Binaural Monitoring for Live Music Performances

E L Í A S Z E A

KTH Computer Science and Communication

Master of Science Thesis
Stockholm, Sweden 2012
Binaural Monitoring
for Live Music Performances

E L Í A S Z E A

DT212X, Master’s Thesis in Music Acoustics (30 ECTS credits)
Degree Progr. in Computer Science and Engineering 270 credits
Royal Institute of Technology year 2012
Supervisor at CSC was Roberto Bresin
Examiner was Sten Ternström

TRITA-CSC-E 2012:070
ISRN-KTH/CSC/E--12/070--SE
ISSN-1653-5715

Royal Institute of Technology
School of Computer Science and Communication
KTH CSC
SE-100 44 Stockholm, Sweden
URL: www.csc.kth.se
To my wonderful family,
fulfilled with such love
and passion for music
Binaural Monitoring for Live Music Performances

Abstract

Current monitoring systems for live music performance rely on having a sound engineer who manipulates sound levels, reverberation and/or panning of the sounds onstage to simulate a spatial rendering of the monitor sound. However, this conventional approach neglects two essential features of the sound field: directivity radiation patterns, and spatial localization; which are naturally perceived by the performers under non-amplified conditions. The present work comprises the design, implementation and evaluation of a monitoring system for live music performance that incorporates directivity radiation patterns and binaural presentation of audio. The system is based on four considerations: the directivity of musical instruments, the Room Impulse Response (RIR), binaural audio with individualized Head-Related Transfer Functions (HRTFs), and motion capture of both the musician’s head and instrument. Tests with musicians performing simultaneously with this system were carried out in order to evaluate the method, and to identify errors and possible solutions. A survey was conducted to assess the musicians’ initial response to the system in terms of its accuracy and realism, as well as the perceived degree of enhancement of the music making experience and the artistic value of the performance. The results point towards further research that might be of interest.
Binaural Övervakning för Musik
Föreställningar

Referat

Acknowledgements

The author would like to thank the people that were essential sources of motivation, support, knowledge and resources during this work:

- Roberto Bresin, Sten Ternström, Damian Murphy, Anders Askenfelt, Anders Friberg, Kjetil Hansen and Julio Walter for the incentives, guidance, assistance and support during this journey
- Gunnar Julin for his receptivity, support and assistance with the contacts for the musicians used in the experiments
- The 27 anonymous musicians for participating in the experiments and evaluation of this work
- Jukka Pätynen and the Virtual Acoustics team from Aalto University, for the valued support in providing the directivity functions of orchestral instruments
- Room Acoustics Team at IRCAM, France, for providing free access to the Listen individualized HRTF databases
- Regents of the University of California, for providing permission to make use of the CIPIC individualized HRTF databases
- My girlfriend Fra, as well as friends and colleagues from all around the world, for the fantastic suggestions, shared ideas, motivation, and "fika" breaks
- My wonderful family, whose unconditional support, artistic professionalism and passionate commitment to their art have encouraged me to think of such challenges in live music performances
Contents

Acknowledgements i

Glossary v
  List of Abbreviations v
  List of Symbols vii

List of Figures ix

1 Introduction 1
  1.1 Overview of the Report 2

2 Background and Theory 3
  2.1 Audio and Signal Processing 3
    2.1.1 Sound Pressure Level 3
    2.1.2 Nyquist-Shannon Sampling Theorem 4
    2.1.3 Analog-to-Digital conversion 5
    2.1.4 Discrete Fourier Transform 5
    2.1.5 Signal Windowing 6
    2.1.6 Signal Convolution 7
    2.1.7 Interpolation Algorithms 7
  2.2 Dynamic Tracking and Auralization 10
    2.2.1 Towards Optimum Performance 10
    2.2.2 Rigid Bodies and Degrees of Freedom 10
    2.2.3 Translation Motion 10
    2.2.4 Rotational Motion 11
  2.3 Musical Acoustics 13
    2.3.1 Spectral Cues and Psychoacoustics 13
    2.3.2 Radiation Directivity 14
  2.4 Room Acoustics Modeling 14
  2.5 Human Sound Localization 15
    2.5.1 Interaural Cues 15
    2.5.2 The Cone of Confusion 16
    2.5.3 Head-Related Transfer Functions 17
    2.5.4 Binaural Sound with HRTF 18
3 Overview of the Method
3.1 The Framework
3.2 Development Tools
3.2.1 Pure Data
3.2.2 ARENA
3.2.3 UDP Network

4 Audio Acquisition with Pd
4.1 Microphones
4.2 MOTU Audio Interface and Drivers
4.3 Audio Drivers for Pd

5 Motion Capture Scheme
5.1 OptiTrack Motion Capture System
5.1.1 Latency and Frame Rate
5.1.2 Spatial Resolution
5.1.3 Calibration
5.1.4 Tracking Volume
5.2 Rigid Bodies
5.2.1 Headphones
5.2.2 Instruments
5.3 The OSC Thread

6 Towards Directivity of Acoustic Instruments with Pd
6.1 Radiation Orientation
6.1.1 Radiation Azimuth
6.1.2 Radiation Elevation
6.2 Computational Algorithms for DFs
6.2.1 Directivity Database
6.2.2 Trilinear Interpolation of SPL
6.2.3 Bilinear Interpolation of DFs
6.3 Hann Windowing
6.4 Adaptive Gain Filter

7 Room Acoustic Model with Pd
7.1 The Room Impulse Response
Glossary

List of Abbreviations

- **3-D**: Tridimensional
- **A/D**: Analog-to-Digital
- **AGF**: Adaptive Gain Filter
- **ASIO**: Audio Stream Input/Output
- **CB**: Critical Bands
- **CD**: Compact Disc
- **D/A**: Digital-to-Analog
- **DF**: Directional Function
- **DFT**: Discrete Fourier Transform
- **DOF**: Degrees of Freedom
- **DRR**: Direct-to-reverberant Ratio
- **DSM**: Dynamic Sound Mixing
- **DSP**: Digital Signal Processing
- **DTF**: Directional Transfer Function
- **DVD**: Digital Versatile Disc
- **ERB**: Equivalent Rectangular Bandwidths
- **FFT**: Fast-Fourier Transform
- **FPS**: Frames per second
- **FWM**: Floor-Wedge Monitoring
- **HRIR**: Head-Related Impulse Response
- **HRTF**: Head-Related Transfer Function
- **I/O**: Input/Output
- **IDFT**: Inverse Discrete Fourier Transform
- **IEM**: In-Ear Monitoring
- **ILD**: Interaural Level Difference
- **ITD**: Interaural Time Difference
- **IFFT**: Inverse Fast-Fourier Transform
- **IR**: Impulse Response
- **MCS**: Motion Capture System
- **MIDI**: Musical Instrument Digital Interface
- **MMIO**: Memory-mapped Input/Output
- **MOTU**: Mark Of The Unicorn
- **OSC**: Open Sound Control
- **PC**: Personal Computer
- **Pd**: Pure Data
- **RB**: Rigid body
- **RIR**: Room Impulse Response
- **rms**: Root-mean square
- **SNR**: Signal-to-Noise Ratio
- **SPL**: Sound pressure level
- **SSM**: Static Sound Mixing
- **SVM**: Support Vector Machine
- **UDP**: User Datagram Protocol
- **USB**: Universal Serial Bus
- **VAE**: Virtual Acoustic Environment
List of Symbols

