Synthesis of binaural stimuli for a listening test on room acoustic perception

Technische Universität Berlin
Fakultät I - Gesiteswissenschaften
Fachgebiet Audiokommunikation

Masterarbeit
vorgelegt von Dmitry Grigoriev
Matrikelnummer: 346995

Erstgutachter: Prof. Dr. Stefan Weinzierl
Zweitgutachter: David Ackermann
Datum: 14. März 2017

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

Dmitry Grigoriev

_____

Datum, Unterschrift

Acknowledgements

First of all I want to thank Prof. Weinzierl and David Ackermann for their constant support and advice throughout the whole process of this work. Special thanks to Dr. Alexander Lindau for his mentoring during the beginning stages of this project.

I also want to thank the co-workers Christoph Böhm for his support with setting up the rendering system, Dr. Steffen Lepa for his help with the statistical analysis, Fabian Brinkmann for his help with the WhisPER preparation. As well as the whole team of the audio communication group for their motivating spirit.

Thanks to Vincent Arbisa for his assistance in conducting the comparative study of the replication algorithms.

I kindly thank Prof. Rindel from the Technical University of Denmark for allowing me to use their anechoic orchestra recordings.

In particular, I would like to thank my friends, especially Suzie for being there for me in stressful times and my family across the globe from San Francisco to Aktobe, especially my mother Elena, my father Andrey and my brother Ilya for their loving endless support and encouragement. I couldn't have done it without you.

Abstract

A ground truth investigation on room acoustic perception needs an extensive amount of diverse stimuli. The stimuli set has to represent a wide range of the physical aspects of room acoustics, as well as a variety of audio contents. Provided with anechoic audio material, binaural synthesis of virtually modeled acoustical environments offers an effective way to create such a set of stimuli. To ensure a diversity in the audio content of the stimuli, polyphonic anechoic audio material is a necessity. Since such polyphonic audio material is scarce to the scientific community and the recording of a polyphonic anechoic stimulus (i.e. orchestra or choir) is highly elaborate, synthetic replication of recordings offer a substantially alleviating measure to produce polyphonic anechoic audio material.

In this study, a broad set of binaural stimuli for audio contents speech, trumpet solo and orchestra was created. The recordings of the string instruments of the polyphonic orchestra stimulus were replicated using a novel segmentation track replication (STR) method with a successive phase correction post-processing algorithm to reduce phasing artifacts to simulate a string section sound. The variation of onset, pitch and sound level of the audio content according to their respective distribution is based on a state-of-the-art string section simulation method.

The STR method proposed here and four other track replication methods were tested for their ability to simulate a recorded violin section. Test subjects (n = 23) rated the similarity between the replicated and recorded violin sections in a double-blind triple-stimulus with hidden reference test. We found that the STR method offers similar results as the state-of-the-art track replication methods and can successfully simulate string section sound better than with a traditional chorus effect, in order to achieve a polyphonic anechoic orchestra stimulus.

This work also presents the experimental setup and automated procedure for a follow-up comprehensive study on the subjective qualities of room acoustics. A total of 70 room acoustical environments were created, based on the earlier work of the audio communication group of the Technical University Berlin. All virtual acoustical environments were simulated with correct instrument directivities, including spatial smoothing, pitch weighting and a diffuse-field equalization of each directivity prior to the simulation.

Zusammenfassung

Die experimentelle Untersuchung der raumakustischen Wahrnehmung benötigt eine umfangreiche Anzahl an Stimuli. Der Satz an Stimuli muss eine große Bandbreite an physikalischen raumakustischen Eigenschaften, sowie auch eine Vielfalt an Audioinhalten repräsentieren. Die Synthese von modellierten binauralen Raumimpulsantworten mit nachhallfreiem Audiomaterial bietet eine effektive Methode, um einen solchen Stimulisatz herzustellen. Um die Vielfältigkeit an Audioinhalten zu gewährleisten, benötigt man polyphones nachhallfreies Audiomaterial. Dieses Material steht der wissenschaftlichen Gemeinschaft nur in geringen Mengen zur Verfügung und der Aufwand, polyphone Quellen (z.B. ein Orchester oder Chor) nachhallfrei aufzunehmen, ist außerordentlich groß. Die künstliche Vervielfältigung von Audioaufnahmen stellt daher eine erhebliche Erleichterung für die Produktion von polyphonen nachhallfreien Audiostimuli dar.
Ein breiter Satz an binauralen Stimuli für drei Audioinhalte (Sprache, Trompete als Soloinstrument und Orchester) wurde im Zuge dieser Arbeit erzeugt. Die Simulation der Streichersektionen des polyphonen Orchesterstimulus wurde durch eine Vervielfältigung der Aufnahmen der Streichinstrumente mit einem neuartigen Segmentation Track Replication (STR) Verfahren, sowie einem sukzessiven frequenzspezifischen Phasenkorrekturprozess erzielt, um Kammfiltereffekten bei Addition der Tonspuren vorzubeugen. Die Veränderung des Onsets, der Tonhöhe und der Lautstärke des Audioinhalts, basierend auf einer der neuesten Simulationsmethode einer Streichsektion, erfolgt nach der jeweiligen statistischen Verteilung der drei Größen.
Das STR Verfahren und vier weitere Vervielfältigungsverfahren von Audioaufnahmen wurden nach ihrer Fähigkeit geprüft, eine aufgenommene Violinsektion zu simulieren. Versuchspersonen (n = 23) haben die Ähnlichkeit zwischen künstlich vervielfältigten Violinsektionen und einer aufgenommenen Violinsektion in einem ABC/HR Test bewertet. Die Ergebnisse zeigen, dass das STR Verfahren qualitativ ähnliche Resultate zu den neuesten Vervielfältigungsverfahren erzielt und damit Streichersektionen besser simuliert werden können, als mit einem traditionellen Choruseffekt, um einen polyphonen nachhallfreien Orchesterstimulus zu erzeugen.

Weiterhin wird in dieser Arbeit der Versuchsaufbau und das Versuchsverfahren für eine nachfolgende umfangreiche Studie zu den subjektiven raumakustischen Qualitäten vorgelegt. Aufbauend auf den vorangegangenen Arbeiten des Fachgebiets Audiokommunikation und -technologie der Technischen Universität Berlin wurden insgesamt 70 raumakustische Umgebungen erstellt. Die Simulation dieser virtuellen raumakustischen Umgebungen wurde mit korrekten Richtcharakteristiken der Instrumente vollführt, bei voriger räumlicher Glättung, Tonhöhengewichtung und Diffusfeldentzerrung der einzelnen Richtcharakteristiken.

Contents

1 Introduction **1**

2 State of the Art **4**

3 Methods **7**
3.1 Simulation of the rooms 7
3.2 Instrument Directivity 10
3.3 Anechoic Audiomaterial 15
3.3.1 Phase Vocoder 16
3.3.2 Phase Problems and Scaled Phase Locking Method 19
3.3.3 String Section Sound 21
3.3.4 Segmentation Track Replication Method 22
3.3.5 Postprocessing 25
3.4 Test interface 28
3.5 Experimental Setup 29
3.5.1 Listening environment 29
3.5.2 Headtracker 30
3.5.3 Software 30
3.5.4 PureData patch 31
3.5.5 SoundScape Renderer 31
3.5.6 Ardour 32
3.5.7 JACK Connections 32
3.5.8 Randomised Block Design 33
3.6 Listening Test Procedure 34
3.7 Comparison of Different Replication Methods on String Section Sound . 35
3.7.1 Investigated Replication Methods 35
3.7.2 Aquisition of the Reference Signal and its Distributions 35
3.7.3 Test Method 36

4 Results **39**

5 Discussion **42**

A Appendix **51**

A.1 Program Coding 51
A.1.1 MATLAB code 51
A.1.2 SHELL scripts 51
A.1.3 PureData 51
A.1.4 Ardour, ASD, XML and txt files 52
A.2 Sociodemographic and expertise Survey 52
A.3 Distribution of BR, G and EDT 55
A.4 Room Acoustical Quality Inventory 55
A.5 Room Acoustical Parameters 57

List of Figures

3.1 Source and receivers constellation for the model of the Gewandhaus in Leipzig. Highlighting the positions of the source and receivers, as well as their distance from the floor and the minimum distance d_{min} of double the critical distance r_H from the source to the receiver as the grey circle area. 8

3.2 Orchestra stage plan: The orchestral parts are divided by the colored areas (left: Violin I, 2nd-left: Violin II, mid-green: Viola, right-bottom: Violoncello, right: Double Bass, yellow: Horn, light-blue: Percussion, brown: Flute, white: Oboe, orange: Clarinet, red: Trumpet, beige: Bassoon & Contrabassoon) and every source is represented by a circle (blue: original recordings, red: replicated recordings (see Sec. 3.3.3) 10

3.3 The absorption α_m against the volume V of the 35 selected rooms, with each color representing a different room geometry. 11

3.4 Balloonplots of the spherical harmonics with increasing order n from top (n = 0) to bottom (n = 4), left plots are all the imaginary parts, right plots are all the real parts. Cyan shades representing positive values of the function, and magenta shades representing negative values of the function (Rafaely (2015)) 12

3.5 3D directivity plots of two consecutive tones (F#4 and G4) from a Violin, illustrating how drastically the directivity can change from one pitch to the next. (red: positive values, blue: negative values) 13

3.6 Pitch distributions of Violoncello and Horn in fortissimo over 128 MIDI notes (C0-G10). (blue: normalized pitch distribution, red: probability density function using kernel density estimation, black: discrete probability density according to the respective pitch range of the instrument) 14

3.7 Detected silence onsets (*red*) for segmentation of the recording into tones (first segment) and short tone passages (second and third segment). *blue*: time signal of the recording, *black*: function course of the sound level (y-axis arbitrary units). 23

3.8 Spectra of one tone taken from two different recordings of the same violin (black and blue) and a replication of the first recording (black) with the STR method (red). 26

3.9 WhisPER RAQI Questionnaire GUI interface: each item is prompted separately on a continuous ordinal scale with the item characteristics at each end (scale labels). . . . 28
3.10 Information flow of the listening experiment. (black: audio content, blue: OSC-messages, red: head tracking, green: subject input) . . . 29
3.11 AKG K1000 extra-aural headphones with Razor AHRS head tracker mounted on the rim. . . . 30
3.12 Screenshot of the digital audio workstation Ardour, the audio content for each instrument is placed on separate audio tracks and the playback is looped. . . . 32
3.13 Possible JACK connection, from Ardour to SSR (BRS1), from SSR to headphone equalisation, from headphone equalisation to system. System input leads directly to the output signal on the AKG headphones. (left: outputs, right: inputs) . . . 33
3.14 Excerpt of the Randomized Block Design by Edgar (row: subject session corresponding to the subject ID, column: treatment condition room-content-position, Voice: speech stimulus, Solo: trumpet stimulus, Orch: orchestra stimulus) . . . 33
3.15 Recording setup of the violin player in the anechoic chamber for the comparative test of different replication algorithms. . . . 36
3.16 WhisPER ABC/HR test interface for comparison of different replication methods. Slider position represents similarity of the stimulus signal to the reference signal. . . . 37

4.1 Distribution of the residuals. *left*: Distribution of the residuals accross the predicted values of the model. *right*: Distribution of the residuals with a normal distribution fit. . . . 40
4.2 Ratings for the similarity to the reference signal of all subjects and test conditions with regard to the main effects method (left), ratio (middle) and reverb (right) described in the text. The rated similarity ranges on the interval [0;-1] with 0 equal to "identical" and -1 equal to "different". Mean ratings are indicated by circles. Standard errors are displayed by solid vertical lines. Connecting lines between conditions are provided to improve readability. . . . 40

A.1 PureData patch receiving OSC messages from WhisPER to select the correct session (left), the correct condition (center) and control the playback of the audio (right) . . . 52
A.2 Screen shot of survey on musical and room acoustical expertise and sociodemographic (page 1). . . . 53
A.3 Screen shot of survey on musical and room acoustical expertise and sociodemographic (page 2). . . . 54

A.4 Scatter-plot matrix of room acoustical parameters Bass Ratio BR, Sound Power G and Early Decay Time EDT of the 35 selected rooms (circle: receiver position 1, cross: receiver position 2, colors indicate room geometry). On the diagonal, the axes of the regarded parameters are displayed, indicating which of the three parameters are compared per scatter-plot (i.e. middle-left: y-axis G vs. x-axis BR; bottom-left: y-axis EDT vs. x-axis BR) 55

List of Tables

3.1 Mean, maximum and minimum values for the room acoustical parameters of the selected 35 rooms. 8
3.2 Overview over the instrumentation of the orchestra, 66 sources divided in string, percussion and wind instruments. 9
3.3 Standard deviation for the normal distributions of onset, pitch and amplitude variation taken from Recke (2011). 24
3.4 Means and standard deviation for the normal distributions of onset, pitch and amplitude variation fitted to the measured data from the reference signal. 38

4.1 Results for main effects (method, reverb and ratio) and interactions in the measurement of similarity of replicated signals to the reference signal. 41

A.1 Room Acoustical Quality Inventory (RAQI) in german language. (o: only used for orchestra stimulus; s: only used for speech stimulus; x: not included in Lepa et al. (2017) 56
A.2 Room acoustical parameters, room volumes and critical distances for the 35 selected rooms in frontal position with a single source taken from Ackermann and Ilse (2015) (V: Volume, r_H: Critical Distance, EDT: Early Decay Time, T_{30}: Reverberation Time (30 dB), C_{80}: Clarity, D_{50}: Definition, G: Sound Power Factor , T_s: Centre Time 57

Introduction

The physical model of *room acoustics* can be described by room acoustical parameters defined in DIN 3382-1 (2009). The measurement of these parameters is detailed in DIN 3382-1 (2009) and only requires the necessary measurement equipment. The psychological aspect of room acoustics is an ambiguity unsolved as of today. So far there is no scientific consensus on the definition or measurement of *subjective qualities* that describe a room acoustical environment on a perceptual level.
The intrigue about the perceptual qualities of room acoustics dates back to the beginning of the 20th century when Sabine (1906) examined the most preferred reverberation for piano music in "moderate sized" rooms. During that time and up until the first dummy head recordings by Plenge et al. (1969) all studies involving listening experiments on room acoustics were elaborate in scope and had to be conducted inside the regarded physical rooms.
The perceptual quality of Sabine's study was the preference of a room acoustical environment for a specific audio content. It was measured depending on the reverberation time of the room. Sabine examined five rooms in the same building with different volumes and absorption surfaces. In the study, five to seven test subjects listened to live piano music in each room, while the reverberation in the rooms was changed by adding or removing cushions, until the test subjects agreed on a "satisfactory" sound experience. Evidently this experimental approach has various issues. To begin with, the consensus on what constitutes "satisfactory conditions" of the room can vary between the groups of test subjects. Moreover, the amount of test subjects per room varies from room to room and is not sufficient to provide a general rating. Furthermore the subjective quality of preference of the individual test subjects is compromised by the study's focus on the consensus opinion of the group, rather than the individual. A more precise approach would involve each test subject to undergo the experiment alone. This would result in practical difficulties, however, as the piano musician would have to play the audio stimuli consistently throughout all repetitions of the experiment. An ideal experiment has to be able to be reproduced at any time. Studies using live played audio stimuli will never be able to meet this criteria in the strictest sense. An example for that is the study by Hawkes and Douglas (1971). Hawkes and Douglas examined a selection of subjective qualities proposed by Beranek (1962) by having questionnaires filled out at concerts by different auditoria. The audio stimuli were performed by full

symphonic orchestras with different specifics (e.g. with or without vocal or instrumental soloists or a chorus) and the audio content ranged across a wide spectrum of classical music. Both the audio stimuli and the test subjects in this study are both specific to the room. None of the test subjects were able to give ratings to different rooms. This highlights a problem for all studies involving test subjects rating subjective qualities inside physical rooms. A test subject's journey from one room to another can have an undesirable impact on the subject and skew the results of the experiment, for instance due to the subject's potential exhaustion or other psychological influences. A scientific experiment needs to be conducted in a controlled environment, while minimizing the amount of possible exterior influences on the subject.

A virtual reproduction of the acoustics of different rooms in one singular experimental room could eliminate the above mentioned problems. The virtual reproduction of the room will always be an approximation of the physical room, but over the last round-robins on room acoustical computer simulation, it has been demonstrated that simulation offers a reliable reproduction of the physical room (Vorländer (2008)). If the goal of the simulation is not an attempt to exactly reproduce a physical room, but to investigate a room's acoustical qualities, then the results from this assessment are as valid for the evaluation of the room acoustical qualities, as the results from an exact simulation of the physical room or results taken from the actual physical room would be.
There are various technologies for room acoustical simulation. A selection of different reproduction systems used for room acoustical investigations is presented in the next chapter. This study uses model based simulated binaural room impulse responses (BRIRs) for reproduction by headphones with a head tracker described in Sec. 3.1. & 3.5
Another advantage to using computer simulation is the ability to reproduce the same audio content in every desired room acoustical environment. The only requirement is that the audio content was recorded in an anechoic chamber. If we have the room impulse response (RIR) of the room, which was recorded or modeled with a source in one specific position and the microphone (or receiver) in another specific position, then the convolution of the anechoic recording material with a RIR of the room will result in the anechoic audio material playing in the position of the RIR source and the listener placed in the position of the RIR receiver. Anechoic monophonic audio material is widely available for free use[1]. A considerable challenge is the recording of anechoic polyphonic audio material (i.e. symphonic orchestra, choir, big band, etc.) due to technical difficulties of making a group recording, while requiring the separate audio tracks to be without crosstalk and with a high signal-to-noise ratio. Here audio track replication algorithms offer a substantially alleviating measure, allowing to simulate string sections out of singular recordings.

The present study offers an experimental setup based on room acoustical computer simulation for a ground truth investigation on room acoustical perception. The experiment has to include a variety of audio content stimuli to offer an insight on how different audio

[1] various studies (i.e. Vigeant et al. (2008) or Lokki et al. (2008)), http://www.openairlib.net/anechoicdb, etc.

contents influence subjective room acoustical qualities. A speaker, a solo trumpet and an orchestra were chosen as a compromise between test procedure length and variety of audio content. The simulation process is taken from Ackermann and Ilse (2015) which allows a wide range room selection as well as different source-receiver configurations. Furthermore, the correct instrument directivities are applied to the respective sources in the simulation for a correct spatial emission of the sound. In a preceding task the directivities undergo procedures, such as movement and energy averaging, as well as tone weighting. A more concentrated focus is drawn on the track replication process, where a detailed description of the main instrument of the process - the phase vocoder - is presented, and problems with the phasing effect and countering developments are discussed. A new segmentation track replication method is introduced and compared with current track replication methods and a post-processing method to reduce the phasing effect is introduced. The results of the comparison and post-processing are discussed in chapter 5. The MATLAB based test interface infrastructure for the investigated subjective qualities is presented. The subjective qualities are taken from the results of preceding focus group sessions of room acoustical experts. After a complete overview over each component of the experimental setup, the experimental procedure is presented and discussed.

