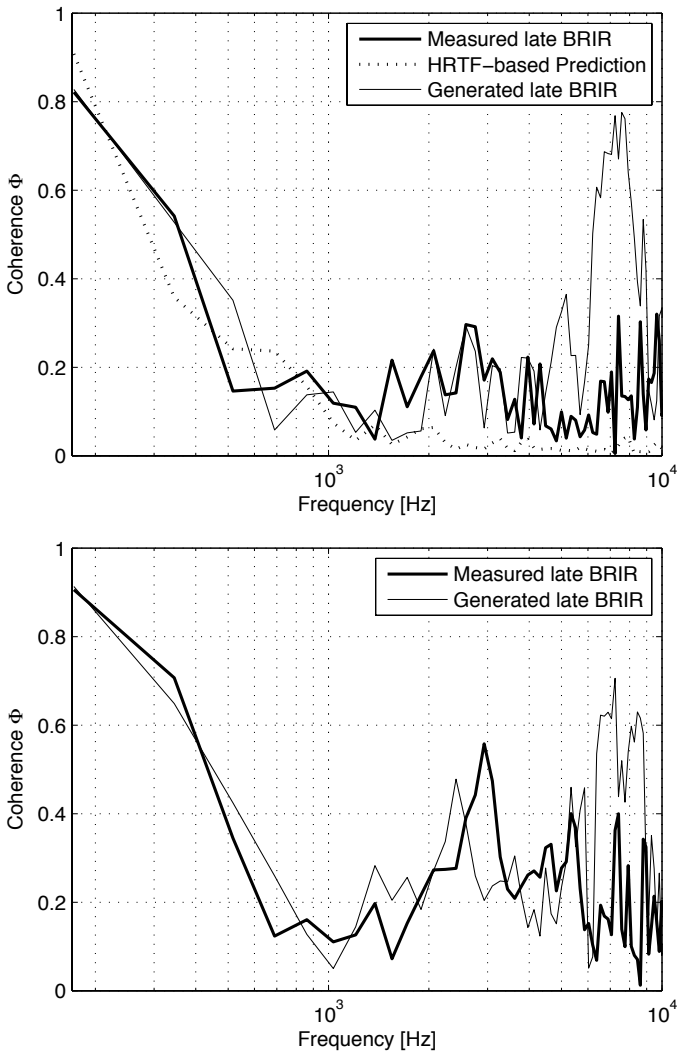

Figure 3.7: Interaural coherence of the reflections part of the same measured and generated BRIRs as in Figure 3.5. **Top panel:** interaural coherence for the diffuse reflections part only, not taking into account the first 150 ms of the BRIR. The dotted line shows the HRTF-based prediction of the coherence of diffuse sound recorded with the artificial head. **Bottom panel:** interaural coherence for the entire reflections part, starting 6 ms after the direct sound. Note that due to the short time Fourier transform approach of the coherence calculation, the frequency resolution of the coherence is smaller than the frequency resolution of the power spectra shown in Figure 3.5.

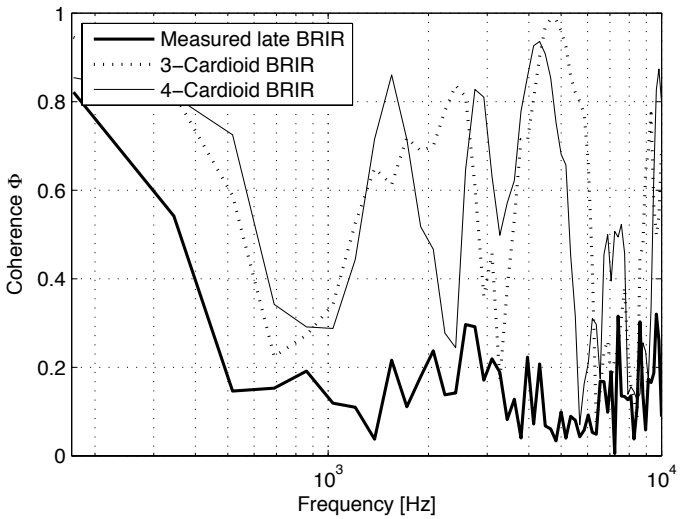

Figure 3.8: Coherence of the measured BRIR and two different BRIRs based on cardioid response decodings of the B-format BRIR. In order to be able to assume a diffuse sound, the first 150 ms of the impulse response are not taken into account. The coherence of the cardioid BRIRs is generally too high and does not follow well the coherence of the measured BRIR.

3- and 4-cardioid BRIRs were used for further investigations.

Figure 3.8 shows the coherence for the late tail of the 3- and 4-cardioid BRIRs and for the reference BRIR (all starting from 150 ms after the direct sound, for fair comparison with the top panel in Figure 3.7). The coherence of the cardioid BRIRs is generally too high, and doesn't follow the curve of the coherence of the measured BRIR above 1 kHz.

Figure 3.9 shows the directional responses as described in (3.11) used for the B-format decoding generating the late BRIRs. For simplicity only the responses for the left channel are shown.

The measured and modeled BRIRs are shown in Figure 3.10. As can be seen in the zoomed portion of the waveform, the early reflections are reproduced well, despite the fact that only the linear B-format decoding was applied and no HRTF for the specific direction of the early reflection was used. The good result can be explained because the linear decoding uses directional responses with maxima to the left for the left channel and to the right for the right channel, as can be seen in Figure 3.3. This is similar to the directional responses of the ears, which have their maxima between 60° and 90° to the left and to the right of the median plane [Shaw, 1974; Middlebrooks et al., 1989]. Therefore it can be expected that the ILDs for early reflections will be reproduced to some extent.

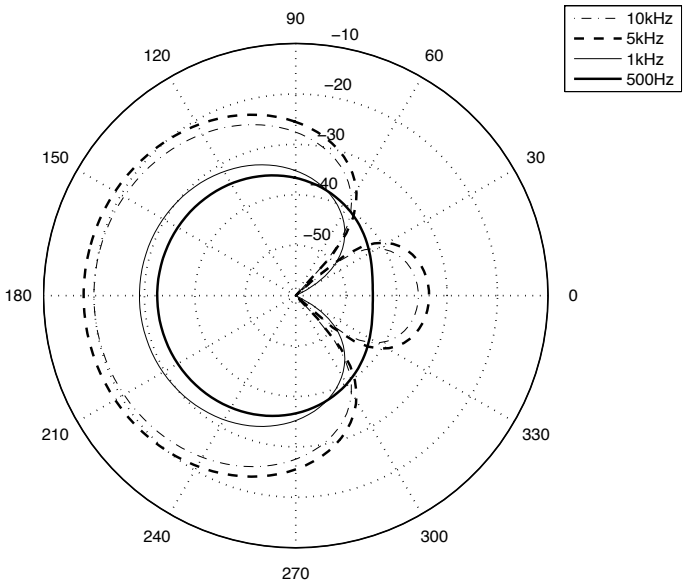

Figure 3.9: Directional responses of the linear decoding of the late B-format RIR for the left channel, for different frequencies, in decibels.

3.4 Subjective evaluation

A subjective test was conducted to show that the proposed method produces high-quality BRIRs comparable to recorded BRIRs and that the proposed method performs better than a conventional method to obtain BRIRs from B-format RIRs (linear cardioid decoding of the B-format and convolution with HRTFs applied). Informal listening showed that the decoding with 3 cardioids performed better than the decoding with 4 cardioids. In order to reduce the number of stimuli, only the decoding with 3 cardioids was used in the subjective test. We have asked both experienced listeners and naive listeners to take part in our subjective test.

3.4.1 Stimuli

In order to test the different BRIRs in different conditions, we applied the BRIRs to 6 different speech excerpts and 6 different dry recordings of musical instruments. The length of the speech excerpts was between 4 and 7 seconds and the length of the music recordings was between 3 and 4 seconds. BRIRs for the azimuth angles $-30°$, $0°$, and $30°$ and elevation $0°$ were used. Furthermore, two excerpts of stereo music recordings were presented using the $-30°$ and $30°$ BRIRs simultaneously. A list of all excerpts is given in Table 3.1.

We chose the sounds with the aim of using natural sounds similar to those that may be used in potential 3D audio applications. Speech and music seemed reasonable

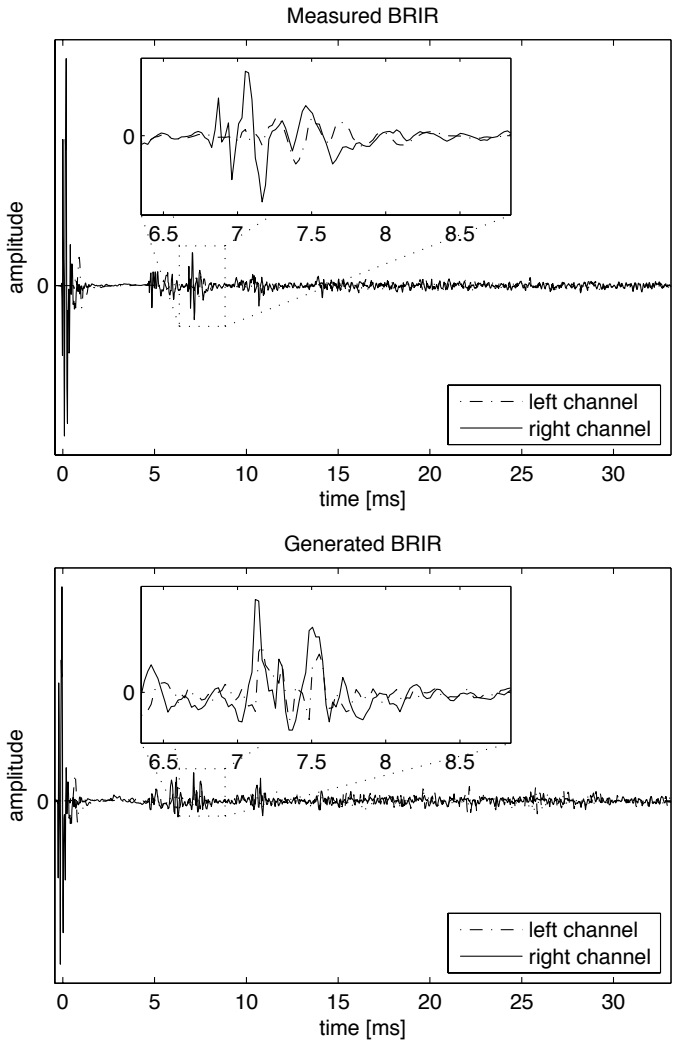

Figure 3.10: Waveforms of measured and generated BRIRs with zoom on an early reflection from the left. **Top panel:** measured BRIR. **Bottom panel:** generated BRIR. As can be seen from the zoomed early reflection, the linear B-format decoding produces approximately correct level differences.

choices in this context.

Each excerpt was convolved with four different “BRIRs" for the assigned direction: the measured BRIR, the generated BRIR, the 3-cardioid BRIR, and a colored (lowpass-filtered) HRTF.

Table 3.1: List of audio excerpts for subjective test. Bold face font indicates that an item was used as a training item.

Excerpt	BRIR angle(s)
English speech, male	0°
English speech, female	30°
French speech, male	30°
French speech, female	−30°
German speech, male	−30°
German speech, female	0°
Electric guitar	30°
Pop Drum	−30°
Oud	0°
Synthesizer	−30°
Shaker	0°
Electric bass	30°
Irish folk (instrumental)	−30°, 30°
Choir	−30°, 30°

3.4.2 Subjects and test setup

We asked nine persons to participate in the test. Five of the subjects were experienced listeners and four of them were naive listeners. They carried out the test with an automated subjective test software. The subjects used high-quality headphones (Sennheiser HD 600 and Sennheiser HD 25). The listeners were instructed to set the volume level to their preferred level.

3.4.3 Test method

A MUSHRA [Rec. ITU-R BS.1116.1, 1997] type subjective test using a relative grading scale was conducted. The subjects were asked to grade the similarity between the reference (the recorded BRIR) and the other BRIRs relative to three difference aspects: spatial aspects, coloration, overall similarity. A hidden reference was used to test the reliability of the subjects, as well as an "anchor" which consisted of the HRTF, and was expected to obtain marks close to "very different".

Figure 3.11 shows the graphical user interface of the subjective test software. The subjects were presented with four play buttons and four sliders to judge the stimuli. Furthermore there was a play button and a frozen slider for the reference. The subjects could switch between the stimuli at any time while the sound instantly faded from one BRIR to the other.

The test software showed written instructions on the computer screen before the test started. The test contained the 14 excerpts listed in Table 3.1, three of which were used as training items (one speech excerpt, one instrument excerpt, and one

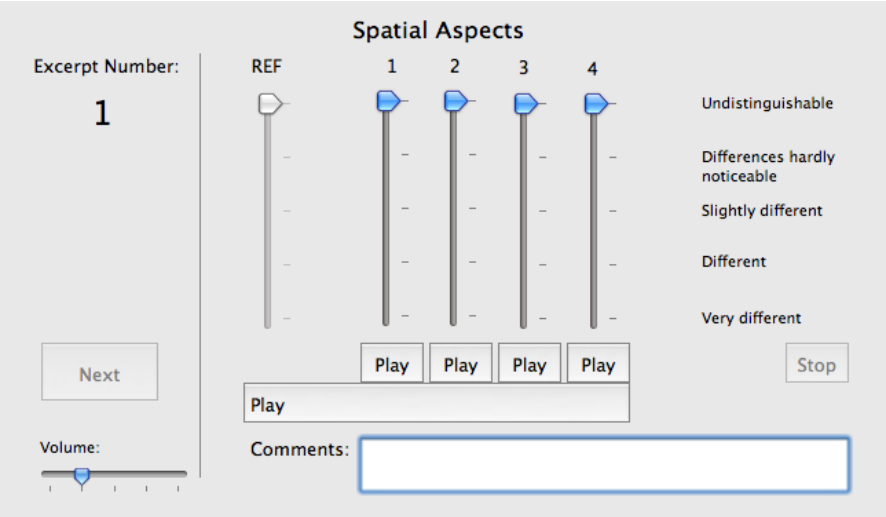

Figure 3.11: Graphical user interface of the subjective test software. The frozen slider to the right corresponds to the reference while the four sliders to the left correspond to the other methods (including the hidden reference).

stereo music excerpt). The excerpt and method order were randomized differently for each subject.

The duration of the test session varied between the listeners due to the freedom to repeat the stimuli as often as requested. Typically the test duration was between 30 min and 1 h.

3.4.4 Results

The results averaged over all subjects and 95% confidence intervals are shown in Figures 3.12 (single instrument music), 3.13 (speech), and 3.14 (stereo music). As can be seen from these graphs, the proposed method produces BRIRs that are significantly more similar to the reference BRIR than the cardioid based BRIRs in all cases.

The average rating for the overall similarity of the proposed method with the reference was in all of the cases between "indistinguishable" and "differences hardly noticeable". We conclude that for the average listener our method produces BRIRs that are hard to distinguish from measured BRIRs.

The samples that were used for the listening test not always covered the whole frequency range. In particular the speech samples had most of their energy below 2 kHz. Two of the musical instrument samples had spectra extending to 10 kHz (and above): the pop drum sample and the shaker sample. Because there was a strong deviation in the coherence above 4 kHz (see Figure 3.7), special attention was paid to these two samples. The averaged results for these two samples only are shown in Figure 3.15. For the proposed method, the results did not deviate significantly from the averaged results for all the musical instrument samples shown in Figure 3.12. It

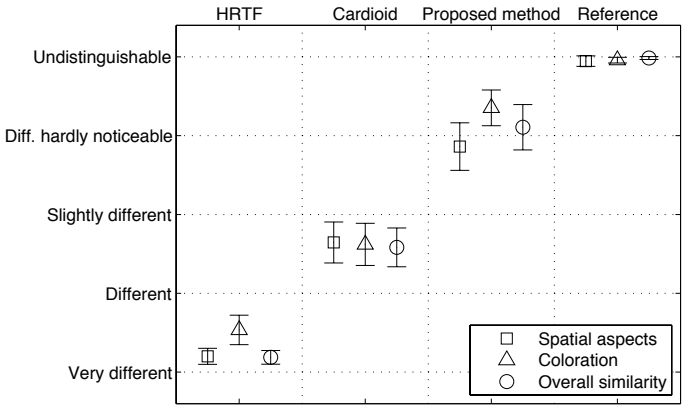

Figure 3.12: Average results for all subjects for all single instrument music stimuli, showing 95% confidence intervals.

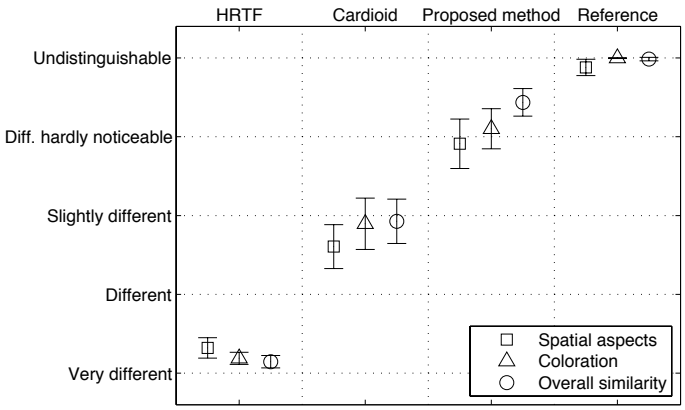

Figure 3.13: Average results for all subjects for the speech stimuli, showing 95% confidence intervals.

may be concluded that the observed deviation of the coherence above 4 kHz does not significantly influence the perception of the wide-band sounds convolved with BRIRs generated with the proposed method.

When comparing the results of the individual listeners, we observed that the main difference between the different listeners was in their overall sensitivity to deviations from the reference BRIR. Some listeners judged the proposed method as almost undistinguishable and the cardioid based method slightly different, while others found the proposed method to be slightly different and the cardioid based method different to

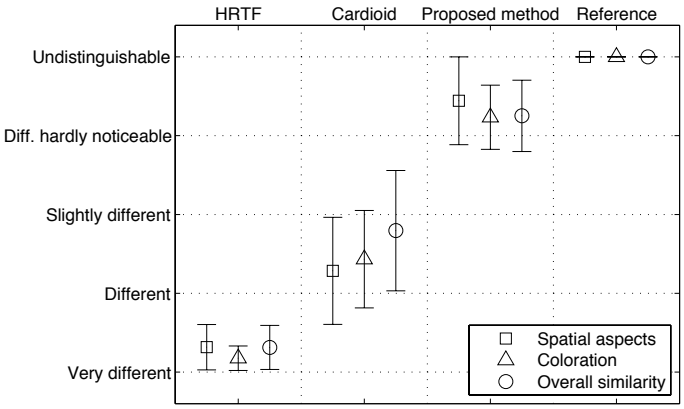

Figure 3.14: Average results for all subjects for the stereo music stimuli, showing 95% confidence intervals.

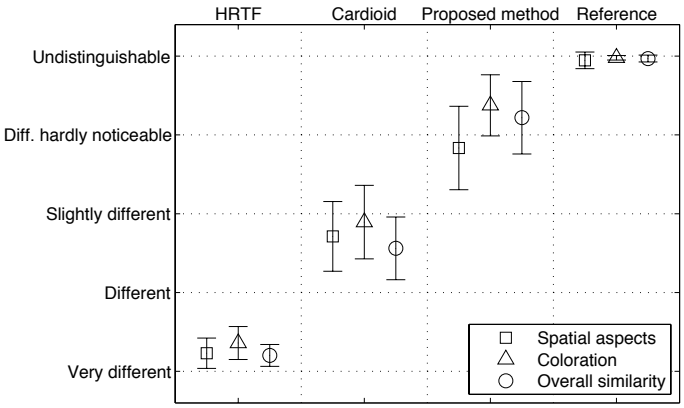

Figure 3.15: Average results for all subjects for the pop drum sample and the shaker sample, showing 95% confidence intervals.

very different. The main difference between the experienced listeners and the naive listeners was that the naive listeners' results tended to have larger confidence intervals. The means were similar for both groups of listeners. The individual results for the overall similarity aspect of the single instrument samples are shown in Figure 3.16 (for the experienced listeners) and in Figure 3.17 (for the naive listeners).

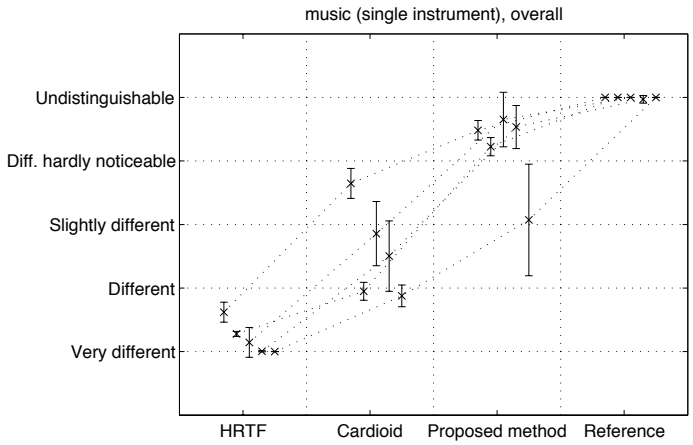

Figure 3.16: Individual results for the experienced listeners for the overall similarity aspect of the single instrument samples, showing 95% confidence intervals.

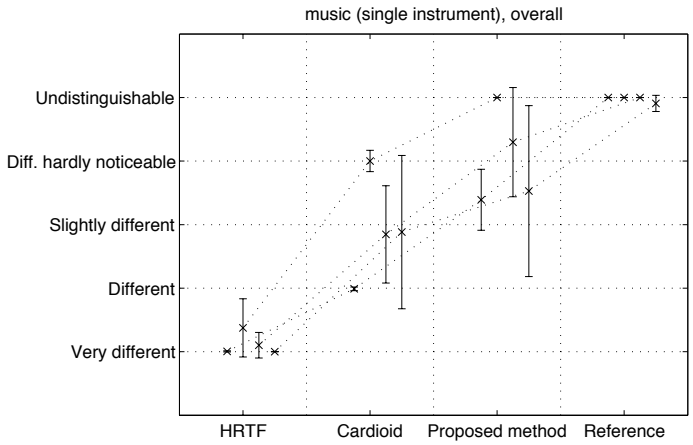

Figure 3.17: Individual results for the naive listeners for the overall similarity aspect of the single instrument samples, showing 95% confidence intervals.

3.5 Conclusions

A technique was proposed to process B-format room impulse responses (RIRs) and head related transfer functions (HRTFs) to obtain a set of binaural room impulse responses (BRIRs), individualized to the same head and torso as the used HRTFs.

This enables conversion of different HRTF sets to BRIR sets for different rooms with only a need for measuring each room with a B-format microphone. The synthesis of the BRIRs is done differently for direct sound and diffuse sound. The direct sound is extracted from a B-format RIR and its direction of arrival is estimated. It is then filtered with the HRTF corresponding to its direction of arrival to generate the direct sound in the BRIR. The late (diffuse) BRIRs are generated by using a linear combination of the B-format signals, chosen at each frequency such that the spectral and interaural cues are the same as for the true BRIRs.

The BRIRs generated with the proposed method were compared to measured reference BRIRs. The comparison has shown that with respect to the spectra and the frequency-dependent interaural coherence, the BRIRs generated with the proposed method are very close to the reference BRIR up to 3 kHz. It may be argued that the interaural coherence above this frequency is less relevant and that rather the coherence of the envelopes is to be considered [van de Par and Kohlrausch, 1995; Bernstein and Trahiotis, 1996]. Also the waveforms of the early reflections are relatively similar, which can be explained because the linear decoding method uses directional responses similar to the directional responses of the human ear. Therefore, even though the linear decoding is based on the assumption that the B-format recording contains only perfectly diffuse sound, i.e. a hypothesis which is true for late reflections, but not for the early reflections, it approximates the ILD of the early reflections in the measured BRIR.

There are known limitations of the method proposed here. The coherence of the generated BRIR does not match the reference BRIR above approximately 4 kHz. There is some coloration around 200 Hz which may be due to the abrupt transition between the direct sound processing and the linear diffuse sound decoding. A better method of separating the direct sound from the rest of the B-format RIR could help solving this issue.

A subjective test was performed, which showed that the differences in spatial aspects and coloration, and the overall similarity of the generated BRIRs to the reference BRIRs are hardly noticeable. The proposed method also performed significantly better than a conventional method of generating BRIRs from B-format RIRs using cardioid responses extracted from the B-format RIR to which the corresponding HRTFs were applied.

Chapter 4

Artificial Binaural Reverberation Using Coherence Matching

4.1 Introduction

The convolution of a dry source signal with a binaural room impulse response (BRIR) produces a signal that, when played back using headphones, gives the listener the impression of hearing the source in the room where the BRIR was recorded. While applying BRIRs as FIR filters is feasible even in real time due to efficient implementations such as [Gardner, 1995], recording a high number of BRIRs to simulate any combination of listener and source positions in a room is not practicable.

Artificial binaural reverberators [Jot et al., 1995, 2006; Borss and Martin, 2009] can overcome the BRIR recording problem by simulating BRIRs based on a model and can also reduce the computational complexity of convolving a signal with a BRIR (or rather an approximation thereof). While most approaches to binaural reverberation include modeling early reflections using tapped delay lines and HRTFs, the goal here is to present two types of reverberators that are based on a different approach to modeling early reflections.

As has been shown in Chapter 2, replacing the entire tail of a BRIR (i.e. everything except for the direct sound) by noise filtered to match the spectrum, the reverberation time and the interaural coherence as functions of frequency leads to a good approximation of the original BRIR. Therefore the first reverberator presented here does not accurately model early reflections, but models their impact on the interaural coherence as a function of frequency. This reverberator is implemented as a simple extension of the monaural Jot reverberator [Jot, 1992].

Chapter 2 also shows that some perceptual differences between a recorded binaural room impulse response and a modeled room impulse response could not be explained only in terms of interaural coherence, spectra, and reverberation times. Individual early reflections are likely to cause these perceptual differences. The second reverberator adds to the first reverberator an additional feedback delay network to simulate an

infinity of early reflections, while reproducing the delays and amplitudes of first and second order reflections as predicted by the image source model [Allen and Berkley, 1979].

The two approaches for implementing binaural reverberators are described in Sections 4.2 and 4.3, respectively. Experimental results comparing the two methods are presented in Section 4.4 and conclusions are drawn in Section 4.5.

4.2 Simple binaural reverberation using coherence matching

In 1991, Jot [Jot and Chaigne, 1991] proposed a general method to design reverberators based on feedback delay networks, allowing to control the spectrum of the impulse response as well as the reverberation time as a function of frequency. In this section a binaural reverberator is proposed as an extension of the Jot reverberator as described in [Jot, 1992], allowing to match the interaural coherence as a function of frequency to the coherence of the reverb tail of a reference BRIR. The proposed binaural reverberator plus HRTFs for the direct sound can be used as a method for simulating BRIRs with very low computational complexity.

4.2.1 Concept

To enable separate processing of the direct sound and the reverberation tail of BRIRs, the BRIRs are decomposed into a direct sound and a tail part,

$$\begin{aligned} b_{\mathrm{L}}(n) &= b_{\mathrm{L,direct}}(n) + b_{\mathrm{L,tail}}(n) \\ b_{\mathrm{R}}(n) &= b_{\mathrm{R,direct}}(n) + b_{\mathrm{R,tail}}(n)\,, \end{aligned} \tag{4.1}$$

where n is the discrete time index, the direct sound parts are denoted $b_{\mathrm{L,direct}}$ and $b_{\mathrm{R,direct}}$, and the remaining parts are denoted $b_{\mathrm{L,tail}}$ and $b_{\mathrm{R,tail}}$. Note that $b_{\mathrm{L,direct}}$ and $b_{\mathrm{R,direct}}$ are equivalent to HRTFs if the first reflection arrives sufficiently late (later than approximately 3 ms after the direct sound).

The proposed reverberator implements the early parts of the BRIR, $b_{\mathrm{L,direct}}$ and $b_{\mathrm{R,direct}}$, as FIR filters while the late parts, $b_{\mathrm{L,tail}}$ and $b_{\mathrm{R,tail}}$, are generated using a specially designed reverberator to avoid the computational complexity which would arise from directly convolving sound signals with the entire BRIR, which can be tens of thousands samples long.

The BRIR tails are modeled as

$$\begin{aligned} \hat{b}_{\mathrm{L,tail}}(n) &= h_L \star (u \star r_1 + v \star r_2)(n) \\ \hat{b}_{\mathrm{R,tail}}(n) &= h_R \star (u \star r_1 - v \star r_2)(n)\,, \end{aligned} \tag{4.2}$$

where $\star$ denotes linear convolution, $r_1(n)$ and $r_2(n)$ are uncorrelated impulse responses having the desired frequency-dependent reverberation time of the BRIR, $u(n)$ and $v(n)$ are used to perform the coherence matching, and $h_L(n)$ and $h_R(n)$ adjust the

spectrum of the output signal to match the spectrum of the left and right channels of the reference BRIR.

The Fourier transforms of $h_L(n)$ and $h_R(n)$ can be calculated as

$$\begin{aligned} H_L(\omega) &= \sqrt{\frac{P_L(\omega)}{T_r(\omega)}} \\ H_R(\omega) &= \sqrt{\frac{P_R(\omega)}{T_r(\omega)}}, \end{aligned} \tag{4.3}$$

where $T_r(\omega)$ is the frequency-dependent reverberation time as defined in [Jot, 1992]. Note that $h_R(n)$ and $h_R(n)$ correspond to the tone correction filter $t(z)$ defined in [Jot, 1992].

In the following is described how to determine the filters $u(n)$ and $v(n)$. Note that these filters need to have a low number of coefficients in order to achieve low computational complexity. In an FIR implementation, $u(n)$ and $v(n)$ may be truncated to suit eventual complexity requirements and in an IIR implementation, a low filter order may be chosen, depending on the desired accuracy of the coherence matching.

The desired filters in the frequency domain are written as

$$\begin{aligned} U(\omega) &= \sqrt{\frac{1+\Phi(\omega)}{2}} \\ V(\omega) &= \sqrt{\frac{1-\Phi(\omega)}{2}}, \end{aligned} \tag{4.4}$$

where $\Phi(\omega)$ is the coherence of the reference BRIR tail as a function of frequency.

Supposing that $r_1(n)$ and $r_2(n)$ are un-correlated, it can be shown from (4.2) and (4.4) that the coherence between $\hat{B}_{\mathrm{L,tail}}(\omega)$ and $\hat{B}_{\mathrm{R,tail}}(\omega)$ is $\Phi(\omega)$.

In order to be able to implement the late BRIR as defined in (4.2) efficiently, an artificial reverberator generating two uncorrelated reverb signals having the same frequency-dependent reverberation time is needed. The design of such an artificial reverberator yielding the desired reverberation signals $r_1(n)$ and $r_2(n)$ is described in Section 4.2.3.

4.2.2 Estimating the parameters from a reference BRIR

To obtain the filters $h_L(n)$, $h_R(n)$, $u(n)$ and $v(n)$, estimates of the left and right BRIR tail power spectra ($P_L(\omega)$ and $P_R(\omega)$) and of the coherence $\Phi(\omega)$ between the left and right BRIR tails are needed. In the following it is explained how to estimate these parameters, given the reference BRIR tails.