- $\beta$: A certain acoustic instrument
- $\mu$: Amount of frequency partials excluding the fundamental
- $\theta$: Azimuth angle
- $\theta_b$: Binaural azimuth angle
- $\theta_{\text{dir}}$: Radiation azimuth angle
- $\psi$: Elevation angle
- $\psi_b$: Binaural elevation angle
- $\psi_{\text{dir}}$: Radiation elevation angle
- $dB_{iM,\beta}$: Sound pressure level for the $i$-th partial of the signal of instrument $\beta$ relative to the head of musician $M$
- $f_c$: Center frequency
- $f_{i\beta}$: Frequency component of the $i$-th partial of instrument $\beta$
- $f_k$: $k$-th nominal one-third octave frequency band
- $f_{\text{max}}$: Highest frequency component in the spectrum
- $f_s$: Sampling frequency
- $G_{\text{rms}_{i\beta}}$: Normalized FFT coefficients of adaptive gain filter for the $i$-th partial of instrument $\beta$
- $L_p$: Sound Pressure Level in decibels
- $M$: Amount of musicians
- $N$: Length of a finite sequence
- $p_0$: Reference pressure
- $p_{\text{rms}}$: Effective pressure (rms)
- $P'$: Translation motion
- $PC1$: Personal Computer running ARENA software
- $PC2$: Personal Computer running Pure Data
- $PC3$: Personal Computer running OSCNatNetClient
- $R'$: Rotational motion
- $S$: Spatial matrix of the stage plot
- $T_s$: Sampling period
- $V_t$: Tracking volume
- $x_{R_M}(t)$: Right-channel playback for musician $M$
- $x_{R_M,\beta}[n]$: Right-channel signal for musician $M$ relative to instrument $\beta$
- $x_{dir_{M,\beta}}[n]$: Directional signal of instrument $\beta$ relative to the head of musician $M$
- $x_{RIR_M,\beta}[n]$: Reverberant signal of instrument $\beta$ relative to the head of musician $M$
- $x_{L_M}(t)$: Left-channel playback for musician $M$
- $x_{L_M,\beta}[n]$: Left-channel signal for musician $M$ relative to instrument $\beta$
- $Y$: Playback matrix
List of Figures

2.1 A certain signal $x(t)$. Top: time domain representation. Bottom: frequency domain representation. Taken from [37]. ................................. 4
2.2 Hann window function of length $N$. Taken from [56]. ......................... 7
2.3 Grid depicting bilinear interpolation of unknown point $P(x, y)$. Taken from [54]. ............................... 8
2.4 Cube depicting trilinear interpolation of unknown point $F[x_i, y_i, z_i]$. ............... 9
2.5 Euler rotation axis $\hat{e}$ and rotation angle $\phi$. ................................. 11
2.6 Euler angles $\phi, \theta$ and $\psi$ on the coordinate system $(x, y, z)$. Taken from [55]. ... 12
2.7 An Equivalent Rectangular Bandwidth centered in $f_c$, depicting neighbor partials $f_n$ and $f_m$, and their difference in hertz $\Delta f_{n,m}$............................................... 14
2.8 The structure of a room impulse response. Taken from [24]. ....................... 15
2.9 Interaural Time Difference, taken from [34]. ................................. 16
2.10 The cone of confusion. Taken from [47]. ................................. 17
2.11 Measurements of individual and generalized HRTF. **Top-left:** HRTF magnitude plot. **Bottom-left:** HRTF phase plot. **Right:** HRIR of around 120 discrete samples. Taken from [34]. ................................. 18
2.12 Block diagram of binaural monitoring with static sound mixing. .................. 20
2.13 Block diagram of binaural monitoring with dynamic sound mixing. ............... 21
2.14 Generalized Floor-Wedge Monitoring scheme for a playback channel. .......... 22
2.15 Generalized In-Ear Monitoring scheme for a playback channel. .................. 23
3.1 High-level structure diagram of the binaural monitoring system. .................. 26
3.2 Diagram of the UDP network. $PC_1$ runs ARENA as a UDP server, $PC_3$ runs an OSC client, and $PC_2$ runs Pure Data as a UDP receiver. ................................. 27
4.1 Block diagram of audio acquisition scheme. ........................................ 29
5.1 Block diagram of motion capture scheme. ........................................ 33
5.2 OSC thread for rigid body data used in the UDP network. ....................... 35
6.1 Block diagram of directivity radiation scheme. .................................... 37
6.2 A certain orientation $(\theta_{dir}, \psi_{dir})$ between the head of a musician $M$ relative to an instrument $\beta$. The blue arrow and the yellow circle are the frontal axis and the median plane of the instrument, respectively. ................................. 38
6.3 Cello radiation table in Pd for 28 FFT bins and a microphone positioned at
(108°, -11°). ................................................................. 41
6.4 Block diagram of digital signal processing. ........................................... 42
6.5 Two-dimensional grid depicting four known DTFs from the database and an
unknown function \( H_{ij}[k, \theta_{ij}, \psi_{ij}] \). .................................................. 44
6.6 Pd code to generate four-overlapping Hann window functions. .......... 45
7.1 Block diagram of the room acoustic model. ................................. 47
8.1 Pythagorean triangle in the x-z plane depicting binaural azimuth angle \( \theta_b \). The
blue arrow is the frontal axis of the musician’s head. ......................... 50
10.1 Audio settings window in Pure Data. ................................................. 55
10.2 Measurement of acquisition latency with Pure Data. ......................... 56
11.1 Results for perceived quality obtained in the questions of the survey for experi-
mental tests. ........................................................................... 62
11.2 Results for perceived quality obtained in the questions of the survey for listening
tests. ................................................................................... 63
A.1 Musical score used in the experimental tests (1/2). .......................... 77
A.2 Musical score used in the experimental tests (2/2). ......................... 78
A.3 Survey for experimental tests (1/3). .................................................. 79
A.4 Survey for experimental tests (2/3). .................................................. 80
A.5 Survey for experimental tests (3/3). .................................................. 81
A.6 Survey for listening tests (1/2). .......................................................... 82
A.7 Survey for listening tests (2/2). .......................................................... 83
B.1 Network patch in Pure Data showing the UDP connection. PC1 corresponds
to "Studio 2" and PC2 corresponds to "Studio 1" ................................. 85
B.2 MoCap-to-IP patch in Pure Data showing the code for the OSC thread used in
the network. .......................................................................... 86
B.3 Rigid-body patch in Pure Data showing the OSC messages routed into the six
degrees of freedom received from the network. ................................. 87
B.4 Directivity-Instruments patch in Pure Data showing the code blocks for the
radiation database, the adaptive gain filter, FFT analysis and trilinear interpo-
lation. ...................................................................................... 88
B.5 RIR-model patch in Pure Data showing the Room Impulse Response, and the
code blocks of the external partconv~ and the direct-to-reverberant sound ratio. 89
B.6 Binaural-model patch in Pure Data showing the code blocks for angular com-
pensation, the external cw_binaural~ and the HRTF dataset used (CIPIC, subject 003). 90
B.7 Interface for experimental tests with Pd. Directivity of instruments is excluded. 91
B.8 Interface for binaural monitoring with Pd. Directivity of instruments is included. 92
C.1 Omnidirectional microphone Behringer ECM 8000. 93
C.2 Pickup microphone Brüel and Kjær 4021 mounted in the bridge of the violin. 94
C.3 MOTU Traveler mk3 multi-channel audio interface. 95
C.4 A Natural Point OptiTrack infrared camera of the ARENA motion capture system. 96
C.5 The rigid body constructed and mounted in one of the headphones Sennheiser HDR-220. 96
C.6 The rigid body constructed and mounted in the violin. 97
C.7 The rigid body constructed and mounted in the bassoon. 97
C.8 Perspective of the room where the experiments were carried out (1/2). 98
C.9 Perspective of the room where the experiments were carried out (2/2). 99
Chapter 1

Introduction

The amplified playback provided to the musicians in a live music performance is based either on In-Ear Monitoring (IEM) with the use of earplugs, or Floor-Wedge Monitoring (FWM) with power amplifiers. A sound engineer may manipulate sound levels, reverberation and/or panning of the sounds on stage to simulate a spatial rendering of the playback. However, these traditional methods neglect two essential collaterals of the music-making experience: directivity radiation patterns and spatial localization. We could hypothesize that, since these elements are naturally perceived by the human ear under non-amplified conditions in live music performances, they may become relevant and valuable for the interpretation of the performer and the audience alike. A motivation for this argument can be found in commercial applications using wave-field synthesis with spatialized feedback for the audience in concert halls [1].