State of the Art

21st century studies on subjective qualities of room acoustics have mainly been using loudspeaker arrangements for reproduction of room acoustical environments. Berg and Rumsey (2006) examined how different recording techniques influenced the spatial quality of an audio system. Using recordings of six different audio contents, from solo singer to symphonic orchestra and environmental recordings, with varying 1-5 channel microphone recording techniques to offer a wide range of the spatial dimension, the reproduction was realized with a five channel loudspeaker arrangement, three speakers (left, center, right) in front of the subject and two speakers to the sides positioned 110 degrees from the front axis, all speakers located in 2m distance from the subject. This setup allows for an exploratory investigation to gather data on spatial quality, but its ability to reproduce room acoustical environments for ground truth investigation is questionable. Pätynen et al. (2009) created a "loudspeaker orchestra" for studies of concert halls which was tested with in-situ listening. 24 loudspeakers, consisting of three different models of Genelec (17 x 1029A, 5 x 8030A, 2 x 1032A) were arranged approximate to a typical symphonic orchestra in American seating. Each loudspeaker reproduced the sound of one instrument of the orchestra. The instruments were recorded in an anechoic chamber, however, there were not enough recordings of the string instruments to accurately represent a full orchestral setting, so the strings were amplified to achieve the desired balance between the orchestral instruments. The typical blending of the sound of multiple string players was difficult to create, due to the small amount of speakers and lack of different string recordings. The overall impression was that the orchestra sound was "too thin". Another difficulty consisted in the major difference of the directivity of the loudspeakers and the directivities of the represented instruments. The in-situ listening comments suggested the strings lacked in brightness, which was improved by turning the loudspeakers of the strings so they faced the auditorium. These observations indicate that a correct directivity is desired to represent the simulated instrument.

Lokki et al. (2012) further developed the loudspeaker orchestra to 33 loudspeakers in total. This enhanced loudspeaker orchestra was utilized to measure 9 concert halls with the same listening position to undergo an exploratory study examining subjective qualities on room acoustical perception. The recording of the loudspeaker orchestra in the concert halls was done with a 6-channel intensity probe, retrieving first order B-format

impulse responses. For the reproduction in the listening laboratory the recordings underwent a spatial impulse response rendering (SIRR) algorithm, dividing the B-format impulse responses into individual impulse responses to be reproduced by a 14-channel spatial sound reproduction system. Furthermore, each recording of the string sections was replicated with modulation in amplitude, pitch and note onset (more on the replication process in Sec. 3.3.3) resulting in a string section simulation (Pätynen et al. (2011)). In a different study Pätynen and Lokki (2010) used the loudspeaker orchestra with 34 sources to investigate the differences between in-situ recordings of two concert halls recorded similar to the above described study and auralizations of the same concert halls simulated with modeled BRIRs with the Odeon Software (ODEON (2010)). The reproduction system was a loudspeaker array consisting of eight Genelec 1029A loudspeakers, evenly distributed on a circumfence around the subject in 2 m distance to the subject in an anechoic chamber. The auralizations underwent an overall sound level correction at each source point and the results of the listening tests indicate that the auralizations are comparable to the real room recordings.

In-situ binaural recordings of rooms offer a useful tool for room acoustical research. Lokki and Järveläinen (2001) compared binaural in-situ recordings with auralization with modeled BRIR. The in-situ recordings were done with real-head recordings similar to (Møller et al. (1996)). For the auralization the DIVA system was used (Savioja et al. (1999)). The results show that a natural sounding virtual auditory environment for simple room geometry is possible with drawbacks in the auralization method, which needs further improvement.

The first approaches in creating this experimental setup for examining subjective qualities in concert halls was done by Lehmann and Wilkens (1980) using dummy head recordings with the correct head-related recording and reproduction technique. The dummy head's were created with "as much resemblance to a human head as possible" (Kürer et al. (1969)). Neumann KM 83 in ear condenser microphones were used for recording and "[a] special kind of electrodynamic headphone arrangement [...] built by Sennheiser" was used for the reproduction. The dummy head was stationary during the recordings, thus only the frontal direction was recorded. This stationary 2-channel recording results in problems with localization, overestimation of reverberance (Kürer et al. (1969)) and confusion of frontward and backward incidents (Weinzierl (2008)).

Weitze et al. (2002) directly compared binaural in-situ recordings of performances in two mosques and a byzantine church with measured BRIRs of the rooms convolved with the anechoic recordings of the performances. The results show a good resemblance between binaural in-situ recordings and auralizations for perceived reverberation, distance and the 3D experience. No listening tests were made for these assumptions, however further comparison between binaural in-situ recordings and auralizations have shown it is rather difficult to devise a good method for comparison of these two (Saher et al. (2006)). The in-situ recordings appear more natural compared to the auralizations, differing in subjective descriptors of 'sense of space', 'reverberance' and 'timbre'. This allows for a possible re-creation of the real environment according to the mentioned descriptors with a level between 'slightly different' and 'rather different'.

The perceptual differences in the arrangement of the sources in the auralization was investigated by Vigeant et al. (2008), where a symphonic orchestra was seated in two typical orchestral formations and a randomized one. The anechoic recordings of the orchestra were done for each instrument with a 5-channel recording system. Four microphones were placed to the front, left, right and back of the musician and the 5th microphone was placed above the musician. For the auralization a BRIR was calculated for each source and convolved with the corresponding anechoic recording and later mixed together for the final auralization mix. This was done for the frontal one channel recording and for the five channel recording separately. The results show that the one vs five channel comparison provides the ability to distinguish between different seating arrangements for the orchestra, thus indicating the influence of the directivity of the sources on the ability to detect the arrangement of sources within the auralization.

Nilsson and Ekman (2009) examined binaural recordings and auralization with modeled BRIR in psychoacoustical tests. The binaural recordings were done in an ordinary classroom and an auditorium. The auralization system developed by Kajastila et al. (2007) uses prioritized beam tracing and is a real-time acoustic simulation module. Both the recording and the modeling were done for only the frontal direction, with no dynamic binaural reproduction during the listening experiments. The results show the subjects ability to discriminate between binaural recordings and auralization with modeled BRIR, although not being able to identify which of the auralizations was using recorded and which modeled BRIRs. This indicates that modeled BRIR auralizations can be perceived as a "realistic" alternative to recorded BRIR auralizations. Lindau and Weinzierl (2012) have shown that recorded BRIR auralizations satisfy a strict plausibility test in comparison with real stimuli. In addition Brinkmann et al. (2014) tested dynamic binaural synthesis for authenticity in comparison with real stimuli using speech and noise stimuli in a highly sensitive ABX test. Here subjects predominantly failed to reliably distinguish the simulated stimuli from the real stimuli for speech signals, but were able to recognize the simulated noise stimulus in all cases. Therefore, dynamic binaural synthesis offers to be a useful method in creating virtual room acoustical environments for this study with speech and music stimuli.

Methods

As of today scientifically established room acoustical parameters are only of physically measurable nature and can be found in DIN 3382-1 (2009). But the physical attributes of the room alone are not sufficient to describe the perception of room acoustics (Weinzierl and Vorländer (2015)). The properties of the audio content and source, as well as the personal preference and experience of the listener have an influence on the perceived room impression. Therefore a listening test for room acoustical perception has to offer not only a variety of different rooms, but also various audio contents and source-receiver constellations to be tested in these rooms. So far room acoustical studies on psychological attributes of room acoustical perception have fallen short of being able to test the attributes for their relation to external variables (physical attributes, personal experience, source properties, etc), their reliability across time and individuals, their ability to distinguish between rooms and the item difficulty of the attributes (Weinzierl and Vorländer (2015)). In order to satisfy these demands, 35 virtually modeled rooms were simulated, each with two source-receiver configurations, offering 70 different room acoustical environments (Grigoriev et al. (2016)).
The following chapter describes the room simulation involving the source and receiver positions and room selection. The calculation of the spatially smoothed and pitch weighted directivities is presented. The choice of the anechoic audio material and a new approach for track replication for simulation of string sections is introduced. The design and control of the listening test is described, as well as the test procedure. Finally a comparative study between different track replication algorithms is presented.

3.1 Simulation of the rooms

In order for the test subject to experience different room acoustics in one experimental room, an experimental setup must be created where the subject is able to listen to the same audio content reproduced in different rooms without having to physically move from one room to the next. Different methods using virtual acoustics have been presented in Chapter 2. The dynamic binaural synthesis offers a highly plausible simulation of room acoustical environments (Lindau et al. (2007), Lindau and Weinzierl

(2012)). Hereby a head tracker follows the listener's head position and communicates to the auralization software which corresponding binaural room impulse response (BRIR) to convolve with the anechoic audio material. The BRIRs can be either recorded in the physical rooms or simulated using models of the rooms. This study uses BRIRs of modeled rooms created by Ackermann and Ilse (2015).

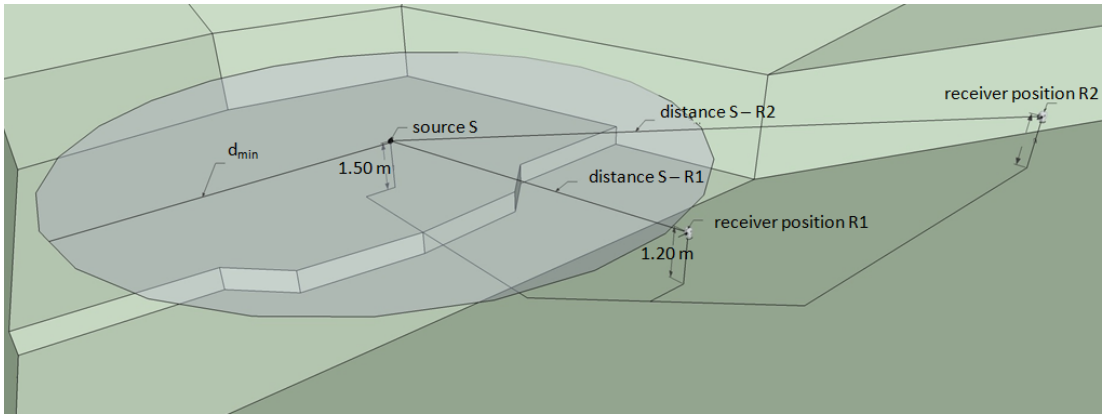

Figure 3.1: Source and receivers constellation for the model of the Gewandhaus in Leipzig. Highlighting the positions of the source and receivers, as well as their distance from the floor and the minimum distance d_{min} of double the critical distance r_H from the source to the receiver as the grey circle area.

In their study the authors created geometrical models of the rooms in SketchUp (CAD software) while specifying various surfaces of the room with absorption levels for future simulation. The simulation of the BRIRs is done with a real-time auralization plug-in for architectural design and education called RAVEN, developed by Schröder (2011) at the Institute of Technical Acoustics at the RWTH Aachen. Here, the Animation Module is used by simulating dirac impulses at the source position and recording the simulated sound at the receiver position, using head related transfer functions (HRTFs) of the FABIAN head and torso simulator (Lindau and Weinzierl (2007)). Only the azimuth orientation is recorded with an angular resolution of 2° resulting in 180 BRIRs per receiver position.

Based on DIN 3382-1 (2009) and DIN 3382-2 (2008) the source position for speaker and solo trumpet is chosen to represent the typical position for the use of the regarded

	V [m^3]	**A** [m^2]	$\mathbf{r}_H$ [m]	**EDT** [s]	**T30** [s]	$\mathbf{C}_{80}$ [dB]	$\mathbf{D}_{50}$	**G** [dB]
Mean	10052	3631	3,52	1,69	2,06	2,69	0,50	7,90
Max	43790	10512	8,43	4,85	7,08	11,19	0,86	18,91
Min	166	215	0,87	0,46	0,56	-6,62	0,1	-1,47

Table 3.1: Mean, maximum and minimum values for the room acoustical parameters of the selected 35 rooms.

room. The acoustical center of the source was placed 1.5 m above the floor. The two receiver positions are placed in the auditorium at least 2 m apart from each other and their acoustical centers 1.2 m above the floor. The minimum distance between source and receiver is double the critical distance r_H to ensure the sound pressure level of the reverberant sound to be higher than that of the direct sound. The source is in the median plane of the receiver at 0° azimuth orientation of the receivers. The elevation orientation is constant at 0°.
The 66 orchestra sources have their acoustic centres 1.2 m above the ground and are placed according to the seating arrangement of the Berlin German Symphonic Orchestra (see Fig. 3.2). The 66 sources are divided into the various orchestral instruments and can be seen in Tab. 3.2. The stage plan of the orchestra is not changed between rooms and the receiver positions are identical to the above mentioned solo source-receiver constellations.

Table 3.2: Overview over the instrumentation of the orchestra, 66 sources divided in string, percussion and wind instruments.

Group	Instrument	Amount
Strings	Violin I	12
	Violin II	11
	Viola	10
	Violoncello	9
	Double Bass	8
Percussion	Timpani	1
	Triangle	1
Winds	Flute	2
	Oboe	2
	Bassoon	2
	Contrabassoon	1
	Horn	3
	Trumpet	2
	Clarinet	2

Ackermann and Ilse (2015) offered a selection of 49 different room models, 35 of which were chosen for this study by cluster analysis on the room acoustical parameters Bass Ratio BR, Sound Power Factor G, Lateral Fraction Cosine J_{FL} and Reverberation Time T_{60}. Fig. 3.3 shows the selected rooms' distribution along the quantities absorption, volume and the room geometry. Offering a wide variety of rooms in these three degrees of freedom. The resulting variance in the three room acoustical parameters Bass Ratio BR, Sound Power G and Early Decay Time EDT can be seen in A.3. Tab. 3.1 offers an overview over mean, maximum and minimum values of the room acoustical parameters over the selected rooms. A detailed table of the room acoustical parameters of each

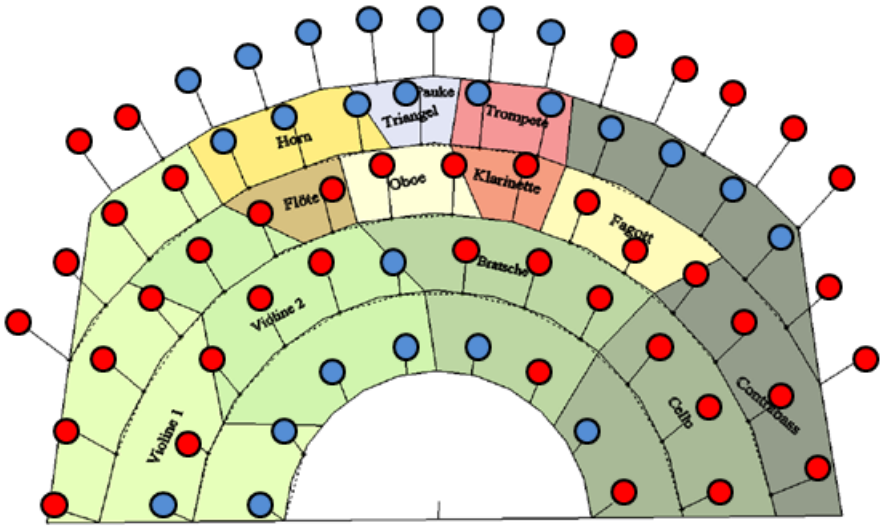

Figure 3.2: Orchestra stage plan: The orchestral parts are divided by the colored areas (left: Violin I, 2nd-left: Violin II, mid-green: Viola, right-bottom: Violoncello, right: Double Bass, yellow: Horn, light-blue: Percussion, brown: Flute, white: Oboe, orange: Clarinet, red: Trumpet, beige: Bassoon & Contrabassoon) and every source is represented by a circle (blue: original recordings, red: replicated recordings (see Sec. 3.3.3)

simulated room is given in Tab. A.5. For the multisource simulation of the orchestra only 25 of the 35 rooms are used because the orchestral stage plan could not fit onto every stage.

3.2 Instrument Directivity

Every musical instrument has its unique directivity, which describes the amount of sound energy emitted in a specific spatial direction. Integrating the directivity in the simulation for each respective instrument can only increase the plausibility of the simulation and thus strengthen this experiment's validity. A directivity can be described as a function f(θ, ϕ) on the unit sphere. Every function on the unit sphere can be written as a weighted sum of a set of basis functions called the spherical harmonics $Y_n^m(\theta, \phi)$,

$$Y_n^m(\theta,\phi) = \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}} P_n^m(cos\theta)e^{im\phi} \tag{3.1}$$

where P_n^m are the associated Legendre functions, m the function degree, n the function order, θ the elevation angle and ϕ the azimuth angle (Rafaely (2015)). Due to the

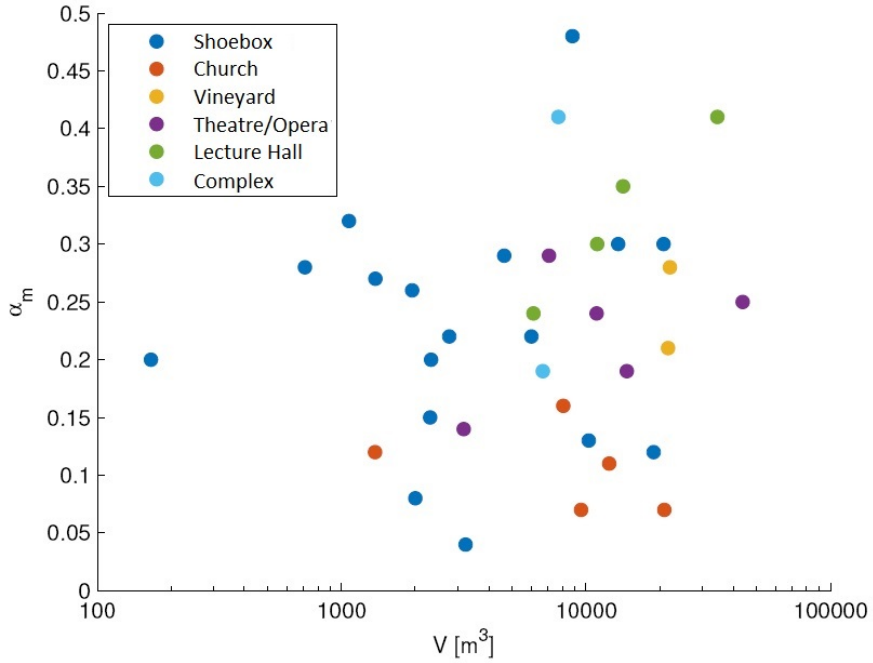

Figure 3.3: The absorption α_m against the volume V of the 35 selected rooms, with each color representing a different room geometry.

phase term the spherical harmonics are complex functions and can be divided into real and imaginary parts. Fig. 3.4 shows how the different imaginary and real parts are responsible for specific characteristics (monopole, dipole, etc.) of the complete function f(θ, ϕ). The associated Legendre functions P_n^m are a specific differentiation of Legendre polynomials P_n.

$$P_n(x) = \frac{1}{2^n n!}\frac{d}{dx^n}(x^2+1)^n \tag{3.2}$$

$$P_n^m(x) = (-1)^m(1-x^2)^{m/2}\frac{d}{dx^m}P_n(x), \quad x \in [-1,1] \tag{3.3}$$

The Legendre polynomials are a complete and orthogonal set of basis functions over the line section x ∈ [-1,1] and are the partial solutions to the Legendre differential equation (Bronstein et al. (2008)). For more detailed information see Rafaely (2015).