A short-time Fourier transform (STFT) is applied to overlapping blocks of the left and right BRIR tails, yielding $B_L(i,k)$ and $B_R(i,k)$, where i and k are the frequency and time indices, respectively. The reasons why an STFT was used are:

- The time-domain filters $u(n)$ and $v(n)$ need to be short in order to achieve low computational complexity. Thus, the reduced frequency resolution of the STFT compared to a single Fourier transform applied to the tail is enough.
- For the estimation of the coherence more than one sample per frequency is needed.

The power spectra and coherence are estimated as

$$\begin{aligned} P_L(i) &= \frac{1}{K}\sum_{k=1}^{K}|B_L(i,k)|^2 \\ P_R(i) &= \frac{1}{K}\sum_{k=1}^{K}|B_R(i,k)|^2 \\ \Phi(i) &= \frac{\Re(\sum_{k=1}^{K} B_L(i,k)B_R(i,k)^\star)}{\sqrt{\sum_{k=1}^{K}|B_L(i,k)|^2\sum_{k=1}^{K}|B_R(i,k)|^2}}, \end{aligned} \tag{4.5}$$

where $|.|$ is the magnitude of a complex number, $\Re(x)$ denotes the real part of x, * denotes the complex conjugate, and K is the number of STFT frames. If needed, the frequency resolution of $P_L(\omega)$ and $P_R(\omega)$ can be modified by interpolating the STFT bins to match the resolution of $T_r(\omega)$ in (4.3).

The top panel in Figure 4.1 shows the left and right power spectra of an example BRIR and the bottom panel shows the interaural coherence of the same BRIR as a function of frequency. The estimated $P_L(\omega)$, $P_R(\omega)$, $\Phi(\omega)$ can be frequency smoothed as much as needed in order to obtain short filters $u(n)$ and $v(n)$.

4.2.3 Design of an uncorrelated two-channel reverberator

The complete reverberator as shown in Figure 4.2 is an extension of the Jot reverberator as described in [Jot, 1992], which was modified in order to produce a second, uncorrelated reverberation channel and to which the filters described in (4.2) were added.

The output signal in the original Jot reverberator is obtained by combining the N channels using a weights vector $\vec{c} = [c_1, c_2, \ldots, c_N]$, leading to an intermediate signal $r_1(n)$ and filtering this signal using a tone correction filter. The idea behind the structure of the reverberator proposed in this paper, as seen in Figure 4.2, is to use a second weights vector $\vec{d} = [d_1, d_2, \ldots, d_N]$ in order to combine the N channels

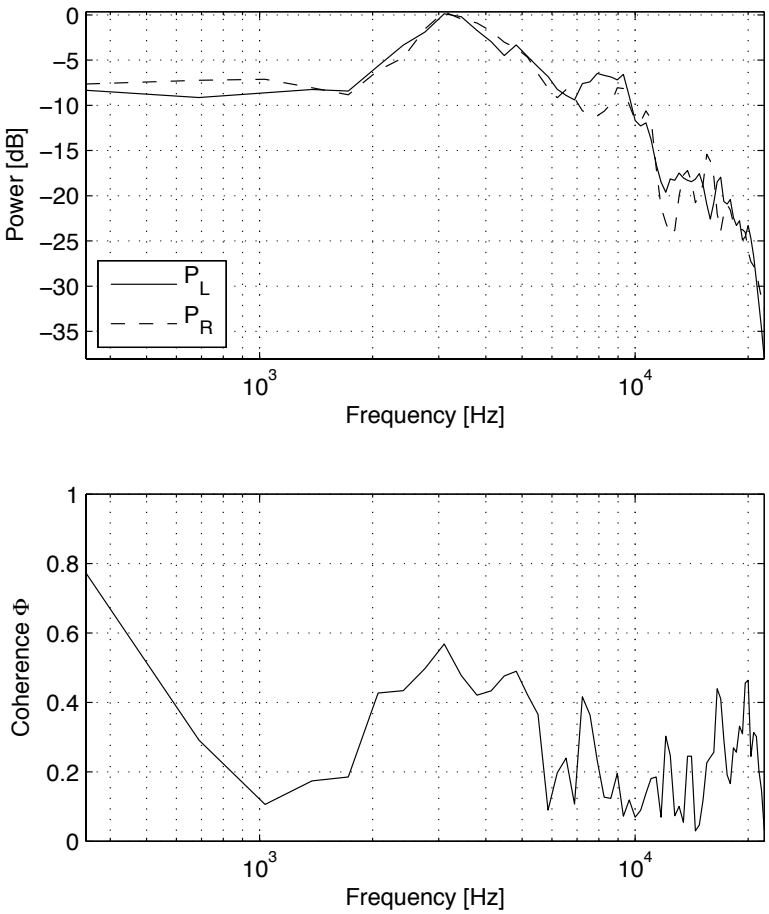

Figure 4.1: Top: power spectrum estimated from example left and right BRIR tails. **Bottom:** corresponding estimated frequency-dependent interaural coherence.

of the reverberator to create an intermediate signal $r_2(n)$ which is not correlated with $r_1(n)$.

Under the hypothesis that the N channels of the reverberator produce uncorrelated Gaussian noise with equal amplitudes, the condition

$$\vec{c} \perp \vec{d} \tag{4.6}$$

is sufficient to assure that $r_1(n)$ and $r_2(n)$ are uncorrelated. In order to have the same energy in $r_1(n)$ and $r_2(n)$, it is necessary to impose also

$$||\vec{c}|| = ||\vec{d}|| \,. \tag{4.7}$$

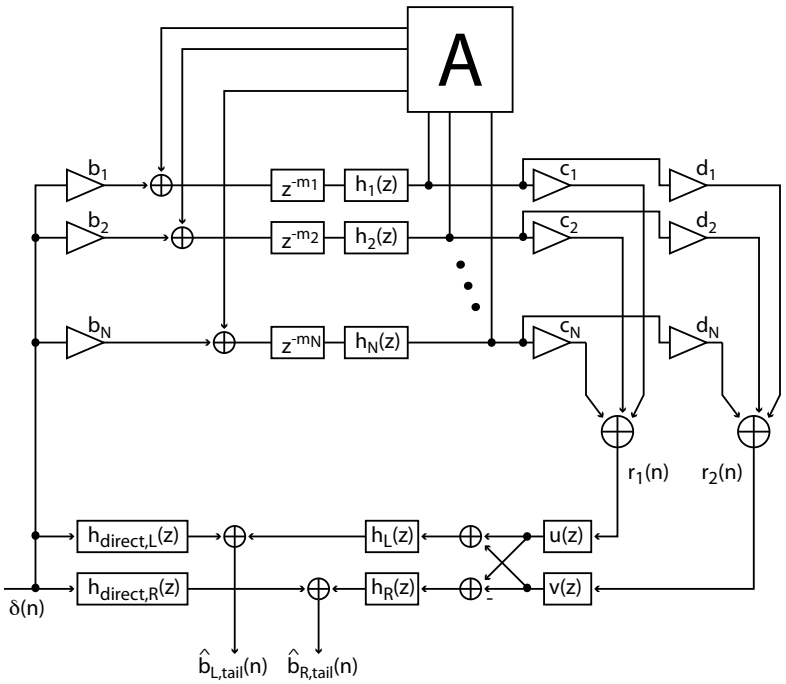

Figure 4.2: Complete block diagram of binaural modified Jot reverberator. A denotes the mixing matrix as defined in [Jot, 1992] and $h_i(z)$ are the filters needed to adjust the frequency-dependent reverberation time. Names of signals were chosen for the case where the input is a single Dirac impulse and the output is the impulse response of the reverberator.

For $N = 6$, a practical example is

$$\begin{aligned} \vec{c} &= [1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1] \\ \vec{d} &= [1 \quad -1 \quad 1 \quad -1 \quad 1 \quad -1] \ . \end{aligned} \tag{4.8}$$

In theory, trivial solutions to (4.6) and (4.7) like $\vec{c} = [100000]$ and $\vec{d} = [010000]$ are possible, but in practice these solutions are undesirable because they lead to a lower reflection density in the first part of the reverberation than the solutions given in (4.8).

The intermediate signals $r_1(n)$ and $r_2(n)$ are processed as defined in (4.2), where the filters $u(n)$, $v(n)$, $h_L(n)$ and $h_R(n)$ are implemented as FIR filters obtained by taking the inverse Fourier transforms of $U(\omega)$, $V(\omega)$, $H_L(\omega)$ and $H_R(\omega)$ as defined in (4.3) and (4.4). In practice, it is possible to truncate the obtained FIR filters, which makes the implementation more efficient, but in return leads to a loss of fre-

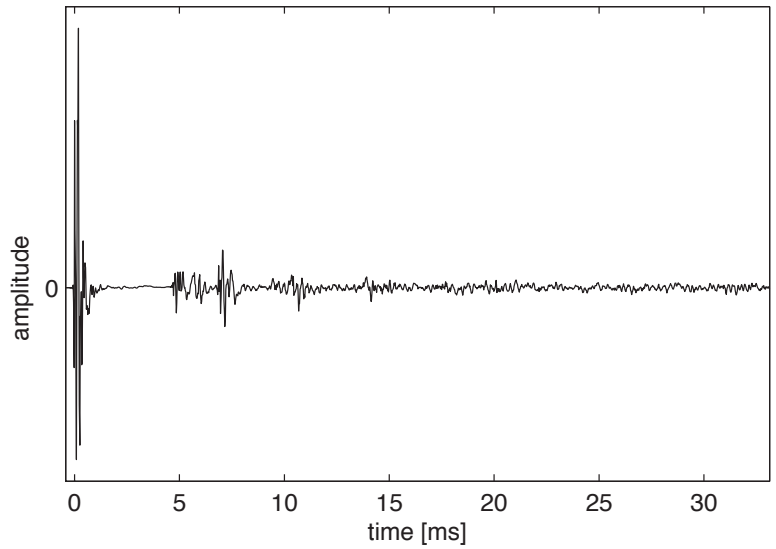

Figure 4.3: Waveform of a measured BRIR. For simplicity only one channel is shown.

quency resolution in the approximation of the left and right spectra as well as in the approximation of $\Phi(\omega)$.

To obtain the output signal of the binaural reverberator, the outputs of the filters $h_L(n)$ and $h_R(n)$ are added to the input signal convolved with the left and right channels of the HRTF corresponding to the direction of the direct sound.

4.3 Implementation based on a geometrical model

As has been shown already in Chapter 2, the presence of distinct early reflections may have an impact on the perception not explainable in terms of interaural coherence and spectral cues only, possibly due to the influence of early reflections on the early lateral energy fraction [Begault, 1992], which is not taken into account in an approach based only on coherence. Therefore an implementation of a binaural reverberator modeling also early reflections is preferable over a one that does not model early reflections, in particular for applications where a high quality is necessary and where a slightly higher computational complexity is not a problem.

While the idea of separating a binaural reverberator into an early reflections part and a late reverb part has been proposed many times, e.g. in [Jot et al., 1995], the approach presented here uses two feedback delay networks with highly overlapping impulse responses, rather than simulating only a limited number of early reflections and to use a reverberator only for the late BRIR tail. While at first glance the choice of using two feedback delay networks in parallel may seem strange, it is easily justified.

By looking at the waveform of a BRIR (see Figure 4.3) it can be noticed that there is never a clearly definable time instant where the early reflections end and where the late reverb starts. There is a certain amount of diffuse (or "ambient") reverberation present almost from the beginning of the BRIR and the distinct reflections never "stop", but become weaker and denser such that at some point they are not distinguishable from diffuse reverberation anymore.

The main difference between the diffuse reverberation and the distinct reflections is that they contribute to the interaural coherence in different ways. Early reflections coming from the front or the back of the listener arrive at the same time at the two ears and will add up coherently, leading to a high interaural coherence (see also Chapter 2). Early reflections from the side will not lead to such a high interaural coherence, at least not across all frequency bands. Diffuse reverberation on the other hand will contribute to a frequency dependent interaural coherence similar to the interaural coherence curve that is characteristic for diffuse sound [Cook et al., 1955]. Therefore, in order to be able to reproduce a realistic interaural coherence, it is necessary to use two separate reverberators whose impulse responses which have different interaural coherence curves. By applying HRTFs corresponding to the directions where the distinct reflections come from, the interaural coherence due to the distinct reflections will be modeled implicitly. The coherence matching for the diffuse sound is performed by the reverberator structure that has already been used in Section 4.2.

4.3.1 Room model

The reverberator modeling the distinct reflections is based on the concept that each channel of a feedback delay network can be interpreted as an object where sound is reflected (in general this corresponds to a wall) [Stautner and Puckette, 1982]. In the following an example is made for a 4-channel reverberator modeling the reflections occurring at the 4 walls of a rectangular room (i.e. the floor and ceiling reflections are not taken into account). In practice, the floor and ceiling reflections should be modeled too since they are likely to be perceptually relevant [Rakerd and Hartmann, 1985], but for simplicity in illustrating the concept, they are not treated here.

Figure 4.4 shows the positions of the sound source and the listener as well as the points where the sound is reflected according to the image source model. These four points correspond to the four channels of the reverberator. The output of each channel is convolved with the HRTFs for the direction from which sound from the corresponding reflection point arrives at the listener's head. Figure 4.5 shows the possible reflections paths between the reflecting points. There are reflection paths possible between all pairs of points, but not between one point and itself. From this follows that if a mixing matrix is defined as in [Jot, 1992], it will have zeros on its diagonal and nonzero elements everywhere else. Note that the delays and amplitude factors in the feedback loop are calculated as a function of the second order image sources. The paths shown in Figure 4.5 are only a visualization of the structure of the feedback loop and their length has no importance.

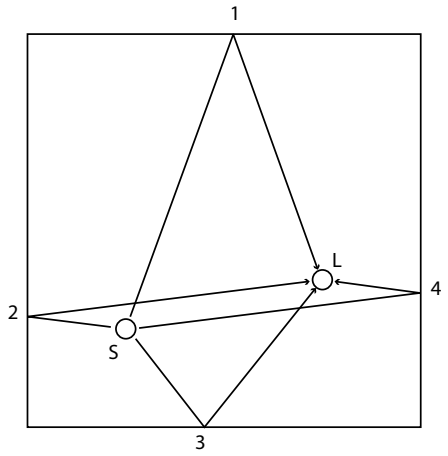

Figure 4.4: Positions of a source (S) and and a listener (L) in a rectangular room and horizontal first order reflection paths. The numbers 1 to 4 indicate the points on the walls where the reflections occur according to the image source model.

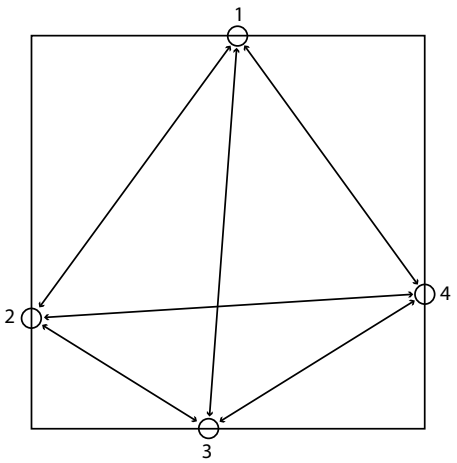

Figure 4.5: Graphic representation of the paths in the feedback loop simulating the reflections in the room shown in Figure 4.4. Note that the delays and amplitude factors in the feedback loop are calculated as a function of the second order image sources and not as a function of the path lengths shown in this figure.

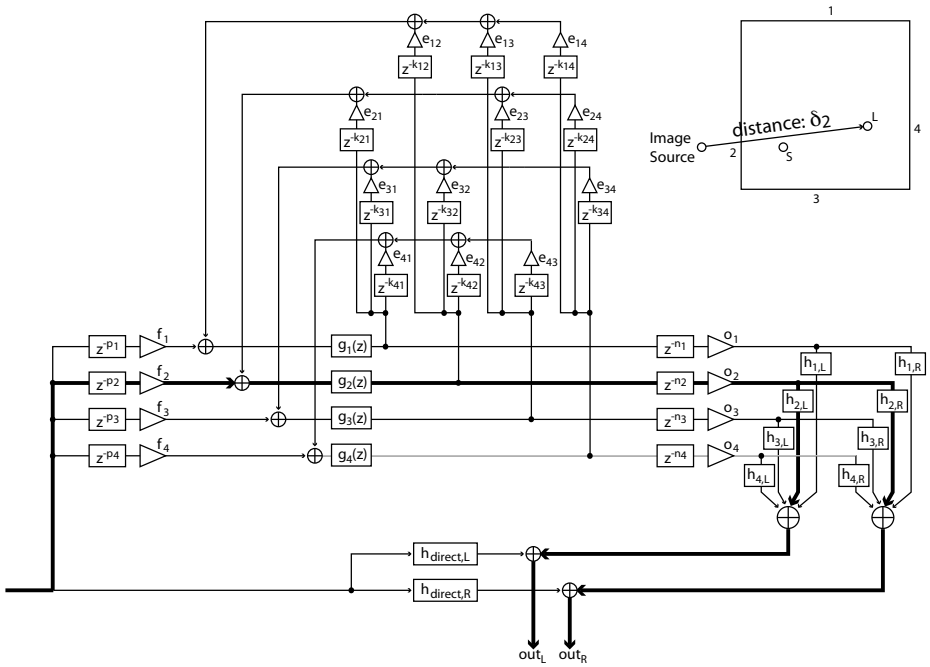

Figure 4.6: Early reflections reverberator with highlighted signal path corresponding to a first order reflection on wall 1. In the top right corner, the position of the image source modeled by the highlighted path is shown.

4.3.2 Structure of the early reflections reverberator

Figure 4.6 shows the early reflections reverberator. It contains HRTFs for the direct sound, a feedback loop implementing the feedback paths defined in Figure 4.5, delays and amplitude factors necessary for modeling first and second order reflections, and HRTFs modeling the sound propagation from the reflection points to the listener. Furthermore, filters inside the loop are used to model the frequency dependent absorption associated with the reflection points.

As illustrated in Figure 4.6, a single reflection from wall i is modeled by a delay of $p_i + n_i$ samples, an amplification factor of $f_i o_i$, a filter $g_i(z)$ modeling the frequency-dependent absorption associated with the reflection point, and left and right HRTFs $h_{i,L}(z)$, $h_{i,R}(z)$ for the direction of the reflection point relative to the listener. Assuming that the distance of the image source corresponding to the echo path for a first order reflection at wall i is δ_i, the following equations must hold for the delays and the amplitude factors in the corresponding signal path:

$$p_i + n_i = \delta_i / c \tag{4.9}$$

$$f_i \cdot o_i = 1/\delta_i \ , \tag{4.10}$$

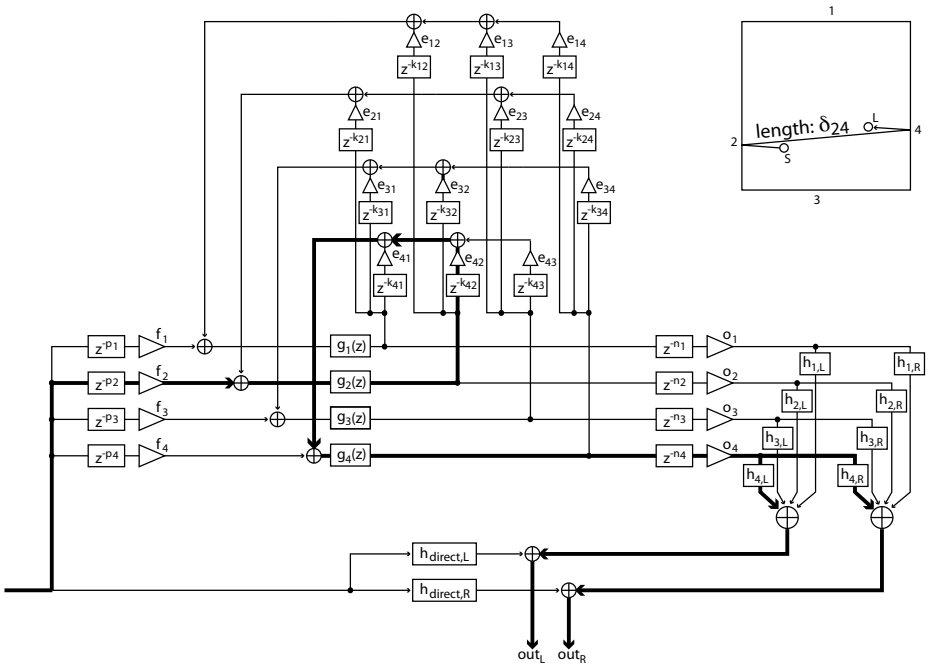

Figure 4.7: Early reflections reverberator with highlighted signal path corresponding to a second order reflection on opposite walls 2 and 4. In the top right corner, the position of the image source modeled by the highlighted path is shown.

where c is the speed of sound and (4.10) follows from the sound pressure as a function of distance for a point source [Allen and Berkley, 1979].

A double reflection at two walls facing each other – first at wall i, then at wall j – can be modeled by a path passing once through the feedback loop, as shown in Figure 4.7. Considering δ_{ij} to be the distance between the the listener and the image source corresponding to a reflection first at wall i then at wall j, the following equations must hold:

$$p_i + k_{ji} + n_j = \delta_{ij}/c \tag{4.11}$$

$$f_i \cdot e_{ji} \cdot o_j = 1/\delta_{ij} \; . \tag{4.12}$$

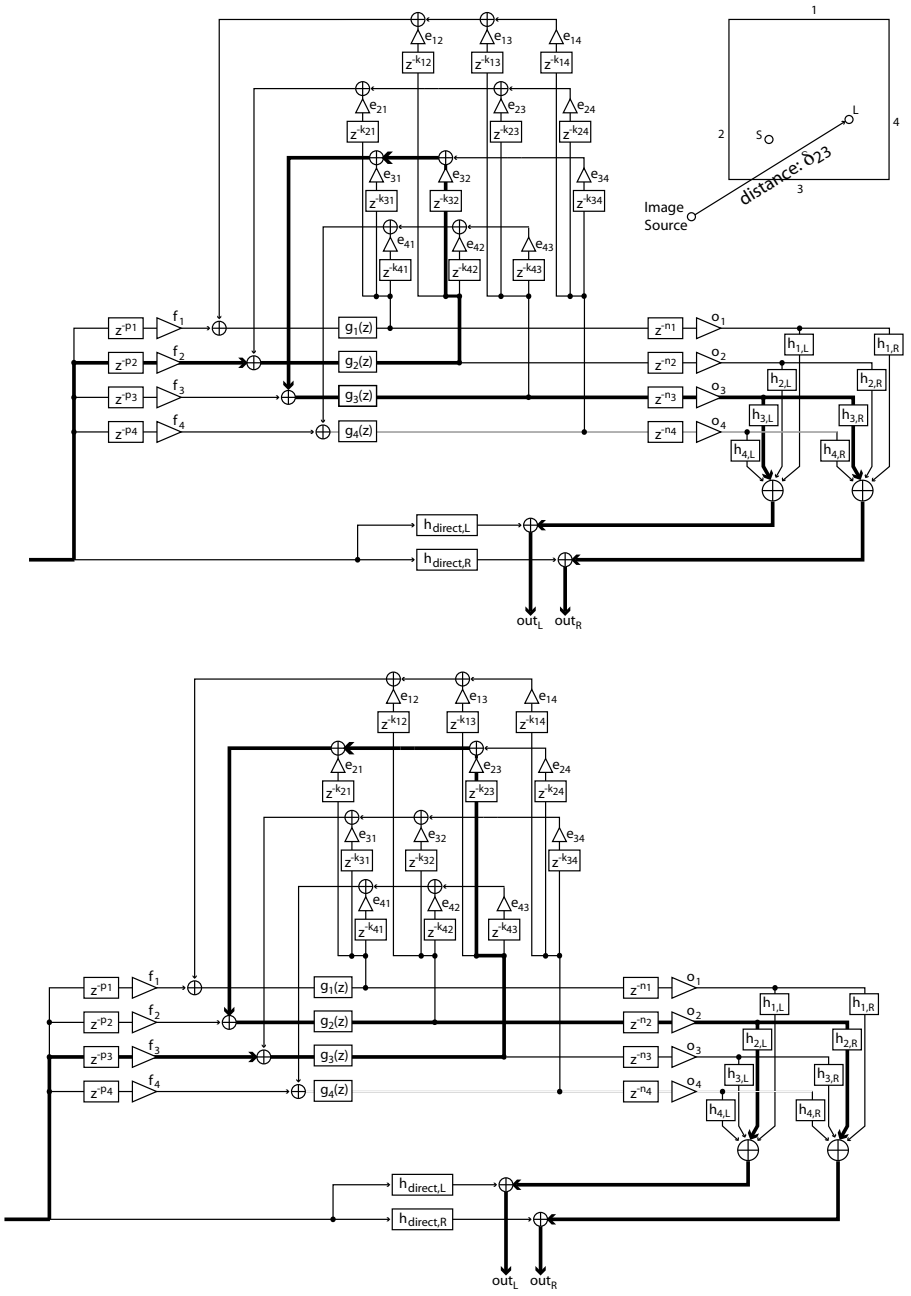

Figure 4.8: The two signal paths in the early reflections reverberator corresponding to a second order reflection on adjacent walls 2 and 3. In the top right corner, the position of the image source modeled by the highlighted paths is shown.

Note that in this case the reflected sound will be convolved with the same HRTF as the sound reflected only at wall j, which is a good approximation for the correct direction when the listener is far from wall j.

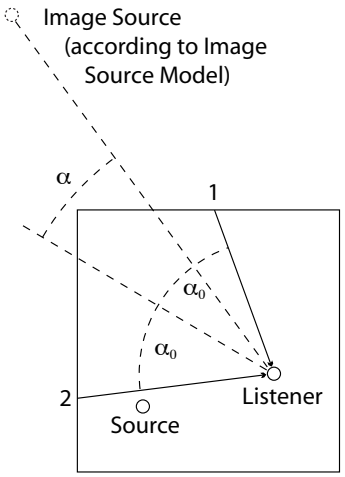

Figure 4.9: Positions of image source according to image source model and angles used for amplitude panning between channels.

For the case of a double reflection at adjacent walls i and j, two signal paths exist in the reverberator, as illustrated in Figure 4.8. The signal of the first path passes through the delays p_i, k_{ji}, and n_j and is convolved with the HRTF of channel j, while the signal of the other path passes through the delays p_j, k_{ij}, and n_i and is convolved with the HRTF of channel i. By choosing the same delay and suitable amplitude factors for the two channels, it is possible to simulate the correct direction of the image source using amplitude panning. Considering $\delta_{ij} = \delta_{ji}$ as the distance between the image source and the listener, the corresponding equations are:

$$p_i + k_{ji} + n_j = \delta_{ij}/c \tag{4.13}$$
$$p_j + k_{ij} + n_i = \delta_{ji}/c \tag{4.14}$$
$$f_i \cdot e_{ji} \cdot o_j = a_{ij}/\delta_{ij} \tag{4.15}$$
$$f_j \cdot e_{ij} \cdot o_i = a_{ji}/\delta_{ji}\ , \tag{4.16}$$

where a_{ij} and a_{ji} are amplitude panning factors which can be calculated using the following equation system [Bernfeld, 1973; Bennett et al., 1985].

$$\frac{a_{ij} - a_{ji}}{a_{ij} + a_{ji}} = \frac{\tan\alpha}{\tan\alpha_0} \tag{4.17}$$
$$a_{ij}^2 + a_{ji}^2 = 1\ , \tag{4.18}$$

where α is the angle between the direction of the image source and the bisecting line between the directions associated with channels i and j (negative if closer to the

direction of i and positive if closer to the direction of j) and α_0 is the angle between the directions associated with channels i and j divided by two. An example for the angles α and α_0 in the case where $i = 2$ and $j = 1$ is given in Figure 4.9.

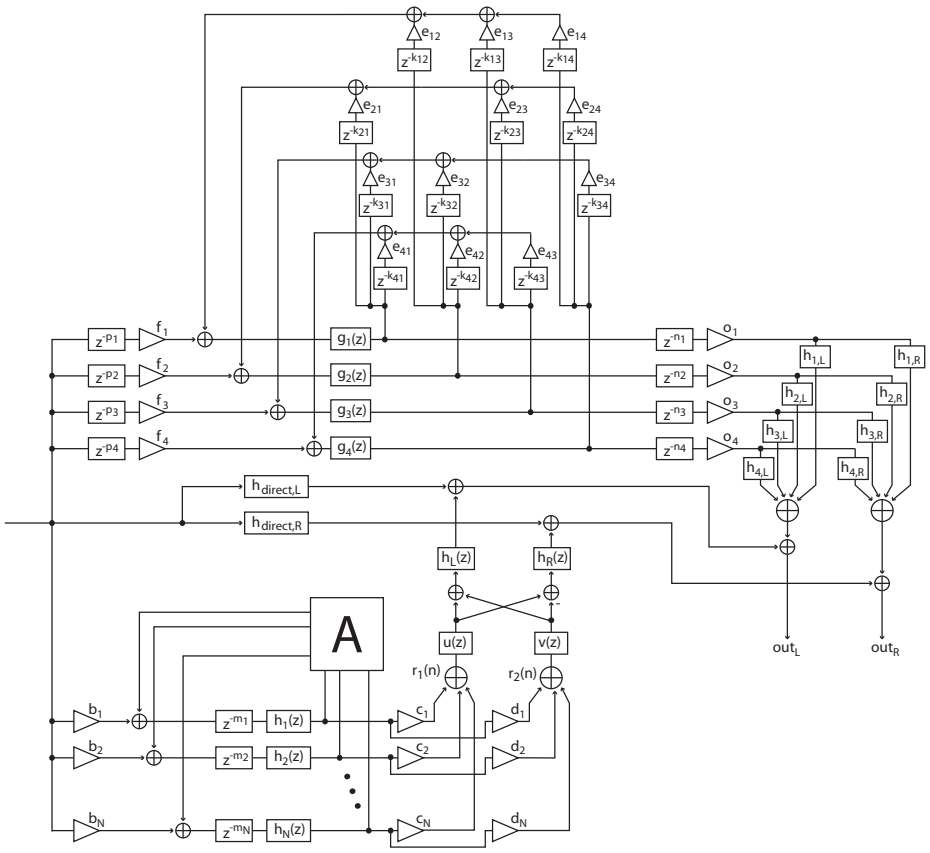

Figure 4.10: Binaural reverberator composed of the early reflections reverberator shown in Figure 4.6 and a diffuse sound reverberator having the structure shown in Figure 4.2.

4.3.3 Calculating parameters from the room model

While in the previous section constraints for the delays and amplitude factors were introduced, it is a priori not evident if a solution exists and how the parameters n_i, p_j, k_{ij}, o_i, f_j, and e_{ij} $(i, j \in \{1, 2, 3, 4\})$ can be calculated in practice.