Still, musical interpretation is a complex process in which musical expression, creativity, imagination and communication interact and determine the artistic value of the performance. If technology is introduced in such process, it should be for the sake of enhancing the enjoyment of the musical experience itself, given that some musicians and/or audiences are not attracted to the use of technology.

It becomes then reasonable to think that, under amplified conditions in live music performances, musicians would like to hear their own performance as radiating from their own instruments and voices, as well as that the playback conveys their spatial location and movements on stage. Additionally, acoustic instruments produce a great variety of unique sounds that sometimes suffer from the use of conventional monitoring schemes [57].

The main research question in this thesis work is therefore to overcome the lack of interaction and realism in traditional monitoring technologies by designing a new immersive music environment in which the spatialization and the directional radiation of the sounds on stage become clearly conveyed. Can binaural audio, directivity radiation and motion capture technology be used in the production of live music and enhance the artistic value of the performance?

The main goal of such solution is to provide the musicians with a natural music-
making experience when performing under amplified conditions. Clearly, such an augmentative approach will be necessary only when the performance is carried out in a room with poor acoustics that requires amplification.

A second limitation of the present project was the number of simultaneous musicians. Given the complexity of the signal processing algorithms and network communications, very high computational power was needed to use the system with three (3) musicians. The estimated audio latencies were then between 25 and 35 ms. However, for two (2) simultaneous performers the latency was 20 ms. Further details are presented in Chapter 10.

1.1 Overview of the Report

Chapter 2 presents the fundamental theory as well as previous and background research relevant to the presented study. In Chapter 3, a preliminary overview of the method is presented, providing the architecture, software tools and block diagrams. The solution for real-time audio acquisition with Pd is covered in Chapter 4, remarking the microphone selection, the audio interface and drivers. In Chapter 5, the chosen framework for motion capture of the musician’s head and the instruments is presented, pointing out the elements of such architecture and tests done. Directional modeling of acoustic instruments is presented in Chapter 6, followed by interpolation methods and signal processing algorithms to make use of a radiation database of orchestral instruments. In Chapter 7, room acoustic modeling with Pd is described, pointing out the resources used and the chosen solution. Chapter 8 presents the method for binaural rendering with Pd, individualized and generalized HRTF databases. The spatial audio playback is presented in Chapter 9, justifying the headphone selection and the tridimensional mix. An analysis of the method is described in Chapter 10, presenting computational power and delay measurements as well as the subjective evaluation of the system with the survey. Discussion of the results is presented in Chapter 11. Future work is devised in Chapter 12, leading to the conclusions obtained with the study in Chapter 13.
Chapter 2

Background and Theory

2.1 Audio and Signal Processing

The human auditory frequency range, also known as the auditory bandwidth, is approximately 20 Hz to 20 kHz. Any audio signal the human ear is capable of perceive is then encountered within this frequency region. Audio instrumentation and processing techniques therefore take into account these considerations. In this way, there are several concepts that are detailed in the following sections, essential to understand the audio acquisition from the analog to the digital world; as well as signal processing techniques and algorithms that are used in the present study.

2.1.1 Sound Pressure Level

There are two conventional and reciprocal signal domains: time and frequency. These two representations are related through the Fourier Transform and entirely characterize the signal with temporal and spectral quantities. At the moment it is essential to present the definition of sound pressure level (SPL), and Fourier will appear again later.

When referring to the frequency domain, the term "spectrum plot" is often found, depicted in the bottom of figure 2.1. The horizontal and the vertical axis of the spectrum correspond to the frequency in hertz (Hz) and the SPL in decibels (dB).

In audio, SPL is defined as the quotient between the effective (rms) sound pressure of the signal \( p_{\text{rms}} \) and a reference pressure \( p_0 \) (usually 20\( \mu \)Pa). The definition of SPL is shown in equation 2.1, where \( L_p \) is in dB.

\[
L_p = 10 \log_{10} \left( \frac{p_{\text{rms}}^2}{p_0^2} \right) = 20 \log_{10} \left( \frac{p_{\text{rms}}}{p_0} \right)
\]  

(2.1)

Audio transducers measure the sound pressure of the air, converting it into an electric signal. Once such an analog signal is obtained, a sampling procedure is applied in order to transform the signal into a sequence of numerical samples.
2.1.2 Nyquist-Shannon Sampling Theorem

Around 1920, Harry Nyquist discovered a sampling criterion when working together with Claude Shannon on information theory [35, 45]. Nyquist realized that to perfectly reconstruct a waveform once sampled, as well as achieve distortionless transmission, the sampling frequency $f_s$ (the inverse of the sampling period $T_s$) has to be at least two times the bandwidth $B$ of the signal. However, $B$ is not necessarily the auditory bandwidth. Instead, $B$ is the maximum frequency component $f_{max}$ found in the spectrum of the signal.

$$f_s \geq 2B = 2f_{max} \quad (2.2)$$

By looking at figure 2.1 and disregarding the noise level, Nyquist states that $x(t)$ has to be sampled with at least $f_s = 2$ Hz. Nowadays, one of the most common sampling frequencies used for audio is 44.1 kHz, given that the effective bandwidth of the waveforms $B$ is around 22 kHz; which is somewhat larger than the auditory bandwidth.

Nowadays, audio interfaces typically provide sampling rates of 44.1 kHz, 48 kHz, 88.2 kHz, 96 kHz, 176.4 kHz and 192 kHz. Intuitively we may realize that the higher the sampling rate is, the more accurate the reconstruction of the signal will be.
2.1. AUDIO AND SIGNAL PROCESSING

2.1.3 Analog-to-Digital conversion
Also known as A/D, it is the digitization process of a sampled waveform. The
theory underlying is based on using digital bits to represent the samples of the
signal, dividing the voltage range into discrete levels that correspond to the bit
depth and linearity of the converter. A/D conversion therefore has a fine but finite
resolution, usually in the order of µvolts, that will produce a quantization error $q_e$,
which is a random signal that alters the original waveform. In principle, higher bit
depth allows lower resolution, thus $q_e \approx 0$. Audio interfaces present depths of 16-,
20- or 24-bit, which are the standard recording qualities of CDs and DVDs.

2.1.4 Discrete Fourier Transform
The dynamic time response of a system is known as the impulse response (IR). This
function, usually labeled as $h(t)$, entirely characterizes linear and time-invariant
(LTI) systems. In practice, systems are considered as LTI in order to calculate
the output $y(t)$ according to any input signal $x(t)$ through a mathematical method
called convolution.

Such model in time has its reciprocal representation in the frequency domain,
called transfer function $H(\omega)$. Particularly, for cases regarding discrete samples,
such a transfer function is referred to as $H[k]$, which is the Discrete Fourier Trans-
form (DFT) of a discrete series of numbers $h[n]$, with $n = 0, 1, ..., N - 1 \in \mathbb{R}$ [36].
However for such a transform to be computed, the signal has to be a finite sequence
of values.