Hence the directivity function f(θ, ϕ) yields:

$$f(\theta,\phi) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} f_{nm}Y_n^m(\theta,\phi) \tag{3.4}$$

where f_{nm} are the weights to the corresponding spherical harmonics $Y_n^m(\theta, \phi)$. So a

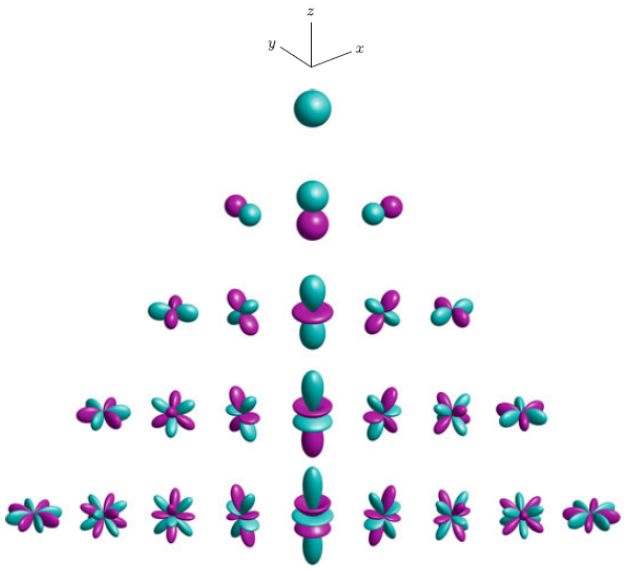

Figure 3.4: Balloonplots of the spherical harmonics with increasing order n from top (n = 0) to bottom (n = 4), left plots are all the imaginary parts, right plots are all the real parts. Cyan shades representing positive values of the function, and magenta shades representing negative values of the function (Rafaely (2015))

vector of weights f_{nm} can represent the spatial function of the directivity of a specific frequency. A note played on an musical instrument consists of more than one frequency so we will need multiple vectors of weights to describe the full spectral information. Since a note's main spectral information can be described by the amplitude of the fundamental tone and its overtones.
The static directivities for each note were obtained by recordings in the anechoic chamber of the TU Berlin (Pollow et al. (2010)), resulting in 16 to 56 directivities per instrument, depending on the respective instrument's pitch range. The order n = 5 was provided by the data. To represent a plausible performance of a musician, the movement of the source as well as the pitch that is being played must be taken into account. Depending on the pitch that is being played by the instrument, the directivities change dramatically (Fig. 3.5). Ideally, the directivity should change dynamically in the auralization, but the directivity is already incorporated in the RAVEN simulation making this ideal solution not possible. Since the directivity data is given per pitch, we can include the pitch variation by collecting all pitch directivities into one single directivity per instrument while weighing each pitch directivity from a representative pitch distribution. The pitch data from Quiring and Weinzierl (2016) of the symphonies of L. v. Beethoven are used as a representative pitch distribution. Two pitch distributions for violoncello and horn are shown in Fig. 3.6. A direct weighing with the values from the normalized pitch distribution would lead to extreme variations in the weighted directivity. The probability density function is hence evaluated using kernel density estimation. The discrete probability density function for the respective pitch range of each instrument is used for weighing the directivities. A static directivity has

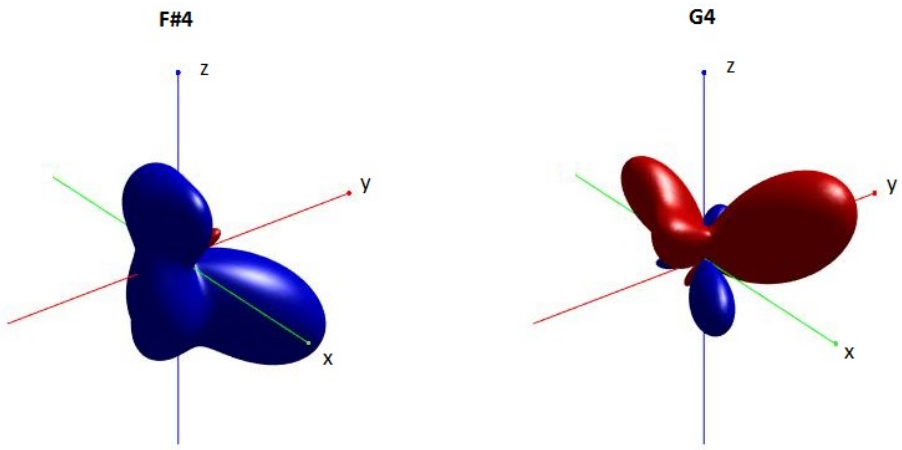

Figure 3.5: 3D directivity plots of two consecutive tones (F#4 and G4) from a Violin, illustrating how drastically the directivity can change from one pitch to the next. (red: positive values, blue: negative values)

in some specific directions an output of zero and other specific directions a maximum output that could cause undesirable audible artifacts due to lack of sound energy or interference while generating the BRIRs. The averaging over the movement by rotation of the directivity around its center would lead to a smoothing of the directivity characteristics. For this purpose, rotational data is taken from motion tracking recordings of musicians playing classical instruments (Steger et al. (2015)). The rotations are done using so-called Euler angles (α, β, γ), first by counter-clockwise rotation of the z-axis by angle γ, then by counter-clockwise rotation of the y-axis by angle β and finally by counter-clockwise rotation of the z-axis by angle α (Rafaely (2015)). A rotation can be calculated with the Euler rotation matrices R_z and R_y:

$$\boldsymbol{x'} = \boldsymbol{R_z(\alpha) R_y(\beta) R_z(\gamma) x} \tag{3.5}$$

$$\boldsymbol{R_z(\alpha)} = \begin{bmatrix} \boldsymbol{cos\alpha} & \boldsymbol{-sin\alpha} & \boldsymbol{0} \\ \boldsymbol{sin\alpha} & \boldsymbol{cos\alpha} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{1} \end{bmatrix} \tag{3.6}$$

$$\boldsymbol{R_y(\beta)} = \begin{bmatrix} \boldsymbol{cos\beta} & \boldsymbol{0} & \boldsymbol{sin\beta} \\ \boldsymbol{0} & \boldsymbol{1} & \boldsymbol{0} \\ \boldsymbol{-sin\beta} & \boldsymbol{0} & \boldsymbol{cos\beta} \end{bmatrix} \tag{3.7}$$

Each pitch weighted directivity was rotated every 100 ms according to each instrument's rotation data from Steger et al. (2015) respectively and averaged over all frames, resulting in pitch weighted and spatially smoothed directivities. Finally the directivities

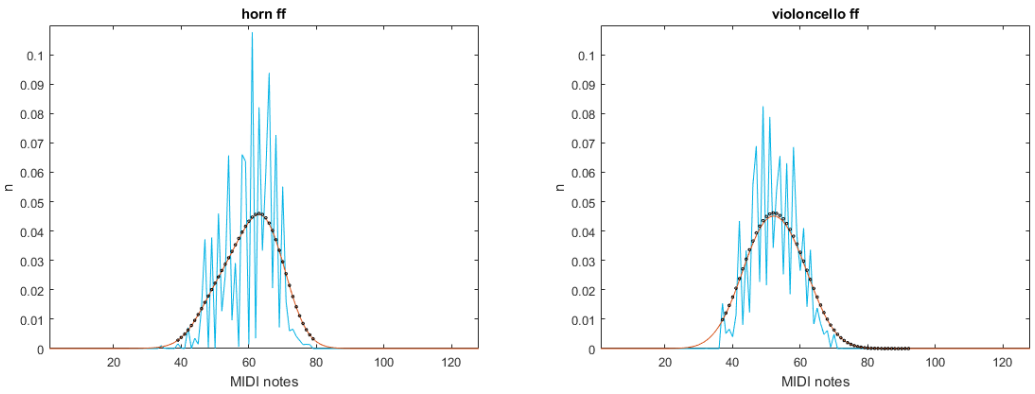

Figure 3.6: Pitch distributions of Violoncello and Horn in fortissimo over 128 MIDI notes (C0-G10). (blue: normalized pitch distribution, red: probability density function using kernel density estimation, black: discrete probability density according to the respective pitch range of the instrument)

were processed with an energetic averaging, similar to a diffuse-field equalization for a microphone. This step is necessary due to the inaccurate information about the microphone position during the recording of the audio material. The instruments' directivity for the position of the microphone (θ_{rec}, ϕ_{rec}) during the recording is inevitably integrated in the recording. Therefore, the directivity used in the simulation must be referenced (equalized) to the position (θ_{rec}, ϕ_{rec}). Since directivities change drastically in their characteristics over short changes in θ and ϕ, inaccurate referencing can cause extreme enhancement or decreasing in the referenced directivity. This presents itself in the auralization as if the audio signal was processed with notch filters. Therefore, the directivity function f(θ, ϕ, t) for the third octave t is averaged by

$$f_{eq}(\theta, \phi, t) = \frac{f(\theta, \phi, t)}{E_{diff}(t)} \tag{3.8}$$

with the diffuse energy E_{diff}(t) of the third octave t,

$$E_{diff}(t) = \sqrt{\int_0^{2\pi} \int_0^{\pi} f(\theta, \phi, t) d\theta d\phi} \tag{3.9}$$

with dθdϕ as the area weight, which depends on the sphere grid resolution when calculating with discrete values for f(θ, ϕ, t). The regarded bandwidth of f(θ, ϕ, t) rises with higher third octave t, resulting in a stronger reduction of the amplitude of the directivity at higher frequencies. This is similar to the frequency response of diffuse-field equalized microphones.

3.3 Anechoic Audiomaterial

The auralization of the sound material in the simulated rooms requires anechoic recordings. That means the recordings consist only of the direct sound from the source and ideally no reflections or reverb from the recording room. Both the speaker and trumpet anechoic recordings were done in the anechoic chamber of the Technical University of Berlin (TU Berlin) by the audio communication group of the TU Berlin. The speaker audio material is a speech of Cicero's third Catiline Oration. The trumpet audio content is "Trumpet Voluntary" by Jeremiah Clarke and "Cellosuite Nr. 1 Gigue" by J. S. Bach.
Acquiring anechoic material for the orchestra auralization is a much more difficult task compared to monophonic material. Currently, there are very few free accessible polyphonic anechoic audio recordings available to the scientific community. Lokki et al. (2008) offer recordings of Beethoven's 7th symphony I. movement, Mozart's aria Donna Elvira from the opera Don Giovanni, Bruckner's 8th symphony II. movement and Mahler's 1st symphony IV. movement. Preliminary listening tests showed that the recordings are have an insufficient signal-to-noise-ratio to be used for binaural reproduction. Lokki et al. applied their recordings on the aforementioned "loudspeaker orchestra" where the noisiness possibly didn't present itself as strong. The only other polyphonic anechoic orchestral recording was done by Vigeant et al. (2008). The material is Brahms' Symphony No. 4, 3rd movement, and Mozart's Symphony No. 40 in g minor, 1st movement. These recordings are not as noisy and are more eligible for this study[1], however, the complete audio material had to be elaborately revised and edited by a sound engineer, since the musicians were playing predominantly out of time and the recordings included unnecessary noise artifacts. Vigeant et al. recorded all instruments separately, where the musician was watching a silent video recording of the conductor, while listening via headphones to a MIDI recording of the same piece.
Vigeant et al. (as well as Lokki et al.) offer between one and three recordings of the same string instrument section. A typical complete classical orchestral instrumentation consists of twelve 1st violins, eleven 2nd violins, ten violas, nine violoncellos and eight double basses. This makes a realistic representation of a complete orchestra with just the given recordings impossible. Identical multiplication of the recordings with spatial distribution on the stage did not give the impression of a "real" string section. Rather, it represented a loud string quartet. In order to create a realistic representation of complete string sections, a simulation of a string section is needed.

In the following section, the phase vocoder is presented as the technical instrument for temporal modification of the original recording. Following that, the different possible approaches to simulating a string section are shown and the proposed segmentation track replication (STR) method is described.

[1] A quasi anechoic recording of a full opera orchestra of Puccini's "O mio babbino caro" was recorded by D'Orazio et al. (2016). Ackermann et al. (2017) produced anechoic recordings for three out of four movements of the 8th symphony of L. v. Beethoven with full orchestration and historical instruments. These recordings, however, were not available during the development of this study.

3.3.1 Phase Vocoder

The phase vocoder was introduced by Flanagan and Golden in 1966. Its application as a tool to separately analyze and influence the temporal and frequency information of an audio signal was described by Dolson in his tutorial (Dolson (1986)). In basic terms, the phase vocoder models the input signal as a sum of a number I(t) of sinusoids with time-varying amplitudes A_i(t) and instantaneous phase ϕ_i(t) (Laroche and Dolson (1999)):

$$x(t) = \sum_{i=1}^{I(t)} A_i(t) e^{j\phi_i(t)} \tag{3.10}$$

where ϕ_i(t) is:

$$\phi_i(t) = \phi_i(0) + \int_0^t \omega_i(\tau) d\tau \tag{3.11}$$

with ω_i(t) as the instantaneous frequency of the ith sinusoid.

The phase vocoder process can be divided into three stages, the *analysis* stage, the *modification* stage and the *resythesis* stage. The analysis stage allows the division of the signal into its temporal and spectral parts. The modification stage applies the desired changes in the temporal and/or spectral domain. Finally the resynthesis stage reassembles the modified parts into a cohesive audio signal. The analysis of the signal is achieved by applying a Short Term Fourier Transform (STFT) on the audio signal. The STFT divides the audio signal into short time segments and applies the Fast Fourier Transformation (FFT) on these windows.

The audio signal is thus divided into segments windowed around time instants t_a^u equally separated by a constant analysis hop factor R_a and can be described successively with integer u as $t_a^u = u \cdot R_a$. A FFT is calculated on every t_a^u over a Hanning window h(n) centered around the time instant t_a^u. The resulting STFT $X(t_a^u,\Omega_k)$ depends on the time instants t_a^u and the center frequency of the kth frequency bin of the vocoder $\Omega_k = \frac{2\pi k}{N}$ with N as the size of the discrete Fourier transform. If x is the original signal we can write $X(t_a^u,\Omega_k)$ as

$$X(t_a^u, \Omega_k) = \sum_{n=-\infty}^{\infty} h(n) x(t_a^u) e^{-j\Omega_k n} \tag{3.12}$$

The resynthesis stage involves a similar approach, by defining resynthesis time instants $t_s^u = u \cdot R_s$ with constant sythesis hop factor R_s. A short segment of the signal y_u(n) is retrieved by inverse Fast Fourier Transformation (IFFT) of the synthesis STFT $Y(t_s^u,\Omega_k)$. Each y_u(n) is multiplied by an optional window w(n) and summed together to the resulting resythesized audio signal y(n):

$$y(n) = \sum_{u=-\infty}^{\infty} w(n - uR_s) y_u(n - uR_s) \tag{3.13}$$

$$y_u(n) = \frac{1}{N} \sum_{k=0}^{N-1} Y(t_s^u, \Omega_k) e^{j\Omega_k n} \tag{3.14}$$

Temporal modification is achieved when $R_s \neq R_a$. Therefore the factor of prolonging or reducing the original audio signal in time is given by $\alpha = \frac{R_s}{R_a}$. The key to successfully retrieving a modified cohesive audio signal is to calculate the phase of $Y(t_s^u,\Omega_k)$. The phase of two consecutive frames of $Y(t_s^u,\Omega_k)$ can be calculated with the instantaneous frequency of the kth channel $\omega_k(t_a^u)$:

$$\angle Y(t_s^u) - \angle Y(t_s^{u-1}) = (uR_s - (u-1)R_s)\omega_k(t_a^u) \tag{3.15}$$

To retrieve $\omega_k(t_a^u)$ we have to go through the process of the *unwrapping* of the phase. Hereby the phase difference between two consecutive frames of the input signal $\angle X(t_a^u,\Omega_k)$ - $\angle X(t_a^{u-1},\Omega_k)$ is used to determine the instantaneous frequency $\omega_k(t_a^u)$ of a nearby sinusoid in channel k. But first we need to establish how the phase $\angle X(t_a^u,\Omega_k)$ is related to the instantaneous phase $\phi_i(t_a^u)$. Going back to the beginning of this chapter we declared that the phase vocoder treats the input signal as a sum of sinusoids (Eq. 3.10). Under the assumption that the amplitude and frequency of the sinusioids vary slowly over time, the amplitude and instantaneous phase can be approximated as:

$$A_i(t_a^u + n) \approx A_i(t_a^u) \tag{3.16}$$

$$\phi_i(t_a^u + n) \approx \phi_i(t_a^u) + \omega_i(t_a^u) \cdot n \tag{3.17}$$

and further the STFT of the input signal (Eq. 3.12) can be calculated as:

$$X(t_a^u, \Omega_k) = \sum_{i=1}^{I(t_a^u)} A_i(t_a^u) e^{j\phi_i(t_a^u)} H(e^{j(\Omega_k - \omega_i(t_a^u))}) \tag{3.18}$$

with $H(e^{j\omega})$ being the FFT of the analysis window h(n).
If the length N of the FFT is big enough, we can safely assume there is only one sinusoid I per channel k, so we are interested in one instantaneous frequency $\omega_I(t_a^u)$. After renaming I with k, we further realize the difference between Ω_k and $\omega_k(t_a^u)$ is bound by the cutoff frequency ω_h of the window,

$$|\Omega_k - \omega_k(t_a^u)| \leq \omega_h \tag{3.19}$$

Reducing (Eq. 3.10) to:

$$x(t) = A_k(t)e^{j\phi_k(t)} \tag{3.20}$$

and the STFT (Eq. 3.18) to:

$$X(t_a^u, \Omega_k) = A_k(t_a^u)e^{j\phi_k(t_a^u)}H(e^{j(\Omega_k - \omega_k(t_a^u))}) \tag{3.21}$$

Since the window h(n) is real and does not contribute to the phase of $X(t_a^u,\Omega_k)$ we can assume that $\angle X(t_a^u,\Omega_k)$ equals the instantaneous phase $\phi_k(t_a^u)$ up to an integer multiple of 2π, because of (3.19).
We can now derive the instantaneous frequency $\omega_k(t_a^u)$ from two consecutive frames:

$$\begin{aligned}\angle X(t_a^u, \Omega_k) - \angle X(t_a^{u-1}, \Omega_k) &= \phi_k(t_a^u) - \phi_k(t_a^{u-1}) + 2\pi n \\ &= \omega_k(t_a^u) \cdot (uR_a - (u-1)R_a) + 2\pi n \\ &= \omega_k(t_a^u) \cdot R_a + 2\pi n\end{aligned}$$

To calculate the correct instantaneous frequency we need to figure out what n has to be. Here we use the above mentioned unwrapping of the phase:

$$\angle X(t_a^u, \Omega_k) - \angle X(t_a^{u-1}, \Omega_k) = \Omega_k R_a + (\omega_k(t_a^u) - \Omega_k)R_a + 2\pi n \tag{3.22}$$

Rearranging (Eq. 3.22) we can define the *heterodyned* phase increment $\Delta\Phi_k^u$:

$$\Delta\Phi_k^u = \angle X(t_a^u, \Omega_k) - \angle X(t_a^{u-1}, \Omega_k) - \Omega_k R_a \tag{3.23}$$

We can write (Eq. 3.22) as,

$$\Delta\Phi_k^u - 2\pi n = (\omega_k(t_a^u) - \Omega_k)R_a \tag{3.24}$$

The right side of the Eq. 3.24 we can combine with (3.19):

$$|(\Omega_k - \omega_k(t_a^u))R_a| < \omega_h R_a \tag{3.25}$$

and we can safely assume $\omega_h R_a < \pi$ since R_a is the analysis hop factor and evolve (Eq. 3.24) to:

$$
\begin{aligned}
|\Delta\Phi_k^u - 2\pi n| &= |(\omega_k(t_a^u) - \Omega_k)R_a| \\
&< \omega_h R_a \\
&< \pi
\end{aligned}
$$

This inequality confines the integer n we can replace it with the pricipal determination of *heterodyned* phase increment $\Delta_p\Phi_k^u$:

$$
|\Delta_p\Phi_k^u| < \pi \tag{3.26}
$$

And finally aquiring the instantaneous frequency $\omega_k(\mathrm{t}_a^u)$:

$$
\omega_k(t_a^u) = \Omega_k + \frac{1}{R_a}\Delta_p\Phi_k^u \tag{3.27}
$$

Now we can calculate the *phase propagation* formula from (Eq. 3.15):

$$
\angle Y(t_s^u, \Omega_k) = \angle Y(t_s^{u-1}, \Omega_k) + \omega_k(t_a^u)R_s \tag{3.28}
$$

3.3.2 Phase Problems and Scaled Phase Locking Method

The phase propagation formula (Eq. 3.28) offers for a sinusiod of a constant frequency to overlap coherently for successive short-time signals. Laroche and Dolson call this the "horizontal phase coherence". The bigger challenge is to ensure "vertical phase coherence" - phase coherence across the frequency channels in a given synthesis frame. If both vertical and horizontal coherence are not ensured, the calculated phase of $\mathrm{Y}(\mathrm{t}_s^u)$ of successive frames can add up in a way that the retrieved signal y(n) (Eq. 3.13) will have fluctuations of the harmonics over time. This results in the output signal acquiring a "phasiness" in its sound.
There have been different propositions offered to reduce phasiness (Griffin and Lim (1984)), (Nawab et al. (1983)) and (Puckette (1995)), but here we will concentrate on the phase locking methods proposed by Laroche and Dolson (1999) - specifically on *scaled phase locking*, since it offers the best results.
The phase locking technique relies on the hypothesis that the phase of neighboring channels in a small bandwidth is related in some way to the channel with the most amount of energy in that bandwidth. In other words, the phase of the channels close to peaks in the regarded STFT frame are connected to the phase of the peak. So instead of applying the phase propagation formula (Eq. 3.28) to all phases in the frame, we apply it only to the peaks of the frame. The phase of the rest of the channels are calculated through the phase locking method that is described as follows.