In the case of the delays, evaluating (4.9) and (4.11) for all possible combinations of i and j leads to the linear system

$$\underbrace{\begin{bmatrix}
1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0\\
1&0&0&0&0&1&0&0&0&0&0&1&0&0&0&0&0&0&0&0\\
1&0&0&0&0&0&1&0&0&0&0&0&0&0&1&0&0&0&0&0\\
1&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&1&0&0&0\\
0&1&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&0&1&0&0&0&0&0&0&0&0&1&0&0&0&0\\
0&1&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&1&0&0\\
0&0&1&0&1&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&1&0&0&0&0&0&0&1&0&0&0&0&0&0&0\\
0&0&1&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&1\\
0&0&0&1&1&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&1&0&0&0&0&0&0&0&1&0&0&0&0&0&0\\
0&0&0&1&0&0&1&0&0&0&0&0&0&0&0&1&0&0&0&0
\end{bmatrix}}_{M_1} \cdot \underbrace{\begin{bmatrix}
p_1\\p_2\\p_3\\p_4\\n_1\\n_2\\n_3\\n_4\\k_{12}\\k_{13}\\k_{14}\\k_{21}\\k_{23}\\k_{24}\\k_{31}\\k_{32}\\k_{34}\\k_{41}\\k_{42}\\k_{43}
\end{bmatrix}}_{K} = \frac{1}{c} \underbrace{\begin{bmatrix}
\delta_1\\\delta_2\\\delta_3\\\delta_4\\\delta_{12}\\\delta_{13}\\\delta_{14}\\\delta_{21}\\\delta_{23}\\\delta_{24}\\\delta_{31}\\\delta_{32}\\\delta_{34}\\\delta_{41}\\\delta_{42}\\\delta_{43}
\end{bmatrix}}_{B_1} , \tag{4.19}$$

where c is the speed of sound, δ_i and δ_{ij} are the distances between the listener and the image source mirrored at wall i, and at walls i and j, respectively. The 16×20 matrix M_1 describes which delays in the reverberator contribute to which signal path and B_1 is a vector containing the delays corresponding to the different δ_i and δ_{ij}.

This system can be solved by calculating the Moore-Penrose pseudoinverse $\widetilde{M_1}^{-1}$ of M_1 and multiplying it with the vector B_1 [Penrose, 1956]:

$$K = \widetilde{M_1}^{-1} \cdot B_1 \ . \tag{4.20}$$

It is worth noting that $\widetilde{M_1}^{-1}$ does not depend on the room size, nor the listener position nor the source positions. Therefore this matrix is the same for all reverberators and can be stated explicitly:

$$\widetilde{M_1}^{-1} = \frac{1}{10} \begin{bmatrix}
5&0&0&0&1&1&1&-1&0&0&-1&0&0&-1&0&0\\
0&5&0&0&-1&0&0&1&1&1&0&-1&0&0&-1&0\\
0&0&5&0&0&-1&0&0&-1&0&1&1&1&0&0&-1\\
0&0&0&5&0&0&-1&0&0&-1&0&0&-1&1&1&1\\
5&0&0&0&-1&-1&-1&1&0&0&1&0&0&1&0&0\\
0&5&0&0&1&0&0&-1&-1&-1&0&1&0&0&1&0\\
0&0&5&0&0&1&0&0&1&0&-1&-1&-1&0&0&1\\
0&0&0&5&0&0&1&0&0&1&0&0&1&-1&-1&-1\\
-5&-5&0&0&2&1&1&8&-1&-1&-1&1&0&-1&1&0\\
-5&0&-5&0&1&2&1&-1&1&0&8&-1&-1&-1&0&1\\
-5&0&0&-5&1&1&2&-1&0&1&-1&0&1&8&-1&-1\\
-5&-5&0&0&8&-1&-1&2&1&1&1&-1&0&1&-1&0\\
0&-5&-5&0&-1&1&0&1&2&1&-1&8&-1&0&-1&1\\
0&-5&0&-5&-1&0&1&1&1&2&0&-1&1&-1&8&-1\\
-5&0&-5&0&-1&8&-1&1&-1&0&2&1&1&1&0&-1\\
0&-5&-5&0&1&-1&0&-1&8&-1&1&2&1&0&1&-1\\
0&0&-5&-5&0&-1&1&0&-1&1&1&1&2&-1&-1&8\\
-5&0&0&-5&-1&-1&8&1&0&-1&1&0&-1&2&1&1\\
0&-5&0&-5&1&0&-1&-1&-1&8&0&1&-1&1&2&1\\
0&0&-5&-5&0&1&-1&0&1&-1&-1&-1&8&1&1&2
\end{bmatrix} . \tag{4.21}$$

Using the same approach, one can also calculate values for the amplitude factors o_i, f_j, and e_{ij} ($i, j \in \{1, 2, 3, 4\}$). However, in order to obtain a linear system, the equations (4.10), (4.12), and (4.15) need to be reformulated. By taking the logarithm on both sides, the products become sums, e.g. (4.10) becomes

$$\log(f_i) + \log(o_i) = -\log(\delta_i)\,, \tag{4.22}$$

and the reformulated equations result in a linear system as shown below:

$$\underbrace{\begin{bmatrix}
1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0\\
1&0&0&0&0&1&0&0&0&0&0&1&0&0&0&0&0&0&0&0\\
1&0&0&0&0&0&1&0&0&0&0&0&0&0&1&0&0&0&0&0\\
1&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&1&0&0&0\\
0&1&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&0&1&0&0&0&0&0&0&0&0&1&0&0&0&0\\
0&1&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&1&0&0\\
0&0&1&0&1&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&1&0&0&0&0&1&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&1\\
0&0&0&1&1&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&1&0&0&0&0&0&0&0&1&0&0&0&0&0&0\\
0&0&0&1&0&0&1&0&0&0&0&0&0&0&0&1&0&0&0&0\\
0&0&0&0&0&0&0&0&1&1&1&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&1&1&1&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1
\end{bmatrix}}_{M_2} \cdot \underbrace{\begin{bmatrix}
\log(f_1)\\ \log(f_2)\\ \log(f_3)\\ \log(f_4)\\ \log(o_1)\\ \log(o_2)\\ \log(o_3)\\ \log(o_4)\\
\log(e_{12})\\ \log(e_{13})\\ \log(e_{14})\\ \log(e_{21})\\ \log(e_{23})\\ \log(e_{24})\\
\log(e_{31})\\ \log(e_{32})\\ \log(e_{34})\\ \log(e_{41})\\ \log(e_{42})\\ \log(e_{43})
\end{bmatrix}}_{A} = -\underbrace{\begin{bmatrix}
\log(\delta_1)\\ \log(\delta_2)\\ \log(\delta_3)\\ \log(\delta_4)\\
\log(\delta_{12}/a_{12})\\ \log(\delta_{13})\\ \log(\delta_{14}/a_{14})\\
\log(\delta_{21}/a_{21})\\ \log(\delta_{23}/a_{23})\\ \log(\delta_{24})\\ \log(\delta_{31})\\
\log(\delta_{32}/a_{32})\\ \log(\delta_{34}/a_{34})\\ \log(\delta_{41}/a_{41})\\
\log(\delta_{42})\\ \log(\delta_{43}/a_{43})\\ \log(\nu)\\ \log(\nu)\\ \log(\nu)\\ \log(\nu)
\end{bmatrix}}_{B_2}\,. \tag{4.23}$$

Compared to the system in (4.19), that the system in (4.23) has four more equations, as can be seen from the last four rows of M_2. These equations are needed to control the stability of the recursive loop, i.e. the behavior of the mixing matrix E, in particular the behavior of its powers E^n. E can be written as:

$$E = \begin{bmatrix}
0 & e_{12} & e_{13} & e_{14}\\
e_{21} & 0 & e_{23} & e_{24}\\
e_{31} & e_{32} & 0 & e_{34}\\
e_{41} & e_{42} & e_{43} & 0
\end{bmatrix}. \tag{4.24}$$

It is not possible to express the constraint that E should be a unitary matrix in the equation system in (4.23). However, simple constraints on the product of amplitude factors can be made in this system, allowing to efficiently control the decay of the early reflections. A possible set of constraints is

$$e_{12}e_{13}e_{14} = \nu \tag{4.25}$$
$$e_{21}e_{23}e_{24} = \nu \tag{4.26}$$
$$e_{31}e_{32}e_{34} = \nu \tag{4.27}$$
$$e_{41}e_{42}e_{43} = \nu\,. \tag{4.28}$$

If all nonzero coefficients in E had the same value, ν should be chosen to be $1/3^3 = 1/27$ in order to ensure the stability of the recursive loop. In practice, where this assumption does not hold, ν needs to be reduced, a suitable value being $1/29$. However, depending on the room that is to be simulated, an even smaller ν may be chosen, leading to a faster decay of the early reflections with an order higher or equal to 3.

The values of o_i, f_j, and e_{ij} $(i, j \in \{1, 2, 3, 4\})$ can be calculated again using the Moore-Penrose pseudoinverse of M_2. Even though the 20×20 matrix M_2 is square, it is necessary to use the pseudoinverse method because M_2 has only rank 19 and therefore does not have an exact inverse. Like M, M_2 does not depend on the room nor on listener or source positions, which means that $\widetilde{M_2}^{-1}$ can be calculated in advance.

4.3.4 Using a Jot reverberator for modeling the diffuse reverberation

The diffuse sound is modeled using a reverberator with the same structure as the one presented in Section 4.2. The only difference is that while the reverberator in Section 4.2 is normally designed to model the properties of a normal reverb tail, in particular also the increase of echo density over time, the diffuse sound reverberator is designed to model only the diffuse part of the reverberation and should therefore have a high echo density from the beginning. This is achieved by using a large number of channels and a mixing matrix that was found suitable for this application (see Appendix A). Since the diffuse reverberation is supposed to be evenly distributed in the whole room and also over all directions of propagation, the diffuse sound reverberator is designed once as a function of the room and does not need to be modified as a function of the position or orientation of the source or the listener.

While the implementation of the diffuse sound reverberator uses the same coherence matching filters as the implementation in Section 4.2, the mixing matrix, the number of channels and the delays are different. A sparse matrix composed from the elements of 9 3×3 unitary matrices arranged to ensure fast mixing between channels (denoted U_{3f} in Appendix A) was used. This results in a reverberator with 27 channels whose recursive loop can be implemented with 200 multiply operations per output sample. The delays are chosen to be mutually prime numbers with a mean of 800 samples.

However, there are also many other possible choices for the mixing matrix which allow efficient implementations, such as Hadamard matrices, or sparse matrices with nonzero elements taken from Hadamard matrices, as is discussed in Appendix A.5.

4.4 Results

The method in Section 4.2 was applied to a BRIR measured in a lecture hall at EPFL and a binaural reverberator with $N = 6$ was designed. The frequency dependent interaural coherence of the reverberator was calculated using (4.5) and compared to the interaural coherence of the reference BRIR. As can be seen in Figure 4.11, the interaural coherence of the reverberator matches well the interaural coherence of

the reference BRIR. However, the reverberator's interaural coherence is in general slightly higher. This effect proved to be systematic also for reverberators which were designed based on different BRIRs. A possible explanation is that the signals $r_1(n)$ and $r_2(n)$ are not completely uncorrelated, therefore leading to a higher correlation at the output. In this case it would be possible to compensate the correlation between $r_1(t)$ and $r_2(t)$ by introducing correction terms into (4.4).

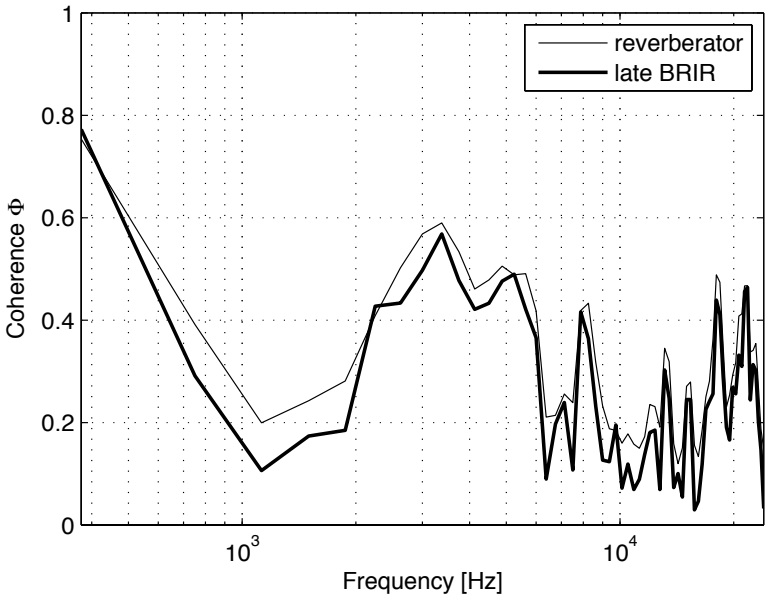

Figure 4.11: Coherence of reference BRIR tail and of artificial reverberator. The coherence of the reverberator's impulse response follows the coherence of the reference late BRIR closely. However, a systematic bias towards higher coherence in the reverberator's impulse response can be observed. Unwanted correlation between $r_1(n)$ and $r_2(n)$ may be the cause of this increase in coherence.

Even though no extensive psychoacoustic tests have been performed, informal listening has lead to the conclusion that a binaural reverberator with frequency dependent coherence matching performs better with respect to the goal of creating a realistic spatial image than a binaural reverberator with uncorrelated $\hat{b}_{\mathrm{L,tail}}(n)$ and $\hat{b}_{\mathrm{R,tail}}(n)$, and also better than a binaural reverberator with an average, frequency-independent coherence between $\hat{b}_{\mathrm{L,tail}}(n)$ and $\hat{b}_{\mathrm{R,tail}}(n)$, which is in line with the results of Chapter 2.

The advanced reverberator consisting of two parallel feedback delay networks was compared with a measured BRIR for its time-frequency interaural coherence. Figure 4.12 shows that the combined reverberator is capable of modeling the increased interaural coherence found at the beginning of the BRIR and that this is due to the contribution of the early reflections reverberator, while the interaural coherence of

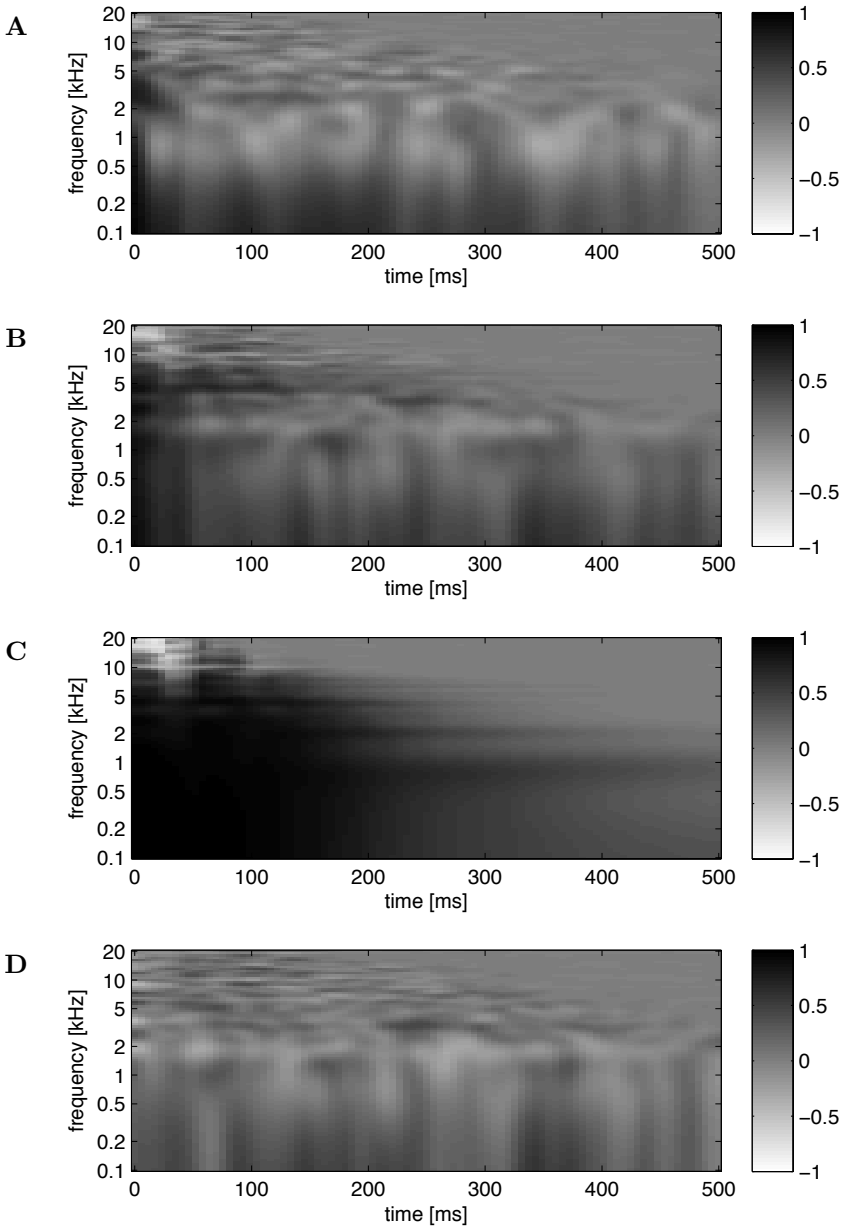

Figure 4.12: Time-frequency interaural coherence of impulse responses. **A:** measured BRIR, **B:** combined reverberator, **C:** distinct reflections reverberator, **D:** diffuse reverberator.

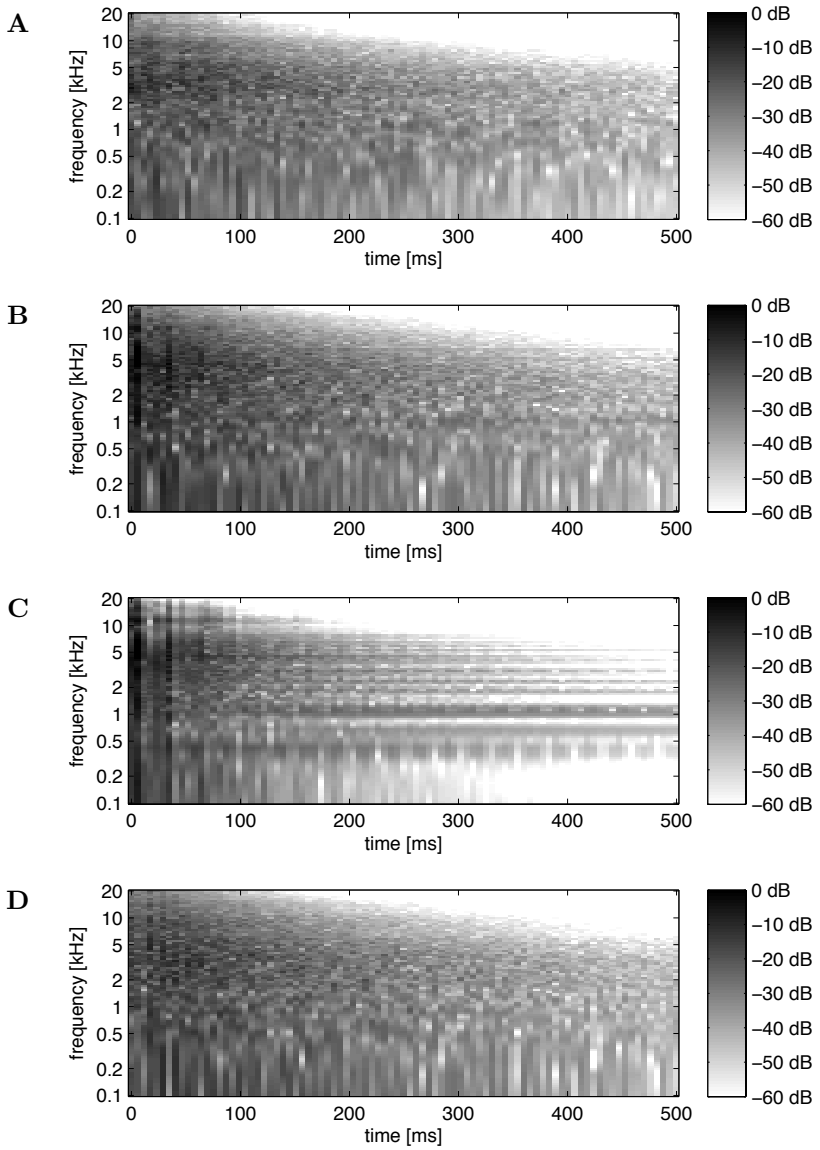

Figure 4.13: Time-frequency spectra of impulse responses. **A:** measured BRIR, **B:** combined reverberator, **C:** distinct reflections reverberator, **D:** diffuse reverberator. Note that the distinct reflections reverberator produces undesired spectral modes (or "ringing", represented by the horizontal lines in plot C), which are covered by the output of the diffuse reverberator in the combined reverberator.

the diffuse reverberator's impulse response has no significant dependence on time. Figure 4.13 shows the time-frequency spectra of a measured BRIR and the impulse responses of the combined reverberator, the early reflections reverberator and the diffuse reverberator. It can be observed that the distinct reflections reverberator produces undesired spectral modes (or "ringing", represented by the horizontal lines in plot C), which are covered by the output of the diffuse reverberator in the combined reverberator.

4.5 Conclusions

A binaural reverberator was developed, modeling the temporal overlap between early reflections and diffuse reverberation found in measured BRIRs. The implementation contains two parallel reverberators, one for early reflections and one for diffuse reverberation, both based on feedback delay networks. The early reflections reverberator models the delays, amplitudes and angles of arrival of the 12 first and second order reflections in an efficient way, using only 4 HRTFs. Because of the use of a feedback network, infinitely many reflections are produced, but only first and second order reflections approximate the reflections predicted by the image source model [Allen and Berkley, 1979]. The second order image sources reflected by two adjacent walls are reproduced using amplitude panning between two HRTFs.

In order to calculate the parameters of the early reflections reverberator, two linear equation systems need to be solved, which is potentially computationally intensive and could pose a problem in a real time implementation. However, because the matrices representing the left hand side of the equation system do not depend on the listener or source positions, their pseudo-inverses need to be calculated only once and the parameters of the reverberator can be obtained by calculating the distances of the first and second order image sources and by performing two matrix multiplications. Assuming that it is sufficient to update the parameters of the early reflections reverberator 50 times per second for moving sources, this represents virtually no overhead.

The diffuse sound reverberator is a modified Jot reverberator matching the interaural coherence of a binaural recording of diffuse sound [Menzer and Faller, 2009a]. In order to obtain a dense diffuse reverberation starting shortly after the direct sound, a high number of channels and short delays were chosen. In order to avoid a high computational complexity, a sparse feedback matrix is used, using the design presented in [Menzer and Faller, 2010b].

Furthermore, a simplified reverberator was introduced, omitting the early reflections part and matching the overall interaural coherence as a function of frequency, which is influenced by the early reflections and which was shown to be a major perceptual cue for binaural reverberation (see Chapter 2). This reverberator can be used as a computationally very efficient way of implementing a binaural reverberator.

Chapter 5

Stereo-to-Binaural Conversion Using Coherence Matching

5.1 Introduction

Stereo signals recorded with coincident microphones or mixed in a studio using amplitude panning are suitable for playback with a stereo loudspeaker setup. In this case, amplitude panning will correctly reproduce the physical measures sound pressure and particle velocity for the desired direction of arrival of sound at the sweet spot [Bauer, 1961; Bernfeld, 1973; Bennett et al., 1985]. However, when playing back the same signals with headphones, important binaural cues are not correctly reproduced. In a free field situation (and in general also for direct sound in reverberant environments) there are a direct relationship between interaural level difference (ILD) and the interaural time difference (ITD) [Gaik, 1993] as well as direction-related spectral cues. Amplitude panning on the other hand changes only the ILD and does not produce any ITD or spectral cues. As a result, amplitude panning produces unnatural binaural cues when used for headphone playback.

Furthermore, diffuse sound recorded with coincident microphones or generated using a stereo reverberator generally does not have the frequency-dependent interaural coherence that a binaural recording of diffuse sound would have. A coincident stereo microphone recording of diffuse sound has an interaural coherence that does not depend on the frequency (assuming that the directional responses of the microphones do not depend on frequency), while the theoretical interaural coherence of a binaural recording follows a sinc-function-like curve [Cook et al., 1955]. Since it was shown [Menzer and Faller, 2009b] that the frequency-dependent interaural coherence is very important for the perception of a natural reverberation, one may conclude that the interaural coherence of the diffuse sound is another unnatural perceptual cue in coincident stereo recordings and artificially reverberated stereo mixes.

When playing back stereo signals with headphones it is desirable to provide the listener with natural binaural cues, rather than with the unnatural cues contained in the original stereo signal optimized for loudspeaker playback. One possible way to achieve this goal is to simulate a stereo setup in a good listening room by applying the appropriate binaural room impulse responses (BRIRs) to the left and right channel of

the stereo signal. This approach is simple, provides realistic binaural cues and does not introduce nonlinear artifacts into the signal. However, simulating a particular listening environment using BRIRs inevitably changes certain aspects of the signal. In particular, since most stereo signals contain reverberation of some sort, adding the reverberation of the BRIRs used for the rendering will lead to an unnatural reverberation composed of a reverberant signal played back in a reverberant environment. Applying BRIRs will increase the reverberation time and will add early reflections not present in the original signal. This changes the timbre of the original recording and can lead to undesired situations when for example a recording made in a very small room is played back using BRIRs recorded in a big listening room or vice-versa.

It may be argued that with a real stereo playback system comes always the effect of the room in which the loudspeakers are placed and that it may be desirable to simulate this effect also with headphone playback. However, with an actual stereo setup, the listener knows in which room he or she is and has a certain expectation of how the playback will sound. With headphones however, the listener is likely to be used to simple stereo playback, so the expectation is rather that there is *no* additional reverberation. Furthermore, with a simulated stereo setup, the listener has no way of gathering prior knowledge about the room and he or she does not know to which extent the reverberation is due to the recording itself or to the simulated room.

It is therefore preferable to be able to add natural binaural cues to a stereo recording without adding the effect of a particular listening environment. In this chapter a method is presented that separates a stereo signal into a coherent and a diffuse part and adds natural interaural cues to both parts by changing the original signals as little as possible, in particular without changing reverberation times or adding early reflections.

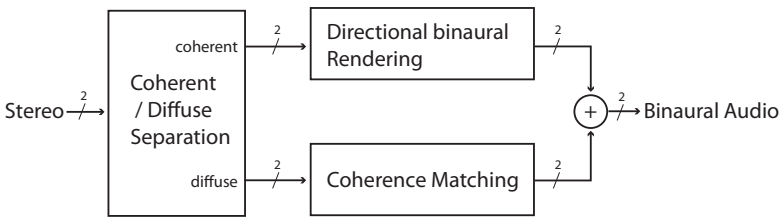

Figure 5.1: Schematic representation of the proposed stereo to binaural conversion method.

Figure 5.1 shows the schema of the proposed method. A stereo signal is separated into coherent and diffuse parts, to which binaural cues are added by a directional binaural rendering (for the coherent part) and coherence matching (for the diffuse part). The rendered signals are added together and can be played to the listener using headphones.

A similar method has been proposed before [Goodwin and Jot, 2007b]. This method does not implement explicit coherence matching of the diffuse sound, but

proposes the use of an (unspecified) decorrelation filter.

In order to precisely match the coherence of the diffuse part using the frequency-dependent coherence synthesis method presented in Section 2.3, it is important to be able to start from two signals that are decorrelated (i.e. have an interchannel coherence close to zero). Therefore, in the design of the proposed stereo to binaural conversion method, particular attention was paid to the separation of the original stereo signal into coherent and diffuse parts. Since other separation methods do not guarantee decorrelated channels, a custom separation algorithm was developed that guarantees the diffuse signal (in sum/difference representation) to be decorrelated in every frequency band. As the decorrelation comes from the design of the separation algorithm, there is no need for using decorrelators (which would change the reverberation time). An example of a separation algorithm that is not suitable for the proposed diffuse sound rendering can be found in [Breebaart and Schuijers, 2008], where the diffuse sound is extracted as an out of phase residual, i.e. a signal that has an inter-channel coherence of -1. The algorithm presented in [Faller, 2006] was also taken into consideration, but finally a new algorithm was developed based on a coincident stereo microphone signal model, and was specifically designed to produce signals suitable for the binaural rendering of the coherent and diffuse sound. Further alternatives for the separation of a stereo signal into coherent and diffuse parts are presented in [Goodwin and Jot, 2007a] and [Merimaa et al., 2007].

For the rendering of the coherent sound two different methods were examined, a simple method where HRTFs are applied to the left and right coherent signals (simulating a stereo loudspeaker setup in anechoic conditions), the other one estimating the directions of the sound sources in a time-frequency representation and applying HRTFs for each estimated direction. Both methods were compared for their advantages and disadvantages.

This Chapter is organized as follows: Section 5.2 introduces the signal model and describes the separation of the stereo signal into coherent and diffuse parts while Section 5.3 shows how binaural cues are added to the coherent and diffuse parts. Section 5.4 presents an application of the proposed method to a stereo music signal and Section 5.5 discusses the results and draws conclusions.

5.2 Separation of a stereo signal into coherent and diffuse parts

5.2.1 Signal model

In the following, a stereo signal $(s_l(n), s_r(n))$ is considered and a method for separating it into a coherent signal $(s_{l,\mathrm{coh}}(n), s_{r,\mathrm{coh}}(n))$ and a diffuse signal $(s_{l,\mathrm{dif}}(n), s_{r,\mathrm{dif}}(n))$ is proposed. The separation method is based on vector geometry, i.e. the signals (or parts of them) are considered as vectors in an N-dimensional space. To distinguish the samples of a signal from the vector representation of the same signal, the sample will always be noted with the time index, i.e. as $x(n)$ and the vector representation without, i.e. as x. The scalar product between two vectors x and y is written as $\langle x\,,\, y\rangle$ and the norm of a vector x as $\|x\|$.

The signal model is based on the assumption that there is only one sound source

at a fixed position. While this is obviously not true for a stereo signal in general, it is assumed to be true for a short timeframe in a single critical band.