Fourier states that an $N$-length sequence $h[n]$ can be represented as a combina-
tion of complex exponentials or equivalently trigonometric functions

$$H[k] = \sum_{n=0}^{N-1} h[n] \cdot e^{-i\frac{2\pi}{N}nk}$$

(2.3)

where $h[n]$ corresponds to the sampled value of $h(t)$ at $t = nT_s$, and $H[k]$ has a
magnitude and phase response shown in equation 2.4

$$|H[k]|^2 = \Re\{H[k]\}^2 + \Im\{H[k]\}^2$$
$$\arg H[k] = \arctan 2(\Im\{H[k]\}, \Re\{H[k]\})$$

(2.4)

where $\Re\{H[k]\}$ and $\Im\{H[k]\}$ are respectively real and imaginary parts of $H[k]$.

The inverse DFT (IDFT) is the way of coming back from the frequency domain
to the time domain. Equation 2.5 shows the inverse transformation from the dis-
crete samples $H[k]$ to the discrete values $h[n]$. In this way, Head-Related Transfer
Functions may be computed as the DFT of the Head-Related Impulse Responses
and vice versa through the IDFT.

$$h[n] = \frac{1}{N} \sum_{k=0}^{N-1} H[k] \cdot e^{i\frac{2\pi}{N}nk}$$

(2.5)
In audio processing, most of the methods are applied to samples of the signals because of the low computational cost and processing time, as well as high efficiency. The computational model used in the present study to calculate the DFT is called fast-Fourier Transform (FFT). As the name suggests, FFT is computationally more efficient, thus real-time implementation can be achieved under very low latency conditions. The IDFT equivalent is commonly known as inverse FFT (IFFT).

We must remember now that the signals have to be a periodic sequence in order to be transformed into the frequency domain, thus a process, called windowing, that satisfies such mathematical restriction is overviewed in the following section.

### 2.1.5 Signal Windowing

Prior to the transformation into the frequency domain via equation 2.3, a window function has to be applied to the signal. Therefore, the incoming sampled signal $x[n]$ is truncated to the length of the window, usually of an interval of $N$ samples. Samples outside this interval are zero-valued. Thus once the incoming signal is windowed to a finite sequence, the DFT can be calculated.

A compromise has to be established in terms of frequency resolution and dynamic range. Usually the term DFT bin is found in discrete signal processing, which refers to components in the frequency spectrum. The resolution of such bins is maximized by means of increasing the length $N$ of the sequences. In audio, the $k$-th bin in Hz is found via equation 2.6

$$bin_k = k \cdot \frac{f_s}{N}$$

We must now consider the resolution in bins of the window function. We can observe that an error will be introduced, which is called frequency deviation $\Delta f_k$ in the present study. Equation 2.7 describes the frequency deviation between a frequency component in the continuous spectrum $f_k$ and the discrete value of $bin_k$.

$$-\frac{f_s}{2N} \leq \Delta f_k \leq \frac{f_s}{2N}$$

In the present study, four overlapping Hann functions are used to window the signal because of low spectral leakage, commonly known as aliasing in sampling theory, as well as enhanced frequency resolution for larger data sets. However, the frequency resolution (see Chapter 6) is not as high as other window functions. Still it was considered as a good and simple approach for coding real-time digital signal processing with Pure Data. Figure 2.2 depicts the time-domain Hann window and equation 2.8 is the mathematical representation $w[n]$ [56].

$$w[n] = 0.5 \left[ 1 - \cos\left(\frac{2\pi n}{N} - 1\right) \right]$$

where $N$ corresponds to the size of the window.
Once in the frequency domain, a mathematical operation called convolution is applied to the signal. In the next section, convolution between the incoming signal $x[n]$ and an IR $h[n]$ is presented.

### 2.1.6 Signal Convolution

Following the theory of LTI systems, the output sequence $y[n]$ can be obtained from the IR of the system $h[n]$ and the input sequence $x[n]$. However, operations in the time-domain usually are more time consuming than in the frequency-domain. Also, given that FFT and IFFT algorithms provide an efficient solution for transformation between the time and frequency domains, typically the convolution follows a multiplication of the DFTs of $x[n]$ and $h[n]$.

$$y[n] = x[n] * h[n] = \frac{1}{\tau} \text{IFFT}\{X[k] \cdot H[k]\}$$  \hspace{1cm} (2.9)

where $\tau$, shown in equation 2.10, is a normalization value that corresponds to the window method applied (four overlapping Hann) to the signal $x[n]$ [27]

$$\tau = 1.5N$$  \hspace{1cm} (2.10)

After computing the IFFT and appropriate normalization, the signal has to be windowed again. Thus in the present study, $y[n]$ is multiplied by the four overlapping Hann windows.

### 2.1.7 Interpolation Algorithms

Let us consider a function of discrete values $F = \{F_0, F_1, ..., F_j, ..., F_m\}$. We know the discrete samples of $F$, however the values in between a given interval $(F_j, F_{j+1})$
are not known. Interpolation is then commonly implemented to retrieve such unknown points [6]. The choice of interpolation method depends on the number of variables, the precision, the efficiency, among many other factors. In this section we will focus on the algorithms with respect to the number of independent variables.

The most basic method is linear interpolation. Let us consider a discrete function of one variable \( F[x] = \{ F[x_0], F[x_1], \ldots, F[x_j], \ldots, F[x_m] \} \), we are interested in computing an unknown value \( F[a] \in I = (F[x_j], F[x_{j+1}]) \). Thus, equation 2.11 describes the computation of \( F[a] \).

\[
F[a] = F[x_j] + \frac{(a - x_j)F[x_{j+1}] - (a - x_j)F[x_j]}{x_{j+1} - x_j}
\]  

(2.11)

However, some of the functions presented in this report depend on more than one variable. Therefore it becomes necessary to understand 2nd and 3rd order interpolation algorithms.

Bilinear interpolation is an extension of linear interpolation to two independent variables. We can treat the problem as a square or grid (see Figure 2.3). Lets consider a discrete function \( F[x,y] \) whose values \( F[x_1, y_1], F[x_1, y_2], F[x_2, y_1] \) and \( F[x_2, y_2] \) are known and respectively referred to as \( F[Q_{11}], F[Q_{12}], F[Q_{21}] \) and \( F[Q_{22}] \). We are now interested in obtaining the unknown value \( F[x_i, y_i] \) depicted in Figure 2.3 as \( P(x,y) \).

![Figure 2.3. Grid depicting bilinear interpolation of unknown point \( P(x,y) \). Taken from [54].](image)

To simplify the formulas, normalization of the lattices of the grid is computed

\[
\begin{pmatrix}
F[Q_{11}] \\
F[Q_{12}] \\
F[Q_{21}] \\
F[Q_{22}]
\end{pmatrix}
= \begin{pmatrix}
F[0,0] \\
F[0,1] \\
F[1,0] \\
F[1,1]
\end{pmatrix}
\]

(2.12)
2.1. AUDIO AND SIGNAL PROCESSING

Then the procedure is done via successive interpolation of one variable after another. Thus following equation 2.11 for the variables $x$ and $y$, we obtain equation 2.13 [54]

$$
F[x_i, y_i] = F[0, 0](1 - x_i)(1 - y_i) + F[0, 1](1 - x_i)y_i + F[1, 0]x_i(1 - y_i) + F[1, 1]x_iy_i
$$ (2.13)

In the same way, trilinear interpolation consists of a 3-successive linear interpolation of a discrete function $F[x, y, z]$. We treat the problem as a cube, where normalization is applied to its lattices. Figure 2.4 depicts the unit cube with the eight normalized lattices $F[0, 0, 0], F[0, 0, 1], ..., F[1, 1, 1]$ and an unknown point $F[x_i, y_i, z_i]$. Again, following equation 2.11 successively for the variables $x, y$ and $z$, we lead to equation 2.14

$$
F[x_i, y_i, z_i] = F[0, 0, 0][1 - x_i][1 - y_i][1 - z_i] + \\
F[1, 0, 0]x_i[1 - y_i][1 - z_i] + \\
F[0, 1, 0][1 - x_i]y_i[1 - z_i] + \\
F[0, 0, 1][1 - x_i][1 - y_i]z_i + \\
F[1, 0, 1]x_i[1 - y_i]z_i + \\
F[0, 1, 1][1 - x_i]y_i z_i + \\
F[1, 1, 0]x_i y_i [1 - z_i] + \\
F[1, 1, 1]x_i y_i z_i
$$ (2.14)

Figure 2.4. Cube depicting trilinear interpolation of unknown point $F[x_i, y_i, z_i]$. 