First search the STFT frame for existing peaks. A peak is defined as a channel k_l whose amplitude is higher than its four nearest neighbors. The bandwidth of phase-locked channels is evidently set by the middle frequencies between successive peak channels, with upper limit $(\Omega_{k_l} + \Omega_{k_{l+1}})/2$ and lower limit $(\Omega_{k_{l-1}} - \Omega_{k_l})/2$.
Now we apply the assumption of the *identity phase locking*, which suggests that phases of $Y(t_s^u)$ are related in the same way as the phases of $X(t_a^u)$. With Ω_{k_l} as the center frequency of the peak channel and Ω_k as the center frequency of the neighboring "locked" channel, the identity in the above set bandwidth is described as:

$$\angle Y(t_s^u, \Omega_k) - \angle Y(t_s^u, \Omega_{k_l}) = \angle X(t_a^u, \Omega_k) - \angle X(t_a^u, \Omega_{k_l}) \tag{3.29}$$

This method works especially well for frequency-stationary signals, but usually the peaks of the input signal vary in its frequency over time. For example if a violin slides up or down on the played string to change the pitch of the played tone. That means a peak in channel k_0 at time frame u-1 changes to channel k_1 at the next time frame u. So the phase connection between successive time frames has to be changed already in the unwrapping process of the phase (Eq. 3.22) and change (Eq. 3.23) to:

$$\Delta\Phi_{k_1}^u = \angle X(t_a^u, \Omega_{k_1}) - \angle X(t_a^{u-1}, \Omega_{k_0}) - \Omega_{k_1} R_a \tag{3.30}$$

leading to the corrected phase propogation formula:

$$\angle Y(t_s^u, \Omega_{k_1}) = \angle Y(t_s^{u-1}, \Omega_{k_0}) + \omega_{k_1}(t_a^u) R_s \tag{3.31}$$

Now the remaining question is to determine which channel the peak Ω_{k_1} in frame u has evolved from in frame u-1. It is safe to assume that the peak stayed within the same "bandwidth of interest". So the phase of the peak k_1 in frame u is connected to the peak where k_1 was closest to. This develops (Eq. 3.29) to the scaled phase locking equation:

$$\angle Y(t_s^u, \Omega_k) = \angle Y(t_s^u, \Omega_{k_l}) + \beta(\angle X(t_a^u, \Omega_k) - \angle X(t_a^u, \Omega_{k_l})) \tag{3.32}$$

with β being the phase scaling factor and $\beta = 2/3 + \alpha/3$ showing best results (Laroche and Dolson (1999)).

Finally we can summarize the phase calculation procedure:

1. Locate peaks in STFT frame u,
2. For the peak channel, find corresponding peak in frame u-1, take the same channel if no peak was detected in frame u-1 or closest peak if more than one peak was in the bandwidth of interest in frame u-1, calculate (Eq. 3.31) for the peak channel,

3. For the surrounding channels around the peak, calculate (Eq. 3.32)
4. Repeat steps 1-3 for remaining peaks in STFT frame u,
5. Go to next frame u+1 and repeat.

3.3.3 String Section Sound

Dolson used the phase vocoder to analytically investigate the differences in solo violin sound and ensembles of violins playing in unison (Dolson (1986)). In his comparison he established that by looking at the frequency distribution in the ensemble sound, there is a pronounced pulsing in the amplitude of the harmonics, with an irregular beat but increasing fluctuations with higher frequencies. Furthermore, the typical vibrato of a solo violin - a sinusoidal pitch variation of 1% at a low frequency between 5 and 7 Hz - is not traceable in the ensemble sound. So a kind of "smearing" and amplitude modulation of the frequency spectrum of solo violin sound is being produced by an ensemble of violins. Dolson's findings have not been sufficiently tested as of today. In some tests, it was found that amplitude modulation of partials alone are not sufficient to simulate all aspects of an ensemble sound especially with the lack of a model for the modulation function (Kahlin and Ternström (1999)).

A simple solution (often used in the industry) to create an ensemble sound is to use the chorus effect implementation as described in Zölzer (2002). It consists of a delay line with a low frequency oscillator driving the delay time parameter. A periodically varying time delay (5 to 10 ms) of an audio signal will result in a periodical pitch variation causing the vibrato effect. If multiple copies of the same audio content are applied on this delay line with 10 to 25 ms delay time and a random modulation type, it will create the commonly known chorus effect. This method, however, did not realize the desired string section sound, since the resulting signal acquired phasing sound artifacts due to the combination of the direct signal and the slightly delayed signals, making the result sound too "artificial". Pätynen et al. (2011) have realized a simulation of a string section by analyzing the performance of an orchestra. The temporal differences between string players were obtained by detecting the onset of the notes with contact microphone recordings. The resulting temporal differences distribution is then applied on a single instrument recording to create the desired simulated section sound. Therefore, to achieve the desired result of an ensemble sound, the differences between instruments within the same instrument section have to be simulated more precisely. A closer look at a string section shows differences in the onset (time instant a note is played), pitch, sound level and timbre of the played note (Recke (2011)). The first three attributes are defined in the time and frequency domain. The timbre aspect of the instrument sound cannot be quantified as easily as the other attributes, however, every repeated available recording adds more timbral information to the summarized signal. Every additional timbre information should hence improve the perceived impression of an ensemble sound, more on this topic see Sec. 3.3.5. If a recording with sufficient timbral information and no errors in the recording is manipulated according to the distributions

along the first three attributes (onset, pitch, sound level), it would be possible to create as many replications of the original recording as needed, thus delivering a complete set of anechoic recordings for the orchestra auralization.

3.3.4 Segmentation Track Replication Method

The idea how to adjust the onset, pitch and amplitude of a string instrument recording in the Segmentation Track Replication (STR) method is based on the work of Pätynen et al. (2011). The initial recordings from Vigeant et al. (2008) consist of three recordings of the each 1st and 2nd violin and only one recording of each instrument of the other string sections (viola, violoncello and double bass). The string sections of a complete orchestra hold twelve 1st violins, eleven 2nd violins, ten violas, nine violoncellos and eight double bass. In order to fill out the remaining "seats" of the orchestra the initial recordings are replicated with slight differences in the onset, pitch and sound level of the recorded audiomaterial. The data for the onset, pitch and sound level distributions are taken from recorded string sections (Recke (2011)).

The main difference between the STR method and the method Pätynen et al. suggested, lies in the implementation of time differences, pitch and amplitude modulation on the initial recordings. Pätynen et al. apply the time differences continuously on each frame of the STFT of the input signal. The random variation of the time differences is acquired using the Metropolis-Hastings algorithm (Chib and Greenberg (1995)). The Metropolis-Hastings sampling offers a random Markov chain that follows a given probability distribution after an initial burn-in period. This should allow for a slow change of time differences to imitate the musician playing slightly out of the average rhythm of the section and slowly catching up with the group (Pätynen et al. (2011)). This implies, however, that every frame of the input signal is equally influenced in the time domain with data from distributions taken from played tones. The pitch change of the replicated signal in the Pätynen et al. method is done on the complete signal. The amplitude modulation is realized similar to the temporal modulation with Metropolis-Hastings sampling while scaling the sum to unity, so the entire sound level stays constant and only the balance between the players varies.

Since pitch variations happen on each tone while the musician is playing, an improvement of the approach would be to apply pitch variation on each tone, especially since the pitch distribution data is acquired from each played note. The sound level variation can also be improved, since an enveloping modulation function could distort the envelope of the tone. The complete envelope of the note should be enhanced or reduced to keep the "natural" sound of the instrument.
These improvements lead to the idea of the STR method: to extract singular tones or tone passages and apply temporal, pitch and sound level variations on them separately. That way the "naturalness" of the sound is kept by preserving the envelope of the tones. Additionally the information used from the distribution data will be applied in a similar way it was retrieved. So the onset, pitch and sound level of the tones from the data will

be applied on the onset, pitch and sound level of the tone in the replication process.

Silence Onset Detection

The first task is to separate the initial recording into individual tones or tone passages. A simple way to approach this problem is to detect time instances within the recording where no tone is being played. An obvious way to detect the silent parts would be to analyze the sound level of the recording. The function *STR_onset.m* (see Sec. A.1.1) is used to extract the positions of the silence onsets in the recording. The function *STR_spl.m* (see Sec. A.1.1) returns the sound pressure level (SPL) of a moving window. The reference sound pressure level is arbitrary, since we do not know the exact circumstances of the recording and more importantly it does not matter. All that is needed is a function course of the recordings' sound level so the minima of the sound level function can be detected. Additional distance of 4000 samples to the next increase in energy, by analyzing the slope of the SPL function assured that the minima was not located directly before the successive tone is played. These time instants of the minima are the positions of the silence onsets (see Fig. 3.7). The window size was set at 500 samples, in order to achieve an adequate resolution to detect the minima. Furthermore, the minimal distance between two consecutive silence onsets was set to 0.1 s (taken from Recke (2011)).

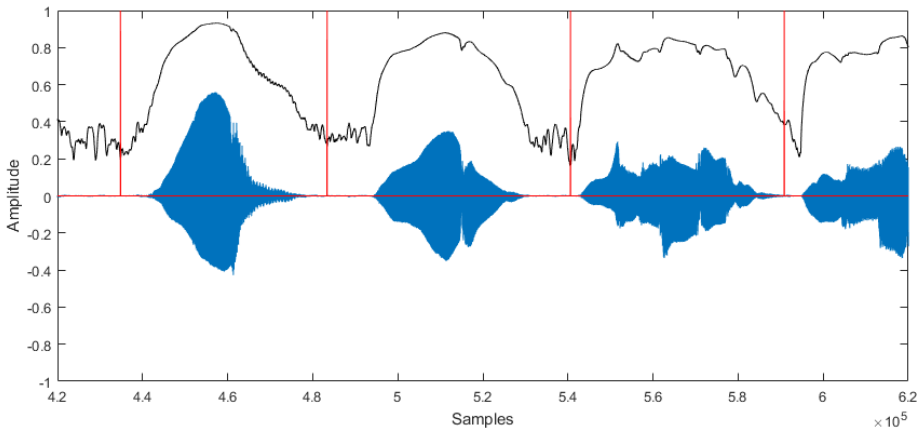

Figure 3.7: Detected silence onsets (*red*) for segmentation of the recording into tones (first segment) and short tone passages (second and third segment). *blue*: time signal of the recording, *black*: function course of the sound level (y-axis arbitrary units).

Temporal, Pitch and Amplitude modulation

As soon as the silence onsets are acquired, the initial recording is spliced into segments ranging from 0.1 s to 3 s. The shorter segments represent individual tones and the longer segments either represent tenuto (sustained note) tones or allegro parts (fast tone passages). The tone passages are not spliced into its individual tones, because the tones are played so fast in successive fashion that a clear abstraction between the beginning and end of successive tones is not possible. In addition, random fluctuation of time differences within an allegro part do not constitute a satisfactory explanation, whereas an argument can be made for the delay or acceleration of the entire allegro part, assuming that musicians perceive the "flow of the music" in terms of the whole passage and not in separate, individual notes while playing an allegro part.
The amount the individual segments are varied in onset, pitch and sound level differences is taken from Recke (2011). Recke recorded string sections and collected data on tone onset, pitch and sound level. Normal distributions were fitted to the provided data and random values from these distribution were used in this study for the variation of onset, pitch and sound level. The distributions' means were equated to zero and the standard deviations are shown in Tab. 3.3. The change in onset and pitch is effected using the phase vocoder described in Sec. 3.3.1. To achieve a displacement of the tone onset, the segments are stretched or squeezed in time, prolonging or shortening the tone played without changing its pitch. The pitch change is done after the time adjustment by resampling the segment by the factor of acceleration or deceleration of the segment F_p,

$$F_p = 2^{\frac{p_{cent}}{1200}} \tag{3.33}$$

where p_{cent} is a pitch difference in cents randomly chosen from the pitch distribution.

For each segment the STFT is calculated and the time vector of the STFT is changed by a factor F. F consists of the pitch factor F_p and time factor F_t,

$$F = F_t + F_p = \Delta t \frac{t_{min}}{t_{seg}} + 2^{\frac{p_{cent}}{1200}} \tag{3.34}$$

with Δt as the time displacement value randomly chosen from the onset distribution, scaled by the quotient of the shortest length of a tone (t_{min} = 0.1 s) divided by the length of the segment t_{seg}, assuring that the maximal onset variation of individual tones within longer segments does not exceed the maximal onset displacement in the onset

Table 3.3: Standard deviation for the normal distributions of onset, pitch and amplitude variation taken from Recke (2011).

	Onset [s]	Pitch [ct]	Amplitude
STD	0.048	13.9	0.4

distribution.
The adjusted time vector has the complete length of the original time vector of the segment but with sample length equal to F. The resynthesis stage described in section 3.3.1 is implemented in *STR_timediff_scale.m* (see Sec. A.1.1). Here the new STFT frames are calculated with the scaled phase locking method (see section 3.3.2) using the adjusted time vector to prolong ($F < 1$) or shorten ($F > 1$) the original STFT frames. The desired pitch shift is accomplished by resampling the segment by the factor F_p. Finally the amplitude of the segment is adjusted by a random value according to the sound level difference distribution calculated from the data in Recke (2011). The procedure is repeated for the following segment with the time displacement value having an opposite sign to the previous segment to avoid an accumulating delay or acceleration over multiple segments.
Furthermore, each segment is aligned by $\Delta t/2$ forward or backward in time, depending on whether the segment was prolonged or shortened in time. In cases when two consecutive segments are too far apart for the signal to be connected and avoid audible cracks, the segments' beginning and end are faded out and in respectively with a quadratic cosine (or sinus) function and zeros are added to the time distance between the segments. In the end all segments are consolidated and we receive the desired onset, pitch and sound level adjusted replication of the original recording.

3.3.5 Postprocessing

After adding up the replicated recordings to receive the string section sound, it is possible to hear *phasing* artifacts in the sum output. The phasing effect appears when the amplitude of narrow bands in the frequency spectrum of the audio signal are modulated over time (Zölzer (2002)). The effect can be compared to adding a notch filter with a slowly varying center frequency over time to the audio signal, resulting in phase cancellations or enhancements, audible as a sweeping effect. The STR method (as well as the Pätynen et al. method) use onset, pitch and sound level modulation for the replication process. To understand this phenomenon in the context of the string section simulation, spectra of different recordings of the same tone played by the same instrument and their sum spectrum are compared over time. Two artificially replicated recordings have been modified in onset, pitch and sound level. Therefore, their frequency amplitude information differs only in a slight pitch shift and amplitude gain but the form and time evolution of their tone spectra is almost identical (see Fig. 3.8). Since the phases are not identical, the addition of the signals can cause phase cancellation or enhancement in certain amplitudes of the sum spectrum that vary over time, creating the sweeping character of the phasing effect.

If we compare the spectra of two different physical recordings of the same tone of the same instrument, we can see the spectra differ more strongly than just in the slight pitch shift and amplitude gain (see Fig. 3.8). All these other differences affect various other more complex attributes of a tone like "noisiness" or "sharpness". They can all be unified by the description of timbre. The differences of timbral aspects of two

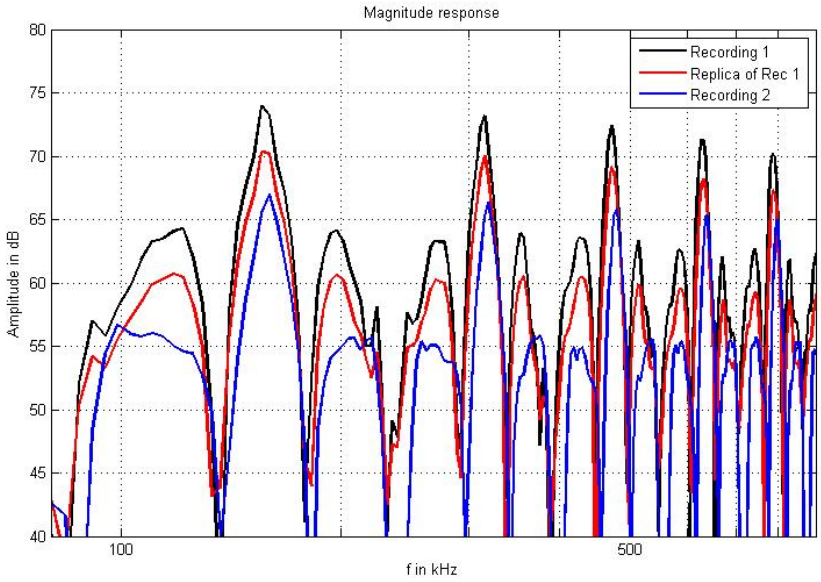

Figure 3.8: Spectra of one tone taken from two different recordings of the same violin (black and blue) and a replication of the first recording (black) with the STR method (red).

physical recordings of the same tones of the same instrument are the reason why no phasing effect is heard with natural recordings. Since the form and time evolution of the spectra of the tones is so different, the amplitudes of the sum spectrum do not vary in a systematic way like with the replicated recordings and therefore no phasing effect is heard. The modulation of timbre, however, is a highly difficult task, since a quantitative and satisfactory model for timbre is yet to be found. Pätynen et al. tried influencing changes in timbre by applying multiple filters to influence the Dünnwald parameters[2] with no success (Pätynen et al. (2011)). In this study, various approaches of time dependent filters, distortion of the signals or noise addition were tested but with little success. Temporal fluctuating filters did not achieve a systematic reduction of the phasing effect. Distortion and noise methods achieved a complete reduction of the phasing effect but deteriorated the quality of the recording. So instead of trying to simulate timbral aspects of the recordings, a postprocessing approach that tries to reduce the amplitude variation of the sum of the signals by selective manipulation of the phases of the separate signals was investigated.