The original stereo signal is assumed to be the sum of the coherent and the diffuse signals:

$$s_l(n) = s_{l,\mathrm{coh}}(n) + s_{l,\mathrm{dif}}(n) \tag{5.1}$$

$$s_r(n) = s_{r,\mathrm{coh}}(n) + s_{r,\mathrm{dif}}(n)\ , \tag{5.2}$$

where the coherent part is assumed to contain only a single signal distributed to the two channels by amplitude panning

$$s_{l,\mathrm{coh}}(n) = \alpha_l c(n) \tag{5.3}$$

$$s_{r,\mathrm{coh}}(n) = \alpha_r c(n)\ , \tag{5.4}$$

and where the diffuse part is assumed to be composed of 3 independent signals: one that only appears in the left channel, one that appears only in the right channel, and one that is common to both channels:

$$s_{l,\mathrm{dif}}(n) = d_l(n) + d_c(n) \tag{5.5}$$

$$s_{r,\mathrm{dif}}(n) = d_r(n) + d_c(n) \tag{5.6}$$

Furthermore it is assumed that the power of d_l and d_r is equal:

$$\|d_l\|^2 = \|d_r\|^2\ . \tag{5.7}$$

In practice this will hold only on long timescales, so the precise formulation of this property is

$$\lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} d_l(n) d_l(n) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} d_r(n) d_r(n)\ . \tag{5.8}$$

The signals d_l, d_r and d_c are assumed to be orthogonal to each other:

$$d_l \perp d_r \perp d_c \perp d_l\ , \tag{5.9}$$

which will also hold only on long timescales and can be precisely defined as:

$$\lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} d_l(n) d_r(n) = 0 \tag{5.10}$$

$$\lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} d_r(n) d_c(n) = 0 \tag{5.11}$$

$$\lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} d_c(n) d_l(n) = 0\ . \tag{5.12}$$

Note that because orthonormal transforms preserve the inner product, the orthogonality of d_l, d_r and d_c is preserved under an orthonormal transform $\mathcal{T}\{\}$, i.e.

$$\mathcal{T}\{d_l\} \perp \mathcal{T}\{d_r\} \perp \mathcal{T}\{d_c\} \perp \mathcal{T}\{d_l\} \tag{5.13}$$

This means that the orthogonality property can be used also in a transformed domain as long as an orthonormal transform is used (e.g. DFT, MDCT, or an STFT with non-overlapping rectangular windows). Problems could in principle arise when a non-orthonormal transform is used (e.g. STFT with overlapping windows). However, in practice the method described in this chapter was actually implemented using an STFT with overlapping windows and produced convincing results.

Furthermore we may assume that also the coherent signal $c(n)$ is orthogonal to all diffuse signals: $c \perp d_l$, $c \perp d_r$, and $c \perp d_c$.

In order to be able to separate the coherent from the diffuse signals, it is favorable to consider the sum and the difference of the two stereo channels:

$$\begin{aligned} s_+(n) &= s_l(n) + s_r(n) &&= (\alpha_l + \alpha_r)c(n) + d_l(n) + d_r(n) + 2d_c(n) && (5.14)\\ s_-(n) &= s_l(n) - s_r(n) &&= (\alpha_l - \alpha_r)c(n) + d_l(n) - d_r(n) && (5.15) \end{aligned}$$

in other terms,

$$\begin{aligned} s_+(n) &= \alpha_+ c(n) + d_+(n) && (5.16)\\ s_-(n) &= \alpha_- c(n) + d_-(n) && (5.17) \end{aligned}$$

where

$$\begin{aligned} \alpha_+ &= (\alpha_l + \alpha_r) && (5.18)\\ \alpha_- &= (\alpha_l - \alpha_r) && (5.19)\\ d_+(n) &= d_l(n) + d_r(n) + 2d_c(n) && (5.20)\\ d_-(n) &= d_l(n) - d_r(n)\ . && (5.21) \end{aligned}$$

It can be shown easily that $d_+ \perp d_-$:

$$\begin{aligned} \langle\, d_+ \,,\, d_- \,\rangle &= \langle\, d_l + d_r + 2d_c \,,\, d_l - d_r \,\rangle && (5.22)\\ &\overset{d_l \perp d_c \perp d_r}{=} \langle\, d_l + d_r \,,\, d_l - d_r \,\rangle && (5.23)\\ &\overset{d_l \perp d_r}{=} \langle\, d_l \,,\, d_l \,\rangle - \langle\, d_r \,,\, d_r \,\rangle && (5.24)\\ &\overset{\|d_l\|^2 = \|d_r\|^2}{=} 0 && (5.25) \end{aligned}$$

Furthermore, because c is orthogonal to d_l, d_r, and d_c, we can assure also that

$$d_+ \perp c \perp d_- \qquad (5.26)$$

5.2.2 Separation algorithm

The proposed separation is based on the sum and difference signals defined in the previous section

$$\begin{aligned} s_+(n) &= s_l(n) + s_r(n) &&= \alpha_+ c(n) + d_+(n) && (5.27)\\ s_-(n) &= s_l(n) - s_r(n) &&= \alpha_- c(n) + d_-(n) && (5.28) \end{aligned}$$

and the orthogonality properties

$$d_+ \perp c \perp d_- \perp d_+ \,. \tag{5.29}$$

So the task of separating coherent and diffuse sound has become the task of approximating $\alpha_+ c(n)$, $\alpha_- c(n)$, $d_+(n)$ and $d_-(n)$ given $s_+(n)$ and $s_-(n)$.

This signal model has only coherent sound from a single direction. In practice, this is of course not true. However, it is possible to make the simplifying assumption that in a time-frequency representation, for any given timeframe and any given critical band, there is coherent sound only from one direction, an assumption also underlying the cue selection model presented by [Faller and Merimaa, 2004]. So while the separation is entirely based on vector geometry and could be performed directly on time-domain signals, in practice it should be performed on the coefficients of a time-frequency transform. In each timeframe, the coefficients for each critical band are grouped together as a vector and processed separately.

In the following the vectors representing one critical band in one timeframe are represented using uppercase letters, i.e. S_+, S_-, etc. The signal model therefore becomes

$$S_+ = \alpha_+ C + D_+ \tag{5.30}$$
$$S_- = \alpha_- C + D_- \tag{5.31}$$

with the orthogonality property

$$D_+ \perp C \perp D_- \perp D_+ \,. \tag{5.32}$$

It may seem strange to produce three orthogonal signals from only two input signals. However, this is similar to what a decorrelator does (i.e. producing N orthogonal signals from only one input signal). The difference is that the method presented here aims at directly producing decorrelated signals without adding a reverberation tail. But similarly to a decorrelator based on linear filters, where choices must be made for the impulse responses, in this problem there is not only one solution: since we are given two vectors and we want to obtain three orthogonal vectors, this is an underdetermined problem with an infinity of valid solutions. It is therefore impossible to solve this problem without making some assumptions.

One may notice that the separation problem may be reduced to the problem of finding an orthogonal basis $\{e_C, e_+, e_-\}$ defining the directions of the coherent part C and the diffuse parts D_+ and D_-, respectively, and then projecting S_+ onto e_C and e_+ to obtain $\alpha_+ C$ and D_+, respectively and S_- onto e_C and e_- to obtain $\alpha_- C$ and D_-, respectively. Since S_+ must lie in the plane spanned by e_C and e_+ and S_- must lie in the plane spanned by e_C and e_-, given e_C, S_+ and S_- it is possible to find e_+ and e_- by performing the first step of the Gram-Schmidt orthonormalization. So the only time where an assumption is necessary is when e_C is defined. This may seem a very limited amount of freedom, but for vectors of length N there are actually $N-2$ degrees of freedom in defining e_C.

The only assumption made in the proposed algorithm is that e_C should be as close as possible to the sum or the difference of S_+ and S_-, whichever is bigger. Note that $S_+ + S_- = 2S_l$ and $S_+ - S_- = 2S_r$. Using S_l or S_r as a first guess for the coherent

sound is an easily justifiable choice: since we assume that the diffuse sound is present with the same energy in the left and the right channel, the channel with more energy is expected to be the channel with more coherent sound.

The complete algorithm to separate S_+ and S_- into coherent and diffuse parts is currently implemented as a recursive random search. First, $\widetilde{S_-}$ is defined as either S_- or $-S_-$ in order to maximize $S_+ + S_-$:

$$\widetilde{S_-} = \begin{cases} S_- & \text{if } \|S_+ + S_-\| \geq \|S_+ - S_-\| \\ -S_- & \text{otherwise} \end{cases} . \tag{5.33}$$

In the following, an iterative method for generating an orthogonal basis $\{e_C, e_+, e_-\}$ is presented. A starting point for the coherent basis vector $e_C^{(0)}$ is calculated as

$$e_C^{(0)} = \frac{S_+ + \widetilde{S_-}}{\left\| S_+ + \widetilde{S_-} \right\|} . \tag{5.34}$$

In each step of iteration (the step being identified by u in the following), the diffuse basis vectors $e_+^{(u)}$ and $e_-^{(u)}$, are calculated such that $e_+^{(u)}$ and $e_C^{(u)}$ span a plane in which S_+ lies and $e_-^{(u)}$ and $e_C^{(u)}$ span a plane in which $\widetilde{S_-}$ lies:

$$e_+^{(u)} = \frac{S_+ - \left\langle e_C^{(u)} , S_+ \right\rangle e_C^{(u)}}{\left\| S_+ - \left\langle e_C^{(u)} , S_+ \right\rangle e_C^{(u)} \right\|} \tag{5.35}$$

$$e_-^{(u)} = \frac{\widetilde{S_-} - \left\langle e_C^{(u)} , \widetilde{S_-} \right\rangle e_C^{(u)}}{\left\| \widetilde{S_-} - \left\langle e_C^{(u)} , \widetilde{S_-} \right\rangle e_C^{(u)} \right\|} . \tag{5.36}$$

The goal of the iterative algorithm is to find an e_C as close as possible to $e_C^{(0)}$ such that $\{e_C, e_+, e_-\}$ form an orthonormal basis. Because by design, e_+ and e_- will be orthogonal to e_C, to measure the quality of the basis, only $\langle e_+ , e_- \rangle$ needs to be evaluated and must be reasonably close to 0 to stop the iteration.

At each iteration, N random unit vectors $\widehat{e_C}_v^{(u+1)}$ ($v \in \{1, \ldots, N\}$) close to $e_C^{(u)}$ are generated and the resulting values $\widehat{q}_v^{(u+1)} = \left| \left\langle \widehat{e_+}_v^{(u+1)} , \widehat{e_-}_v^{(u+1)} \right\rangle \right|$ are calculated. The vector $\widehat{e_C}_{v'}^{(u+1)}$ that produces the smallest value $\widehat{q}_{v'}^{(u+1)}$ is chosen to become $e_C^{(u+1)}$.

The iteration is continued until $\left| \left\langle e_+^{(u+1)} , e_-^{(u+1)} \right\rangle \right|$ drops below a given threshold (e.g. 10^{-3}). N is chosen as a function of the dimension of S_+ and S_-, normally as 10 times the number of elements of these vectors.

We can assume that after M iterations an acceptable basis

$$\{e_C, e_+, e_-\} = \left\{ e_C^{(M)}, e_+^{(M)}, e_-^{(M)} \right\}$$

has been found. Therefore we can obtain the vectors

$$\alpha_+ C = \langle e_C , S_+ \rangle e_C \tag{5.37}$$
$$\alpha_- C = \langle e_C , S_- \rangle e_C \tag{5.38}$$
$$D_+ = \langle e_+ , S_+ \rangle e_+ \tag{5.39}$$
$$D_- = \langle e_- , S_- \rangle e_- , \tag{5.40}$$

which allows to obtain the coherent vectors for the left and right channel $S_{l,\mathrm{coh}}$ and $S_{r,\mathrm{coh}}$ as well as the diffuse vectors $S_{l,\mathrm{dif}}$ and $S_{r,\mathrm{dif}}$:

$$S_{l,\mathrm{coh}} = \frac{1}{2}(\alpha_+ C + \alpha_- C) \tag{5.41}$$
$$S_{r,\mathrm{coh}} = \frac{1}{2}(\alpha_+ C - \alpha_- C) \tag{5.42}$$
$$S_{l,\mathrm{dif}} = \frac{1}{2}(D_+ + D_-) \tag{5.43}$$
$$S_{r,\mathrm{dif}} = \frac{1}{2}(D_+ - D_-) \ . \tag{5.44}$$

After processing all blocks in the transformed domain, the corresponding time-domain signals $s_{l,\mathrm{coh}}(n)$, $s_{r,\mathrm{coh}}(n)$, $s_{l,\mathrm{dif}}(n)$, and $s_{r,\mathrm{dif}}(n)$ are calculated using the appropriate inverse transform $\mathcal{T}^{-1}\{\}$.

5.3 Binaural rendering

5.3.1 Rendering the coherent sound

In principle, there are many different possibilities to render a the coherent part of a stereo signal with plausible interaural cues. The most obvious way would be to apply HRTFs to the left and the right channel and therefore to simulate a stereo loudspeaker setup in anechoic conditions. This method is a viable option, but it has three problems: first, using regular HRTFs leads to strong spectral modifications, which is contrary to the goal of changing the stereo signal only as little as necessary. Second, simulating a stereo setup leads to an additional comb filter effect for sources panned to the middle of the sound stage because in a stereo setup each ear receives sound from the left and from the right loudspeaker with different delays. Third, the simulation of a stereo setup limits the maximum angles left and right between which sources can be placed. While a stereo setup with loudspeakers at $\pm 45°$ may still be acceptable, $\pm 90°$ (i.e. one loudspeaker to the left of the listener and one to the right) is definitely not a viable option.

Because in a direct comparison, due to the coloration introduced by HRTFs, most people prefer the original sound over HRTF-processed sound, it is preferable to use diffuse field equalized HRTFs [Larcher et al., 1998]. A coherent sound rendering using just two diffuse-field equalized HRTFs was implemented and the result is discussed in Section 5.4.2. Recent research [Merimaa, 2009] has shown that more extensive equalization of the HRTFs is possible without destroying the localization. Applying such a technique is likely to improve the quality of the HRTF based coherent sound rendering. However, informal listening indicates when using HRTFs for $\pm 60°$ the sources

at 0° sound unnaturally close to the head. On the other hand, in the ±45° case, the sound scene seems to be narrower than in the original signal. It therefore cannot be expected to obtain completely convincing results by simply applying HRTFs, even if advanced equalization techniques are used.

A more sophisticated way of rendering the coherent sound is to estimate the directions of arrival in a time-frequency representation of the signal and to apply the interaural level and time difference cues as well as spectral cues corresponding to the estimated direction. The proposed method corresponds to the one presented in [Breebaart and Schuijers, 2008].

First, the angle of arrival is estimated for each critical band in each time frame of the STFT representation of the left and right channels of the coherent signal. The estimation is based on the energy in the left and the right channel using the following formula:

$$\varphi(i,b) = \varphi_{\max}\left(1 - \frac{2\left\|S_{l,\mathrm{coh}}(i,\vec{k}_b)\right\|}{\left\|S_{l,\mathrm{coh}}(i,\vec{k}_b)\right\| + \left\|S_{r,\mathrm{coh}}(i,\vec{k}_b)\right\|}\right) \tag{5.45}$$

where $\vec{k}_b$ is the frequency bin range corresponding to the critical band b and $S_{l,\mathrm{coh}}(i,\vec{k}_b)$ and $S_{r,\mathrm{coh}}(i,\vec{k}_b)$ denote the vector corresponding to the timeframe i and critical band b of the left and right channels, respectively. If all of the energy is in the left channel, the estimated angle is $-\varphi_{\max}$, if all the energy is in the right channel, the estimated angle is $\varphi_{\max}$ and in between, a linear interpolation is applied. In fact, for a setup of two coincident cardioid microphones pointing at $-45°$ and $+45°$, respectively, the estimation method described above with $\varphi_{\max} = 128.7°$ (value obtained using linear regression) is a good approximation of the actual positions of the sound sources (see Figure 5.2). But because the goal is to reproduce a stereo signal played back with headphones as faithfully as possible, while adding plausible binaural cues, the natural choice is $\varphi_{\max} = 90$: if the signal has energy only in the left channel, it is perceived only at the left ear, i.e. at a position roughly corresponding to $-90°$.

The result of the angle estimation can be seen in Figure 5.3. It may be noticed that the estimation is very noisy. Therefore, to avoid artifacts, it is necessary to de-noise the angle estimation. This can be done very effectively by applying a median filter over three timeframes and three critical bands. The de-noised angle estimation can be seen in Figure 5.4. Informal listening suggests that the median filtering significantly reduces the audible artifacts in the rendered stereo signal.

In order to efficiently implement the application of HRTFs corresponding to the estimated angles, the effect of applying the HRTFs was simulated directly in the STFT representation of the signals. It has been shown that the main perceptual cues of HRTFs can be expressed by a simple delay and amplitude factors for the left and the right channel for each critical band [Breebaart and Kohlrausch, 2001a]. For each block in the STFT domain for which the angle of arrival was estimated, it is sufficient to look up the amplitude and the delay in a table and to apply both directly in the STFT domain:

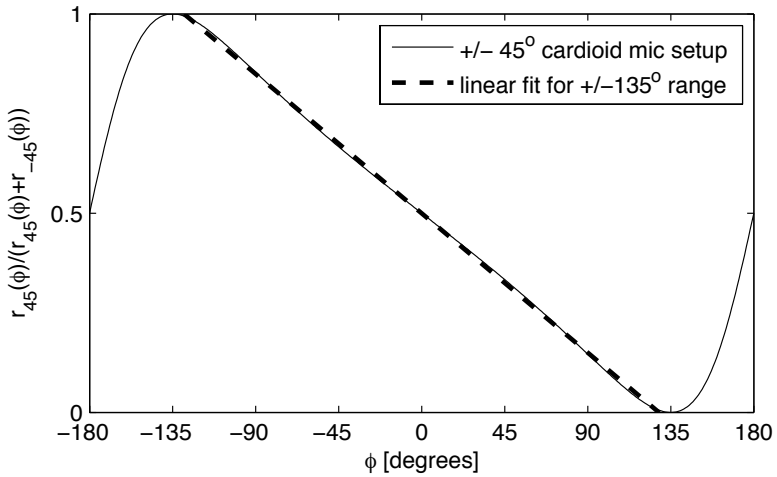

Figure 5.2: Proportion of left directional response of a coincident cardioid microphone setup as a function of angle and linear fit between $-135°$ and $135°$.

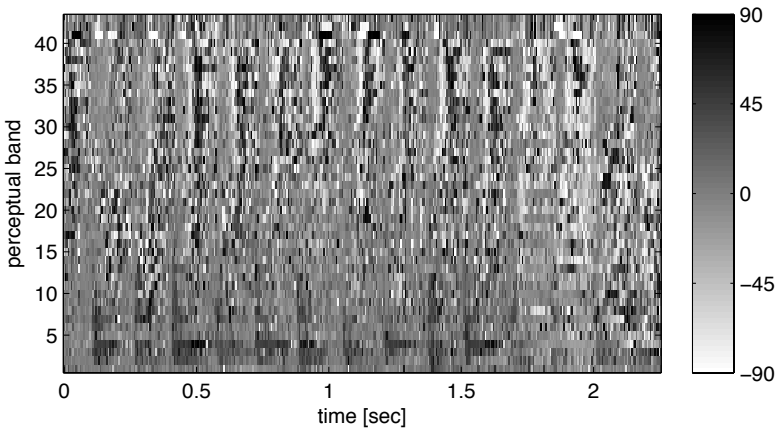

Figure 5.3: Angles estimated from coherent sound before de-noising.

$$C_l(i,k) = S_{l,\mathrm{coh}}(i,k)e^{-jk\pi\tau(\varphi(i,b))/N}A_l(b,\varphi(i,b))\frac{\left\|S_{l,\mathrm{coh}}(i,\vec{k}_b)\right\| + \left\|S_{r,\mathrm{coh}}(i,\vec{k}_b)\right\|}{2\left\|S_{l,\mathrm{coh}}(i,\vec{k}_b)\right\|} \tag{5.46}$$

$$C_r(i,k) = S_{r,\mathrm{coh}}(i,k)e^{jk\pi\tau(\varphi(i,b))/N}A_r(b,\varphi(i,b))\frac{\left\|S_{l,\mathrm{coh}}(i,\vec{k}_b)\right\| + \left\|S_{r,\mathrm{coh}}(i,\vec{k}_b)\right\|}{2\left\|S_{r,\mathrm{coh}}(i,\vec{k}_b)\right\|}$$

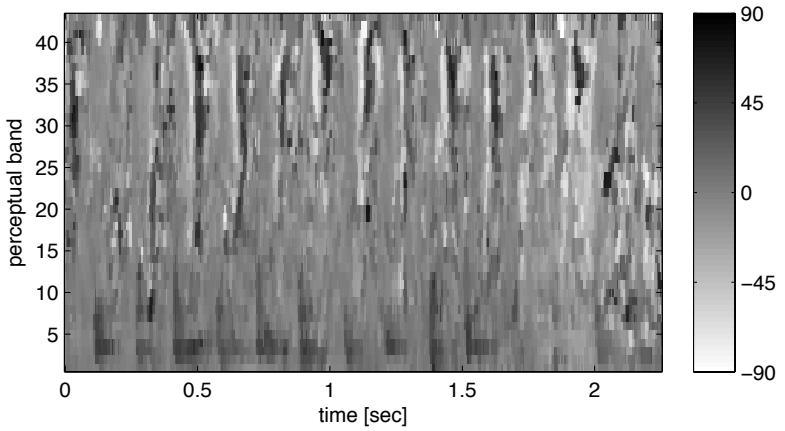

Figure 5.4: Angles estimated from coherent sound after de-noising using a median filter.

where $\tau(\varphi(i,b))$ is the interaural time difference corresponding to the estimated angle, N is the FFT size, and $A_l(b,\varphi(i,b))$ and $A_r(b,\varphi(i,b))$ are the amplitude factors extracted from the left and right HRTFs, respectively, evaluated for the critical band b and the angle $\varphi(i,b)$. Notice that the terms $e^{-jk\pi\tau(\varphi(i,b))/N}$ and $e^{jk\pi\tau(\varphi(i,b))/N}$ correspond to (circular) delays by $\tau(\varphi(i,b))/2$ and $-\tau(\varphi(i,b))/2$, respectively. This means that one half of the ITD measured from the HRTF set is applied to the left channel and the other half is applied to the right channel. The ITD as a function of the azimuth angle can be seen in Figure 5.6.

A_l is shown in Figure 5.5. The data for A_l and A_r was obtained from the diffuse field equalized version of the MIT KEMAR HRTF set [Gardner and Martin, 1994] with an angular resolution of 5° and has been further normalized such that

$$\frac{1}{72}\sum_{x=0}^{71} A_l(b,5x) = 1\ . \tag{5.47}$$

This normalization improved the spectrum of the rendered coherent sound, which can be seen in Figure 5.12. Without the normalization, larger deviations from the original spectrum occur, in particular in the low frequencies.

Finally, the rendered coherent sound can be calculated using the appropriate inverse transform $\mathcal{T}^{-1}\{\}$:

$$c_l(n) = \mathcal{T}^{-1}\{C_l(i,k)\} \tag{5.48}$$
$$c_r(n) = \mathcal{T}^{-1}\{C_r(i,k)\}\ . \tag{5.49}$$

5.3.2 Rendering the diffuse sound

The goal of the diffuse sound processing is to obtain a stereo signal with the same magnitude time-frequency spectrum as the extracted diffuse sound and an interau-

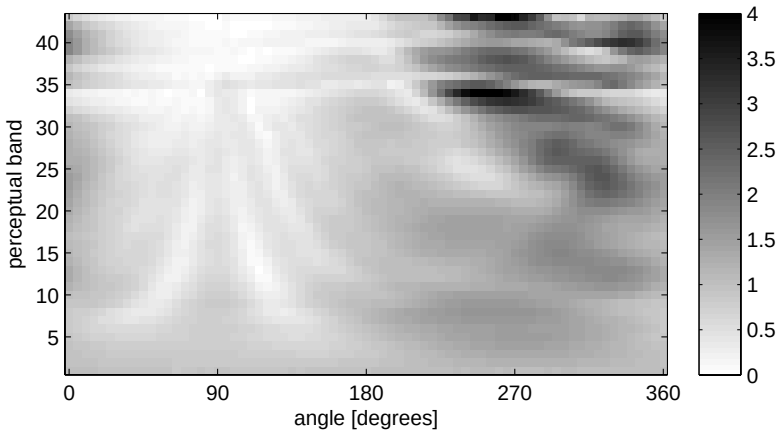

Figure 5.5: Amplitude data extracted from HRTF set. The amplitude data was normalized separately in each critical band in order to obtain a mean of 1. For simplicity, only the left channel is shown. The data for the right channel is identical except for a reversal of the angles.

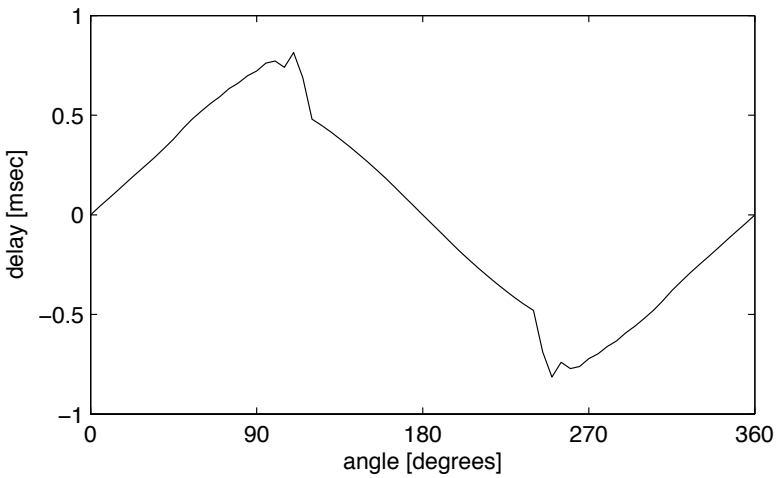

Figure 5.6: Interaural time delay as a function of azimuth angle of the sound source (data extracted from HRTF set).

ral coherence matching a given frequency-dependent interaural coherence curve (e.g. calculated from an HRTF set).

Given two sets of time-frequency coefficient $D_+(i,k)$ and $D_-(i,k)$ and a frequency-

dependent coherence $\Phi(i)$, the same method used to obtain coherence matched noise signals in Section 2.3 of this thesis (Chapter 2) can be applied to obtain two sets of time-frequency coefficients $\widetilde{D_l}(i,k)$ and $\widetilde{D_r}(i,k)$ such that the corresponding time-domain signals $\widetilde{d_l}(n)$ and $\widetilde{d_r}(n)$ have the desired frequency-dependent interaural coherence $\Phi(i)$:

$$\widetilde{D_l}(i,k) = a(i,k)D_+(i,k) + b(i,k)D_-(i,k) \tag{5.50}$$

$$\widetilde{D_r}(i,k) = a(i,k)D_+(i,k) - b(i,k)D_-(i,k) \tag{5.51}$$

where

$$\begin{aligned} a(i,k) &= \sqrt{\frac{P_-(i,k)^2(1+\Phi(i,k))}{P_+(i,k)^2(1-\Phi(i,k)) + P_-(i,k)^2(1+\Phi(i,k))}} \\ b(i,k) &= \sqrt{1-a(i,k)^2} \\ &= \sqrt{\frac{P_+(i,k)^2(1-\Phi(i,k))}{P_+(i,k)^2(1-\Phi(i,k)) + P_-(i,k)^2(1+\Phi(i,k))}}\,, \end{aligned} \tag{5.52}$$

where $P_+(i,k) = \mathcal{S}\{|D_+(i,k)|^2\}$ and $P_-(i,k) = \mathcal{S}\{|D_-(i,k)|^2\}$ and $\mathcal{S}$ is a smoothing operator defined by

$$\mathcal{S}\{H(i,k)\} = \sum_{m=-l}^{l} w(m)H(i,k+m)\,, \tag{5.53}$$

where $w(m)$ is a set of $2l+1$ weights for the moving average.

To ensure that the spectrum of the rendered diffuse sound matches the spectrum of the extracted left and right diffuse sound, a spectral matching is performed:

$$D_l(i,k) = \widetilde{D_l}(i,k)\frac{|\mathcal{S}\{S_{l,\mathrm{dif}}(i,k)\}|}{\left|\mathcal{S}\{\widetilde{D_l}(i,k)\}\right|} \tag{5.54}$$

$$D_r(i,k) = \widetilde{D_r}(i,k)\frac{|\mathcal{S}\{S_{r,\mathrm{dif}}(i,k)\}|}{\left|\mathcal{S}\{\widetilde{D_r}(i,k)\}\right|} \tag{5.55}$$

Finally, the rendered diffuse sound can be calculated using the appropriate inverse transform $\mathcal{T}^{-1}\{\}$:

$$d_l(n) = \mathcal{T}^{-1}\{D_l(i,k)\} \tag{5.56}$$

$$d_r(n) = \mathcal{T}^{-1}\{D_r(i,k)\}\,. \tag{5.57}$$

Finally, the rendered binaural signal can be calculated as

$$r_l(n) = c_l(n) + d_l(n) \tag{5.58}$$

$$r_r(n) = c_r(n) + d_r(n)\,. \tag{5.59}$$

5.4 Results

The algorithms described above were implemented in Matlab and applied to an excerpt of Jean-Michel Jarre's "Oxygène 4".

5.4.1 Analysis of the separated coherent and diffuse signals

The spectrograms of the original signal as well as the extracted coherent and diffuse parts are shown in Figure 5.7, while Figure 5.8 shows the interaural coherence of the original, coherent and diffuse stereo signals as a function of time and frequency calculated using the method described in Equation (2.6) in Chapter 2. It can be seen that, compared to the original signal, the interaural coherence of the coherent signal is increased while the interaural coherence of the diffuse signal is decreased. The same observation can be made quantitatively in Figure 5.9, which shows the interaural coherence of the same signals as a function of frequency only calculated using the method described in Equation (2.4). It can be seen that the interaural coherence of the coherent part is close to 1 above 1.5 kHz and the interaural coherence of the diffuse sound varies between 0 and -0.4.

Finding a negative interaural coherence of the diffuse sound may be surprising since this case was not accommodated for in the signal model in Equations (5.5) and (5.6). However, a negative interaural coherence is not explicitly avoided in the separation algorithm and will occur in the case where

$$\|D_-\|^2 > \|D_+\|^2 \ , \tag{5.60}$$

as can be easily derived by calculating the sign of the scalar product of the vectors representing the left and right diffuse signals:

$$\begin{aligned}
\langle\, S_{l,\mathrm{dif}}\,,\, S_{r,\mathrm{dif}}\,\rangle &= \left\langle \frac{1}{2}(D_+ + D_-)\,,\, \frac{1}{2}(D_+ - D_-) \right\rangle && (5.61)\\
&= \frac{1}{4}\,\langle\, D_+ + D_-\,,\, D_+ - D_-\,\rangle && (5.62)\\
&\overset{D_+ \perp D_-}{=} \frac{1}{4}\left(\langle\, D_+\,,\, D_+\,\rangle - \langle\, D_-\,,\, D_-\,\rangle\right) && (5.63)\\
&= \frac{1}{4}\left(\|D_+\|^2 - \|D_-\|^2\right) && (5.64)\\
&\overset{(5.60)}{<} 0\ . && (5.65)
\end{aligned}$$

In principle the signal model could be easily adapted to accommodate for negative coherence of the diffuse part, too, by introducing, besides the case described in Equations (5.5) and (5.5), a second case:

$$s_{l,\mathrm{dif}}(n) = d_l(n) + d_c(n) \tag{5.66}$$

$$s_{r,\mathrm{dif}}(n) = d_r(n) - d_c(n)\ . \tag{5.67}$$

However, for the rest of the processing, the interaural coherence between $s_{l,\mathrm{dif}}(n)$ and $s_{r,\mathrm{dif}}(n)$ is not relevant since the diffuse sound rendering algorithm only relies on the interaural coherence between $d_+(n)$ and $d_-(n)$ to be 0. Figure 5.10 shows that this is verified in practice with good accuracy at all frequencies.