9
2.2 Dynamic Tracking and Auralization

Auralization, commonly referred to as the creation of virtual acoustic environments (VAEs), is an emerging research field that has become popular in the last decades. Surround sound, binaural audio, wave-field synthesis, among others; have converged nowadays into what we consider as immersing the listener in a spatial audio experience [57].

In 1999, Savioja and co-workers introduced an auralization framework with orchestral instruments synthesized with MIDI [44]. Savioja’s study dealt with the directional radiation of the musical instruments, an RIR model and binaural, transaural and/or multichannel synthesis. Movements of the immersed listener were tracked in order to achieve full spatial exploration of the acoustic environment.

Other studies found that there are three fundamental perceptual cues that determine the accuracy of a dynamic tracking mechanism of a certain VAE. Taken into account such parameters, an optimum auralization can be performed and full spatial exploration can be achieved [43].

2.2.1 Towards Optimum Performance

In order to render a satisfactory and efficient VAE, there are three basic parameters to consider: latency, frame rate and smallest spatial measurement [43]. Time delays and latencies of the tracking system for optimum performance was found to be 29 ms. Furthermore, the necessary update rate obtained in the same study was about 60 Hz (equivalently 60 FPS). And the last value found was of about 1° for the minimum spatial resolution the tracking mechanism has to satisfy.

It is enough to say that by satisfying these three conditions, an adequate performance can be achieved with dynamic tracking systems for auralization. In Chapter 5, the motion capture scheme is presented with further details regarding these variables.

2.2.2 Rigid Bodies and Degrees of Freedom

The concept of rigid body (RB) is useful for the dynamic tracking of objects. A mechanism is known as a composition of RB members. Each of these bodies determine the kinematic properties of the mechanism [46]. Therefore, in the case of dynamic tracking, motion capture of RBs is commonly implemented, providing six degrees of freedom (DOF) on a fixed coordinate system \((x_0, y_0, z_0)\). The DOF of such objects are divided into two categories: translation motion and rotational motion.

2.2.3 Translation Motion

In such movement, the RB moves from one point in space to another. There is a vector \(P_0 \in \mathbb{R}^3\) representing the initial position \((x_0, y_0, z_0)\) and a translation
2.2. DYNAMIC TRACKING AND AURALIZATION

vector $P_\Delta \in \mathbb{R}^3$ represented by $(x_\Delta, y_\Delta, z_\Delta)$ [9]. Therefore, translation motion $P'$ is described by equation 2.15.

$$P' = (x_0 + x_\Delta, y_0 + y_\Delta, z_0 + z_\Delta)$$  \hspace{1cm} (2.15)

2.2.4 Rotational Motion

In some cases, motion capture mechanisms make use of unit quaternions to represent rotational movements. This notation has numerical advantages in stability, simplicity and robustness in comparison to traditional Euler angles. Thus, rotational motion of a RB can be described with an axis, also known as Euler rotation axis $\hat{e} \in \mathbb{R}^3$, shown in equation 2.16 [26]

$$\hat{e} = [e_1, e_2, e_3]$$  \hspace{1cm} (2.16)

Let us consider $\hat{e}$ and $\phi$ as the Euler rotation axis and angle respectively, depicted in figure 2.5; thus the definition of a quaternion $Q \in \mathbb{R}^4$ is shown in equation 2.17

$$Q = \left[ \hat{e} \cdot \sin(\frac{\phi}{2}), \cos(\frac{\phi}{2}) \right]$$

$$= \left[ e_1 \sin(\frac{\phi}{2}), e_2 \sin(\frac{\phi}{2}), e_3 \sin(\frac{\phi}{2}), \cos(\frac{\phi}{2}) \right]$$  \hspace{1cm} (2.17)

![Figure 2.5. Euler rotation axis $\hat{e}$ and rotation angle $\phi$.](image)

Quaternion components are also represented by $q_1, q_2, q_3, \text{ and } q_4 \in \mathbb{R}$. Therefore, in equation 2.18, expressions for individual components of the quaternion in function of Euler rotation axis and angle are found

$$\begin{pmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{pmatrix} = \begin{pmatrix} e_1 \sin(\frac{\phi}{2}) \\ e_2 \sin(\frac{\phi}{2}) \\ e_3 \sin(\frac{\phi}{2}) \\ \cos(\frac{\phi}{2}) \end{pmatrix}$$  \hspace{1cm} (2.18)
CHAPTER 2. BACKGROUND AND THEORY

These components are correlated, and a fundamental condition shown in equation 2.19 has to be satisfied [26]

\[ |Q| = 1 \Rightarrow q_1^2 + q_2^2 + q_3^2 + q_4^2 = 1 \]  

(2.19)

Once a quaternion has been defined, it becomes essential to present the conversion between quaternions and Euler angles. Equation 2.20 describes the matrix transformation from a quaternion \( Q \) to the set of Euler angles \( \phi, \theta \) and \( \psi \).

\[
\begin{pmatrix}
\phi(Q) \\
\theta(Q) \\
\psi(Q)
\end{pmatrix} = \begin{pmatrix}
\arctan\left(\frac{2(q_1q_2 + q_3q_4)}{1 - 2(q_1^2 + q_2^2)}\right) \\
\arcsin\left(2(q_1q_3 + q_2q_4)\right) \\
\arctan\left(\frac{2(q_1q_4 + q_2q_3)}{1 - 2(q_3^2 + q_4^2)}\right)
\end{pmatrix}
\]  

(2.20)

where \( \phi, \theta \) and \( \psi \) correspond to roll, pitch and yaw respectively, depicted in figure 2.6

![Figure 2.6. Euler angles φ, θ and ψ on the coordinate system (x, y, z). Taken from [55].](image)

In this way, there will be an initial vector \((\phi_0, \theta_0, \psi_0)\) and a rotation vector \((\phi_\Delta, \theta_\Delta, \psi_\Delta)\), that will both define rotation motion \(R'\) shown in equation 2.21

\[ R' = (\phi_0 + \phi_\Delta, \theta_0 + \theta_\Delta, \psi_0 + \psi_\Delta) \]  

(2.21)
2.3 Musical Acoustics

The sound radiation of a certain acoustic instrument depends inherently on its spectral properties, physical material and dimensions, resonant frequencies; as well as the orientation and position from a given listening point in space. Such properties of instruments become relevant to the present work in terms of musical acoustics and psychoacoustics. In particular we will be concerned with frequency analysis of musical signals that will be overviewed in Chapter 6. In the following sections, literature and theory relevant to the present study are presented.

2.3.1 Spectral Cues and Psychoacoustics

Our perception of the sound of an instrument corresponds largely to the resultant frequencies within a rich and complex spectrum. In this way, two complementary psychoacoustic considerations are now presented: spectral masking and auditory filter banks.