Selective Phase Correction

For the phase correction, separate amplitude peaks are analyzed. One peak of the sum spectrum of two replicated signals can vary in its amplitude due to the phase relation

[2]frequency bands that describe timbre characteristics of the violin (Buen (2007))

of each replicated signal at that specific frequency. As explained above their amplitude spectra are very similar $A_{rep1} \approx A_{rep2}$, so the amplitude of the sum spectrum A_{sum} of two almost identical peaks depends on the sum of the phases ϕ_{sum} at the regarded frequency. If

$$\phi_{sum} = (n + \frac{1}{2})\pi \tag{3.35}$$

with n from a set of integers the amplitude A_{sum} is at its minimum, this is called phase cancellation. If

$$\phi_{sum} = n\pi \tag{3.36}$$

with n from a set of integers the amplitude A_{sum} is at its maximum and a phase enhancement is obtained. A low frequency oscillation of the amplitude peak and its surrounding peaks of the sum spectrum over time will result in the phasing effect. In order to avoid this, oscillating peaks of the sum spectrum need to be detected. Here the STFT of the sum and of each replicated signal is taken. To reduce computational cost we are only interested in peaks of the spectrum. A peak is defined as the frequency bin, the amplitude A_{peak} of which is higher than the six neighboring frequency bins and at least 20% as high as the maximum amplitude in the regarded STFT frame. The oscillation detection is achieved by comparing peaks at the same frequency bin. If the sum peak amplitude $A_{p,sum}$ varies over the five consecutive frames differently than the peaks of each replicated signal $A_{p,r1}$ and $A_{p,r2}$, as well as $A_{p,sum} < \max(A_{p,r1}, A_{p,r2})$ then the amplitude of the sum peak is being influenced by the phase relation of the separate signals. Now the phase at the regarded frequency bin of one of the replicated signals ϕ_{r2} is adjusted by $\Delta\phi$ so that the new sum peak amplitude at frequency bin p is

$$A'_{p,sum} = f \cdot (A_{p,r1} + A_{p,r2}) \tag{3.37}$$

The factor f is the phase relation factor, with $f \in [0, 1]$. If $f = 0$ then we achieve maximum phase cancellation, if $f = 1$ then we achieve maximum phase enhancement. After preliminary listening tests, $f = 0.65$ showed the best results in reducing the phasing effect. This procedure is repeated for all peaks in one frame and repeated for all following frames. A further reduction in the phasing effect occurred by applying the same phase $\Delta\phi$ to the phase of the two neighboring frequency bins of the adjusted signal $\phi_{p-1,r2}$ and $\phi_{p+1,r2}$, when using a small frame length of the STFT ($N = 2048$). The choice of the STFT frame length has an influence on two factors. A larger STFT frame length needs more computational cost, however, it increases the frequency resolution. This allows the algorithm to detect more distinctive peaks in the spectrum, but less fluctuations since a larger time frame is analyzed. A STFT length of $N = 2048$ showed most reduction of the phasing effect in preliminary tests.

3.4 Test interface

In order to test how different rooms are acoustically perceived, a set of attributes that describe room acoustical perception is required. Since the 1950s, there has been a number of attempts on distinguishing these psychological parameters, however as described in the beginning of the chapter, none have established a complete and satisfactory set of psychological room acoustical parameters. In this study we use a selection of 46 out of 50 attributes to describe room acoustical properties called Room Acoustical Quality Inventory (RAQI). These RAQI items were taken from the results of a focus group of experts in room acoustics, organized by the audio communication group of the TU Berlin and can be seen in the Appendix (Sec. A.4). The 50 items can be separated in 7 categories: Difference, Timbre, Geometry, Room, Time, Dynamics, Artifact, General. So far the technical vocabulary exists only in German. The selection of the attributes goes beyond the scope of this study.

The experiment interface is a MATLAB based test controlling environment called *WhisPER* (Ciba et al. (2009)). To create the RAQI questionnaire on WhisPER various functions have to be written. These functions are based on the functions for the SAQI vocabulary, a psychological measurement instrument for simulated acoustical environments (Lindau et al. (2014)). First, functions for the item names, definitions, categories, question phrases and scale labels are defined. These definitions end up in *edit test sections* where the test procedure can be modified. For the question order, a randomization

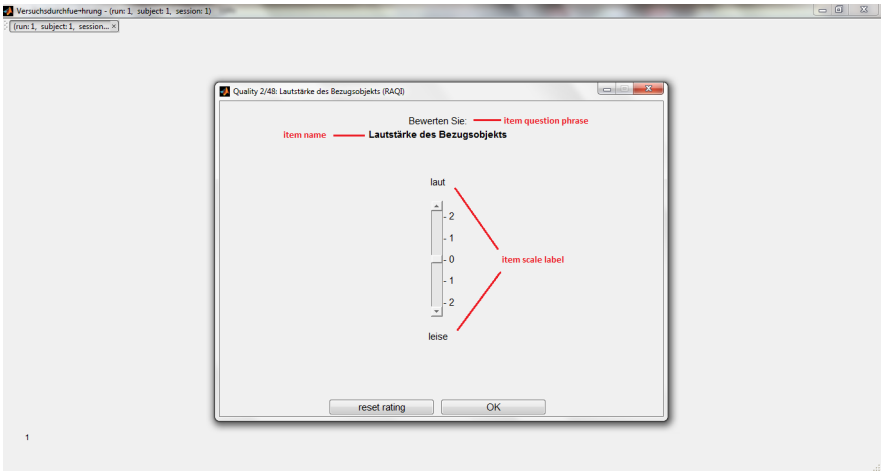

Figure 3.9: WhisPER RAQI Questionnaire GUI interface: each item is prompted separately on a continuous ordinal scale with the item characteristics at each end (scale labels).

of the order of items within a category and a randomization of the order of categories is chosen. The GUI interface of the RAQI questionnaire can be seen in Fig. 3.9. Each item is rated on a quasi continuous ordinal scale. In the results vector, the rating is

scaled on the interval [-1,1] and a resolution of $1 \cdot 10^{-4}$.

3.5 Experimental Setup

This section describes the soft- and hardware used in the experiment as well as the setup and environment of the experiment and the experiment's randomized block design. The information flow of the test procedure can be seen in Fig. 3.10. The rating input of the subject in WhisPER is done on a laptop in a sound-insulated room. The audio rendering is done on a separate computer.

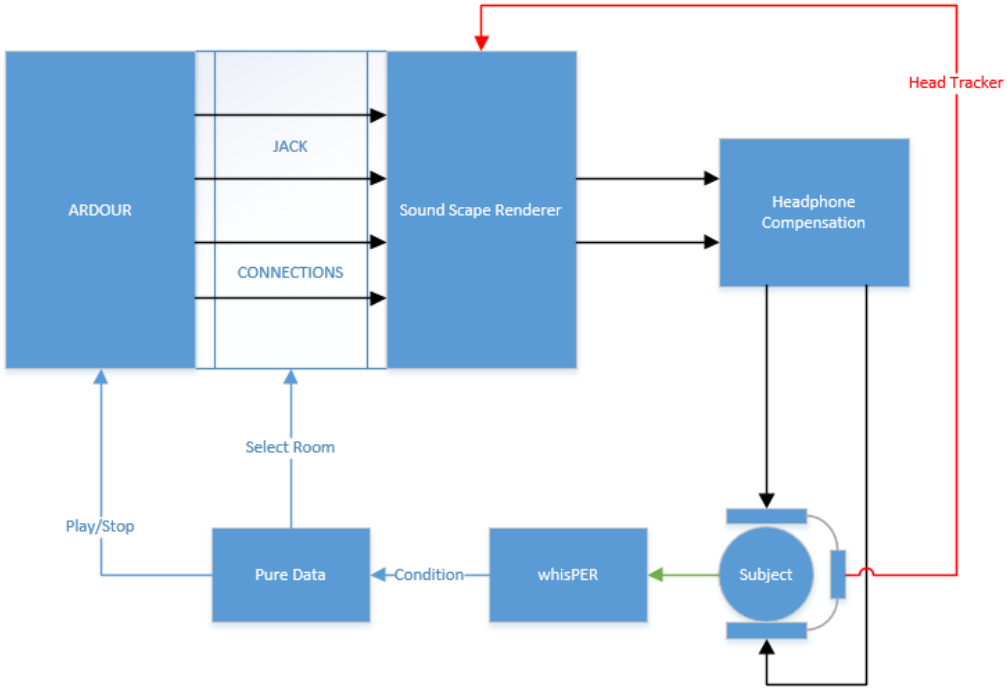

Figure 3.10: Information flow of the listening experiment. (black: audio content, blue: OSC-messages, red: head tracking, green: subject input)

3.5.1 Listening environment

The subject is situated in a small sound-insulated room. After passing the Lake People headphone amplifier G109-P the audio signal is produced with *AKG K1000* extra-aural headphones (see Fig. 3.11). These headphones have been successfully tested for their

adequacy for binaural reproduction (Møller et al. (1995)). The head tracker is centered on the rim of the headphones above the subject's head (see Fig. 3.11).

Figure 3.11: AKG K1000 extra-aural headphones with Razor AHRS head tracker mounted on the rim.

3.5.2 Headtracker

The rendering software receives the head orientation of the subject from the Razor AHRS[3] head tracker via USB and convolves the complete BRIR with the audio content. The Razor AHRS head tracker hardware is based on the *9DOF Razor IMU* board and uses a gyroscope, an accelerometer and a magnetometer to locate the orientation of the tracker. It is affected by the earth's magnetic field as well as surrounding magnetic influences from other devices and must be precisely calibrated before usage.

3.5.3 Software

The communication between WhisPER and the rendering computer is done via Open Sound Control (OSC), a protocol designed for realtime communication between computers, sound synthesizers and other multimedia devices (Wright and Freed (1999)). The OSC message from WhisPER is received and all further communication between the different software used for the experiment is accomplished with a PureData patch.

[3]https://github.com/ptrbrtz/razor-9dof-ahrs/wiki/Tutorial

3.5.4 PureData patch

PureData (PD) is an open source visual programming language developed by Miller Puckette (Puckette (1996)). Using *import mrpeach* and *import osc* the PD version *PD extended* can receive OSC messages. Depending on the route message the patch initializes either the loading of the session or the selection of the condition. One session consists of 14 stimuli conditions that are represented by 14 sets of BRIRs loaded in the SoundScape Renderer (SSR). The PD Patch for the experimental setup can be seen in Sec. A.1.3.

3.5.5 SoundScape Renderer

The SoundScape Renderer is a realtime spatial audio rendering software that allows binaural rendering with the a head tracker (Geier et al. (2008)). Each BRIR set consists of and is labeled by a combination of the *room*, *audio content* and *position*. The room is labeled as a two digit number that corresponds to the room numbers in Ackermann and Ilse (2015) and involves the BRIR information of the room described in Sec. 3.1. The audio content is labeled as 1 for speech, 2 for trumpet and 3 for orchestra, according to the content described in Sec. 3.3 and contains the BRIR information of the location of the sound sources simulated with their correct directivities described in Sec. 3.2. The position is labeled 1 for the first frontal position and 2 for the second position and involves the BRIR information of the receiver position as described in Sec. 3.1.

The SSR information is loaded with a *SSR.asd* file created during the session loading (see Sec. A.1.2). With the ASD file the BRIRs for each source-receiver pair are loaded into the SSR. The BRIRs of all source-receiver pairs that are chosen in the audio routing (see Sec. 3.5.7) are convolved with each other by the SSR to create the complete BRIR of the virtual acoustic simulation. The BRIR source-receiver pair can be loaded as WAV files, but since this format requires significant working memory space, the available memory (32 GB) runs out after loading less than half of the orchestral instruments. Therefore, the WAV format BRIRs are first converted in SOFA format (AES (2015)). This format separates the direct signal and early reflections of the BRIR from the statistical reverb of the BRIR and creates two SOFA files. One SOFA file contains the BRIR of all directions without the statistical reverb and one file with only the frontal direction of the statistical reverb. The direction information of the statistical reverb is unnecessary since the sound pressure coming from the reverberation of the room has no specific direction (Möser (2007)). The separation criterion, as to the calculation of the time instant for the transition from early reflections to statistical reverb was taken from Lindau et al. (2010). This way the file size of the BRIR source-receiver pair is reduced by 200 to 300 times and the internal memory of the rendering computer is not overloaded. Per source-receiver pair, two BRIR files have to be loaded, resulting in 132 BRIR files per orchestra stimulus and 2 BRIR files per single source stimulus. The SSR experiences rendering issues when more than 420 source-receiver BRIRs are loaded, resulting in audible artifacts (crackle) due to CPU overload. Therefore, only

three different orchestra stimuli can be loaded per SSR session.

3.5.6 Ardour

The anechoic audio content described in Sec. 3.3 is placed on different audio tracks of the digital audio workstation Ardour[4]. One track is reserved for speech and the trumpet signal respectively and 66 tracks are used for the orchestra signals. The length of all tracks is adjusted so that a global loop can be placed (see Fig. 3.12). Playback is controlled with the PD patch via OSC messages sent to Ardour. The Ardour file is included in the Sec. A.1.4.

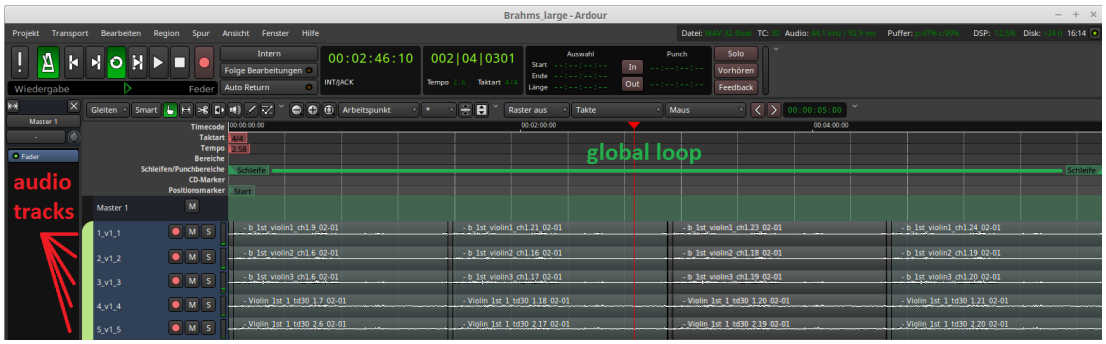

Figure 3.12: Screenshot of the digital audio workstation Ardour, the audio content for each instrument is placed on separate audio tracks and the playback is looped.

3.5.7 JACK Connections

Since the rooms are rated successively, the condition selection by the PD patch is done by routing the audio correctly. This is done using the JACK Audio Connection Kit[5], which allows audio applications to communicate with each other and with audio hardware. One possible JACK connection is shown in Fig. 3.13. The audio content goes from Ardour to SSR, where each audio track in Ardour is connected to the corresponding sound source in the SSR. From the SSR the two headphone channels are connected to the Headphone Compensation. The headphone equalization is done with Jconvolver[6], a nearly latency free FFT convolution engine. The resulting two channels from the Headphone Compensation finally go to the system output - resulting in signals that the subject hears on the headphones. The JACK connections are set up during the loading of the session (see Shell scripts in Sec. A.1.2) in XML files (see Sec. A.1.4). For each

[4]https://ardour.org/
[5]http://www.jackaudio.org/
[6]http://kokkinizita.linuxaudio.org/linuxaudio/

condition one XML file is created and can be triggered using *jmess* commands (see Sec. A.1.3).

Figure 3.13: Possible JACK connection, from Ardour to SSR (BRS1), from SSR to headphone equalisation, from headphone equalisation to system. System input leads directly to the output signal on the AKG headphones. (left: outputs, right: inputs)

3.5.8 Randomised Block Design

Recapitulating from Sec. 3.1 the amount of treatment conditions for both single source stimuli is 2 audio contents X 2 receiver positions X 35 rooms = 140 conditions. The amount of orchestra conditions is 1 audio content X 2 receiver positions X 25 rooms = 50 conditions, resulting in a total of 190 treatment conditions. With 14 rooms to be tested per subject we allow each condition to be tested 14 times. Therefore we can divide the necessary sample of 190 subjects into groups (blocks) of 14 subjects for the further experiment design. A Experimental Design Generator And Randomiser[7] (Edgar) was used to create a Randomized Block Design. The condition combination of room-content-position is chosen as the treatment factor. The above mentioned block size of 14 is chosen as the block factor. The Randomized Block Design allows a higher estimate of treatment effects if the variability within blocks is less than the variability between blocks. The future experimenter has to ensure that the variability of subjects within one block stays the same or as similar as possible over all blocks. An excerpt of the resulting experiment design is shown in Fig. 3.14.

ID	1	2	3	4	5	6	7	8
1	Room31 Solo Pos1	Room21 Solo Pos2	Room34 Voice Pos2	Room18 Orch Pos2	Room28 Voice Pos1	Room12 Solo Pos1	Room5 Voice Pos1	Room8 Voice Pos1
2	Room1 Solo Pos1	Room8 Solo Pos1	Room22 Voice Pos2	Room24 Voice Pos2	Room8 Solo Pos1	Room20 Voice Pos1	Room9 Orch Pos1	Room9 Orch Pos2
3	Room13 Orch Pos2	Room7 Solo Pos2	Room33 Voice Pos1	Room20 Orch Pos1	Room33 Solo Pos1	Room17 Solo Pos1	Room12 Orch Pos1	Room31 Voice Pos2
4	Room11 Solo Pos2	Room11 Orch Pos2	Room5 Solo Pos1	Room15 Solo Pos2	Room5 Voice Pos1	Room12 Solo Pos2	Room7 Orch Pos1	Room27 Solo Pos2
5	Room4 Orch Pos2	Room23 Orch Pos1	Room23 Solo Pos1	Room22 Solo Pos1	Room13 Voice Pos2	Room22 Solo Pos1	Room13 Voice Pos1	Room4 Voice Pos2

Figure 3.14: Excerpt of the Randomized Block Design by Edgar (row: subject session corresponding to the subject ID, column: treatment condition room-content-position, Voice: speech stimulus, Solo: trumpet stimulus, Orch: orchestra stimulus)

[7] http://www.edgarweb.org.uk/

3.6 Listening Test Procedure

The test starts by setting up the communication between the input laptop and the rendering computer. When starting the WhisPER test a subject ID is selected. This subject ID corresponds to a specific session defined in the text files *exp_ rooms.txt, exp_ pos.txt* and *exp_ cont.txt* (see Sec. A.1.4). The text files contain the information about the conditions' order and setup created by the Edgar algorithm (see 3.5.8). The information from the text files is used to create the SSR.asd file to open the SSR session, as well as the XML files for the correct JACK connections.