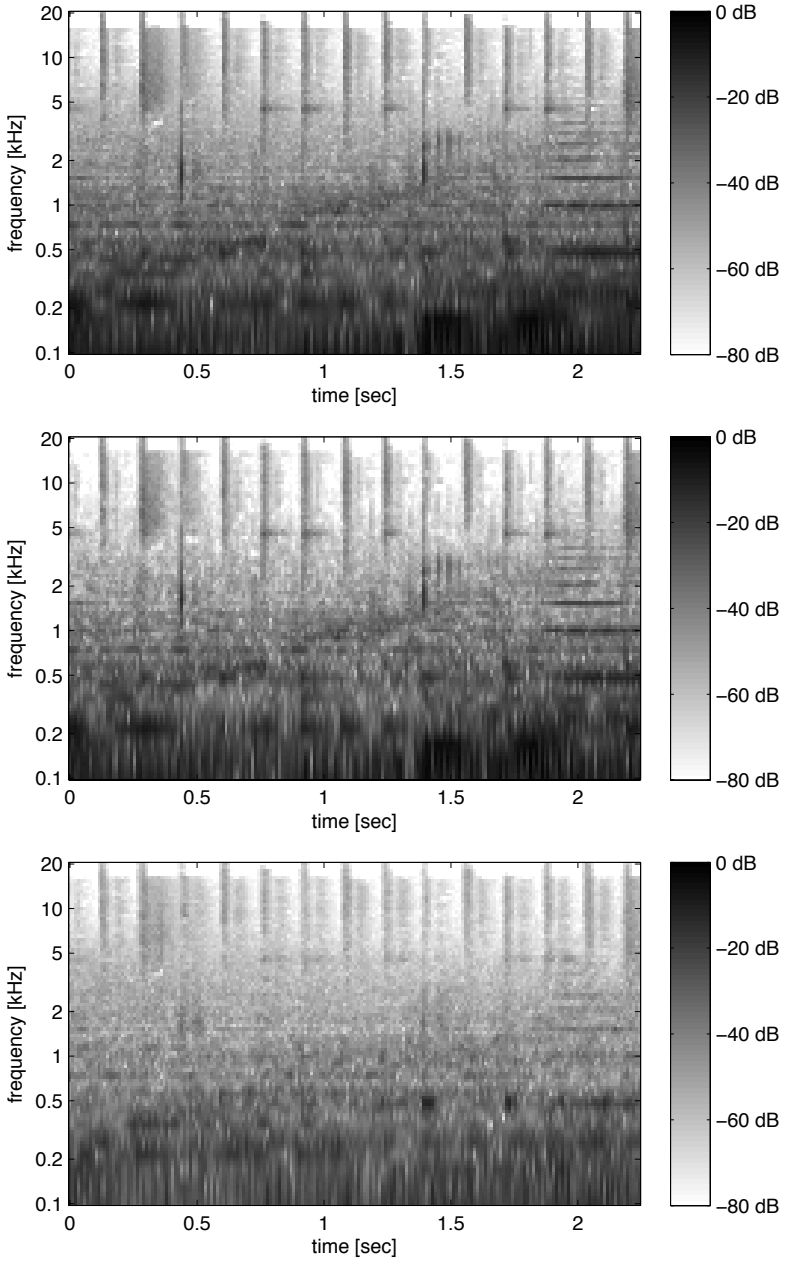

Figure 5.7: Magnitude time-frequency spectral coefficients (only left channel shown). **Top panel**: original stereo signal. **Middle panel**: coherent part. **Bottom panel**: diffuse part.

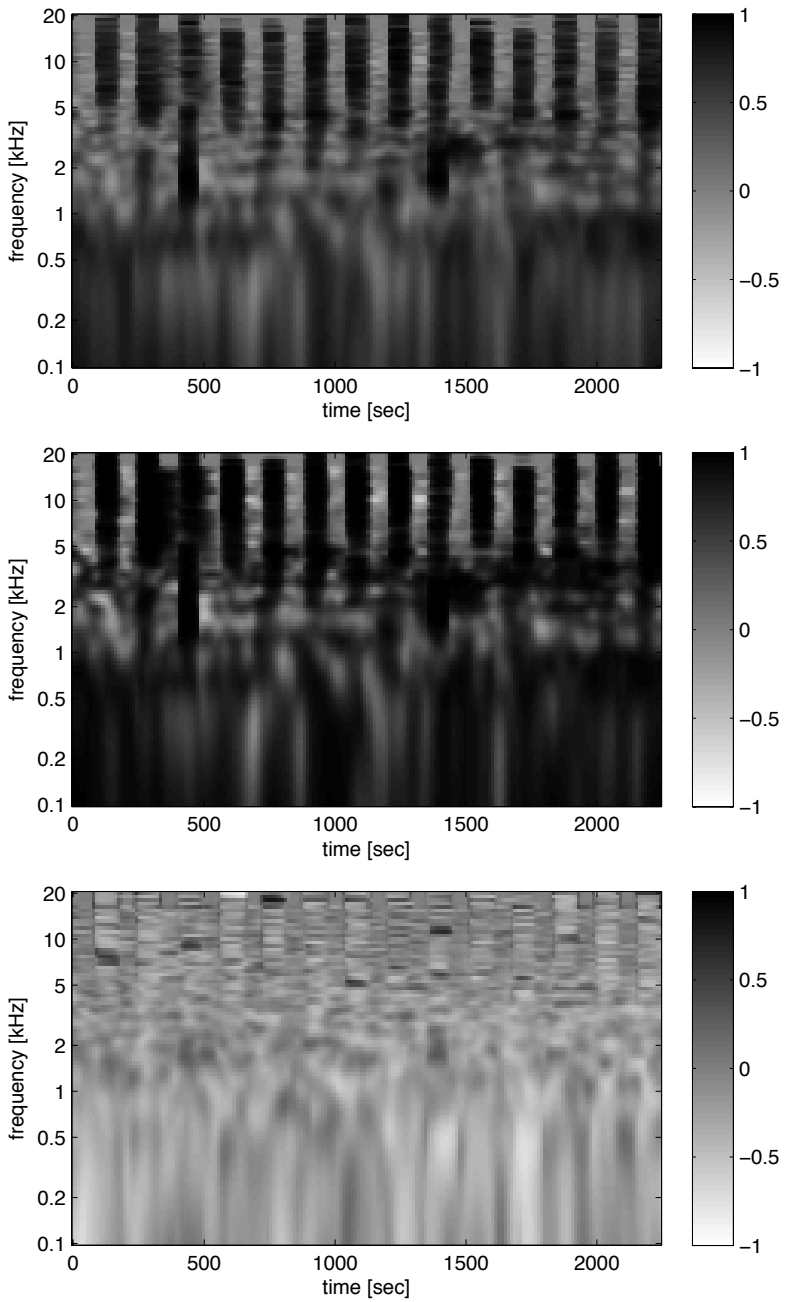

Figure 5.8: Interaural coherence as a function of time and frequency. **Top panel**: original stereo signal. **Middle panel**: coherent part. **Bottom panel**: diffuse part.

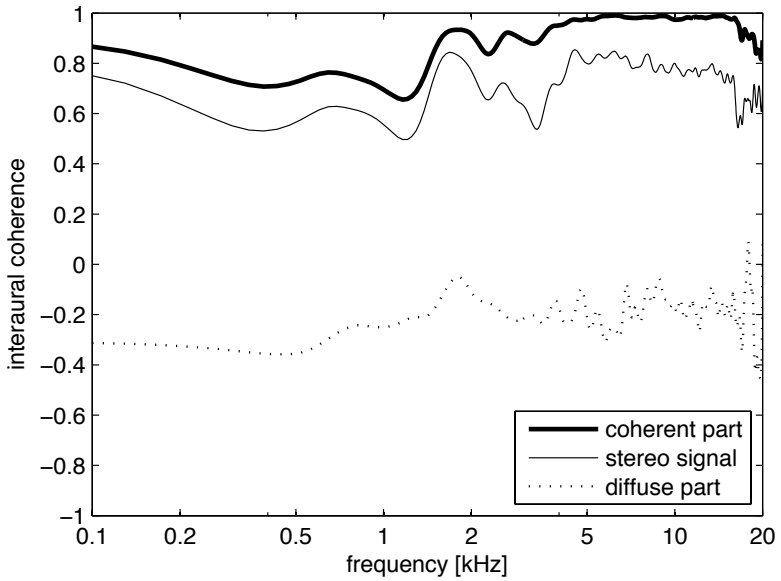

Figure 5.9: Interaural coherence as a function frequency for the original stereo signal and the coherent and diffuse parts.

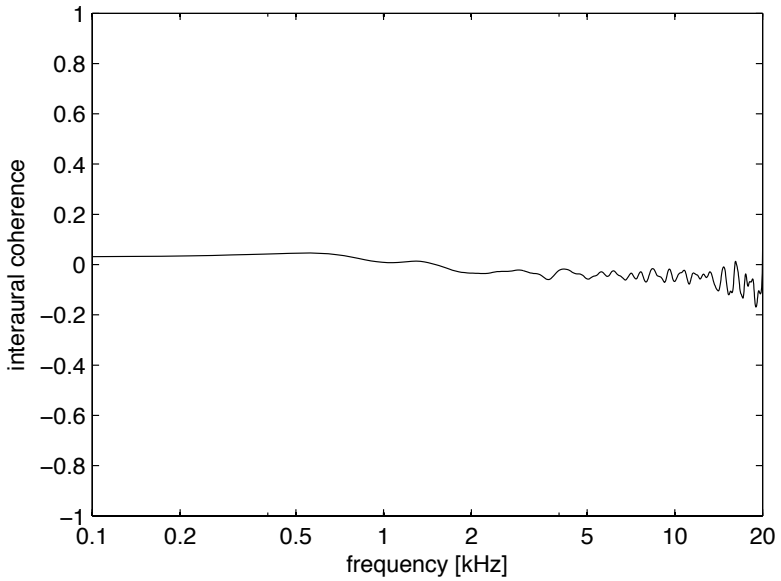

Figure 5.10: Interaural coherence between $d_+(n)$ and $d_-(n)$.

5.4.2 Analysis of the rendered coherent signals

Figure 5.11 shows the spectrum of a coherent signal rendered using diffuse field equalized HRTFs, compared to the spectrum of the original coherent signal. The top panel shows the spectrum obtained by simulating a setup with loudspeakers at $\pm 45°$ and while in the bottom panel a setup with speakers at $\pm 60°$ is simulated. It can be seen that the spectral differences with the original coherent sound become bigger in the $\pm 60°$ case.

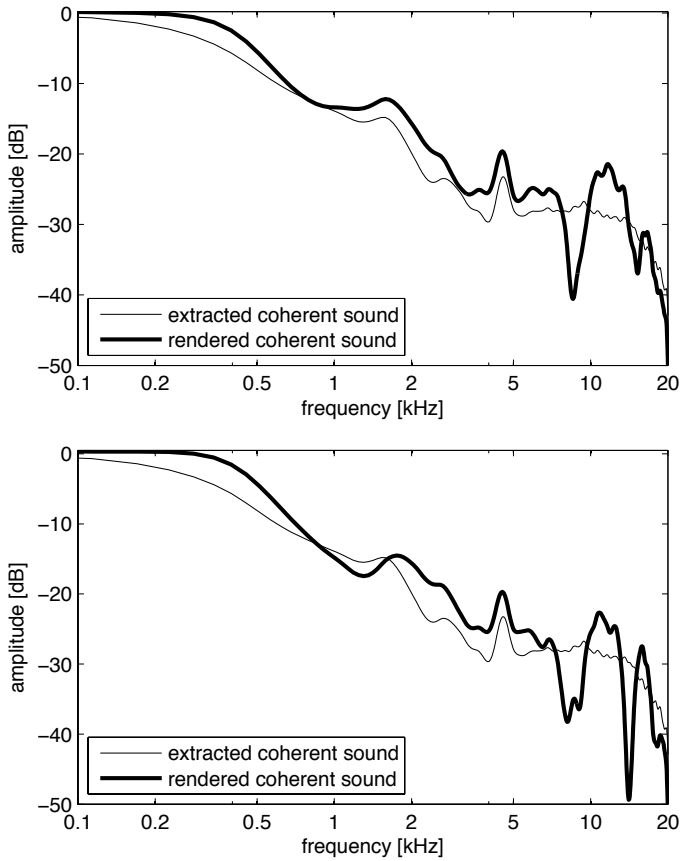

Figure 5.11: Spectrum of extracted and rendered coherent sound using only two diffuse-sound normalized HRTFs (for simplicity, only left channel is shown). **Top panel:** HRTF angles used: $\pm 45°$. **Bottom panel:** HRTF angles used: $\pm 60°$.

Figure 5.12 shows the spectrum of a coherent signal rendered using the more advanced rendering method described Section 5.3.1. By comparing Figure 5.12 to Figure 5.11, it can be observed that the advanced method leads to a reduced coloration

of the rendered signal.

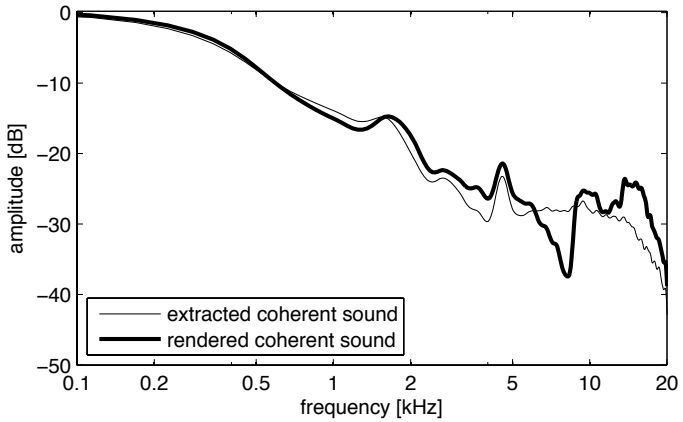

Figure 5.12: Spectrum of extracted and rendered coherent sound using the advanced rendering method (for simplicity, only left channel is shown).

5.4.3 Analysis of the rendered diffuse signal

Figures 5.13 and 5.14 show that the interaural coherence of the rendered diffuse sound roughly matches the interaural coherence calculated for perfect diffuse sound using the CIPIC HRTF set [Algazi et al., 2001] and that the spectral matching with the extracted diffuse sound works very well. The spectral matching is important because a spectral mismatch could make artifacts of the coherent / diffuse sound separation become audible that are not audible when the spectral matching works well (considering that the coherent and diffuse parts are designed to perfectly add up to the original signal).

5.4.4 Analysis of the rendered binaural signal

It is interesting to see what is the total effect of the binaural rendering on the interaural coherence and the spectrum. Since it means evaluating the combined effect of two completely different rendering algorithms, there is no specific expected result, except that the modification of the spectrum should be as small as possible.

Figure 5.15 shows the interaural coherence of the original stereo signal and the rendered binaural signal. It can be seen that the generally high coherence of the stereo signal is slightly increased up to ca. 300 Hz and significantly decreased above ca. 3 kHz.

Figure 5.16 shows the spectrum of the original stereo signal and the rendered binaural signal. By comparing to Figures 5.12 and 5.14 it can be seen that the overall spectral modifications introduced are slightly bigger than those introduced by the coherent and diffuse rendering alone (at least up to 5 kHz). It is probable that the

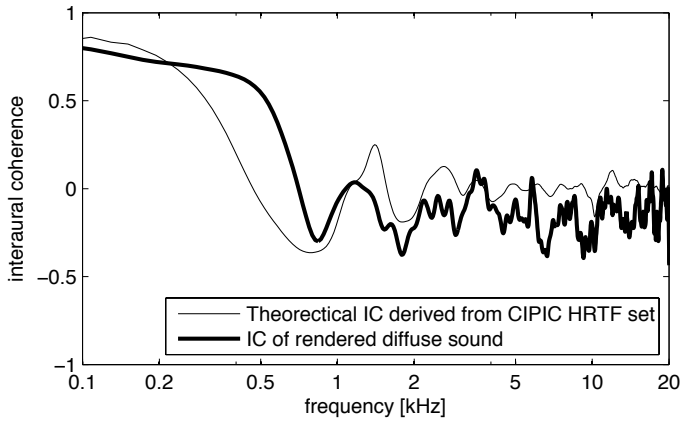

Figure 5.13: Theoretical interaural coherence for perfectly diffuse sound calculated from CIPIC HRTF set and measured interaural coherence for rendered diffuse sound.

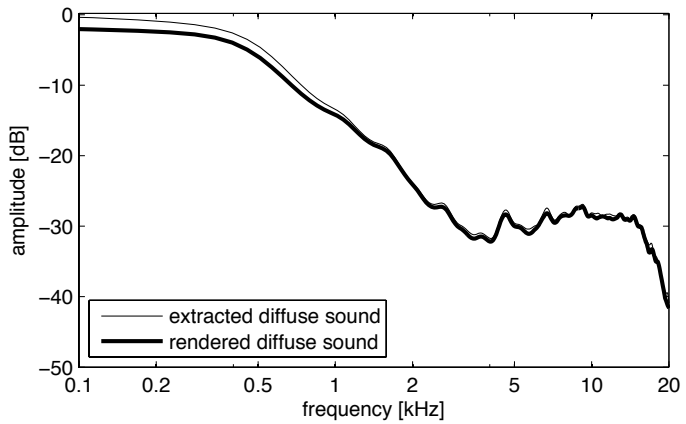

Figure 5.14: Spectrum of extracted and rendered diffuse sound (for simplicity, only left channel is shown).

phase shifts introduced by the coherent sound rendering are the reason for this effect. However, all the differences of the original and the rendered spectra are below 10 dB, which is probably acceptable as long as these differences are perceptually meaningful (i.e. caused by the use of HRTFs).

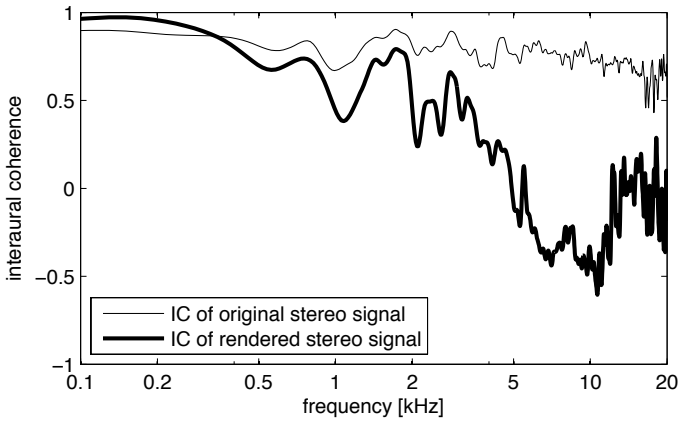

Figure 5.15: Interaural coherence of the original stereo signal and the rendered binaural signal.

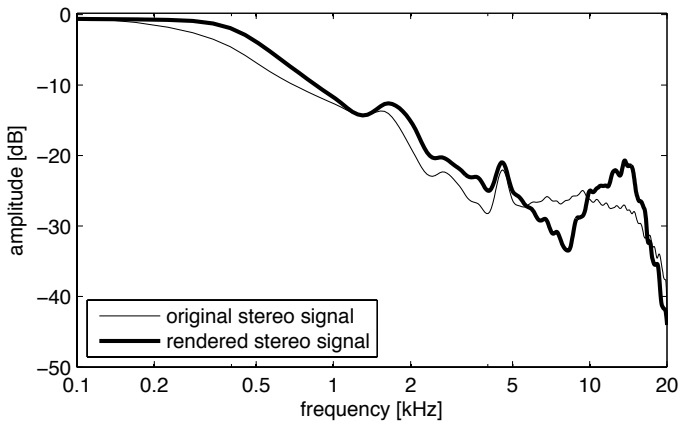

Figure 5.16: Spectrum of the original stereo signal and the rendered binaural signal (for simplicity, only left channel is shown).

The rendered binaural signal was compared to the original stereo signal in an informal listening test using headphone playback. Even though the perceptual differences between the two signals are – by design – relatively subtle and some listeners had problems distinguishing the two versions, most listeners reported that switching from the rendered version, there "seems to be something missing" in the signal. In may be concluded that the rendered signal is perceived as more natural than the original stereo signal.

5.5 Conclusions

When playing back stereo signals using headphones, unnatural binaural cues are presented to the listener. Binaural rendering of stereo signals allows to add realistic binaural cues to a stereo signal and to externalize the perceived sound sources. However, rendering stereo signals using head related transfer functions (HRTFs) yields often only limited spatial impression and externalization. When binaural room impulse responses (BRIRs) are used as opposed to HRTFs, spatial impression and externalization are usually improved. But the drawback of using BRIRs is that the spatial impression of the specific room corresponding to the BRIRs is imposed on the stereo signal. The stereo signal's inherent room related cues, such as reverberation time, are changed. For different music styles and different tastes different BRIRs may be required for optimal results.

A technique was proposed for rendering stereo signals with binaural cues, avoiding the drawbacks of imposing a specific room on the stereo recording, but with the other advantages associated with the use of BRIRs versus HRTFs. The stereo signal is converted into a coherent (dry) stereo signal and a diffuse stereo signal. An HRTF-based rendering is applied to the direct sound stereo signal. The ambient stereo signal is processed such that the frequency-dependent coherence of the ambient sound mimics a listener in a diffuse sound field, i.e. the ambience is rendered with physically motivated binaural cues.

While in the literature various techniques can be found for separating stereo signals into direct and ambient signals, also here such a technique is proposed. This technique has the feature that when direct and ambient signals are added one obtains the original stereo signal. Furthermore, the sum and difference ambient channels are orthogonal, making it possible to implement frequency-dependent coherence matching with linear time-independent filters.

Chapter 6

Conclusions

6.1 Thesis summary

In this thesis perceptual properties of binaural room impulse responses (BRIRs) were studied and applications thereof were developed. Different methods for modeling BRIRs were investigated in Chapter 2 and the interaural coherence as a function of frequency was found to be the perceptually most relevant interaural property of BRIR tails. Methods for analyzing the interaural coherence of audio signals and for synthesizing signals with a given time- and frequency-dependent interaural coherence were developed. Furthermore, it was shown that the head orientation of the listener in a room with a fixed sound source influences the interaural coherence only in the early part of the BRIR, but not in the late part. Two subjective tests proved the perceptual relevance of the time- and frequency-dependence of interaural coherence. Algorithms making use of these insights were presented in the subsequent chapters.

A method for generating BRIRs from B-format room impulse responses (RIRs) and a set of head-related transfer function (HRTFs) was proposed in Chapter 3. The proposed method allows to measure the room related properties and head related properties of BRIRs separately, reducing significantly the amount of measurements necessary for obtaining BRIRs for different rooms and different persons. A novel feature of the proposed method is that a BRIR with correct spectral and coherence cues is obtained using a linear, frequency-dependent decoding of the B-format RIR. A subjective test indicated that the computed BRIRs are perceptually similar to corresponding directly measured BRIRs.

In Chapter 4, two efficient binaural reverberators were proposed, reproducing the main perceptual cues of binaural reverberation. The first reverberator is a simple, computationally efficient extension a Jot reverberator, implementing coherence matching. The second reverberator simulates also early reflections with binaural cues. This is achieved by adding a feedback delay network for the early reflections in parallel to a diffuse reverberator having the same structure as the previously introduced simple binaural reverberator. The early reflections reverberator is designed to reproduce the first and second order reflections correctly and inherently also produces an infinite number of higher-order reflections, enabling a natural transition from early reflections to late reverberation. The impulse response of the combined reverberator

has an interaural coherence that closely resembles the interaural coherence of a BRIR in time-frequency domain.

A method for adding realistic binaural cues to a stereo signal was presented in Chapter 5, having the property of maintaining all the other cues such as direct to reverberant sound ratio, reverberation time, and early reflections. The stereo signal is first separated into coherent and diffuse parts which are rendered separately by adding the perceptually relevant binaural cues to each part while maintaining the spectral and reverberation related cues as far as possible. Informal listening suggests that stereo signals rendered using this method are perceived by many listeners as more natural than the original stereo signals.

Different methods for designing efficient unitary mixing matrices for Jot reverberators were studied in Appendix A, with a particular emphasis on applications to diffuse reverberation and decorrelation. A tradeoff between effective mixing among channels and the number of multiply operations per channel and output sample was found and efficient solutions for different scenarios were derived.

6.2 Potential applications

While the insight on the perceptual relevance of frequency dependent interaural coherence may benefit many binaural audio algorithms, two methods with a potential for commercial applications have been proposed and studied in detail in this thesis: a novel structure for binaural reverberators and a technique for adding binaural cues to stereo signals.

Binaural reverberation can be applied in various fields, such as video games, telecommunications (e.g. teleconferencing), music production, movie scoring, and even in cognitive neuroscience research [Menzer et al., 2010]. In this thesis, two methods have been proposed, suitable for delivering high quality binaural reverberation at a low computational complexity. These methods may allow to improve the quality of binaural audio in interactive applications with constraints on computational complexity, such as video games or teleconferencing applications.

Stereo to binaural conversion has a potential mainly in mobile music and video players (including mobile telephones with such playback capabilities). Since most music and movie sound tracks exist in a stereo version but not in a binaural version, adding binaural cues to the stereo signal in real time during playback is desirable. This thesis describes a method for performing such a stereo-to-binaural conversion taking into account the different perceptual cues relevant for direct sound and for diffuse sound.

The research on efficient mixing matrices presented in Appendix A is particularly suitable for the design of high-quality decorrelators and may lead to a novel approach to multi-channel decorrelation, with applications to rendering or simulating diffuse sound in multi-channel loudspeaker systems.

6.3 Further research questions

While all methods presented in this thesis were proven to work with real-life input signals, further work is needed in order to obtain algorithms suitable for commer-

cial applications. Furthermore, there may be potential applications of the presented methods to fields not yet explored. For example, the method for separating coherent from diffuse sound from Chapter 5 may also be applicable to signals other than stereo signals, which may lead to a technique for B-format to binaural conversion.

The binaural reverberators presented in Chapter 4 should be implemented in real time to prove their computational efficiency in practice and the rendering of realistic scenarios is necessary in order to verify their suitability for dynamic 3D audio applications.

Appendix A

Sparse Unitary Matrices for Diffuse Jot Reverberators

A.1 Introduction

In 1991, Jot and Chaigne [Jot and Chaigne, 1991] presented a reverberator based on the feedback delay network structure introduced by [Stautner and Puckette, 1982] and proposed a systematic method for calculating the parameters of the reverberator. Figure A.1 shows the feedback loop of a four-channel Jot reverberator, containing a delay element and a filter in each channel and amplification and summing elements assuring the mixing between channels. To simplify the analysis, the amplification

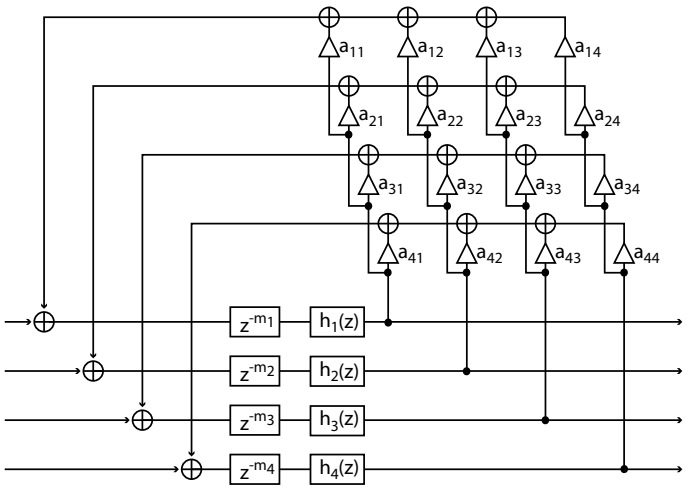

Figure A.1: Feedback loop of a 4-channel Jot reverberator.

factors are normally represented as the so-called mixing matrix:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} .$$

The mixing matrix A is crucial for the stability of the feedback loop and it was proposed by [Jot, 1992] to use unitary feedback matrices, which is a sufficient condition for keeping the total power of the signals in the feedback loop constant when no filters are present. The frequency-dependent reverberation times can therefore be easily controlled by the filters in the loop.

In practice however, not only the power conservation matters, but also mixing capability of the matrix is important, i.e. the capability to spread power from one channel to all the other channels. While for example an N-by-N identity matrix would be a perfectly valid mixing matrix, it does not have any mixing capability and will reduce the Jot reverberator to the first stage of a Schroeder reverberator [Schroeder, 1962], i.e. N parallel comb filters. A matrix with high mixing capability has a low crest factor, as defined in (A.1). The minimum achievable crest factor is 1, in which case all elements must have the same magnitude.

The requirements for mixing matrices vary depending on the application of the reverberator. A Jot reverberator designed to model the increase in echo density found in measured room impulse responses may not require a very efficient mixing matrix, but rather a mixing matrix that leads to the desired increase in echo density.

The aim of this study was to find mixing matrices for decorrelators and diffuse sound reverberators. A decorrelator is a reverberator that implements two or more short and statistically independent reverb tails while a diffuse sound reverberator simulates a room impulse response from which the direct sound and the early reflections have been removed. Diffuse sound reverberators and decorrelators both require a high mode density, in order not to introduce coloration to the signal, and a high echo density to make the reverberation sound smooth. For decorrelators it is also crucial that a high echo density is reached quickly because the reverberation tail is typically very short.

Because the mode density is directly related to the total delay length [Jot and Chaigne, 1991] and a rapid increase in echo density implies short average delays, a high number of channels is required for a decorrelator or a diffuse sound reverberator. In practice it may be desirable to have 20 to 40 channels to make the reverberator sound good. For such high numbers of channels, random N-by-N unitary mixing matrices are computationally very expensive and should be avoided.

To reduce the computational complexity, the use of Hadamard matrices has been proposed before [Jot, 1997], which allows to implement mixing matrices with a crest factor of 1 using only $N \log_2 N$ operations. However, for a 32-by-32 matrix, $\log_2 N = \log_2 32 = 5$, and therefore the implementation of the Hadamard matrix needs $5N = 160$ operations. The goal of this research is to study mixing matrices that can be implemented using even less operations, and matrix structures have been proposed that can be implemented with $\frac{4}{3}N$ to $5N$ operations, regardless of N.

It must be mentioned that, besides the already mentioned Hadamard matrix, several other special cases of mixing matrices are known to have highly efficient im-

plementations [Jot, 1997]. Contrary to most of these cases, which rely on elements of the matrix having the same magnitude, the approach chosen here is different (and to some extent orthogonal to the same-magnitude approach) and imposes that the majority of elements in the mixing matrix is zero, i.e. that the matrix is sparse. This may seem contradictory to the goal of achieving efficient mixing between channels, but it needs to be considered that an impulse fed to one of the channels will go many times through the mixing matrix before its amplitude becomes negligible. It is possible to design a sparse unitary matrix U such that U^n has only nonzero elements for a small n, meaning that after passing n times through the mixing matrix, an impulse in an arbitrary channel will have spread to all other channels.