Spectral masking establishes a threshold in dB between the frequency components (partials) in the spectrum. A partial whose magnitude relative to the spectrum’s maximum is below this threshold will not significantly be perceived.

\[ 20 \log_{10} \left( \frac{|X[k_{\text{max}}]|}{|X[k_i]|} \right) < \gamma, \quad i = 0, 1, \ldots, N - 1 \] (2.22)

where \( |X[k_{\text{max}}]| \) and \( |X[k_i]| \) are the Fourier amplitudes of the loudest and the i-th DFT bins respectively found in the spectrum when computing N-length FFT. In research studies, the value of \( \gamma \) has been found in the range of 40 and 60 dB [17].

Following equation 2.22, a relationship between the amplitudes of the DFT bins can be obtained via equation 2.23

\[ |X[k_i]| > |X[k_{\text{max}}]| \cdot 10^{\gamma/20}, \quad i = 0, 1, \ldots, N - 1 \] (2.23)

The second psychoacoustic criterion is related to the auditory filters. Some studies refer to them as Critical Bands (CB) [59, 21]. Others call them Equivalent Rectangular Bandwidths (ERB) [31]. In the present study, we will refer to them as ERBs as functions of center frequency \( f_c \)

\[ ERB_N(f_c) = 24.7(0.00437f_c + 1) \] (2.24)

The detectability of partials in complex signals has been studied, pointing towards an approximate threshold for differences in hertz between different tones. A detectability accuracy of partials of about 75% is found when they are separated by a difference of at least 1.25 \( ERB_N \) [32]. Looking at Figure 2.7, this principle means that \( \Delta f_{n,m} \) should be at least 1.25 times \( ERB_n(f_{cn}) \); which leads to equation 2.25

\[ \Delta f_{n,m} > 30.875(0.00437f_{cn} + 1), \quad n \neq m \] (2.25)

where \( n \) and \( m = 0, 1, \ldots, N - 1 \), and \( f_{cn} \) is the center frequency of the \( ERB \) where the partial \( f_n \) is found.
2.3.2 Radiation Directivity

Extensive research has been carried out on the directional radiation patterns of musical instruments [30, 7, 52, 53, 39], among others; obtaining databases of discrete-sampled radiation patterns such as in [38].

Pätynen constructed a database of radiation patterns of acoustic instruments with 22 microphones positioned in a tetrahedral configuration surrounding the musician; thus responses are found for one-third octave-band frequencies of typical woodwind, brass and string instruments [38]. Averaged data for different tones was computed in order to provide a generalized directivity function of the instruments, depending on three variables: azimuth angle $\theta$, elevation angle $\psi$ and frequency $f$. The output of this function is the SPL corresponding to the instrument obtained from each microphone in the array, in 28 frequency bands (or FFT bins).

In the present study, sound radiation of acoustic instruments is modeled as a function of frequency, and azimuth and elevation angles: $H_{rad}(f, \theta, \psi)$. Thus for a given frequency component $f_i$ in a certain listening point in space $P$, trilinear interpolation via equation 2.14 is computed to estimate the SPL within the eight nearest discrete values measured in the database [38]. This method is presented in detail in Chapter 6.

2.4 Room Acoustics Modeling

In this report room acoustics will not be deeply described. Instead, databases of impulse responses and applications for binaural rendering of audio will be overviewed.

One of the reasons of choosing binaural sound is the fact that the room acoustic response of the place where the live performance is carried out does not extremely
2.5. HUMAN SOUND LOCALIZATION

affect the signal processing chain; given that, normally, omnidirectional near-field microphones are used. Thus, any RIR can be used for modeling a virtual environment where the musicians are performing.

The structure of an RIR, depicted in Figure 2.8, consists of three fundamental elements: direct sound, early reflections and late reverberations. Such a time-domain representation is important to understand, although we would like to pay more attention to the frequency domain.

Taking advantage of FFT algorithms, a finite-length RIR can be convolved with an audio signal in real-time in order to immerse the listener in a certain room. Free access, under some conditions, to some useful databases and RIRs can be obtained in several websites [48, 16, 29, 49], among others. Using these open-source resources, an efficient FFT algorithm for RIR convolution is implemented in the present study. It is described in greater detail in Chapter 7.

2.5 Human Sound Localization

We now recall the seemingly trivial fact that the human ear has the natural ability to perceive with impressive accuracy the spatial location and motion of a sound source. In the present study, when referring to non-monitored situations, localization and movements of the instruments are natural components of the music-making process. Thus the need arises for including these elements into amplified conditions. In the following sections, sound localization will be overviewed remarking relevant aspects and features of the human auditory system.

2.5.1 Interaural Cues

There are four cues that allow our auditory system to locate a sound source in space: interaural cues, pinnae differences, head movements and sight. However interaural
cues are sufficient to render a sound in space [34, 14]. These cues are divided into Interaural Time Difference (ITD) and Interaural Level Difference (ILD).

ITD, shown in equation 2.26 [34], corresponds to the difference in time, resulting in phase shifting, between the sound waves that arrive at each ear.

\[ ITD = \frac{r}{c}(\theta + \sin \theta) \]  
(2.26)

where \( r \) is half of the ear-to-ear measure, \( \theta \) is the incidence angle of the sound wave and \( c \) is the speed of sound.

In the same way, ILD describes the sound level difference between the sound waves that reach each ear.

![Interaural Time Difference](source)

**Figure 2.9.** Interaural Time Difference, taken from [34].

However, there is a frequency threshold that divides the prominency of each cue, which means that in some cases ILD or ITD will not provide the correct information. Such approximate threshold is presented in the condition 2.27

\[
\text{Prominent cue} \begin{cases} 
\text{ITD}, & \text{when } f < 700\text{Hz} \\
\text{ILD}, & \text{when } f > 700\text{Hz}
\end{cases}
\]  
(2.27)

### 2.5.2 The Cone of Confusion

We now observe that there might be points in space with the same ITD, because sound waves from different sources could reach the ears with equal phase shifting. Thus we may expect that confusion might occur when localizing the sound sources. This identity is, in fact, referred to as the cone of confusion and is depicted in Figure 2.10. However, a healthy auditory system is still able to identify and resolve such conditions.
2.5. HUMAN SOUND LOCALIZATION

2.5.3 Head-Related Transfer Functions

HRTFs are mathematical functions that model the ILDs, ITDs, spectral differences as well as head shadowing, pinnae shaping, torso and shoulder diffractions [34]. Such functions are measured by placing one microphone in each ear of a mannequin’s head, in the case of generalized datasets; or of an individual’s head, in the case of individualized databases. Typically a loudspeaker emits sound waves towards the listening point at frequencies covering all the auditory bandwidth. This procedure is repeated for several points of a spherical coordinate system \((r, \theta, \psi)\), where \(r\) is kept constant (usually to 1 meter).

The spatial resolution of the HRTFs is a crucial factor for optimum auralization, along with the minimum spatial measure the tracking mechanism can do. Intuitively we can note that the higher the resolution is, either more measurements have to be done on the spherical plane or higher computational cost and power is needed for smooth interpolation among discrete values.

As mentioned in the previous sections, HRTF is the DFT of the HRIR, both depicted in Figure 2.11

It is worth pointing out that every individual will have an unique HRIR according to his/her anthropomorphic data, e.g. head dimensions, torso width, pinnae shapes, among many others [15]. In the very beginning of the research, the author was considering to include anthropomorphic measurements along with a Support Vector Machine classification (SVM) algorithm of multiple features [23] within individualized HRTF databases. However such model, described in Chapter 8, was not included in the final code.