Before initializing the test session, a training session starts and the subject listens to four different stimuli conditions. First, the speech stimulus in room G4 (no. 40) with the lowest reverberation time (T_{30} = 0.49 s) in seating position 1 is compared to the speech stimulus in room Kammersaal 2 with the highest reverberation time (T_{30} = 7.08 s) in seating position 1. Then the speech stimulus in room Sejong Concert Hall (no. 47) with the least Sound Power Factor (G = -1.47 dB) in seating position 2 is compared to the orchestra stimulus in room Basilica of Eberbach Monastery (no. 07) with the most Sound Power Factor (G = 10.79 dB) in seating position 1. This allows the subject to imagine the range of sound level and reverberation of the rooms. After the training the main experimental session begins, where the subject has to successively rate 14 different rooms. Each room is rated on 46 RAQI qualities taken from Sec. 3.4. After all 14 rooms are rated, the subject has to fill out an online survey on their own sociodemographic information, musical and room acoustical expertise (see Sec. A.2). A total of 190 test subjects is planned for this experiment. This results in a very prolonged experimental study. To ensure as much stability as possible in the procedure of the experiment, the communication between WhisPER and the rendering computer is optimized to achieve minimal involvement of the experimenter.

As soon as the subject commences the test session by initiating the first room, WhisPER sends an OSC message that triggers the correct JACK connection and playback of the audio content with the PD patch, letting the subject rate all 46 RAQI qualities successively while the audio is playing in a loop. As soon as the test of one room is concluded, WhisPER sends a OSC message to stop the playback. This procedure is repeated until all 14 rooms are rated and the session is closed by an OSC message from WhisPER. This way the experiment is completed without any interaction from the experimenter, except for a short introduction of the test to the subject in the beginning. The only exception is when a subject's session includes more than three orchestra conditions, since the SSR can only load three different orchestra conditions in one session (see Sec. 3.5.5). Therefore, a new SSR session has to be loaded to continue the test after the third orchestra condition. As soon as the third orchestra condition has been rated by the subject, WhisPER sends a OSC message to close the first SSR session and load the second SSR session. The successive SSR sessions are denoted in the text files as ".1", ".2" and ".3" (see Sec. A.1.4). The experimenter has to then take action, since the head tracker has to be re-calibrated with every new loading of a SSR session.

3.7 Comparison of Different Replication Methods on String Section Sound

A comprehensive listening test was conducted to determine which method is best suitable for simulating string sections, in terms of its similarity to a recorded string section in two binaurally reproduced virtual room acoustical environments with varying ratio of original anechoic recordings to its replications. It aimed to investigate if test subjects can distinguish and rate the similarity between a recorded string section and a replicated string section using a double-blind triple-stimulus with hidden reference (ABC/HR) test according to Recommendation BS.1116-1 (1997).

3.7.1 Investigated Replication Methods

We regard five different methods for track replication algorithms. In addition to the proposed STR method and the method proposed by Pätynen et al. (2011) we examine an improved version of the Pätynen method. The improved version uses a state of the art phase vocoder, including the scaled phase locking technique (described in Sec. 3.3.2) with a transient processing method proposed by Röbel (2003). The traditional chorus effect (see Sec. 3.3.3) with a random delay line following a normal distribution, with a modulation depth of 1.3 ms, and low-pass filtered at 3 Hz, was used as the anchor condition. The last algorithm is the Time Domain - Pitch Synchronous Overlap and Add (PSOLA) method developed by Moulines and Laroche (1995). The PSOLA method divides the signal in short overlapping segments and adds or reduces the segments depending on the desired pitch or time duration modification. A more detailed description of the PSOLA method surpasses the scope of this study and can be found in Moulines and Laroche (1995).

3.7.2 Aquisition of the Reference Signal and its Distributions

The replication algorithm's output signal strongly depends on the input signal's recording situation. When replicating signals that were recorded in an anechoic chamber, the reference signal has to be multiple anechoic recordings that were recorded under similar circumstances. Two amateur violin players were invited to the anechoic chamber of the TU Berlin. The violin players were situated in the northern corners of the room and absorption panels were placed next to them to reduce cross-talk (see Fig. 3.15). The sound material was taken from the 1st movement of Beethoven's 8th symphony and the sound engineer acted as conductor. The recording microphone Neumann U89 was placed above the violin player in the approximate direction of the maximum sound energy of the violin's directivity. An eight directivity was chosen to reduce incoming sound from the other violin player. A total of twelve different violin recordings were

Figure 3.15: Recording setup of the violin player in the anechoic chamber for the comparative test of different replication algorithms.

produced and later used as the reference signal.

All investigated methods (with exception of the chorus effect) use distribution of onsets, pitch and volume (only in the STR method) of string sections. Evidently if we want to test the methods' ability to reproduce the reference signal, the method needs to use the distribution of onset, pitch and volume of the reference signal. The onsets were detected using Sonic Visualizer[8] and the pitch distribution was acquired using the YIN algorithm described by de Cheveigné and Kawahara (2002) to get the fundamental frequency of the played tone. The volume distribution corresponds to the sound level differences of the played tones between the twelve recorded violins and was obtained using the $STR\_spl.m$ function described in Sec. 3.3.4. The resulting means and standard deviations for onset, pitch and amplitude are shown in Tab. 3.4.

3.7.3 Test Method

The test environment is taken from the overlying room acoustical experiment (see Sec. 3.5.1), using the same modeled room acoustical simulation (see Sec. 3.1), and binaural reproduction using the SoundScape Renderer. For the test interface the WhisPER (see Sec. 3.4) toolbox for Matlab was used, enabling instantaneous switching between the stimuli. The presentation order of the stimuli was randomized across subjects. All five methods were tested in two room acoustical environments, one environment was

[8] An application for analyzing audio content, developed at the Centre for Digital Music, Queen Mary, University of London. (http://www.sonicvisualiser.org/)

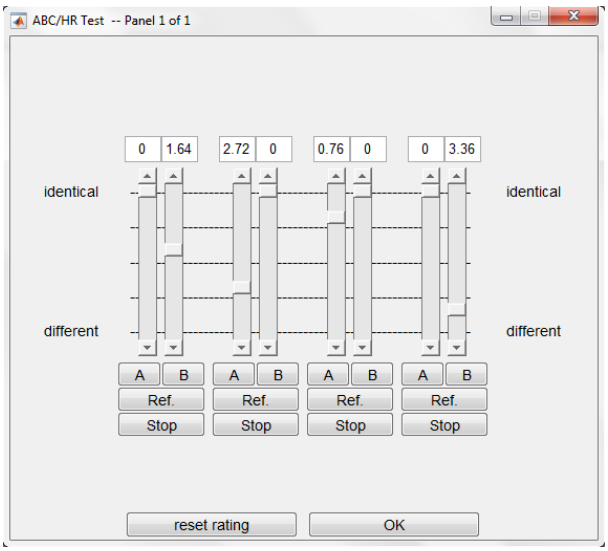

Figure 3.16: WhisPER ABC/HR test interface for comparison of different replication methods. Slider position represents similarity of the stimulus signal to the reference signal.

anechoic - the other was the Gewandhaus in Leipzig with reverberation time $T_{30} = 2{,}27$s. In addition, the methods were tested with varying ratio of the original anechoic signals to the amount of their replications (1:11, 2:5 and 3:3). The test subject is presented with a set of paired stimuli (A and B) and a corresponding reference (Ref) see Fig. 3.16. For each pair of stimuli there is a hidden reference which is chosen randomly, so either stimulus A or stimulus B is identical to the reference.

The subject's fist task is to decide which stimulus is identical to the reference. Their second task is to rate how similar the other stimulus is compared to the reference. Finally the subject has to place the ratings in a hierarchy within the whole set of stimulus pairs. The listening test took approximately 90 minutes, including a training phase to familiarize the subjects with the user interface and stimuli. Subjects were explicitly instructed to listen to the full length of the stimuli and rate the similarity in all aspects of the sound (i.e. sound coloration, timing, etc.). To increase the reliability of the ratings, subjects were free to switch between replicated section and reference section stimuli as often and as fast or slow as was their preference, and compared fifteen conditions in a multi slider rating interface at a time. For orientation, the sliders had as labels, *identical* on top and *different* on the bottom, with a numeric representation of the slider's position on the scale ranging from 0 to 4.00. After the listening test, the subjects filled out a questionnaire on social demographics, music listening habits and experience in music production. 23 subjects (4 female, median age 27) participated in the test. 82% of the participants claimed experience (8 years on average) in music production either as a hobby or professionally. 91% of the subjects listen to music multiple times a week and 9% listen to music once a week on average.

In total, 30 conditions were statistically tested in a three factorial, fully repeated mea-

sures ANOVA (analysis of variance) design, with factors *method* (STR, Pätynen, improved Pätynen, PSOLA, Chorus), *reverb* (no reverb, with reverb) and *ratio* (1:11, 2:5, 3:3). The results of the comparative study can be found in Sec. 4.

Table 3.4: Means and standard deviation for the normal distributions of onset, pitch and amplitude variation fitted to the measured data from the reference signal.

	Onset [s]	Pitch [ct]	Amplitude
Mean	0	0.0342	0
STD	0.022	8.85	0.18

Results

This chapter presents the results of the comparative study between different replication algorithms.

Results of the Comparative Test on Different Replication Methods

Listening test results for all conditions and subjects are shown in Fig. 4.2 by means and standard error for similarity to the reference signal. A Shapiro-Wilk test (Shapiro and Wilk (1965)) showed that the ANOVA requirement of normally distributed model residuals was not met, but a visual inspection implied that the violations occurred from an uneven deviation of the residuals across the range of predicted values for similarity from the ANOVA model. The residuals deviate more strongly at high similarity ratings which can be seen in Fig. 4.1. Their clear distribution around zero with different deviations along the predicted values of the model, indicates that the predicted model is still valid, but high similarity ratings of subjects vary stronger within subjects. A skewed distribution would indicate a false model altogether. Furthermore, a visual inspection of the distribution of residuals indicates a strong similarity to a normal distribution fit (see Fig. 4.1). Sphericity was violated for one main effect and one interaction after Mauchly's test. According to the measure of departure from sphericity ($\epsilon \leq 0.75$) Greenhouse-Geisser corrections (Greenhouse and Geisser (1959)) were used in further evaluation.

The reference signal was recognized in 97.7% of all cases. The cases of false recognition did not correlate in any way with the listening habits or music production experience of the subjects. The main effects, i.e. reverb, ratio and method were significant for the perceived similarity to the reference signal (see 4.1). The method is the most contributing effect compared to ratio and reverb. A slight increase in similarity can be seen with rising ratio ($p \leq .003$ for linear relation). The presence of reverb in the signal offers results with higher similarity ($p \leq .044$ for linear relation). The methods were compared successively (STR to PTA, PTA to PTA impr, PTA improved to PSOLA and

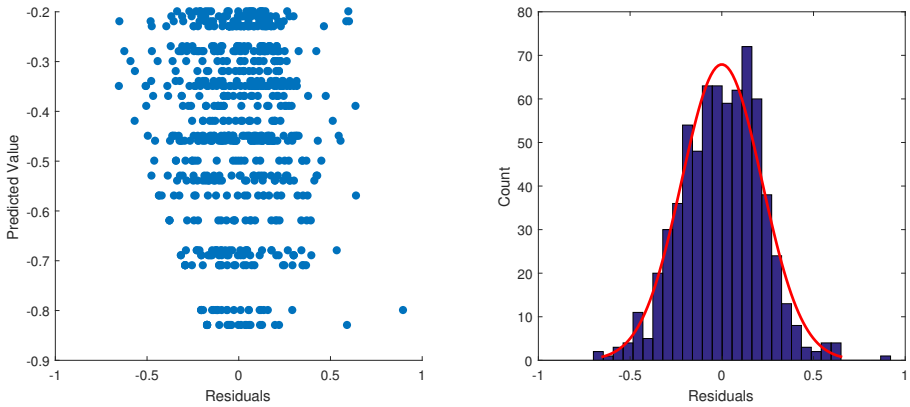

Figure 4.1: Distribution of the residuals. *left*: Distribution of the residuals accross the predicted values of the model. *right*: Distribution of the residuals with a normal distribution fit.

PSOLA to Chorus). Only PTA improved to PSOLA and PSOLA to Chorus showed highly significant differences ($p \leq .001$ for both). The comparison of the STR to PTA and PTA to PTA improved showed no significant differences, but all three showed higher similarity results than PSOLA or Chorus (see Fig. 4.2). The only significant interaction effects were observed for ratio x method and reverb x ratio x method (see 4.1). For the ratio x method interaction significant contrasts were observed for a linear ratio relation when comparing STR to PTA, PTA improved to PSOLA and PSOLA to Chorus ($p \leq .001$, $p \leq .004$ and $p \leq .001$ respectively).

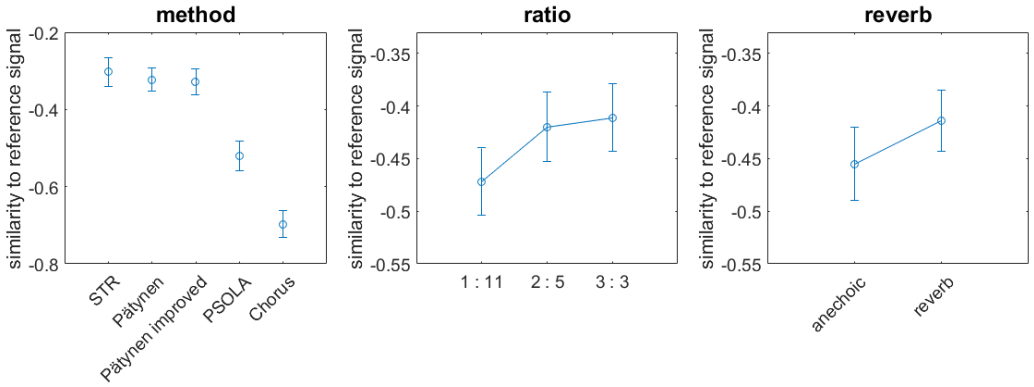

Figure 4.2: Ratings for the similarity to the reference signal of all subjects and test conditions with regard to the main effects method (left), ratio (middle) and reverb (right) described in the text. The rated similarity ranges on the interval [0;-1] with 0 equal to "identical" and -1 equal to "different". Mean ratings are indicated by circles. Standard errors are displayed by solid vertical lines. Connecting lines between conditions are provided to improve readability.

Both the STR and the PTA method showed higher similarity to the reference signal when increasing the ratio from 1:11 to 2:5, but the STR method was rated worse at 3:3 ratio compared to 1:11. Indicating that a higher ratio not always increases similarity to the reference signal and depends on the used replication method. Especially while comparing the PSOLA to the Chorus method an linear increase in similarity with higher ratio is not evident. For the third order interaction, significant contrasts were obtained only for STR to PTA ($p \leq .003$) and PTA improved to PSOLA ($p \leq .002$) at linear ratio and linear reverb relation. Results show the similarity ratings between methods to be more distant from each other in the anechoic situation in comparison to the reverb environment.

Table 4.1: Results for main effects (method, reverb and ratio) and interactions in the measurement of similarity of replicated signals to the reference signal.

	Main Effects			Interactions	
	Method	Reverb	Ratio	Method X Ratio	Method X Ratio X Reverb
η^2	.794	.222	.172	.433	.214
p	.001	.004	.044	.001	.001

Discussion

The presented work offers a method of synthesizing polyphonic binaural stimuli for exploratory studies on room acoustical perception. Based on the work of Ackermann and Ilse (2015) a set of 70 different virtual room acoustical environments were produced in form of binaural room impulse responses for three different audio contents. Two monophonic sources (speaker and trumpet) were simulated in 35 different modeled rooms and one polyphonic source configuration (orchestra with 66 sources) was simulated in 25 of the 35 modeled rooms. All virtual acoustical environments were simulated for two receiver positions.

All sources were simulated with the respective correct directivity that prior to the simulation underwent a spatial smoothing based on motion tracking data of musicians and a pitch weighting of the directivity according to the pitch distributions of Beethoven's nine symphonies. To ensure no further audible interference in the auralization, the directivity was averaged for every third octave by the average energy in the third octave. This resulted in reduced amplitudes of the directivity at higher frequencies, analogous to a diffuse-field equalization of a microphone. The audible interferences in the auralization without diffuse-field equalization seem to be caused by an incorrect referencing of the directivity to the position of the microphone during the recording. An ideal referencing is impossible since the precise position of the microphone is unknown and an equalization at an incorrect point can result in drastic changes in the referenced directivity and finally in audible artifacts similar to a notch filtering. The diffuse-field equalization reduces these artifacts, however, to such an extent that the directivity is heavily reduced in its dominant characteristics (i.e. protruding lobes). A different approach could be to average the directivity function over a surrounding cone area in the recording direction, thus averaging only over the uncertainty of the recording position.

The anechoic audio material for the monophonic content for the speech stimulus and the trumpet solo stimulus was taken from recordings of the audio communication group of the Technical University Berlin. The preliminary anechoic recordings for the polyphonic orchestra stumulus were taken from Vigeant et al. (2008). After elaborate editing of the individual tracks by a sound engineer, the string section was augmented with a novel segmentation track replication (STR) method based on the work of Pätynen et al. (2011). The STR method divides the input signal in short segments and varies these

segments in length, pitch and amplitude according to onset, pitch and sound level distributions taken from orchestra recordings. The resulting replications of the recordings create together a satisfactory string section sound. A total of twelve 1st violins, eleven 2nd violins, ten violas, nine violoncellos and eight double basses were created and together with the wind, brass and percussion instruments offered a total of 66 sources for the polyphonic orchestra stimulus.

A reproducible experimental setup for a study on room acoustical subjective qualities that uses the prepared monophonic and polyphonic stimuli was presented. The setup allows communication of the subjects' input on the test interface to the audio rendering computer. The MATLAB based test interface WhisPER sends OSC messages to the Puredata patch on the rendering computer, which controls the playback with the DAW Ardour, the convolution of the BRIRs with the SoundScape Renderer (SSR) and the condition selection using the JACK Connection Kit.
The orchestra involves 66 sources per stimulus, making it impossible for the SSR to work with WAV format BRIRs, due to working memory overload. The SOFA format divides the signal into a directional direct sound with early reflections and a single directional reverb tail, thereby doubling the amount of sources but reducing the size of the BRIRs by a factor of 200 to 300.
The SSR is also limited on the amount of 431 sources per loaded instance. Which does not allow more than three orchestra stimuli per SSR instance to be loaded. Multiple SSR instances cause the system processor to overload, since every SSR instance processes the input of the head tracker and it is not possible to turn off the convolution process in the unused SSR instances. Another limiting factor of the computing system is the ability of the processor to calculate 132 convolutions for 66 instruments of the orchestra at the same time. A buffer size of 4096 samples is needed so the system is not in overload, while acquiring a noticeable latency of 93 ms as a trade off. However, fast head movements would rather rarely occur in the proposed experiment on room acoustical perception and the experiment supervisor can advise the test subjects to disregard the latency in their rating. An automated script for loading and closing all involved software has been presented in this study, allowing for as many orchestra stimuli to be loaded for the test subject's session, with only minimal action needed from the experiment supervisor (calibration of the head tracker).