Studying the sparsity of U^n gives only an approximative indication on the behavior of the feedback loop. On one side, because the delays in the feedback loop are all different, it is impossible to define a single time instant when all impulses have passed n times through the feedback loop, meaning that U^n does not represent the real spreading of energy from one channel to the other, especially for large n. On the other side, a matrix with only nonzero elements can still behave like a sparse matrix if the magnitudes of the elements are very different (e.g. some elements "stick out"). To gain more detailed information, the crest factor of U^n can be studied. For a matrix A with elements $a_{i,j}$ $(1 \leq i, j \leq N)$, the crest factor is defined as

$$\mathrm{C}(A) = \frac{\max\limits_{i,j} |a_{i,j}|}{\sqrt{\dfrac{\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{N} a_{i,j}^2}{N^2}}} \,. \tag{A.1}$$

However, despite the shortcomings of studying only the sparseness of U^n, this method turned out to give a simple yet useful indication and was therefore used throughout this research.

It should be mentioned also that there is extensive mathematical literature on unitary matrices. However, while it is known how to factorize any unitary matrix into a *series* of sparse unitary matrices [Vetterli and Kovačević, 1995, Section 2.B], little seems to be known about which non-sparse unitary matrices can be expressed as a power of a *single* sparse unitary matrix. The approach for designing suitable sparse unitary matrices presented here is a bottom-up approach, combining small and simple unitary matrices to generate a big unitary matrix with the desired properties.

This appendix is structured as follows: Section describes A.2 the method of evaluating the different sparse matrix types proposed in Section A.3 while Section A.4 presents the results and Section A.5 discusses them. Conclusions are drawn in Section A.6.

A.2 Matrix evaluation

In this research different structures for sparse unitary matrices (denoted U in the following) are proposed and evaluated with respect to different aspects and under different conditions. The aspects are the sparsity of U^n as a function of n, as well

as the time- and frequency-domain density of the impulse responses produced by Jot reverberators using U as the mixing matrix.

Different application conditions were simulated by three different scenarios. In the first scenario, the number of channels (and therefore the matrix size) is constant. Wherever possible, 24 channels were used and 25 was used in the cases where 24 was not possible due to the matrix design. The second and third scenario simulate complexity constraints as they could arise when implementing a diffuse reverberator or a decorrelator in an environment with limited computational resources.

As the measure of computational complexity, the number of multiplications per output sample was chosen. This measure is expected to be roughly proportional to the number of clock cycles per output sample needed for the implementation of the reverberator on a CPU in the case where the multiply operation is much more costly than the add operation (which may be the case with older or low-end CPUs) and also in the case where a multiply-accumulate (MAC) operation exists, which is the case for DSPs and many multimedia-oriented CPUs. Each element in the mixing matrix that is neither 0 nor 1 is supposed to require one multiplication per output sample, as long as U contains at most one element equal to 1 per column. This condition is necessary in order to have a realistic complexity estimate for CPUs with multiply-accumulate and is fulfilled for all matrix types proposed in this study.

Counting the number of elements different from 0 or 1 does not take into account the fact that many matrix types exist that can be implemented in a more efficient way because many nonzero elements have the same magnitude (different from 1). However, such simplifications are not of primary concern for this study since the main focus here is on the structure of the sparse matrices, not on the actual element values, and the nonzero elements are in practice calculated from random parameters, therefore not allowing simplifications based on equal element values. In practice, it is of course possible to design matrices that take advantage of *both* complexity reductions, due to sparseness *and* due to equal magnitudes. This study does not include the equal magnitude approach because it imposes many constraints on the matrices, as often for a given matrix size only few possible matrices are known, which would be contradictory with the approach used here, evaluating a large number of matrices of the same type and taking the mean over the results.

In the reverberator scenario with constrained complexity, the total number of multiplications for the recursive loop was required to be less than or equal to 200, including 4 additional multiplications per channel for the filters modeling the frequency-dependent reverberation times. For testing the matrices in the "fixed size" and the "fixed cost reverberator" scenarios, the reverberation time (RT60) was fixed to 1 s for all frequencies and the delays were mutually prime numbers randomly generated from a Gaussian distribution with a mean of 400 samples and a standard deviation of 300 samples.

In the constrained complexity decorrelator scenario, the total number of multiplications is limited to 100, including 1 additional multiplication per channel (since a decorrelator should have a decaying white noise tail as an impulse response, only one attenuation factor per channel is needed inside the recursive loop). The attenuation factors were calculated to achieve a reverberation time (RT60) of 250 ms and the delays were mutually prime numbers randomly generated from a Gaussian distribution with a mean of 300 samples and a standard deviation of 200 samples.

A.3 Matrix types

In the following, the different matrix types studied in this study are presented. The first two types of unitary matrices were introduced just as a reference and are the two most extreme cases of all possible ways of designing mixing matrices for a Jot reverberator: an identity matrix and a random (non-sparse) unitary matrix. As mentioned before, the goal of this study is to design mixing matrices that have many zero elements and still produce a temporally dense reverb. These conditions are not fulfilled by the two mentioned matrices: the identity matrix does not provide any mixing and the random unitary matrix is not sparse.

Since in an identity matrix only the elements on the diagonal are non-zero, using a N-by-N identity matrix as a mixing matrix will reduce the resulting Jot reverberator to the first stage of a Schroeder reverberator [Schroeder, 1962], i.e. N comb filters in parallel. The relationship between the Jot reverberator and the Schroeder reverberator is discussed in detail by [Jot and Chaigne, 1991].

The non-sparse random unitary matrices can be easily obtained using the singular value decomposition (SVD) of a random matrix. Using a random unitary matrix as the mixing matrix assures a very good mixing because the signal from each channel immediately propagates to all the other channels. However, from the implementation point of view it is the worst possible choice because all elements are nonzero and N^2 multiplications are needed.

The first attempt made to make a mixing matrix with the desired properties was a matrix composed of B blocks of 2–by–2 unitary matrices, arranged to the following structure:

$$U_2(B) = \begin{bmatrix} 0 & 0 & & & 0 & 0 & & 0 & 0 \\ 0 & 0 & & G_2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & & & & 0 & 0 \\ 0 & 0 & 0 & 0 & & G_3 & & 0 & 0 \\ & & \vdots & & & & \ddots & & \\ 0 & 0 & 0 & 0 & 0 & 0 & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & & & G_B \\ & & 0 & 0 & 0 & 0 & & 0 & 0 \\ G_1 & & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

where B is the number of blocks and G_i are Givens rotations

$$G_i = \begin{bmatrix} \cos\alpha_i & -\sin\alpha_i \\ \sin\alpha_i & \cos\alpha_i \end{bmatrix}$$

and the α_i are randomly chosen using a uniform distribution on the interval $[0, 2\pi]$. Implementing an N-by-N matrix of this type requires $2N$ multiplications.

An attempt was also made to design a computationally very efficient matrix requiring only $\frac{4}{3}N$ multiplications to implement a N-by-N matrix. These matrices are composed of B 3–by–3 unitary matrices that are sparse by themselves. The general

structure then looks like this:

$$U_3(B) = \begin{bmatrix} 0 & 0 & 0 & & & & 0 & 0 & 0 & & & & 0 & 0 & 0 \\ 0 & 0 & 0 & & \widehat{U}_2 & & 0 & 0 & 0 & & & & 0 & 0 & 0 \\ 0 & 0 & 0 & & & & 0 & 0 & 0 & \cdots & & & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & & & & & & & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & & \widehat{U}_3 & & & & & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & & & & & & & 0 & 0 & 0 \\ & & & \vdots & & & & & & & \ddots & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & \widehat{U}_B & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & & & \\ & & & 0 & 0 & 0 & 0 & 0 & 0 & & & & 0 & 0 & 0 \\ & \widehat{U}_1 & & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & & & 0 & 0 & 0 \\ & & & 0 & 0 & 0 & 0 & 0 & 0 & & & & 0 & 0 & 0 \end{bmatrix}$$

where $\widehat{U}_i$ are 3x3 unitary matrices of one of the following forms:

$$\widehat{U}_i \in \left\{ \begin{bmatrix} 0 & & \\ 0 & & G_i \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & & \\ 0 & & G_i \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ & & 0 \\ G_i & & 0 \end{bmatrix}, \begin{bmatrix} & & 0 \\ G_i & & 0 \\ 0 & 0 & 1 \end{bmatrix} \right\}$$

and G_i are random Givens rotations. Even though it would have been possible to use special values for the rotation angle (e.g. $\frac{\pi}{4}$), allowing a reduction of computational complexity, this was not done in order to have a fair comparison between matrix structures and also to stay with the most general case, avoiding possible unwanted effects due to one specific set of values.

While $U_2(B)$ and $U_3(B)$ can be considered as valid candidates for good mixing matrices (see discussion), they never become non-sparse, and therefore do not fulfill the goal set above. However, a simple way was found to modify U_2 such that the new matrix U_{21} fulfills the constraint that U_{21}^n should be non-sparse for some finite n by using the following structure:

$$U_{21}(B) = \begin{bmatrix} 0 & & & \\ \vdots & & U_2(B) & \\ 0 & & & \\ 1 & 0 & \cdots & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & & & & 0 & 0 \\ 0 & 0 & 0 & & G_2 & & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & & & & & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & & G_3 & & & 0 & 0 \\ & & & & \vdots & & & & \ddots & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & G_B & \\ 0 & & G_1 & & 0 & 0 & 0 & 0 & & 0 & 0 \\ 0 & & & & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & & & 0 & 0 \end{bmatrix}.$$

The same method can be used also on U_3:

$$U_{31}(B) = \begin{bmatrix} 0 & & \\ \vdots & U_3(B) & \\ 0 & & \\ 1 & 0 \;\cdots\; 0 & \end{bmatrix} = \begin{bmatrix} 0\;0\;0\;0 & & 0\;0\;0 & & 0\;0\;0 \\ 0\;0\;0\;0 & \widehat{U}_2 & 0\;0\;0 & & 0\;0\;0 \\ 0\;0\;0\;0 & & 0\;0\;0 & \cdots & 0\;0\;0 \\ 0\;0\;0\;0\;0\;0 & & \widehat{U}_3 & & 0\;0\;0 \\ 0\;0\;0\;0\;0\;0 & & & & 0\;0\;0 \\ 0\;0\;0\;0\;0\;0 & & & & 0\;0\;0 \\ & \vdots & & \ddots & \\ 0\;0\;0\;0\;0\;0\;0\;0\;0 & & & & \\ 0\;0\;0\;0\;0\;0\;0\;0\;0 & & & & \widehat{U}_B \\ 0\;0\;0\;0\;0\;0\;0\;0\;0 & & & & \\ 0 & & 0\;0\;0\;0\;0\;0 & & 0\;0\;0 \\ 0 & \widehat{U}_1 & 0\;0\;0\;0\;0\;0 & \cdots & 0\;0\;0 \\ 0 & & 0\;0\;0\;0\;0\;0 & & 0\;0\;0 \\ 1\;0\;0\;0 & & 0\;0\;0\;0\;0\;0 & & 0\;0\;0 \end{bmatrix}$$

where $\widehat{U}_i$ are the same sparse 3x3 unitary matrices as in $U_3(B)$.

A systematic way of generating sparse unitary matrices U such that U^n becomes non-sparse for small values of n was found by using the coefficients of B random unitary $m \times m$ matrices and arrange them in such a way on a $Bm \times Bm$ matrix that in the resulting Jot reverberator the signal from channel i is fed to channels $((i-1)m+1 \mod Bm)+1$ to $(im \mod Bm)+1$. For $m = 2$ this means that the output of channel 1 goes to channels 2 and 3, channel 2 to channels 4 and 5, channel 3 to channels 6 and 7, etc.

For $m = 2$ and $B = 3$, the resulting matrix $U_{2f}(3)$ (the subscript f standing for "fast") looks like this:

$$U_{2f}(3) = \begin{bmatrix} 0 & 0 & c_3 & 0 & 0 & c_4 \\ a_1 & 0 & 0 & a_2 & 0 & 0 \\ a_3 & 0 & 0 & a_4 & 0 & 0 \\ 0 & b_1 & 0 & 0 & b_2 & 0 \\ 0 & b_3 & 0 & 0 & b_4 & 0 \\ 0 & 0 & c_1 & 0 & 0 & c_2 \end{bmatrix}$$

where

$$\begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}, \quad \begin{bmatrix} b_1 & b_2 \\ b_3 & b_4 \end{bmatrix}, \quad \begin{bmatrix} c_1 & c_2 \\ c_3 & c_4 \end{bmatrix}$$

are random 2×2 unitary matrices, e.g. random Givens rotations.

For this study, matrices designed in the same way but with $m = 3$, $m = 4$ and $m = 5$ are used and are denoted $U_{3f}(B)$, $U_{4f}(B)$ and $U_{5f}(B)$.

For all matrices except the first two types, versions with randomized column orders have been generated. They are denoted $\widetilde{U_x}$ instead of U_x. An overview of matrix sizes and numbers of multiplications for the different scenarios and matrix types are shown in Table A.1.

	fixed size	reverberator		decorrelator	
	channels	channels	multiplications	channels	multiplications
I	24	50	200	100	100
U_{full}	24	12	192	9	90
U_2	24	32	192	32	96
U_3	24	36	192	42	98
U_{21}	**25**	33	196	33	97
U_{31}	**25**	37	196	43	99
U_{2f}	24	32	192	32	96
U_{3f}	24	27	189	24	96
U_{4f}	24	24	192	20	100
U_{5f}	**25**	20	180	15	90

Table A.1: Channel numbers and multiplications per output sample as a function of matrix design and scenario (the randomized versions have been omitted from this table because they have the same size as their non-randomized counterparts)

A.4 Results

The matrices and the impulse responses generated by using them in a reverberator were examined under four aspects. First the evolution of the matrices (i.e. their different powers) was studied graphically in order to see how they converge to a non-sparse matrix. Then, the number of iterations of the matrix needed to become non-sparse was computed and the time needed for the impulse response to achieve 100% echo density was calculated as well as the standard deviation of the spectrum of the late impulse response (in 1-ERB bands). Because the matrices (except the identity matrix) depend on random values and the (random) delays used in the recursive loop also have an influence on the performance, the measures described above may change as a function of the random numbers used to generate the matrices and the delays. Each case was repeated 100 times for different random number generator seeds and the mean and standard deviations were calculated.

In the following, all illustrations of matrices show their absolute values on a scale from 0 (white) to 1 (black). Using absolute values and white for the value 0 makes it easy to estimate the sparseness of the matrices. Furthermore, the signs of the values do not carry relevant information in this context.

A.4.1 Fixed matrix size

As shown in Table A.1, in the "fixed matrix size" scenario, all reverberators have either 24 or 25 channels. The difference is due to the fact that no single matrix size could be generated by all the design methods, so in the following one should keep in mind that a difference in the results may be due to a difference in the number of channels of 5%.

Figure A.2 shows the evolution (i.e. different powers) of the identity matrix and

a non-sparse random unitary matrix. Both matrix types do not show any qualitative change when taken to higher powers: the identity matrix always stays the same, and a random unitary matrix always stays a random unitary matrix.

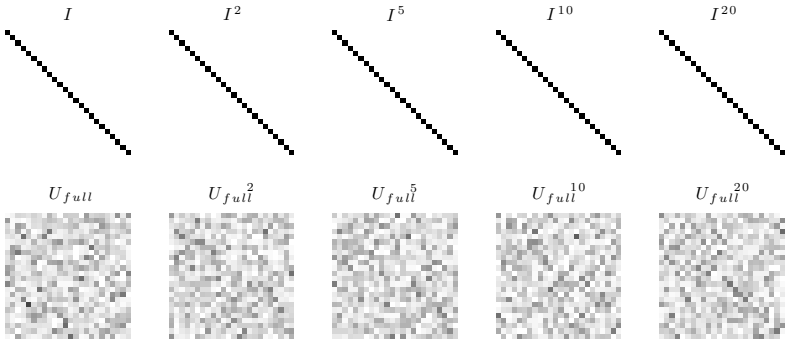

Figure A.2: Evolution of the two most extreme cases of mixing matrices for "fixed size" scenario. **Top:** 24×24 identity matrix. **Bottom:** 24×24 random unitary matrix.

Figure A.3 shows the evolution of matrices generated directly by the different design methods. It can be seen that U_2 and U_3 do not converge, but rather a diagonal "chain" of small (2×2 or 3×3) unitary matrices moves across the matrix. The cases U_{21} and U_{31} both converge, but very slowly (taken into account the approximately logarithmic display of matrix powers). In U_{21}^{20}, a diagonal band of higher values can be distinguished. This is much less the case in U_{31}^{20}, which in turn shows some single high values that "stick out". A value close to 1 would mean that – if all delays were equal – after passing 20 times through the recursive loop, the signal from one channel would predominantly show up in one single (different) channel.

The cases U_{2f} to U_{5f} do not show any such behavior and also converge much more quickly: already after 5 or 10 iterations, these matrices look line a random unitary matrix generated using an SVD. The more nonzero elements the original matrix has, the quicker is the convergence.

Figure A.4 shows instances of the same matrix types, but with randomized column order. It can be noticed that the differences in convergence between U_2 and U_{21} completely disappeared after the randomization. The same holds also for the randomized versions of U_3 and U_{31} in general. However, in this instance of $\widetilde{U}_{31}$ one can see a drawback of randomization: randomizing can actually impair the convergence behavior. Because $\widetilde{U}_{31}$ has two elements equal to 1 on the diagonal, it never converges to a non-sparse matrix.

On the "fast" matrices U_{2f} to U_{5f}, the effect of the column randomization seems to be rather adverse in the short term: the number of iterations needed for achieving complete non-sparsity increases (which is also confirmed by the data in Figure A.5), but in the long term, no significant change can be seen: U_{nf}^{20} and $\widetilde{U}_{nf}^{20}$, $n \in \{2, 3, 4, 5\}$ all look like random unitary matrices.

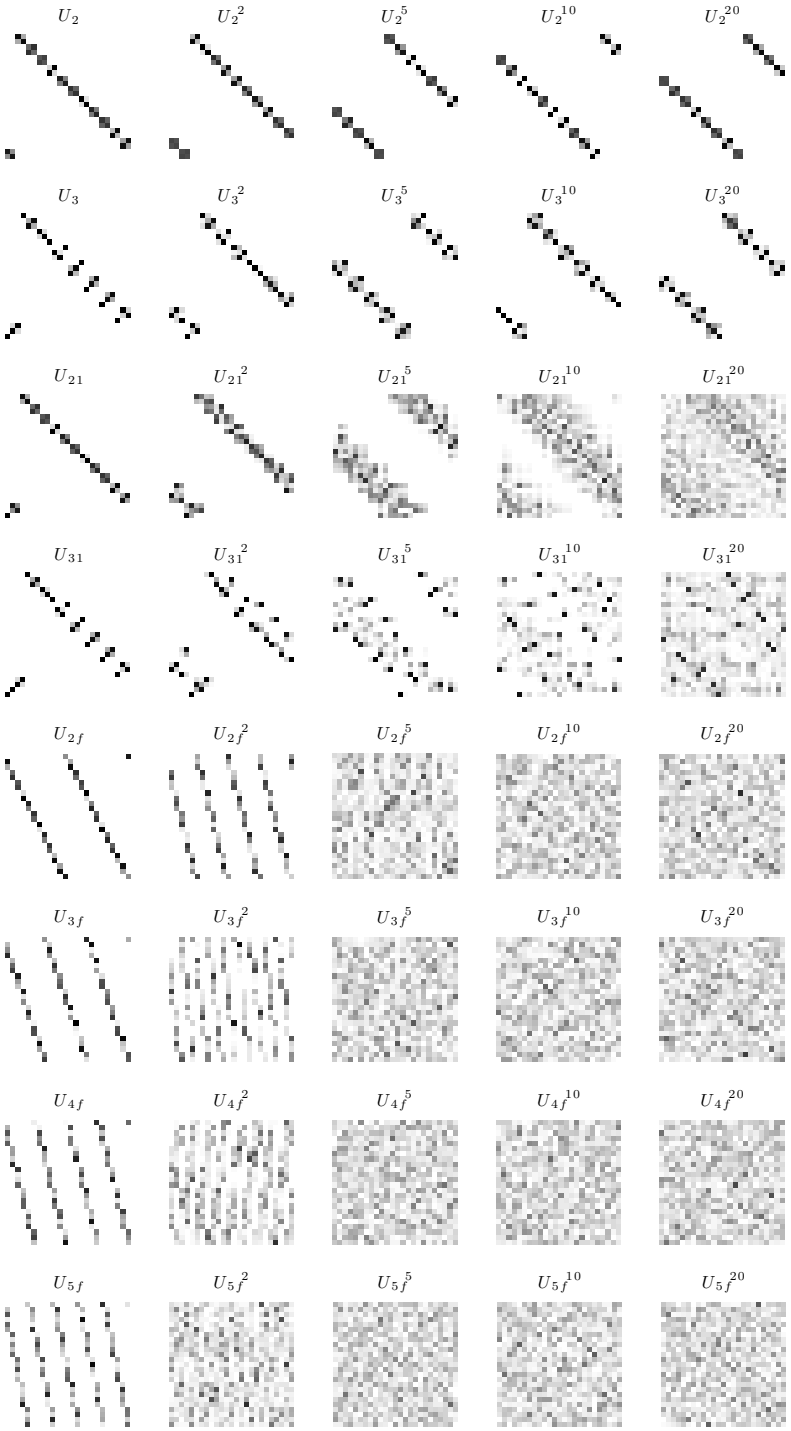

Figure A.3: Evolution of studied sparse mixing matrices for "fixed size" scenario.

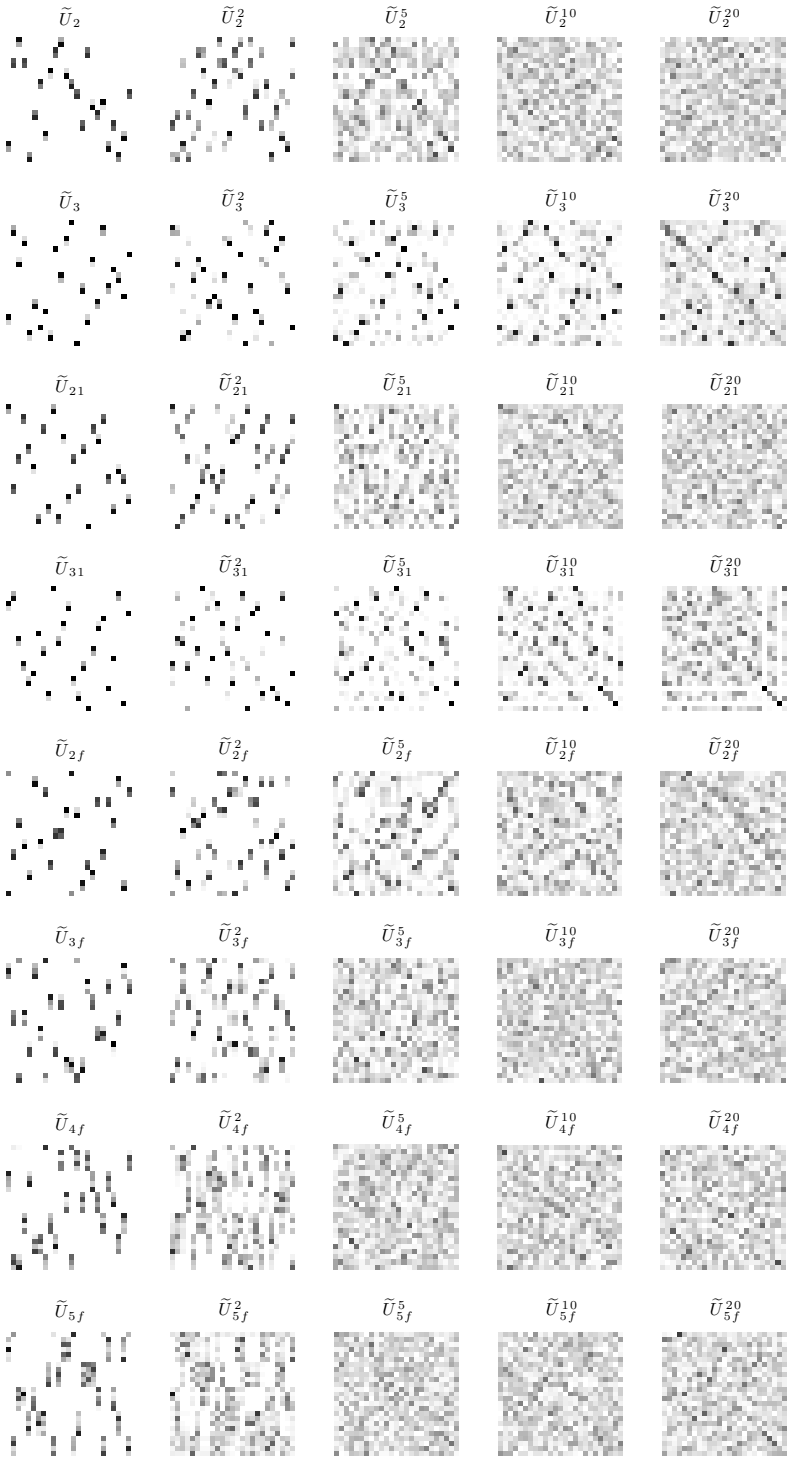

Figure A.4: Evolution of studied sparse mixing matrices with randomized column ordering for "fixed size" scenario.

Figure A.5 shows the number of iterations $k_{\min}$ that were needed to obtain a non-sparse matrix as a function of the matrix type, separately with and without column randomization. The observations made on Figure A.3 are confirmed by the averages: the matrices I, U_2, and U_3 never converge; the "fast" matrices U_{Nf} converge more rapidly than all the others (except for U_{full} of course); randomization makes U_2 and U_3 converge faster, while it slows down the convergence for the those matrices that are "fast" by design. The reason why U_{31} converges much slower than U_{21} is that U_{31} is much more sparse.

Figure A.6 shows the time needed to reach 100% echo density (i.e. non-sparsity of the impulse response). It may be observed that this time is very closely related to the value $k_{\min}$ shown in Figure A.5, with one notable exception: for the time needed to reach 100% echo density, U_2 and U_3 behave like their randomized versions and also like U_{21} and U_{31}. Furthermore, only an insignificant difference between U_2 and U_{2f} can be observed.

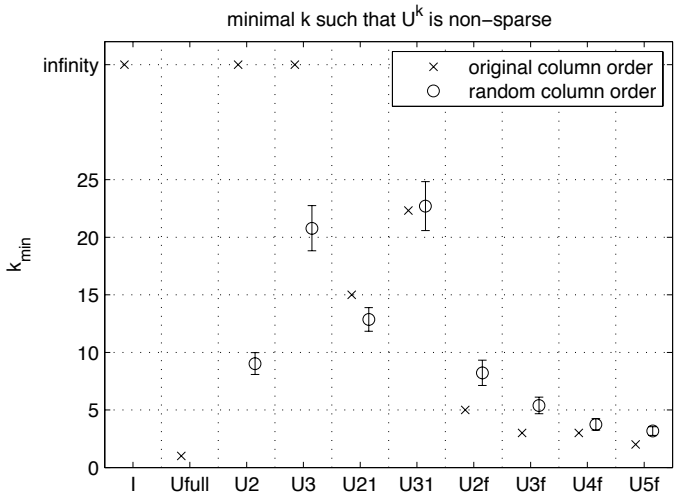

Figure A.5: Number of iterations needed to obtain non-sparse matrix for "fixed size" scenario.

Figure A.7 shows the spectral deviation of the late tail of the reverberators impulse responses. For all matrices, except for the identity matrix, no significant difference between spectral deviations can be observed. This is in line with the finding that the mode density (which is related to the spectral deviation) of a feedback delay network only depends on the total length of delays [Jot and Chaigne, 1991]. Since here the number of channels is always 24 or 25, i.e. varies only by 5%, the average total length of the delays also varies by 5%. That the reverberator using an identity matrix performs significantly worse even though it has the same total delay length as all the other cases may be explained by the fact that it never reaches 100% echo density.

It is interesting to note that even the case U_3 which has no complete mixing and a very sparse matrix performs as well as the other cases with respect to the spectral

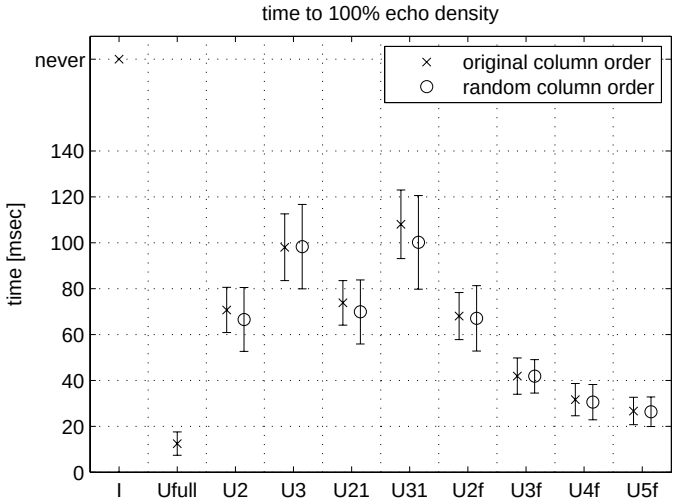

Figure A.6: Time needed to achieve full echo density for "fixed size" scenario.

deviation.

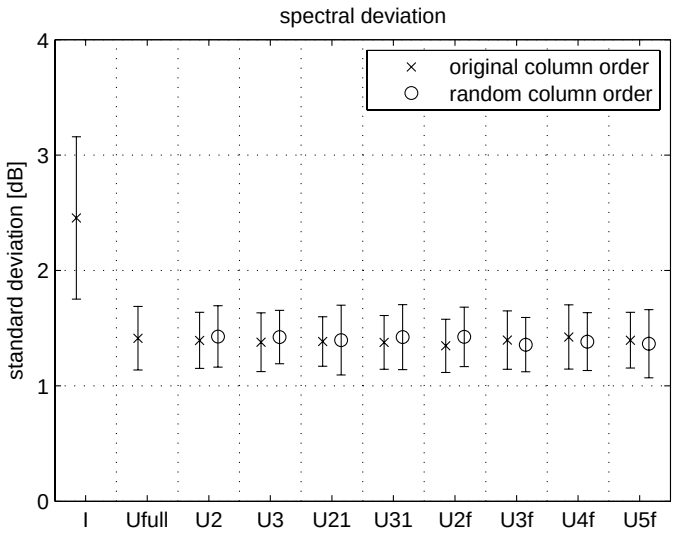

Figure A.7: Spectrum standard deviation in late tail for "fixed size" scenario.

A.4.2 Reverberator scenario

Figure A.8 shows the evolution of the identity matrix and a random unitary matrix for the "reverberator" scenario, where the number of multiplications is limited to 200 and each channel contains a 4-tap FIR filter (thus consuming 4 multiplications per channel, independently of the mixing matrix). It can be observed that this set of constraints leads to large differences in matrix size.

Figures A.9 and A.10 show the evolution of instances of the other matrix types, with original column order and with random column order, respectively. In general, the same observations can be made as in the "fixed size" scenario. Due to the bigger size of the matrices, which makes the convergence of U_{31} and U_{21} very slow, it can be observed well how the random column order improves the convergence behavior in these two cases.