Measurements of HRTF provide then a method (specially when using individual-
ized datasets) for rendering sound sources in space along with headphone (binaural) or loudspeaker (transaural) playback. However, the problem of transaural audio is much more complex, given the well-known crosstalk between the left and right channels respectively for left and right ears; as well as the interaction with the acoustic response of the room where the sound synthesis is carried out. This would require the measurement, design and implementation of an active crosstalk cancellation (XTC) and inverse RIR filters [18].

In the present study the author dealt with binaural sound in order to avoid such complexity, and present a preliminary standard that will provide backwards compatibility with IEM schemes. In the following section, HRTF databases and resources for real-time binaural synthesis are devised.

### 2.5.4 Binaural Sound with HRTF

As previously mentioned above, binaural technology is implemented in the present work to simplify the problem of 3-D sound playback. Free HRTF databases are available to use under some conditions [19, 15, 2], among others; whose spatial resolution is in the order of 5°.

The software used to develop the computational models and algorithms is Pure Data (presented in Chapter 3). An external code (`cw_binaural` [11]) was used to fulfill the need of real-time HRTF decomposition and interpolation, as well as binaural synthesis.

Doukhan made a decomposition of the HRIR into minimum phase and all-pass components prior to the spatial interpolation [11, 57]; fulfilling the entire spherical resolution within the measurements of individualized HRTF databases Listen [2] and CIPIC [15]. Then a convolution algorithm is implemented via equation 2.9.
2.6. SOUND MIXING

this way DFT of an incoming audio signal $x[n]$ is computed, and is multiplied with the DFT of the interpolated HRIR $h[n]$ according to a given listening point with the angles $(\theta, \psi)$, thus providing a binaural image of the sound source in space.

In the present study, according to the theory of dynamic tracking reviewed in Section 2.2 and the variables used by Doukhan, angular compensation of binaural rendering is modeled by means of considering only two head movements: rotation and tipping. Some studies, such as [51, 4], reveal the influence of head movements correlated to sound localization. It has been found that when humans attempt to localize a sound source, they tend to rotate and tip the head with about respectively 45% and 15% probabilities [51]. However, head pivoting (which in Section 2.2 corresponds to roll Euler angle) seems to be barely 4% probable [51]. In this way, the mapping of Euler angles for binaural rendering with HRTF is usually done via azimuth-to-rotation and elevation-to-tipping; disregarding the roll movements which humans do not usually tend to make.

2.5.5 Sound Externalization

Since the very beginning of studies in human sound localization, researchers started to wonder about an issue with headphone reproduction: the so-called 'inside-the-head locateness' (IHL) [4]. Indeed, sound source externalization remains one of the biggest challenges in binaural technology [18, 22].

Externalization of sound images depends on four factors: pinnae shape, reverberation cues, interaural cues and sight. It turns out that room reverberation is the most relevant cue to successfully externalize a sound image from the head [18]. It is fundamental then to include a room response to enhance lateral externalization of sounds from the headphones, which usually suffer from IHL, thus avoid confusion along the front-back listening axis.

2.6 Sound Mixing

In this section, two definitions for sound mixing are presented, following the binaural monitoring scheme: static and dynamic sound mixing. Static mix corresponds to render static binaural signals to each of the performers with the spatial information of the instruments, omitting the use of tracking technologies. In the present study, the author implemented and evaluated the binaural monitoring system under conditions of dynamic sound mixing, which means that every movement of the head and/or the instrument will be used to compensate the spatial rendering in real-time.

2.6.1 Static Sound Mixing (SSM)

In cases where the musicians do not tend to move on stage (academic music), it turns out that one may think that SSM can be easily implemented in real-life situations (e.g. the concert). And even more if the computation of directional radiation patterns is bypassed. It then becomes a single signal processing stage in
the chain, which corresponds to the synthesis of binaural audio with the spatial information of the instruments on stage. Figure 2.12 shows the block diagram of binaural monitoring under conditions of static sound mixing.

Let us consider the rider, also known as stage plot, provided by the musicians to the sound engineer prior to the performance. A two-dimensional grid can be considered, and an $M \times 2$ matrix is defined via equation 2.28, containing the coordinate pairs $(x_i, y_i)$ of the musicians

$$S = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_M & y_M \end{bmatrix}$$

(2.28)

where $i = 1, 2, ..., M$, and $(x_i, y_i)$ corresponds to the (static) spatial localization of the $i$-th musician in the stage plot.

The spatial matrix, $S$, is used to compute the angles for spatial rendering with HRTFs. As a matter of simplification, all the elevation angles, and the azimuth angle of each musician relative to himself/herself can be set to zero degrees. However it could be modified according to the musician's preferences.

If the computation of directional radiation patterns is bypassed, the range of musical sounds that can be monitored with the binaural system is greatly broadened. This means that even electronic-based instruments and human voices can be mixed under conditions of SSM.

### 2.6.2 Dynamic Sound Mixing (DSM)

Situations in which the musicians are likely to move on stage (e.g. rock or pop concert), motivate the hypothesis of including a tracking mechanism to capture the movements of the musicians and their instruments. Dynamic sound mixing
2.7 MONITORING SYSTEMS

comprises the use of motion capture techniques, and compensation of the spatial and directional information of the instruments. Figure 2.13 presents the block diagram of binaural monitoring under conditions of dynamic sound mixing.

Figure 2.13. Block diagram of binaural monitoring with dynamic sound mixing.

On the contrary of having an $M \times 2$ spatial matrix under conditions of SSM, two $M \times 5$ matrices have to be computed according to 5 degrees of freedom provided by the tracking mechanism. These two matrices correspond to the translational position, azimuth angle and elevation angle of the musicians and the instruments, respectively.

The main limitation of binaural mixing is, as with SSM, the directional properties of electronic instruments. Only acoustic instruments whose directional model is known can include the signal processing stage of directional radiation. However it can be easily solved by means of bypassing the computation of directional functions.

In general terms, if directional radiation is not considered, the use and processing of radiation databases is left out from the problem. Hence, any specific instrument could be monitored with binaural feedback, disregarding its particular radiation properties which may or may not be known. However, since the goal of the present study is to recreate the acoustic events that occur under non-amplified conditions (particularly for acoustic instruments), computation of directional radiation in real-time was performed with the use of a radiation database (refer to Section 6.1).

2.7 Monitoring Systems

A sound engineer is employed at live music performances to create the sound mixes required both for the musicians and for the audience. In the following sections, studies regarding monitoring systems of on stage sounds are presented.
2.7.1 FWM vs. IEM

Traditionally, Floor-Wedge Monitoring is the most common technique found on live music performances. Powerful, precise and specialized speakers are used to provide the playback to the musicians. The reproduced audio consists of mono signals sent via one or more speakers on stage individually assigned to each performer. Such monitoring scheme is depicted through a generalized model in Figure 2.14.

![Figure 2.14. Generalized Floor-Wedge Monitoring scheme for a playback channel.](image)

Alternatively, In-Ear Monitoring, as the name suggests, involves the use of earplugs for audio playback. Although musicians are isolated from onstage sounds other than those coming from the mix, IEM has the advantage of being stereo-based, and of providing easy tracking and manipulation of the musician’s sweet-spot. Thus stereo signals can be handled, allowing panning techniques and/or binaural audio to be included into the mix. Furthermore, IEM has became very popular in the last decade due to sound quality and hearing-loss protection; and has the advantage of simple hardware requirements for each musician on stage [3]. Figure 2.15 shows a generalized model for IEM playback.