A comparative study of different replication methods showed that, for a string instrument, the novel method achieves results in similar quality to Pätynen's method (see Fig. 4.2). In this study, five different replication methods were applied to simulate a recorded string section and tested for the perceived similarity to the recorded section in a comprehensive listening test. The results indicate that the method has the strongest influence on the perception of similarity to a real string section, when compared to the other main effects: ratio and reverb. This confirms the assumption that, by applying a more sophisticated method on manipulating the onset, pitch and sound level of a recording, it is possible to more accurately simulate a recorded string section. The perceived dissimilarities occur from audible artefacts after adding the replicated signals together, i.e. phasing effects or timing errors. The frequency and time domain based

methods (STR, Pätynen and improved Pätynen) show better results than the time domain based methods (PSOLA and Chorus), indicating that frequency and time domain based approaches are more suitable for replicating string instruments.
The presence of reverb reduces the perception of the above mentioned artefacts occurring from adding replicated signals, which can be seen in the higher similarity ratings in the reverb environment. The increase in similarity occurring from a higher ratio of original recordings to their replications in the section, confirms the assumption that a broader timbre representation of the instrument in the string section can lead to a closer representation of a recorded string section.
However, not every method achieves higher similarity ratings with higher ratio. Some methods acquired more phasing or timing artefacts with higher ratio. This appears in the ANOVA as an interaction effect between method and ratio. Furthermore, the presence of reverb seems to reduce the influence of the method-ratio interaction, evident in the same behavior of the interaction when comparing the same methods for different reverb environments and more similar results when comparing different methods at the reverb environment compared to the anechoic environment. It could indicate that the perception of similarity to a recorded string section would be better described by more dimensions like the perception of coloration or timing differences in the simulation, to disentangle this interaction. However, these results should be viewed carefully due to the small sample size of 23 test subjects.
The high recognition rate of replica and reference signal shows that the replica signals can easily be detected by the audible artifacts created in the sum signal (sound coloration and timing differences). Furthermore, the observation of the exact timing and pitch in the beginning of the stimulus can also reveal which stimulus is identical with the reference stimulus. To avoid such problems in the future, it is advised to use multiple reference signals of the same musicians and audio content, but never including the same reference in the comparison.
After reassembling the segments in the STR method, the resulting replication of the original recording undergoes a selective phase correction to reduce a phasing effect that occurs when adding both signals together. Through locating fluctuations of the peaks of the spectrum in the sum of both signals, it is possible to adjust the phase of the replicated signal so as to reduce the fluctuations of the sum signal. Informal listening tests showed a reduction of the phasing effect after selective phase correction with STFT frame length $N = 2048$ samples. A quantitative measurement of the phasing effect is required to offer a more clear judgment of the effectiveness of the proposed post-processing algorithm. Although the detection of fluctuation of peaks in the spectrum offers some insight in the detection of the phasing effect, and the selective phase correction allows some reduction of the phasing effect, additional information is needed for a complete reduction of the phasing effect.

In conclusion, the suggested STR method offered results on par with the state-of-the-art. A differentiation of the perception of similarity in perception of sound coloration and time differences could offer a clearer picture on the differences between the replication methods. A bigger sample size would allow for a higher validity of the results. This could be achieved by conducting a listening test on a wider scale (i.e. via the internet) without binaural reproduction, and since the results of this study reveal only a small

influence of the reverb on the perceived similarity, the challenges to achieve a high similarity to a recorded string section more likely lie within the method itself.

The presented experimental setup for investigating a ground truth on room acoustical perception has successfully ran for approximately one year, collecting subjective qualities data from 190 test subjects on the presented binaural stimuli. The results of this study are to be published in the proceedings of the Deutsche Gesellschaft für Akustik in March 2017 (Lepa et al. (2017)).

Bibliography

3382-1, DIN EN ISO (2009): „Akustik - Messung von Parametern der Raumakustik - Teil 1: Aufführungsräume."

3382-2, DIN EN ISO (2008): „Akustik - Messung von Parametern der Raumakustik - Teil 2: Nachhallzeit in gewöhnlichen Räumen."

Ackermann, D.; C. Böhm and S. Weinzierl (2017): „Eine nachhallfreie Orchester-Aufnahme zum Einsatz in virtuellen akustischen Umgebungen." In: *Fortschritte der Akustik: Tagungsband der 43. DAGA, Kiel.* (forthcoming).

Ackermann, D. and M. Ilse (2015): „The Simulation of Monaural and Binaural Transfer Functions for a Ground Truth for Room Acoustical Analysis and Perception (GRAP)." Master Thesis, Fachgebiet Audiokommunkation, Technische Universität Berlin.

AES (2015): „AES69-2015: AES standard for file exchange - Spatial acoustic data file format."

Beranek, L. L. (1962): *Music, Acoustics and Architecture.* Wiley, New York.

Berg, J. and F. Rumsey (2006): „Identification of quality attributes of spatial audio by repertory grid technique." In: , 54(5) S. 365–379.

Brinkmann, F.; A. Lindau; M. Vrhovnik and S. Weinzierl (2014): „Assessing the Authenticity of Individual Dynamic Binaural Synthesis." In: *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, Germany.* S. 62–68.

Bronstein, I. N.; K. A. Semendjajew; G. Musiol and H. Mühlig (2008): *Taschenbuch der Mathematik.* 7. auflage. Verlag Harri Deutsch.

BS.1116-1, Rec. ITU-R (1997): „Methods for the subjective Assessment of small Impairments in Audio Systems including Multichannel Sound Systems."

Buen, A. (2007): „On Timbre Parameters and Sound Levels of Recorded Old Violins." In: *J. Violin Soc. Am., VSA Papers*, 21(1) S. 57–68.

Chib, S. and E. Greenberg (1995): „Understanding the metropolishastings algorithm." In: *American Statistician*, 49(4) S. 327–335.

Ciba, S.; A. Wlodarski and H. J. Maempel (2009): „WhisPER: A new tool for performing listening tests." In: *Audio Engineering Society Convention*, vol. 126. AES.

de Cheveigné, A. and H. Kawahara (2002): „YIN, a fundamental frequency estimator for speech and music." In: *J. Acoust. Soc. Am.*, 111(4) S. 1917–1930.

Dolson, M. (1986): „The phase vocoder: A tutorial." In: *Computer Music Journal*, 10 S. 14–27.

D'Orazio, D.; S. De Cesaris and M. Garai (2016): „Recordings of Italian Opera orchestra and soloists in a silent room." In: *Proc. of International Congress of Acoustics.*

Geier, M.; J. Ahrens and S. Spors (2008): „The SoundScape Renderer: A Unied Spatial Audio Reproduction Framework for Arbitrary Rendering Methods." In: *Proceedings of the 124th AES Convention, Amsterdam.*

Greenhouse, S. W. and S. Geisser (1959): „On methods in the analysis of profile data." In: *Psychometrika*, 24 S. 95–112.

Griffin, D. W. and J. S. Lim (1984): „Signal estimation from modified short-time fourier transform." In: *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32. IEEE, S. 236–243.

Grigoriev, D.; D. Ackermann; S. Pelzer and S. Weinzierl (2016): „Ein psychologisches Messinstrument für die akustische Wahrnehmung von Räumen für Musik und Sprache: Stimulus-Erzeugung." In: *Fortschritte der Akustik: Tagungsband der 42. DAGA, Aachen.*

Hawkes, R. J. and H. S. Douglas (1971): „Subjective acoustic experience in concert auditoria." In: *Acustica*, 24(5) S. 235–250.

Kahlin, D. and S. Ternström (1999): „The Chorus Effect Revisited - Experiments in Frequency - Domain." In: *Proceeding of Euromicro 99, Milano.* IEEE, S. 75–80.

Kajastila, R.; S. Siltanen; Lundén P.; Lokki T. and Savioja L. A (2007): „A distributed real-time virtual acoustic rendering system for dynamic geometries." In: *AES Convention*, vol. 122. AES.

Kürer, R.; G. Plenge and H. Wilkens (1969): „Correct spatial sound perception rendered by a special 2-channel method." In: *37th AES Convention New York.* S. Sonderdruck 666 (H–3).

Laroche, J. and M. Dolson (1999): „Improved Phase Vocoder Time-Scale Modification of Audio." In: *IEEE Transactions on Speech and Audio Processing*, vol. 7(3). IEEE, S. 323–332.

Lehmann, P. and H. Wilkens (1980): „Zusammenhang subjektiver Beurteilungen von Konzertsalen mit raumakustischen Kriterien." In: *Acustica*, 45 S. 256–268.

Lepa, S.; D. Grigoriev and S. Weinzierl (2017): „RAQI - Ein neues psychologisches Messinstrument für die akustische Wahrnehmung von Räumen für Musik und Sprache und seine psychometrische Qualität." In: *Fortschritte der Akustik: Tagungsband der 43. DAGA, Kiel.* (forthcoming).

Lindau, A. (2009): „The Perception of System Latency in Dynamic Binaural Synthesis." In: *Proc. of the 35th DAGA, Rotterdam.* S. 1063–1066.

Lindau, A. and F. Brinkmann (2012): „Perceptual evaluation of head- phone compensation in binaural synthesis based on non-individual recordings." In: *J. Audio Eng. Soc*, 60(1/2) S. 54–62.

Lindau, A.; T. Hohn and S. Weinzierl (2007): „Binaural resynthesis for comparative studies of acoustical environments." In: *Proceedings of the 122th AES Convention*, vol. 41. AES.

Lindau, A.; L. Kosanke and S. Weinzierl (2010): „Perceptual Evaluation of Physical Predictors of the Mixing Time in Binaural Room Impulse Responses." In: *Proceedings of the 128th AES Convention, London.* AES, S. paper no. 8089.

Lindau, A. and S. Weinzierl (2007): „FABIAN-schnelle Erfassung binauraler Raumimpulsantworten in mehreren Freiheitsgraden." In: *Fortschritte der Akustik: Tagungsband d. 33. DAGA*, vol. 33. DAGA, S. 633–634.

Lindau, A. and S. Weinzierl (2012): „Assessing the plausibility of virtual acoustic environments." In: *Acta Acustica united with Acustica*, 98 S. 804–810.

Lindau, A. et al. (2014): „A spatial audio quality inventory (SAQI)." In: *Acta Acust. United Acust.*, 100 S. 984–994.

Lokki, T. and H. Järveläinen (2001): „Subjective evaluation of auralization of physics-based room acoustic modeling." In: *Proceedings of the 7th International Conference on Auditory Display, Espoo, Finland.* S. 26–31.

Lokki, T.; J. Pätynen and V. Pulkki (2008): „Recording of anechoic symphony music." In: *Joint ASA/EAA Meeting, Acoustics'08, Paris.* S. 6431–6436.

Lokki, T.; Y. Pätynen; A. Kuusinen and S. Tervo (2012): „Disentangling preference ratings of concert hall acoustics using subjective sensory profiles." In: , 132(5) S. 3148–3161.

Moulines, E. and J. Laroche (1995): „Non parametric techniques for pitch-scale and time-scale modification of speech." In: *Speech Communication*, 16 S. 175–205.

Möser, M. (2007): *Technische Akustik*, vol. 7. Springer-Verlag Berlin Heidelberg.

Møller, H.; D. Hammershøi; C. B. Jensen and F. Sørensen (1995): „Transfer Characteristics of Headphones Measured on Human Ears." In: , 43(4) S. 203–217.

Møller, H.; C.B. Jensen; D. Hammershøi and M.F. Sørensen (1996): „Using a typical human subject for binaural recording." In: *Jthe 100th Audio Engineering Society (AES) Convention, Copenhagen*. S. paper no. 4157.

Nawab, S. H.; T. Quatieri and J. S. Lim (1983): „Signal reconstruction from short-time fourier transform magnitude." In: *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31. IEEE, S. 986–998.

Nilsson, M. E. and M. Ekman (2009): „Perceptual evaluation of a real time auralization tool." In: *In Euronoise, Edinburgh*. S. 26–28.

ODEON (2010): „ODEON Room Acoustics Software." `http://www.odeon.dk`.

Plenge, G.; R. Kürer and H. Wilkens (1969): „Über die Reproduktion von Hörbildern mit Hilfe eines künstlichen Kopfes." In: *Berichtsheft Tonmeistertagung Hamburg*. S. S.80.

Pollow, M.; G. K. Behler and F. Schultz (2010): „Musical Instrument Recording for Building a Directivity Database." In: *Fortschritte der Akustik: Tagungsband der 36. DAGA*. DAGA.

Puckette, M. S. (1995): „Phase-locked vocoder." In: *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE.

Puckette, M. S. (1996): „Pure Data." In: *International Computer Music Conference, San Francisco*. S. 269–272.

Pätynen, J. and T. Lokki (2010): „Evaluation of concert hall auralization with virtual symphony orchestra." In: *Proceedings of the International Symposium on Room Acoustics, ISRA, Melbourne*. S. 1–9.

Pätynen, J.; S. Tervo and T. Lokki (2009): „A loudspeaker orchestra for concert hall studies." In: , 34 S. 32–37.

Pätynen, J.; S. Tervo and T. Lokki (2011): „Simulation of the violin section sound based on the analysis of orchestra performance." In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), NY*. IEEE, S. 173–176.

Quiring, R. and S. Weinzierl (2016): „Tonhöhenverteilungen im klassischen Orchesterrepertoire." In: *Fortschritte der Akustik: Tagungsband der 42. DAGA*.

Rafaely, B. (2015): *Fundamentals of Spherical Array Processing*. Vol. 8. Springer Verlag.

Recke, T. (2011): „Zum orchestralen Klang." Master Thesis, Fachgebiet Audiokommunkation, Technische Universität Berlin.

Röbel, A. (2003): „A new approach to transient processing in the phase vocoder." In: *Proc. of the 6th Int. Conference on Digital Audio Effects (DAFx-03), London, UK*.

Sabine, W. C. (1906): „The accuracy of musical taste in regard to architectural acoustics." In: *Proceedings of the American Academy of Arts and Sciences*, vol. 42(2). S. pp. 53–58.

Saher, K.; J.H. Rindel and L. Nijs (2006): „Comparisons between Binaural In-situ Recordings and Auralizations." In: *the 120th Convention of Audio Engeneering Society*. S. paper no. 6744.

Savioja, L.; J. Huopaniemi; T. Lokki and R. Väänänen (1999): „Creating interactive virtual acoustic environments." In: , 47(9) S. 675–705.

Schröder, D. (2011): *Physically Based Real-Time Auralization of Interactive Virtual Environments*. Ph.D. thesis.

Shapiro, S. S. and M. B. Wilk (1965): „An analysis of variance test for normality (complete samples)." In: *Biometrika*, 52 (3-4) S. 591–611.

Steger, D.; H. Egermann and S. Weinzierl (2015): „Spielbewegungen von Musikinstrumenten und deren Bedeutung für das Klangergebnis - Ein experimenteller Zugang durch Motion Tracking klassischer Orchesterinstrumente und Auralisation der Bewegungsdaten." In: *Fortschritte der Akustik: Tagungsband der 41. DAGA*.

Vigeant, M.; L. Wang and J. H. Rindel (2008): „Investigations of orchestra auralizations using the multi-channel multi-source auralization technique." In: *Acta Acustica united with Acustica*, 94(6) S. 866–882.

Vorländer, M. (2008): *Auralization - Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*. Springer-Verlag Berlin Heidelberg.

Weinzierl, S. (2008): *Handbuch der Audiotechnik*. Springer-Verlag Berlin Heidelberg.

Weinzierl, S. and M. Vorländer (2015): „Room Acoustical Parameters as Predictors of Room Acoustical Impression: What Do We Know and What Would We Like to Know?" In: *Acoustics Australia*.

Weitze, C.A.; C.L. Christensen; J.H. Rindel and A.C. Gade (2002): „Comparison between In-situ recordings and Auralizations for Mosques and Byzantine Churches." In: *Proc. Joint Baltic-Nordic Acoustical Meeting*. S. 53–57.

Wright, M. and A. Freed (1999): „Open Sound Control: A New Protocol for Communicating with Sound Synthesizers." In: *International Computer Music Conference, Thessaloniki, Greece*. S. 101–104.

Zölzer, U. (2002): *DAFX - Digital Audio Effects*. John Wiley & Sons, Ltd.

Appendix

A.1 Program Coding

All MATLAB scripts, shell scripts, PureData patch, Ardour, XML, ASD and text files used in this work can be found on attached CD-ROM. An overview over the application of the scripts is given here.

A.1.1 MATLAB code

MATLAB scripts were used for:

- the calculation of correct **directivities** (folder: MATLAB\1_Directivities)
- calculation of **STR method** (folder: MATLAB\2_STR_method)
- the development of the **RAQI package for WhisPER** (folder: MATLAB\3_RAQI_whisper)

All descriptions of the individual functions are given in commentary within the code.

A.1.2 SHELL scripts

Shell scripts (folder: Rendering\1_SHELL) initialize and control the rendering computer:

- **Start_RAQI.sh** starts JACK, audio card (hdspmixer) and PD patch
- **RAQI_Session.sh** prepares ASD file for the Sound Scape Renderer and XML files for the JACK Connections (to change between different rooms per SSR session)
- **RAQI_Run.sh** starts SSR and Ardour with prepared files from RAQI_Session.sh

A.1.3 PureData

The PureData patch *RAQI.pd* can be found in folder: Rendering\2_PD.

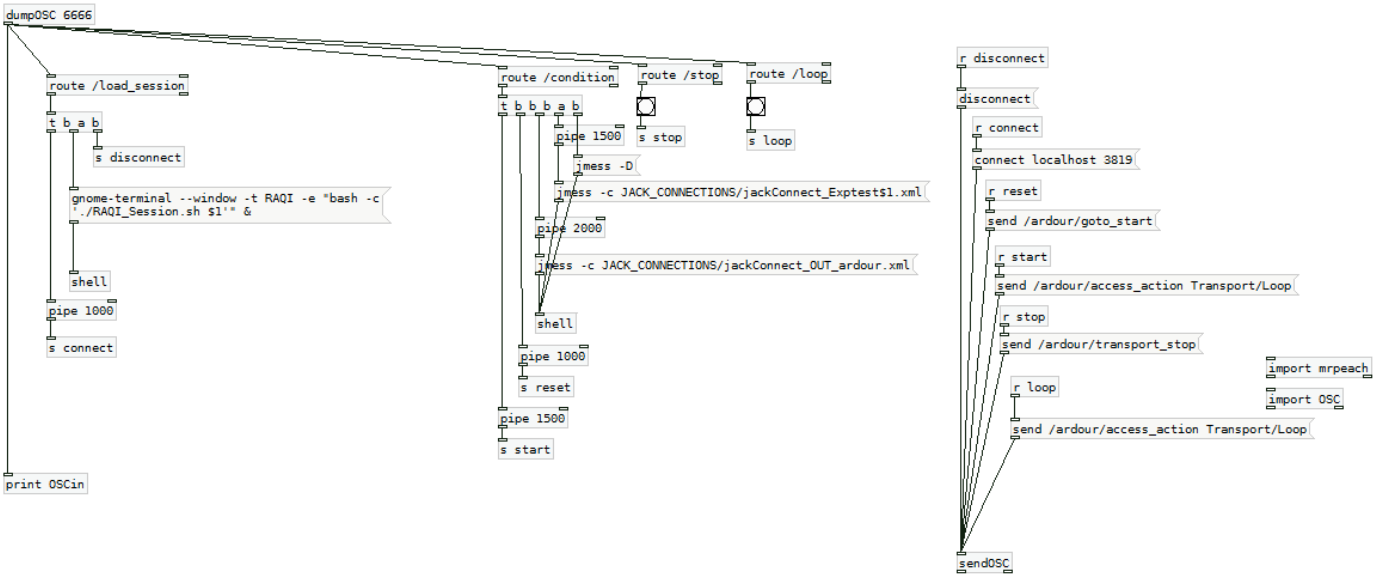

Figure A.1: PureData patch receiving OSC messages from WhisPER to select the correct session (left), the correct condition (center) and control the playback of the audio (right)

A.1.4 Ardour, ASD, XML and txt files

Ardour files can be found in folder: Rendering\3_Ardour.