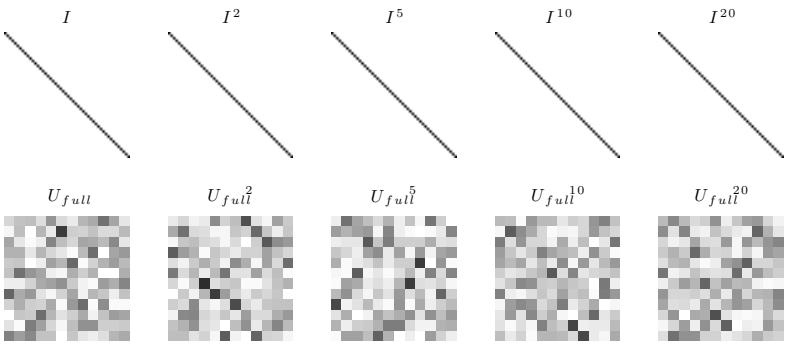

Figure A.8: Evolution of the two most extreme cases of mixing matrices for "reverberator" scenario. **Top:** 50×50 identity matrix. **Bottom:** 12×12 random unitary matrix.

Figure A.11 shows the number of iterations needed for convergence to a non-sparse matrix and confirms the improvement of convergence due to randomized column ordering of U_{31} and U_{21}. The same figure also confirms the degradation of the convergence for the "fast" matrix types U_{2f} to U_{5f}.

Figure A.12 shows the time to reach 100% echo density. The same observations as in the "fixed size" scenario can be made. In particular this figure shows that the improvement in convergence for U_{2f} does not translate in any significant improvement of the time to 100% echo density.

Figure A.13 shows the spectral deviation of the late tail of the reverberators impulse responses. Knowing that the mode density of a reverberator depends on the total delay length, it can be expected that the lowest spectral deviations occur for the reverberators with the highest number of channels. This is true indeed, as the reverberators based on U_2, U_3, U_{21}, U_{31}, and U_{2f} perform best.

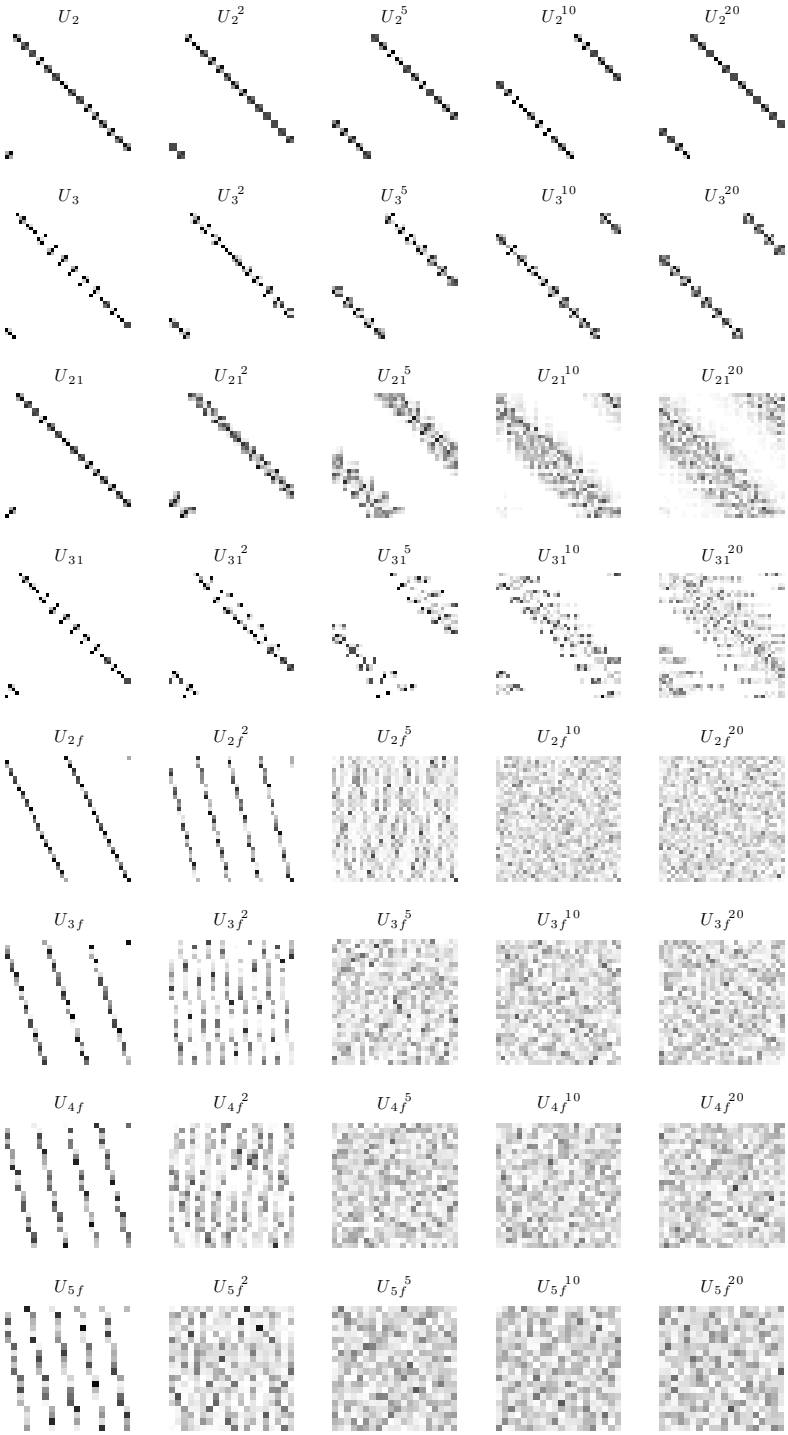

Figure A.9: Evolution of studied sparse mixing matrices for "reverberator" scenario.

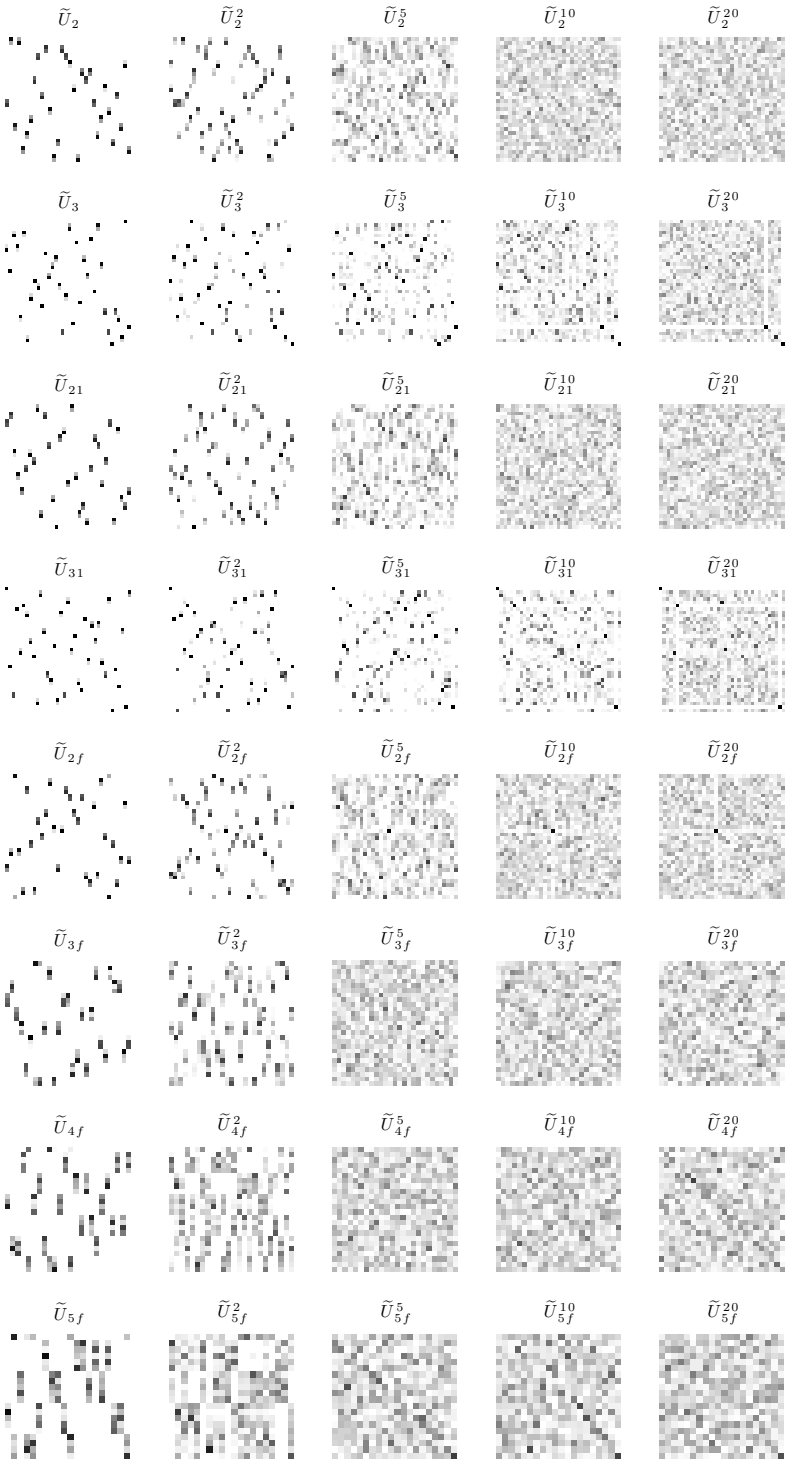

Figure A.10: Evolution of studied sparse mixing matrices with randomized column ordering for "reverberator" scenario.

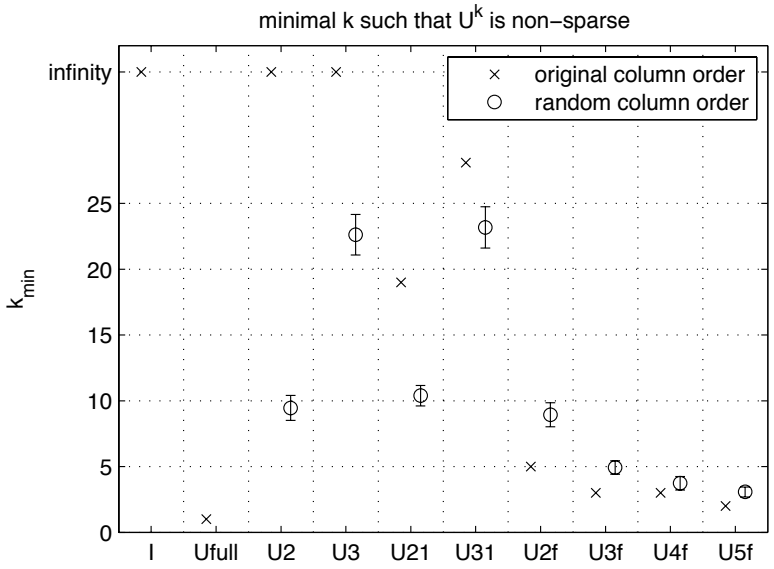

Figure A.11: Number of iterations needed to obtain non-sparse matrix for "reverberator" scenario.

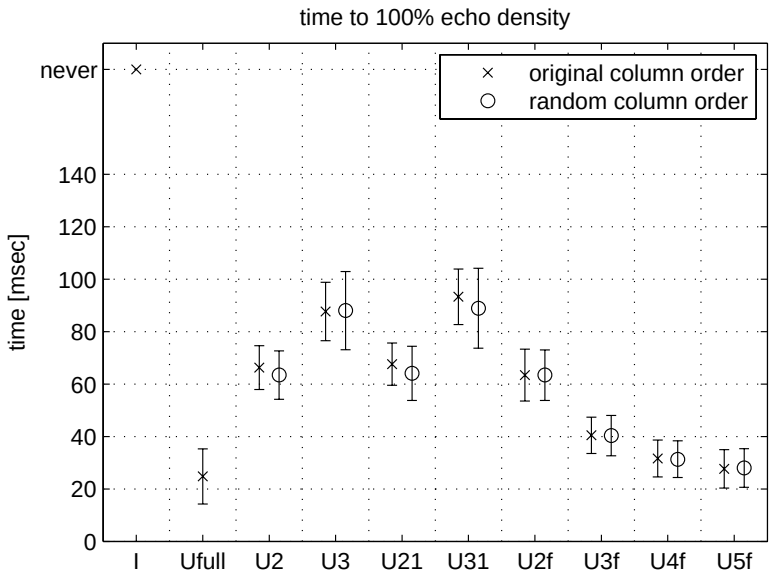

Figure A.12: Time needed to achieve full echo density for "reverberator" scenario.

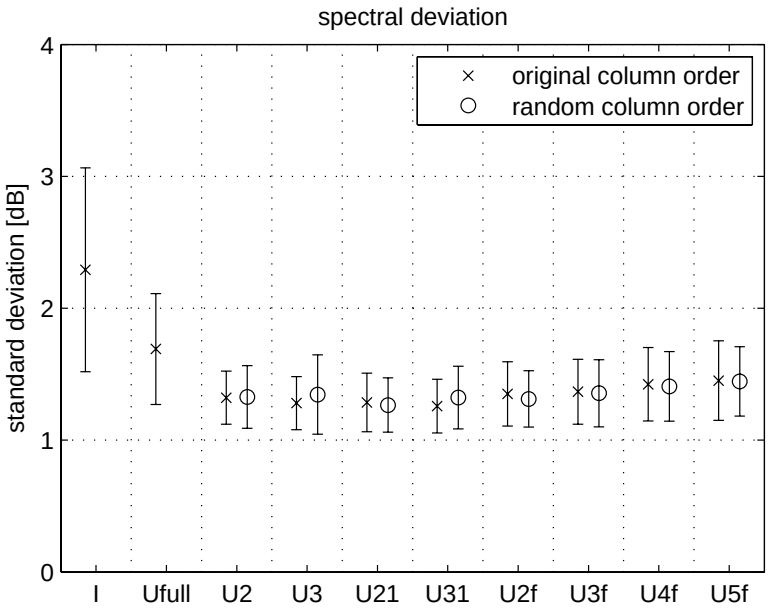

Figure A.13: Spectrum standard deviation in late tail for "reverberator" scenario.

A.4.3 Decorrelator scenario

Figure A.14 shows the evolution of the identity matrix and a random unitary matrix for the "decorrelator" scenario, where the number of multiplications is limited to 100 and each channel contains a single amplifier (thus consuming 1 multiplications per channel, independently of the mixing matrix). It can be observed that this constraint leads to very big differences in matrix size.

Figures A.15 and A.16 show the evolution of instances of the other matrix types, with original column order and with random column order, respectively. In general, the same observations can be made as in the "fixed size" scenario.

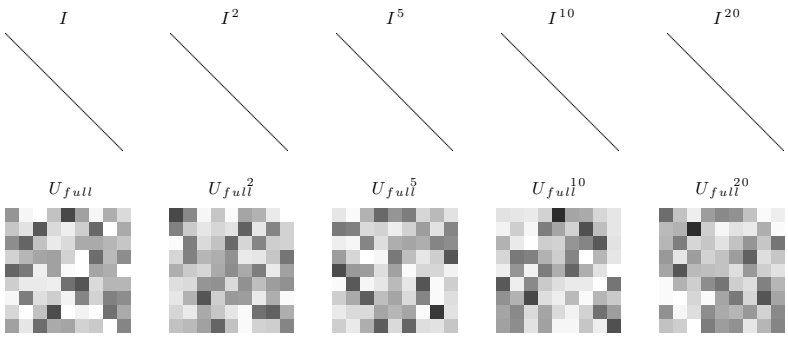

Figure A.14: Evolution of the two most extreme cases of mixing matrices for "decorrelator" scenario. **Top:** 100×100 identity matrix. **Bottom:** 9×9 random unitary matrix.

Figure A.17 shows the number of iterations needed for convergence to a non-sparse matrix and generally confirms the observations made in the "reverberator" case.

Figure A.18 shows the time to reach 100% echo density. The same observations as in the other two scenarios can be made.

Figure A.19 shows the spectral deviation of the whole impulse responses of the decorrelators. The triangles below each error bar are the minimum values found while testing 100 instances of each matrix type. This figure shows that spectral standard deviations nearly as low as 1 dB can be reached with such a decorrelator. It is interesting that the minimum value was reached with U_{2f}. This indicates that fast convergence plays an important role in the design of decorrelators.

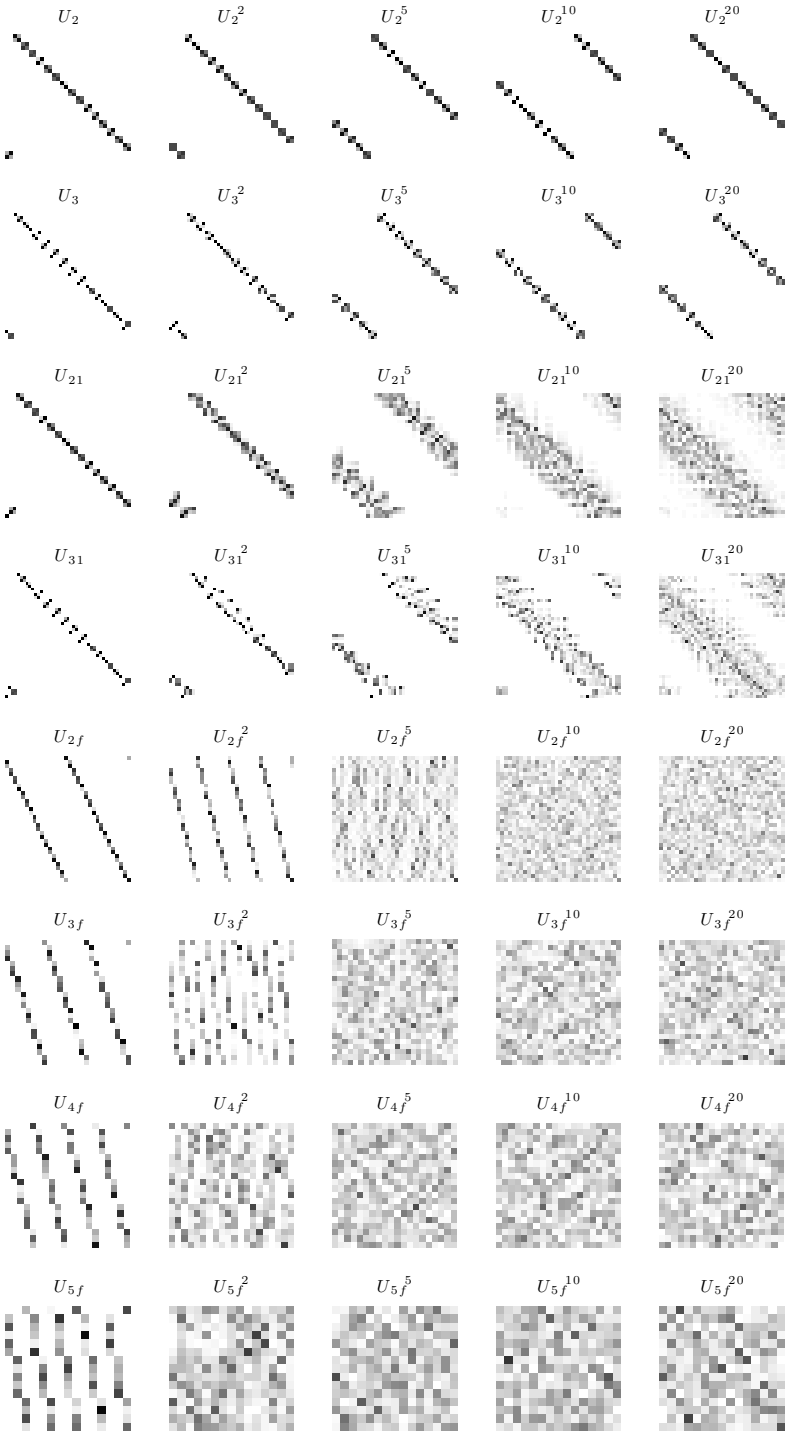

Figure A.15: Evolution of studied sparse mixing matrices for "decorrelator" scenario.

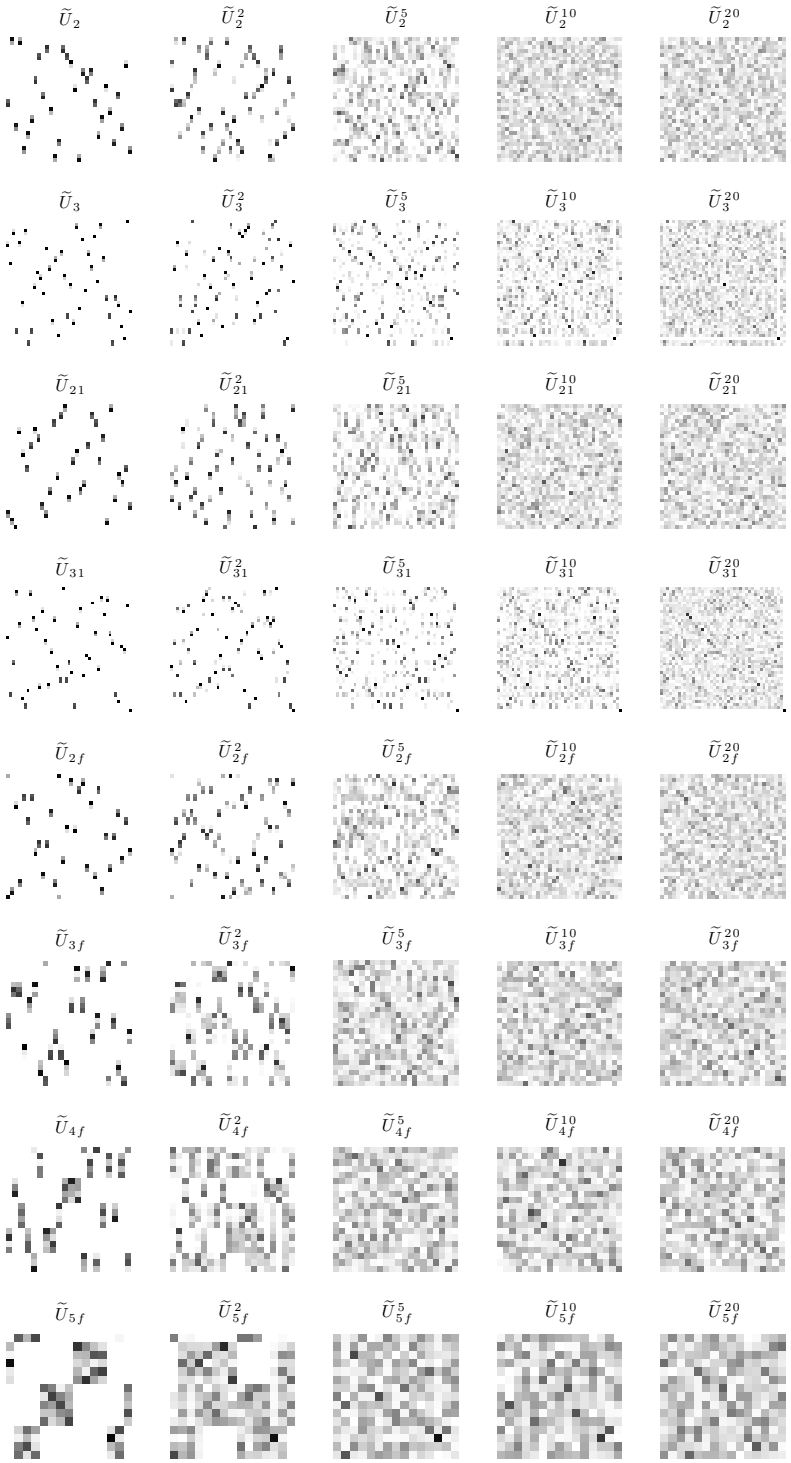

Figure A.16: Evolution of studied sparse mixing matrices with randomized column ordering for "decorrelator" scenario.

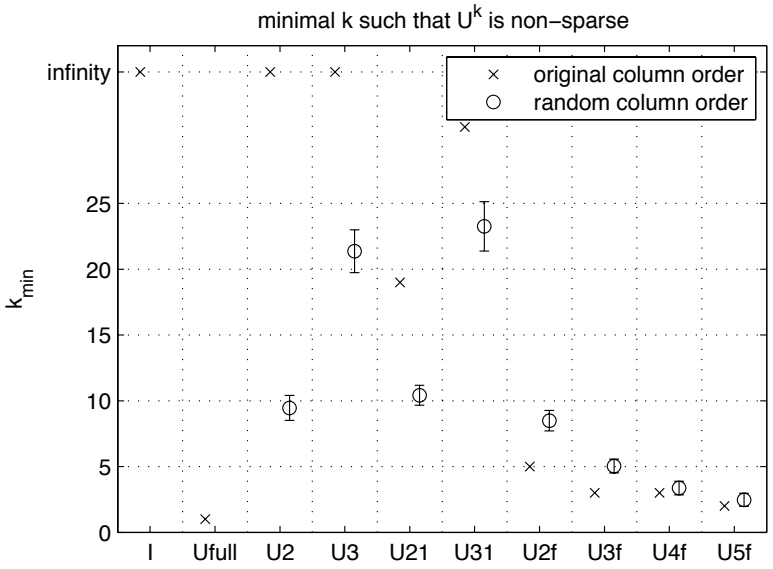

Figure A.17: Number of iterations needed to obtain non-sparse matrix for "decorrelator" scenario.

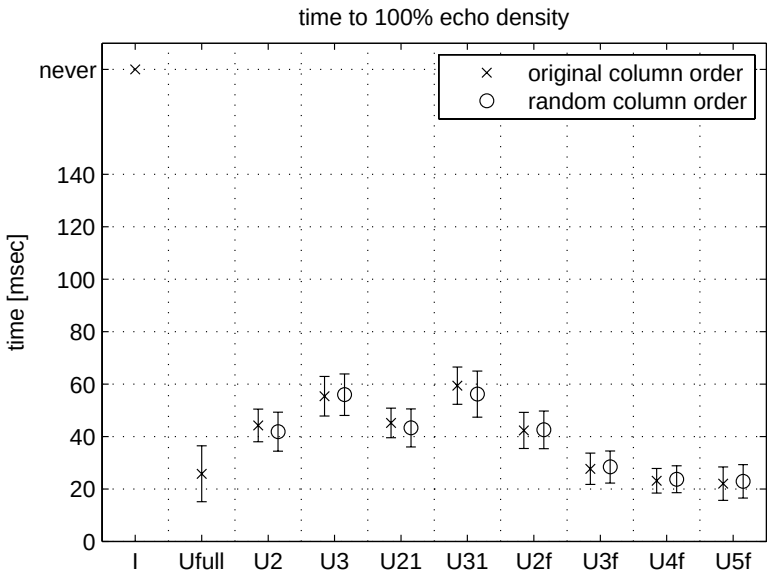

Figure A.18: Time needed to achieve full echo density for "decorrelator" scenario.

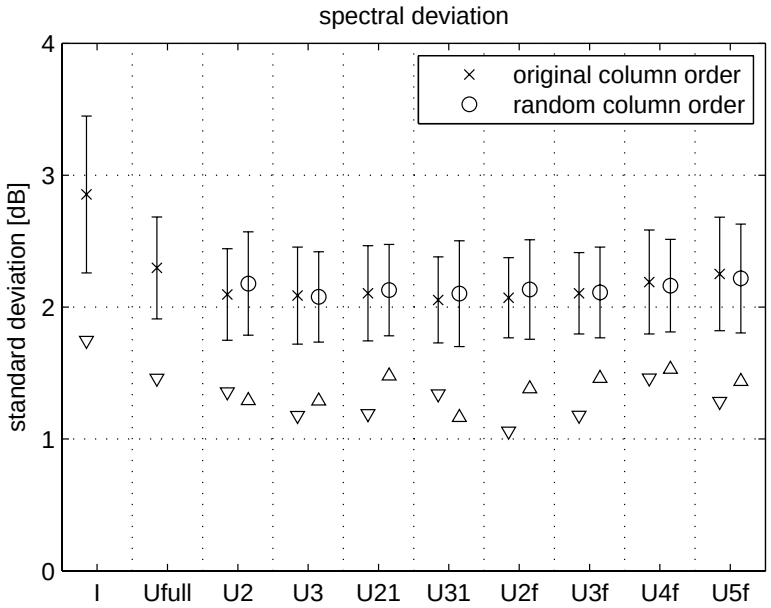

Figure A.19: Spectrum standard deviation of entire impulse response for "decorrelator" scenario. The triangles below the error bars show the minimum values found in 100 random instances.

A.5 Discussion

It was found that the minimum power of the matrix that is non-sparse allows to predict after which time the echo density in the impulse response reaches 100%. However, there is a notable exception because the matrix types U_2 and U_3 never converge, but still produce impulse responses that reach 100% echo density relatively fast. The explanation lies in the fact that the powers of the mixing matrix only indicate how a signal spreads among channels if all delays in the recursive loop are equal. In a real reverberator the delays are normally chosen to be mutually prime and are therefore different.

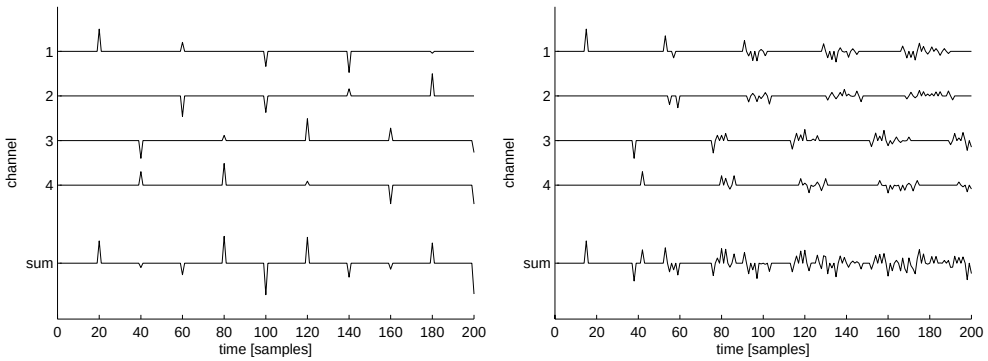

Figure A.20: Response of a four-channel reverberator to a single impulse in the first channel. Matrix type: U_2. **Left:** Equal delays. **Right:** Mutually prime delays.

Figure A.20 shows the signals in a 4-channel reverberator, where channel 1 was excited with a dirac impulse at time 0. The mixing matrix is of type U_2 and has the following structure

$$U_2(2) = \begin{bmatrix} 0 & 0 & b_1 & b_2 \\ 0 & 0 & b_3 & b_4 \\ a_1 & a_2 & 0 & 0 \\ a_3 & a_4 & 0 & 0 \end{bmatrix} .$$

This matrix does not converge to a non-sparse matrix. In the left hand plot the delays are all equal to 20 samples and it can be seen that at each iteration of the recursive loop only two channels are nonzero. This is what the powers of the mixing matrix predict.