ASD files (folder: Rendering\4_SSR) are used by the Sound Scape Renderer. This sythax allows to load the desired BRIRs for auralization.

XML files (folder: Rendering\5_JACK) are used by **jmess**[1] to connect the desired JACK Connections.

The order and combination of conditions, produced by Edgar, are written in txt files (folder: Rendering\1_SHELL) and are used by RAQI_Session.sh.

A.2 Sociodemographic and expertise Survey

The survey on sociodemographic, listening habits and expertise in room acoustics and music can be seen in the following pages.

[1] https://github.com/jcacerec/jmess-jack

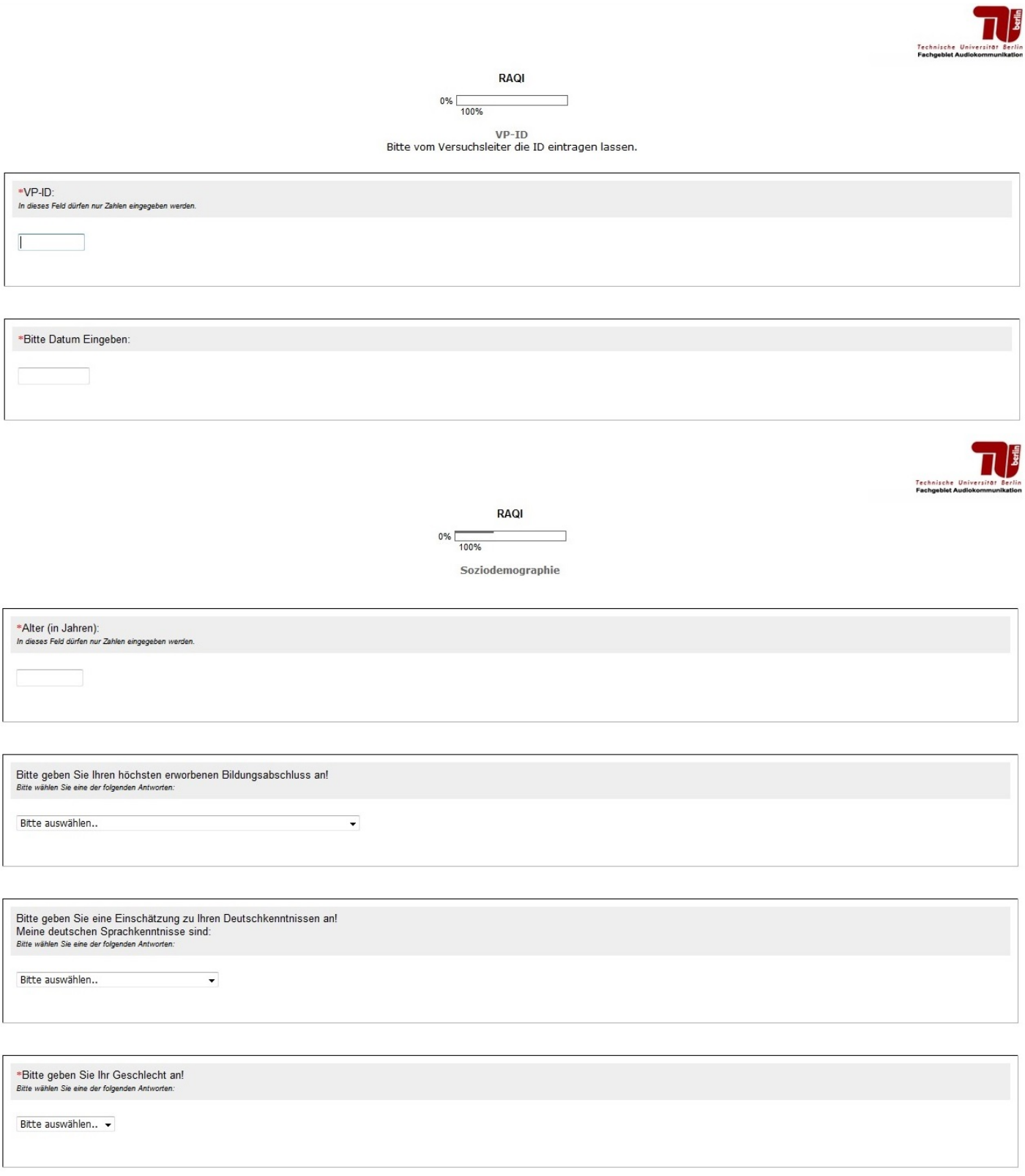

Figure A.2: Screen shot of survey on musical and room acoustical expertise and sociodemographic (page 1).

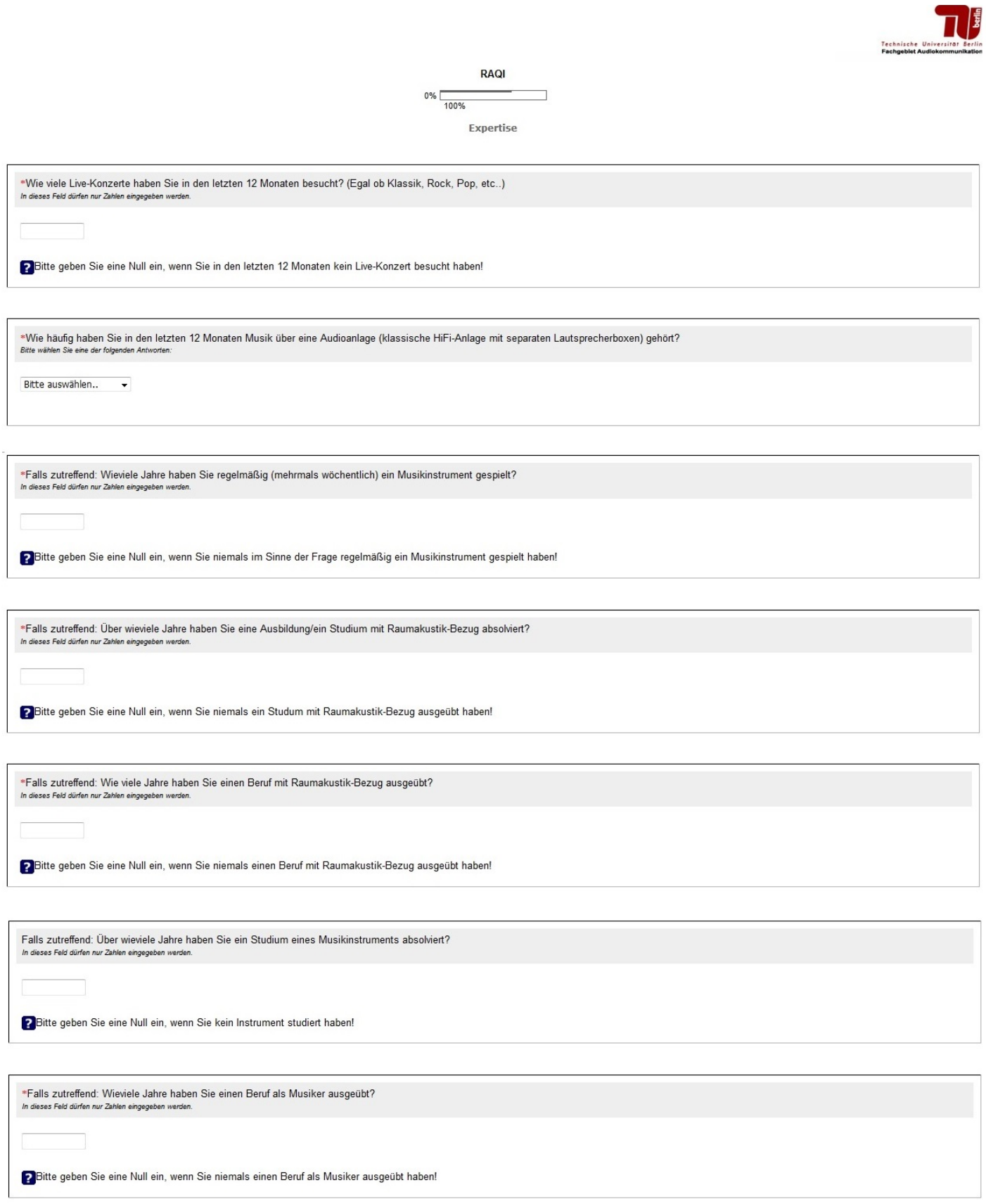

Figure A.3: Screen shot of survey on musical and room acoustical expertise and sociodemographic (page 2).

A.3 Distribution of BR, G and EDT

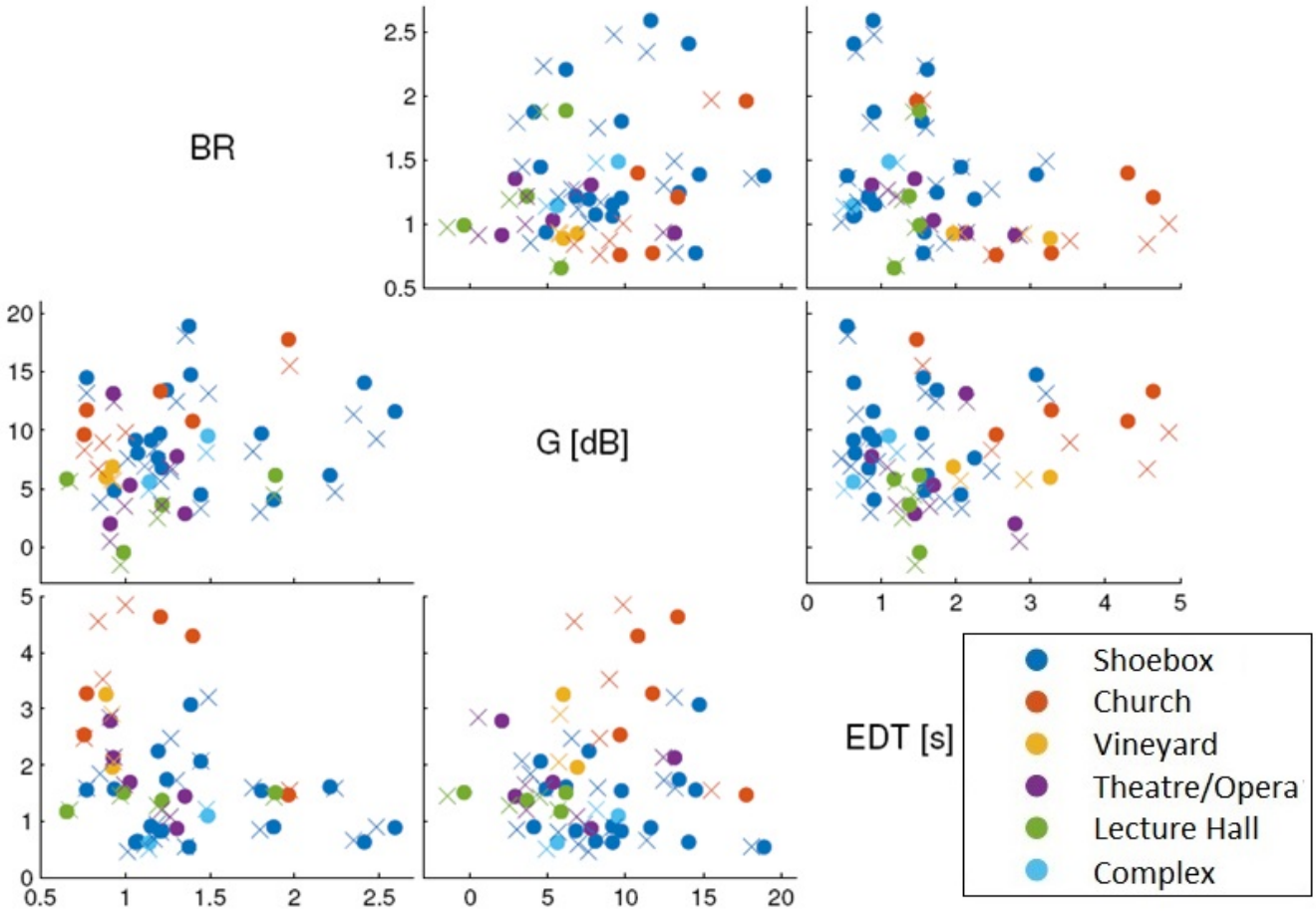

Figure A.4: Scatter-plot matrix of room acoustical parameters Bass Ratio BR, Sound Power G and Early Decay Time EDT of the 35 selected rooms (circle: receiver position 1, cross: receiver position 2, colors indicate room geometry). On the diagonal, the axes of the regarded parameters are displayed, indicating which of the three parameters are compared per scatter-plot (i.e. middle-left: y-axis G vs. x-axis BR; bottom-left: y-axis EDT vs. x-axis BR)

A.4 Room Acoustical Quality Inventory

The detailed Room Acoustical Quality Inventory (RAQI) can be seen in RAQI_Items.xls (folder: RAQI). The following table A.1 offers the item names (in german language) and their corresponding categories.

Table A.1: Room Acoustical Quality Inventory (RAQI) in german language. (o: only used for orchestra stimulus; s: only used for speech stimulus; x: not included in Lepa et al. (2017)

Kategorie / Category	**Attribute / Attributes (only in german language)**
Unterschied / Difference	1.Wahrgenommener Unterschied (x)
Klangfarbe / Timbre	2.Klangfarbliche Ausprägung im Höhenbereich 3.Klangfarbliche Ausprägung im Mittenbereich 4.Klangfarbliche Ausprägung im Tiefenbereich 5.Klangfarbe hell-dunkel 6.Schärfe 7.Rauigkeit 8.Kammfilterartigkeit 9.Nasalität 10.Metallische Klangfarbe 11.Wärme 12.Brillianz 13.Dumpfheit 14.Dröhnen 15.Registerdurchsichtigkeit
Geometrie / Geometry	16.Richtung (x) 17.Distanz 18.Ausdehnung in die Tiefe 19.Ausdehnung in die Breite 20.Ausdehnung in die Höhe 21.Größe 22.Örtliche Zerfallenheit 23.Lokalisationsunschärfe 24.Räumliche Transparenz (o)
Raum / Room	25.Halligkeit 26.Stärke des Nachhalls 27.Dauer des Nachhalls 28.Ungleichmäßigkeit im Nachhallverlauf 29.Umhüllung der Nachhall 30.Echo 31.Flatterecho 32.Offenheit
Zeit / Time	33.Zeitliche Klarheit 34.Ansprechverhalten 35.Reaktionsfreudigkeit
Dynamik / Dynamics	36.Lautstärke 37.Dynamikumfang
Artefakte / Artifact	38.Störgeräusch (x)
Allgemein / General	39.Sprachverständlichkeit (s) 40.Intimität 41.Lebendigkeit 42.Räumliche Präsenz 43.Gefallen 44.Hörsamkeit 45.Leichtigkeit des Zuhörens 46.Lautstärkebalance (o) 47.Globale Balance 48.Klangliche Durchmischung (o) 49.Klangfülle 50.Sonstiges (x)

A.5 Room Acoustical Parameters

Table A.2: Room acoustical parameters, room volumes and critical distances for the 35 selected rooms in frontal position with a single source taken from Ackermann and Ilse (2015) (V: Volume, r_H: Critical Distance, EDT: Early Decay Time, T_{30}: Reverberation Time (30 dB), C_{80}: Clarity, D_{50}: Definition, G: Sound Power Factor , T_s: Centre Time

Room	**V [m^3]**	**r_H [m]**	**EDT [s]**	**T_{30} [s]**	**C_{80} [dB]**	**D_{50}**	**G [dB]**	**T_s [ms]**
Gewandhaus	22051	5.29	1.96	2.27	2.33	0.50	6.90	0.10
Komische Oper	7057	3.88	0.87	1.32	3.82	0.36	7.79	0.07
Bas. of Eberbach Mon.	20924	3.46	4.30	5.33	-0.17	0.45	10.79	0.21
Cl. du Couvent d. C.	8805	4.88	0.90	1.07	7.86	0.71	4.09	0.04
Cultuurzentrum	6097	3.74	1.17	1.23	3.76	0.55	5.85	0.08
Teatre Jean Vilar	7705	5.40	0.62	0.76	8.10	0.82	5.62	0.03
Kammersaal 1	2323	2.30	0.89	1.26	6.36	0.66	11.61	0.05
Kammersaal 2	3217	1.17	3.08	7.08	0.01	0.42	14.75	0.18
Kirche	12453	2.95	3.27	4.19	-2.69	0.27	11.74	0.20
Konzertsaal 1	21659	4.45	3.26	3.20	1.09	0.44	5.99	0.16
Konzertsaal 2	10260	2.97	2.25	3.48	-0.82	0.40	7.64	0.15
Teatro Farnese	43790	6.90	2.79	2.68	0.80	0.52	2.04	0.13
Teatro Olimpico 1	3158	2.08	2.13	2.12	-1.17	0.30	13.16	0.14
Yachiyo-Za	1942	2.70	0.65	0.74	9.73	0.83	8.06	0.03
Murakuni-Za	1376	2.31	0.83	0.74	7.58	0.74	9.72	0.03
Kaho-Gekijo	4629	3.59	0.83	1.04	6.35	0.72	6.79	0.04
Kanamaru-Za	2757	2.76	0.91	1.04	5.16	0.68	9.17	0.05
Houou-Za	1071	2.33	0.63	0.58	11.19	0.86	9.16	0.02
Concertgebouw	20786	5.40	2.06	2.07	1.32	0.47	4.52	0.11
Dortmund	18902	3.53	1.57	4.24	0.02	0.37	4.87	0.11
Elmia	11124	4.68	1.51	1.46	2.70	0.52	6.16	0.08
Eurogress	14196	5.20	1.37	1.51	2.98	0.50	3.66	0.07
Haus fuer Musik	13536	4.82	1.61	1.69	2.51	0.54	6.16	0.08
Kursaal	6656	3.95	1.10	1.23	4.16	0.58	9.53	0.06
Sejong Concert Hall	34480	8.43	1.51	1.42	2.41	0.45	0.40	0.08
Seminar Room HFT616	166	0.87	0.54	0.64	10.15	0.81	18.91	0.03
Jesus Christ Church	8077	2.92	2.54	2.77	1.61	0.49	9.65	0.12
Gewandhaus 1781	2002	1.44	1.74	2.89	1.17	0.44	13.44	0.11
Eglise du College St M.	9541	2.46	4.64	4.55	-2.69	0.30	13.35	0.27
Gulbenkian Hall	11057	4.18	1.70	1.83	4.85	0.62	5.33	0.07
Oper	14695	4.98	1.45	1.68	-0.21	0.46	2.88	0.09
Aula 1	5977	3.38	1.54	1.53	2.15	0.49	9.74	0.09
Frosinone	2301	2.08	1.56	1.55	0.87	0.42	14.50	0.10
Kammermusiksaal	707	1.75	0.63	0.67	8.43	0.76	14.06	0.04
Santa Maria de Melque	1371	1.64	1.47	1.48	3.11	0.54	17.76	0.08