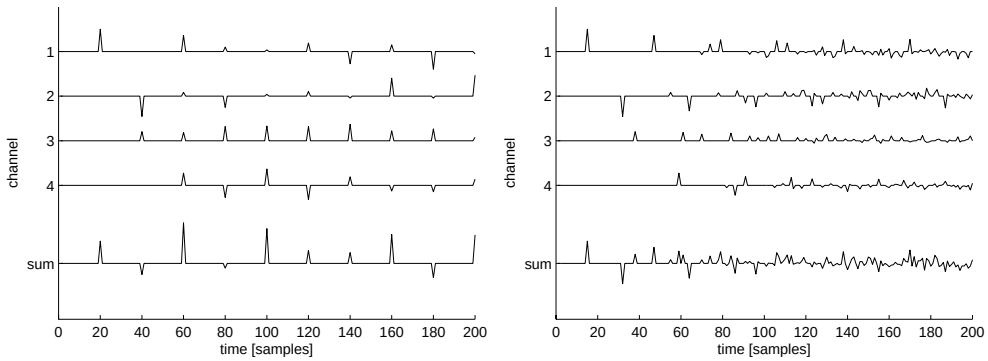

Figure A.21: Response of a four-channel reverberator to a single impulse in the first channel. Matrix type: U_{2f}. **Left:** Equal delays. **Right:** Mutually prime delays.

The left plot of Figure A.21 shows the same signals for a reverberator with a mixing matrix of type U_{2f} which has the following structure

$$U_{2f}(2) = \begin{bmatrix} 0 & b_3 & 0 & b_4 \\ a_1 & 0 & a_2 & 0 \\ a_3 & 0 & a_4 & 0 \\ 0 & b_1 & 0 & b_2 \end{bmatrix},$$

and which has, contrary to U_2, the property that U_{2f}^2 is non-sparse. From 40 samples after the initial impulse – corresponding to two iterations in the recursive loop – all channels are nonzero at each iteration.

So far the behavior of the reverberator follows closely the behavior of the powers of the matrix. However, if one looks at reverberators with mutually prime delays, the behavior of the reverberator is more complicated. On the right side of Figures A.20 and A.21 the signals for reverberators with mutually prime delays are shown. It can be observed that the total number of impulses in the channels as a function of time is roughly the same for both reverberators. The difference is that for the reverberator using U_2 they appear in bursts where at one iteration channels 1 and 2 are active and at the next iterations channels 3 and 4 are active, while for the reverberator using U_{2f} all channels are equally active over time. In fact, both matrices have the effect that each impulse will generate two impulses at the next iteration. The only difference is how the impulses are distributed to the channels, as well as their exact timing. As a result, the evolution of the echo density is similar for both matrices.

While this study focused on the structure of sparse unitary matrices and on composing such matrices from random unitary matrices, in a practical implementation it is desirable to take advantage also of the reduction of computational complexity due to choosing particular values for the elements of the mixing matrix. As an example, a highly efficient mixing matrix can be obtained by composing $U_{4f}(4)$ from 4-by-4 Hadamard matrices with randomized column order. Such a matrix is shown below.

$$U_{4fh} = \begin{bmatrix}
0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 \\
0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1
\end{bmatrix} /2 \, .$$

This matrix has the properties that U_{4fh}^2 is non-sparse and the crest factor $C(U_{4fh}^2)$ is equal to 1. Due to the efficient implementation of Hadamard matrices, the computational cost of this matrix is 32 operations per output sample, i.e. half the cost of a U_{4f} matrix of the same size with random values or a 16-by-16 Hadamard matrix.

A.6 Conclusions

In this study it was shown that using sparse unitary matrices as mixing matrices in Jot reverberators is an efficient way to reduce the computational complexity. Matrices of different types were analyzed visually and by studying how fast the powers of the matrices become non-sparse, i.e. after how many iterations in the recursive loop a signal fed to an arbitrary channel will spread to all other channels. The resulting impulse responses were analyzed with respect to echo and spectral density.

The mixing matrix of a Jot reverberator has a strong influence on how fast the echo density in the impulse response increases. When the goal of a reverberator is to model only diffuse reverberation or to be used as a decorrelator, it is desirable to reach full echo density as soon as possible. Different methods for designing sparse unitary matrices were proposed and evaluated in 3 different scenarios with different constraints. In the first scenario the matrix size was fixed, while in the second and third scenario the number of multiplications was limited. The difference between the second and the third scenario was that the former used parameters suitable for a reverberator and the latter aimed at implementing a decorrelator.

The exact choice of a mixing matrix can be made only in the context of the other design parameters of the reverberators (e.g. memory constraints, constraints on the time when 100% echo density should be reached, etc.) and should also depend on the choice of delays. However, it was found that in the reverberator scenario it is favorable to use a mixing matrix that is sparse and big rather than small and converging fast, while in the decorrelator case fast convergence plays an important

role and particularly good results were obtained with the fast converging matrix types U_{2f}, U_{3f}.

While several types of mixing matrices have been proposed before that can be implemented with a low computational complexity because they contain elements with equal magnitude, this study focused on the complexity reduction in reverberators due to the sparse structure of the mixing matrix, i.e. the distribution of zero and nonzero elements. It could be shown that the advantages of both methods can be combined by choosing the values of the nonzero elements from the previously proposed efficient mixing matrices.

Appendix B

BRIR Measurements

All room impulse measurements for this research were conducted in a lecture hall at our university (ELA 2) which is 10 m wide, 14 m long and whose floor ascends in steps towards the back of the room. The loudspeakers and the microphones were placed in the front of the room, where the height is 4 m.

For all measurements, the same microphone position and the same loudspeaker setup were used. Seven loudspeakers were placed in a vertical plane pointing towards the microphone position. Their elevation angles and distances relative to the microphone position are shown in Table B.1.

All D/A and A/D conversions were done with a MOTU 896HD firewire sound interface at 96 kHz. To measure the impulse responses, a logarithmic sweep signal of 2.5 s length, covering the frequency range between 20 Hz and 48 kHz was used.

Table B.1: Loudspeaker positions relative to the microphone position.

Distance	Elevation
1.2 m	60°
1.5 m	30°
1.5 m	15°
1.5 m	0°
1.5 m	−15°
1.5 m	−30°
1.2 m	−60°

The B-format room impulses were measured using a Soundfield ST350 microphone and the BRIRs were measured with a KEMAR artificial head with torso. The artificial head was put on a remote-controlled turntable in order to measure BRIRs precisely every 5° in azimuth.

The setup was designed such that the first 3 ms of the BRIRs could be used as HRTFs (i.e. no reflections arrive at the microphone position in the first 3 ms after the direct sound). Therefore the measurements yielded at the same time a BRIR set and an HRTF set for 7 elevation angles and 72 azimuth angles.

Figure B.1: Left: Loudspeaker setup with Soundfield ST350 at the microphone position. **Right:** Loudspeaker setup with KEMAR artificial head and torso at the microphone position.

Appendix C

How This Thesis Was Developed

For the sake of keeping a record of how the ideas of this thesis developed, it may be interesting to look at how the ideas behind it evolved. While the thesis was arranged in a linear order starting from the most fundamental results and ending with the most applied algorithms, the actual research was not carried out exactly in this order and included also subjects that do not appear in the thesis.

When the work on this thesis started the idea was to validate the cue selection model [Faller and Merimaa, 2004] by using it to predict new psychoacoustic effects and experimentally verifying the existence of the predicted effects. These efforts resulted in finding an effect similar to the precedence effect, where sound events are incorrectly localized [Menzer and Faller, 2007]. Unlike the case of the precedence effect, narrowband noise bursts (rather than clicks or wideband noise bursts) were used as stimuli. The novelty was that with the observed effect the lead / lag delays are on a much longer timescale (approximately 100 ms vs. 10 ms - 30 ms for the precedence effect) and that no fusion occurs, i.e. that all the sound events can be heard separately. This effect effectively led people to believe that they heard sound events from a different place than where the sound events were actually played from. However, this research was not continued because it became clear that the observed effects were influenced also by still unknown higher-order perceptual processes. In particular the number of sound sources – a parameter not included in any of the models that were studied – had a strong influence on the perception of the positions.

At the same time a collaboration with the Cognitive Neuroscience Lab at EPFL (LNCO) was started to study the perception of the sound of footsteps. Here, the original idea was that hearing one's own footsteps in an unnatural position (e.g. behind oneself or in front) could lead to strange perceptions similar to "out-of-body" experiences. This did not turn out to be true, possibly because the spatial cue of the sound of footsteps coming from below may be relatively weak because the direct sound is to some extent blocked by the body. However, the same study, which was designed around a backpack mounted system for applying BRIRs in real time to the sound of footsteps recorded using microphones attached to the shoes of the subjects, also involved applying different delays to the sound of the footsteps. It turned out that

different delays caused the subjects to walk faster or slower – an effect not foreseen in the original design of the study! A paper describing this effect is currently under revision by the Cognitive Neuroscience journal.

During the development of the software for the footsteps experiment (for which applying recorded BRIRs or applying artificial binaural reverberation were considered as options) several questions were raised, such as "what makes BRIRs sound better than HRTFs?", "do early reflections help localize the direct sound?", or "how can BRIRs be recorded efficiently?". The latter question ultimately led to starting a new direction of research and the development of a method for generating BRIRs from B-format room impulse responses and an HRTF set [Menzer and Faller, 2008, 2010a], thus separating the measurement of room related properties from the measurement of listener related properties, which can be a big advantage if individualized BRIRs for a large set of rooms and a large set of listeners have to be provided. This method is presented in Chapter 3.

During the research on processing B-format RIRs, a method was developed for processing the diffuse part of a B-format signal in order to obtain a stereo signal that matches the frequency-dependent interaural coherence of diffuse sound predicted from an HRTF set. This method proved to be applicable also in other cases. The first such application was presented in [Menzer and Faller, 2009a] and introduced a systematic way of designing binaural reverberators by extending the approach presented in [Jot, 1992] from mono reverberation to binaural reverberation. This was done by applying the frequency-dependent interaural coherence matching method to a modified Jot reverberator. Instead of using an HRTF set to predict the diffuse sound coherence, the coherence was directly calculated from a BRIR and instead of applying the processing to a B-format room impulse response, it was applied to two channels of independent reverberation produced by the reverberator.

The work on binaural reverberation led to the question what the most important binaural cues for binaural reverberation are. It is commonly agreed upon that early reflections are very important binaural cues. However, there were indications that frequency-dependent interaural coherence could be an even more important cue, and even the idea came up that early reflections might be perceived only through their effect on interaural coherence and coloration. While the latter could not be proven, an extensive research on the perception of binaural room impulses (presented in Chapter 2) showed that in fact the frequency dependence of the interaural coherence is one of the main perceptual cues of binaural room impulse responses.

Another application of interaural coherence matching was found in the binaural processing of stereo signals. The interchannel coherence of diffuse sound recorded with a coincident pair of microphones is not the same as the interaural coherence of a binaural recording of diffuse sound (and the same is in general also true for diffuse sound generated by a stereo reverberator). Therefore, when playing back a stereo signal using headphones, unnatural interaural coherence cues are given to the listener. It is therefore desirable to process the diffuse sound contained in a stereo signal such that the frequency-dependent interaural coherence matches the coherence that could be found in a binaural recording in order to provide the listener with plausible and natural coherence cues.

In order to be able to perform such a processing, a separation algorithm was developed to split the stereo signal into coherent and diffuse parts. This separation

algorithm was intended to produce two two-channel signals: a coherent signal and a diffuse signal, such that the sum of the two signals will be the original stereo signal (perfect reconstruction property). Furthermore the coherent signal is supposed to have an interchannel coherence close to 1 in a time-frequency representation (more precisely, in a single timeframe and a single critical band, the two channels should differ only by an amplitude factor) and the diffuse signal should be such that the interchannel coherence of a signal composed of the sum and the difference of the diffuse signal should be zero. The latter was motivated as a property of recordings of diffuse sound using a symmetric coincident pair microphone setup.

Furthermore, an algorithm for processing the coherent sound was developed, similar to the one presented in [Breebaart and Schuijers, 2008]. Combining the separation, the diffuse sound processing and the coherent sound processing, a method for adding realistic binaural cues for headphone playback to a stereo signal while introducing only minimal modifications into the stereo signal was developed. This can be seen as an alternative to rendering stereo signals for headphone playback using BRIRs, which also generates realistic binaural cues but introduces many modifications, such as an increased reverberation time, comb filter effects due to the simulation of two loudspeakers, etc.

Finally, the idea of separating the processing of coherent and diffuse sound was also applied to binaural reverberators. The concept of a reverberator consisting of two separate reverberators for coherent reverberation (i.e. direct sound and early reflections) and diffuse reverberation was developed. Contrary to similar concepts such as the ones described in [Toma et al., 2005], the proposed method is based on two variations of the Jot reverberator [Jot, 1992] and the diffuse reverberation is not limited to late reverberation but present from the beginning of the impulse response. In order to be able to implement such a binaural reverberator efficiently, research on the design of unitary matrices for diffuse sound Jot reverberators was conducted.

In parallel to the research work, several semester projects on near-field HRTFs and on binaural rendering were supervised by the author of this thesis.

Bibliography

V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The CIPIC HRTF Database. In *Proc. Workshop on Appl. of Sig. Proc. to Audio and Acoust.*, New Palz, NY, Oct. 2001. IEEE.

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

A. Avni and B. Rafaely. Interaural cross correlation and spatial correlation in a sound field represented by spherical harmonics. In *Proc. Ambisonics Symposium*, June 2009.

M. Barron. Measured early lateral energy fractions in concert halls and opera houses. *Journal of Sound and Vibration*, 232(1):79 – 100, 2000.

M. Barron and A. H. Marshall. Spatial impression due to early lateral reflections in concert halls: the derivation of a physical measure. *J. of Sound and Vibration*, 77 (2):211–232, 1981.

B. B. Bauer. Phasor analysis of some stereophonic phenomena. *J. Acoust. Soc. Am.*, 33:1536–1539, Nov. 1961.

D. R. Begault. Perceptual effects of synthetic reverberation on three-dimensional audio systems. *J. Audio Eng. Soc.*, 40(11), 1992.

D. R. Begault, E. M. Wenzel, and M. R. Anderson. Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *J. Audio Eng. Soc.*, 49(10), 2001.

J. S. Bendat and A. G. Piersol. *Random data: Analysis and measurement procedures.* John Wiley & Sons, New York, 1971.

J. C. Bennett, K. Barker, and F. O. Edeko. A new approach to the assessment of stereophonic sound system performance. *J. Audio Eng. Soc.*, 33(5):314–321, May 1985.

B. Bernfeld. Attempts for better understanding of the directional stereophonic listening mechanism. In *Preprint 44th Conv. Aud. Eng. Soc.*, Feb. 1973.

L. R. Bernstein and C. Trahiotis. The normalized correlation: Accounting for binaural detection across center frequency. *The Journal of the Acoustical Society of America*, 100(6):3774–3784, 1996.

J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization.* The MIT Press, Cambridge, Massachusetts, USA, revised edition, 1997.

C. Borss and R. Martin. An improved parametric model for perception-based design of virtual acoustics. In *Preprint 35th Int. Conf. Aud. Eng. Soc.*, Feb. 2009.

J. Breebaart and A. Kohlrausch. The perceptual (ir)relevance of HRTF magnitude and phase spectra. In *Proc. 110th AES Convention*, May 2001a.

J. Breebaart and A. Kohlrausch. The influence of interaural stimulus uncertainty on binaural signal detection. *The Journal of the Acoustical Society of America*, 109 (1):331–345, 2001b.

J. Breebaart and E. Schuijers. Phantom materialization: A novel method to enhance stereo audio reproduction on headphones. *IEEE Trans. on audio, speech and language proc.*, 16(8), Nov. 2008.

R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson. Measurement of correlation coefficients in reverberant sound fields. *J. Acoust. Soc. Am.*, 27:1071–1077, 1955.

J. F. Culling, H. S. Colburn, and M. Spurchise. Interaural correlation sensitivity. *The Journal of the Acoustical Society of America*, 110(2):1020–1029, 2001.

P. Damaske and Y. Ando. Interaural crosscorrelation for multichannel loudspeaker reproduction. *Acustica*, 27:232–238, 1972.

C. Faller. *Parametric Coding of Spatial Audio.* PhD thesis, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, July 2004. URL `http://library.epfl.ch/theses/?nr=3062`. Thesis No. 3062.

C. Faller. Multi-loudspeaker playback of stereo signals. *J. of the Aud. Eng. Soc.*, 54 (11):1051–1064, Nov. 2006.

C. Faller and M. Kolundzija. Design and limitations of non-coincidence correction filters for soundfield microphones. In *Preprint 126th Conv. Aud. Eng. Soc.*, May 2009.

C. Faller and J. Merimaa. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *J. Acoust. Soc. Am.*, 116(5): 3075–3089, Nov. 2004.

K. Farrar. Soundfield microphone: Design and development of microphone and control unit. *Wireless World*, pages 48–50, Oct. 1979a.

K. Farrar. Soundfield microphone - 2: Detailed functioning of control unit. *Wireless World*, pages 99–103, Nov. 1979b.

W. Gaik. Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling. *J. Acoust. Soc. Am.*, 94(1):98–110, July 1993.

B. Gardner and K. Martin. Measurements of a KEMAR dummy-head microphone. Technical report, MIT Media Lab, May 1994.

W. G. Gardner. Efficient convolution without input-output delay. *J. Audio Eng. Soc.*, 43(3):127–136, 1995.

W. G. Gardner. Reverberation algorithms. In M. Kahrs and K. Brandenburg, editors, *Applications of Digital Signal Processing to Audio and Acoustics*, chapter 2. Kluwer Academic Publishing, Norwell, MA, USA, 1998.

M. A. Gerzon. Periphony: Width-Height Sound Reproduction. *J. Aud. Eng. Soc.*, 21 (1):2–10, Jan. 1973.

M. M. Goodwin and J.-M. Jot. Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement. In *Proc. ICASSP-2007*, 2007a.

M. M. Goodwin and J.-M. Jot. Binaural 3-d audio rendering based on spatial audio scene coding. In *Proc. 123rd AES Convention*, 2007b.

Holophonic SA. Holophonic sound demos. Lugano, Switzerland, 2006. URL `http://www.holophonic.ch/`.

J. Huopaniemi. *Virtual Acoustics and 3D Sound in Multimedia Signal Processing*. PhD thesis, Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Finland, 1999. Rep. 53.

J.-M. Jot. An analysis/synthesis approach to real-time artificial reverberation. In *Proc. ICASSP-92*, volume 2, pages 221–224, 1992.

J.-M. Jot. Efficient models for reverberation and distance rendering in computer music and virtual audio reality. In *Proc. International Computer Music Conference*, pages 236–243, September 1997.

J.-M. Jot and A. Chaigne. Digital delay networks for designing artificial reverberators. In *Proc. 90th AES Convention*, 1991.

J.-M. Jot, V. Larcher, and O. Warusfel. Digital signal processing issues in the context of binaural and transaural stereophony. In *Preprint 98th Conv. Aud. Eng. Soc.*, Feb. 1995.

J.-M. Jot, A. Philp, and M. Walsh. Binaural simulation of complex acoustic scenes for interactive audio. In *Preprint 121st Conv. Aud. Eng. Soc.*, Oct. 2006.

V. Larcher, J.-M. Jot, and G. Vandernoot. Equalization methods in binaural technology. In *Preprint 105th Conv. Aud. Eng. Soc.*, September 1998.

R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman. The precedence effect. *J. Acoust. Soc. Am.*, 106(4):1633–1654, Oct. 1999.

R. Mason, C. Kim, and T. Brookes. Perception of head-position-dependent variations in interaural cross-correlation coefficient. In *Preprint 126th Conv. Aud. Eng. Soc.*, May 2009.

F. Menzer and C. Faller. Reduced localizability in sequences of narrowband noise bursts. In *Proc. 33. Jahrestagung für Akustik DAGA 2007*, March 2007.

F. Menzer and C. Faller. Obtaining binaural room impulse responses from B-format impulse responses. In *Preprint 125th Conv. Aud. Eng. Soc.*, October 2008.

F. Menzer and C. Faller. Binaural reverberation using a modified jot reverberator with frequency-dependent interaural coherence matching. In *Preprint 126th Conv. Aud. Eng. Soc.*, May 2009a.

F. Menzer and C. Faller. Investigations on modeling brir tails with filtered and coherence-matched noise. In *Preprint 127th Conv. Aud. Eng. Soc.*, October 2009b.

F. Menzer and C. Faller. Obtaining binaural room impulse responses from B-format impulse responses. *IEEE Trans. on Speech and Audio Proc.*, 2010a. in revision.

F. Menzer and C. Faller. Unitary matrix design for diffuse jot reverberators. In *Preprint 128th Conv. Aud. Eng. Soc.*, May 2010b.

F. Menzer, A. Brooks, P. Halje, C. Faller, M. Vetterli, and O. Blanke. Feeling in control of your footsteps: Conscious gait monitoring and the auditory consequences of footsteps. *Cognitive Neuroscience*, 2010. (submitted Nov. 2009).

J. Merimaa. *Analysis, Synthesis, and Perception of Spatial Sound – Binaural Localization Modeling and Multichannel Loudspeaker Reproduction*. PhD thesis, Helsinki University of Technology, 2006.

J. Merimaa. Modifications of HRTF filters to reduce timbral effects in binaural synthesis. In *Proc. 127th AES Convention*, Oct. 2009.

J. Merimaa and V. Pulkki. Spatial impulse response rendering I: Analysis and synthesis. *J. Aud. Eng. Soc.*, 53(12), 2005.

J. Merimaa, M. M. Goodwin, and J.-M. Jot. Correlation-based ambience extraction from stereo recordings. In *Proc. 123rd AES Convention*, Oct. 2007.

J. C. Middlebrooks, J. C. Makous, and D. M. Green. Directional sensitivity of sound-pressure levels in the human ear canal. *The Journal of the Acoustical Society of America*, 86(1):89–108, 1989.

Mo'Vision. Headphone-surround. URL `http://www.headphone-surround.de`.

R. Penrose. On best approximate solutions of linear matrix equations. *Proceedings of the Cambridge Philosophical Society*, 52:17–19, 1956.

V. Pulkki. Virtual sound source positioning using Vector Base Amplitude Panning. *J. Audio Eng. Soc.*, 45:456–466, June 1997.

V. Pulkki and J. Merimaa. Spatial impulse response rendering ii: Reproduction of diffuse sound and listening tests. *J. Aud. Eng. Soc.*, 54(1), 2006.

B. Rakerd and W. M. Hartmann. Localization of sound in rooms, II: The effects of a single reflecting surface. *J. Acoust. Soc. Am.*, 78(2):524–533, Aug. 1985.

Rec. ITU-R BS.1116.1. *Methods for Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Surround Systems.* ITU, 1997. http://www.itu.org.

D. E. Robinson and L. A. Jeffress. Effect of varying the interaural noise correlation on the detectability of tonal signals. *The Journal of the Acoustical Society of America*, 35(12):1947–1952, 1963.

M. R. Schroeder. Natural sounding artificial reverberation. *J. Aud. Eng. Soc.*, 10(3): 219–223, 1962.

E. A. G. Shaw. Transformation of sound pressure level from the free field to the eardrum in the horizontal plane. *The Journal of the Acoustical Society of America*, 56(6):1848–1861, 1974.

Starkey Laboratories Inc. Virtual barbershop. Eden Prairie, Minnesota, USA, 2007. URL `http://www.youtube.com/watch?v=IUDTlvagjJA`.

Starkey Laboratories Inc. Personal Communication, October 2008.

J. Stautner and M. Puckette. Designing multi-channel reverberators. *Computer Music Journal*, 6(1):52–65, 1982.

N. Toma, M. Topa, and E. Szopos. Aspects of reverberation algorithms. *ISSCS 2005.*, 2:577–580 Vol. 2, July 2005. doi: 10.1109/ISSCS.2005.1511306.

S. van de Par and A. Kohlrausch. Analytical expressions for the envelope correlation of certain narrow-band stimuli. *J. Acoustical Society of America Journal*, 98:3157–3169, dec 1995.

M. Vetterli and J. Kovačević. *Wavelets and Subband Coding.* Prentice Hall, Englewood Cliffs, New Jersey, USA, 1995.

Wikipedia. History of video game consoles (seventh generation), 2010a. URL `http://en.wikipedia.org/w/index.php?oldid=356738480`. [accessed 18-April-2010].

Wikipedia. Nintendo DS, 2010b. URL `http://en.wikipedia.org/w/index.php?title=Nintendo_DS&oldid=356527264`. [accessed 18-April-2010].

Wikipedia. Game boy, 2010c. URL `http://en.wikipedia.org/w/index.php?title=Game_Boy&oldid=356458267`. [accessed 18-April-2010].

Wikipedia. Sega Game Gear, 2010d. URL `http://en.wikipedia.org/w/index.php?title=Sega_Game_Gear&oldid=354194979`. [accessed 18-April-2010].

Wikipedia. iPhone, 2010e. URL `http://en.wikipedia.org/w/index.php?title=IPhone&oldid=356165798`. [accessed 18-April-2010].

Wikipedia. iPod Touch, 2010f. URL `http://en.wikipedia.org/w/index.php?title=IPod_Touch&oldid=355851699`. [accessed 18-April-2010].

Wikipedia. PlayStation Portable, 2010g. URL `http://en.wikipedia.org/w/index.php?title=PlayStation_Portable&oldid=356522085`. [accessed 18-April-2010].

H. Zuccarelli. Device for the spatial codification of sounds. *European Patent No. 0 050 100 A2*, April 1982. Filed: Oct. 1981.

H. Zuccarelli. Process for forming an acoustic monitoring device. *United States Patent No. 4,680,856*, Jul. 1987. Filed: Jan. 1986.

Curriculum Vitae

Fritz Menzer

Audiovisual Communications Laboratory
Swiss Federal Institute of Technology Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
fritz.menzer@a3.epfl.ch

Personal

Date of birth	16. 9. 1978
Nationalities	Swiss, German
Civil status	Married to Elham Menzer-Nasr Esfahani

Education

2006-2010	Ph.D. at Audiovisual Communications Laboratory, Swiss Federal Institute of Technology (EPFL): Binaural audio signal processing using interaural coherence matching.
1998-2004	Studies at EPFL, leading to a pre-degree (propédeutique) in physics and a master of science in communication systems. Graduated with an average of 5.8 on a scale from 1 to 6.
2004	Master's thesis on nonlinear modelling of the human vocal folds at Music Technology Group, University of York, UK.
2000	3 months of studies in physics at University of Iceland, Reykjavík.
1991-1998	High school in Frauenfeld, Switzerland. Graduated with a grade average of 5.8 on a scale from 1 to 6.
1994-1995	Exchange year at Dunn School, Los Olivos, California.

Professional Experience

2005-2007	Teacher at University of Art and Design, Lausanne.
2004-2005	Technical and teaching assistant at University of Art and Design, Lausanne.
1998	Six-month internship at Schmidhauser AG, Romanshorn. DSP programming in C and Assembler.
1996	Four-week internship at Sulzer Electronics AG, Winterthur. C programming (video capture on embedded PC platform).

Competitions and Awards

2005	Annaheim Prize for Master Thesis.
1998	Participation in the final round of the Swiss Youth and Science Contest.
1998	Participation in the 29^{th} International Physics Olympiad (Reykjavík, Iceland). Result: 2^{nd} Swiss (131^{st} out of 266 overall).
1997	Participation in the 28^{th} International Physics Olympiad (Sudbury, Canada). Result: 4^{th} Swiss (177^{th} out of 266 overall).

Academic Activities

2007-2009	Regularly supervised EPFL students for audio related semester projects.
2006-2009	Teaching assistant at Audiovisual Communications Laboratory, EPFL (4 semesters Signal Processing for Audio and Acoustics, 1 semester Adv. Signal Processing: Wavelets and Applications).
1999-2009	Member of the Swiss Study Foundation.

Publications

F. Menzer and C. Faller. "Unitary matrix design for diffuse Jot reverberators." In *Proceedings of the 128th AES Convention*, London, 2010, submitted in March 2010 (accepted).

F. Menzer and C. Faller. "Stereo-to-binaural conversion using interaural coherence matching." In *Proceedings of the 128th AES Convention*, London, 2010, submitted in March 2010 (accepted).

F. Menzer and C. Faller. "Investigations on an early reflections free model for BRIRs." *Journal of the Audio Engineering Society*, submitted in Nov. 2009.

F. Menzer, A. Brooks, P. Halje, C. Faller, M. Vetterli, and O. Blanke. "Feeling in control of your footsteps: Conscious gait monitoring and the auditory consequences of footsteps." *Cognitive Neuroscience*, submitted in Nov. 2009 (accepted).

F. Menzer and C. Faller. "Investigations on modeling BRIR tails with filtered and coherence-matched noise." In *Proceedings of the 127th AES Convention*, New York, 2009.

F. Menzer and C. Faller. "Binaural reverberation using a modified Jot reverberator with frequency-dependent interaural coherence matching." In *Proceedings of the 126th AES Convention*, Munich, 2009.

F. Menzer, C. Faller, and H. Lissek. "Obtaining Binaural Room Impulse Responses from B-Format Impulse Responses using Frequency-Dependent Coherence Matching." *IEEE Transactions on Audio, Speech, and Language Processing*, submitted in May 2009 (accepted).

F. Menzer and C. Faller. "Obtaining Binaural Room Impulse Responses from B-Format Impulse Responses." In *Proceedings of the 125th AES Convention*, San Francisco, 2008.

F. Menzer and C. Faller. "Reduced Localizability in Sequences of Narrowband Noise Bursts." In *Proceedings of the Annual Conference of the German Acoustical Society (DAGA)*, 2007.

F. Menzer, J. Buchli, D. M. Howard, and A. J. Ijspeert. "Nonlinear modelling of double and triple period pitch breaks in vocal fold vibration." *Logopedics Phoniatrics Vocology*, 31:36-42, 2006.

F. Menzer, J. Buchli, D. M. Howard, and A. J. Ijspeert. "Nonlinear modelling of double and triple period pitch breaks in vocal fold vibration." In *Proceedings of the 6th Pan European Voice Conference (PEVOC 6)*, 2005.

P. Polotti, F. Menzer, and G. Evangelista. "Inharmonic sound spectral modelling by means of fractal additive synthesis." In *Proc. COST G-6 Conference on Digital Audio Effects (DAFx-02)*, pages 127-132, 2002.

F. Menzer, "Method and apparatus for generating sound signals," *European Patent No. 1 239 453 (A1)*, Sept. 2002, Filed: March 2001.

Other Activities

2008-2009	Giving presentations at high schools to promote studies in computer science.
2006-2009	Coordination of the activities of the Swiss Physics Olympiad in the French-speaking part of Switzerland.
2000	Contribution of a chapter on the physics of sailing for the sailing instructors' manual edited by Swiss Sailing.
1999	Accreditation as a sailing instructor by Swiss Youth + Sport.

Languages

German	native language
English	fluent
French	fluent
Persian	basic
Italian	basic

Computer Skills

Systems	Mac OS X, Linux, Windows
Languages	C, C++, Objective-C, Matlab, Java, Processing, Assembler
Publishing	LaTeX, Illustrator, PHP, MySQL