A comparative study of FWM and IEM technologies regarding the musician’s perception of latency on live music has been carried out [28]. This work intuitively points out that latency was found to be lower when using IEM than when using FWM. Also, we have to remember that earplugs provide hearing protection to the musicians, which is an important feature that FWM does not necessarily address.
2.7. MONITORING SYSTEMS

2.7.2 Automated Monitoring

In 1975, Dugan started to study applications for sound reinforcement of speech zones with multiple microphones [12]. He formulated an automatic and adaptive threshold for active sound reduction, enhancing the SNR by means of gating the SPL of a microphone recording the ambient noise. Such technique is called voice-operated switch (VOX). Years later, Dugan published another article regarding automatic mixing functions for live sound [13]. In this case he argued two considerations: detectability thresholds for rejection to noise in microphones, and automatic master gain and feedback control with maximum gain for each microphone.

Reiss and Gonzalez have studied an automatic gain normalization algorithm for prevention of feedback on stage [42]. Such work has been very useful to automate the gain maximization before feedback, which typically sound engineers do by sporadically checking the gain control in the mixing console.

In the same way, Terrell and Reiss studied an automatic monitor mixing scheme [50]. They formulated a conjunction of matrices, where gains for musicians and instruments are suggested as an objective function for optimum feedback from each of the monitors on stage. Such gains are frequency-independent to avoid higher complexity of the algorithms, as well as many realistic effects of sound propagation and degradation are disregarded. The authors implemented an algorithm for automatic feedback control with maximum gain as well [42].
Chapter 3

Overview of the Method

The present work has four major components: motion capture, directional radiation of musical instruments, room acoustics and binaural audio. In the following sections, the methodology of the study will be overviewed, describing the framework from previous work done in [44] and by the author in [57]. Higher-level block diagrams of the system and the development tools are presented.

3.1 The Framework

The author has designed a framework for spatial rendering of amplified instruments [57] following the methodology used in the research project DIVA, developed by Savioja et al. [44]. The differences from what Savioja did are related to the fact that sound synthesis is performed via MIDI and the musical scores are already known by the system. However the scheme itself is very similar, except for audio acquisition and analysis that in [44] are not implemented.

The framework of the present study focuses on developing real-time playback with spatial rendering of musical instruments for live music performances. The structure mentioned above is composed of several sub-components: audio acquisition, motion capture, digital signal processing (DSP), directivity (radiation) functions, room acoustic responses, binaural audio and headphone playback. A high-level diagram of the system is shown in Figure 3.1.

3.2 Development Tools

Three computers were used in the present study. Two DELL Intel Core Duo 2.0 with Windows 7 Enterprise, which will be mentioned as PC1 and PC2, run respectively the motion capture software ARENA and a Pure Data patch. And one Vaio with Windows XP Service Pack 2, called PC3, runs an OSC client to stream the data incoming from ARENA to Pure Data. In the following sections, Pure Data, ARENA and the UDP network will be described.
3.2.1 Pure Data

Pure Data (Pd) is an open-source software based on graphical language and is a programming tool for real-time audio, video and image processing [41]. It has become a useful and powerful tool thanks to its real-time processing capabilities and code simplicity. Also, it supports UDP network communications, thus Open Sound Control (OSC) messages can be sent among applications, computers and even mobile phones.

In the present study, all of the blocks shown in Figure 3.1, except some parts of the motion capture scheme, are programmed in Pd. A main patch runs in PC2, calling abstractions and additional patches. Later on, every Pd patch will be described in detail, as well as images relevant to the code will be shown.

3.2.2 ARENA

OptiTrack ARENA is a motion capture software from the company Natural Point [40]. It is used in conjunction with a motion capture system which consists of eight (8) infrared cameras connected in a daisy chain configuration. Camera data is retrieved in PC1 via USB.

ARENA allows the user to track the 3-D position of infrared markers. Rigid bodies built with 3 or 4 infrared markers can be captured as well, which is the main application in this study. Data streaming can be forwarded to an IP address specified by the user. Some features can be modified such as the characteristics of the cameras: infrared threshold, exposure, intensity and frame rate. For the time being, the author will just mention that this software along with the infrared cameras satisfy the requirements for optimum dynamic tracking. More details regarding the motion capture system are presented in Chapter 5.
3.2. DEVELOPMENT TOOLS

3.2.3 UDP Network

In order to successfully retrieve the motion capture data from ARENA and read it in Pd, an UDP network was designed by means of the three PCs. As shown in Figure 3.2, PC1 acts as a server, PC2 acts as the receiver and PC3 acts as the client. IP addresses of these PCs and an UDP port were used for running the client.

![Diagram of the UDP network](image)

**Figure 3.2.** Diagram of the UDP network. PC1 runs ARENA as a UDP server, PC3 runs an OSC client, and PC2 runs Pure Data as a UDP receiver.

The client used is called OSCNatNetClient [10]. It is a tool developed to send rigid body data from ARENA through OSC messages. However, some compatibility problems were found with Windows 7 and it was the reason of including a third computer (PC3) to run the client.
Chapter 4

Audio Acquisition with Pd

The scheme for audio acquisition, depicted in Figure 4.1, consists of $M$ microphones and a multi-channel audio interface which digitize and sends the numerical samples to Pure Data. In the present study two musicians used the monitoring system during the experiments, hence $M = 2$. However we will generalize the acquisition scheme for $M$ signals.

![Figure 4.1. Block diagram of audio acquisition scheme.](image)

In the figure above, the acoustic signals $x_1(t), x_2(t), ..., x_M(t)$ are sampled with the interface to obtain $x_1[n], x_2[n], ..., x_M[n]$, for instruments $\beta = 1, 2, ..., M$ respectively. The following sections concern the selection of the microphones as well as the audio interface and drivers used.

4.1 Microphones

The author would like to explain the criteria for selecting the microphones used in the present study. The main arguments are based on the acoustic instruments and the number of musicians involved in the experimental tests.

29
CHAPTER 4. AUDIO ACQUISITION WITH PD

Three microphones were used: one lavalier, one omnidirectional (model Behringer ECM 8000) and one violin pickup (model Brüel and Kjær 4021). The instruments used in the experiments were violin, oboe and bassoon. The lavalier microphone was mounted on the bell of the oboe, the omnidirectional was placed pointing towards the holes of the bassoon, and the pickup was mounted in the bridge of the violin. Appendixes C.1 and C.2 show respectively the omnidirectional microphone, and the pickup microphone used in the violin.

Following some rules of thumb for microphone recording techniques [25, 5], the separation of the mikes was appropriately chosen. A separation of 1.5 m between the microphones was used, and approximately 30-50 cm between each microphone and its corresponding instrument was used; thus satisfying a fundamental condition that is shown in equation 4.1

$$\Delta d_{m_{ij}} \geq 3 \cdot \Delta d_{mi}, \text{ and}$$
$$\geq 3 \cdot \Delta d_{mj}$$

(4.1)

where $\Delta d_{m_{ij}}$ corresponds to the separation among microphones $i$ and $j$ and $\Delta d_{mi}$ ($\Delta d_{mj}$) corresponds to the distance from microphone $i$ ($j$) to the instrument $i$ ($j$) intended to be recorded.

Given that the omnidirectional and the pickup microphones are condenser-type, phantom power was needed to supply the transducers. Therefore, these microphones were connected via XLR cables to the multi-channel audio interface supplying 48 Volts with phantom power. The lavalier microphone was connected to a line-in balanced input. The MOTU interface is presented in the following section.

4.2 MOTU Audio Interface and Drivers

A MOTU Traveler mk3 multi-channel audio interface, depicted in Appendix C.3, is used in the present study [33]. Universal MOTU drivers were installed in PC2 in order to work with the interface and Pure Data [33].

The Traveler mk3 has four balanced analog inputs for microphones (1-4) with preamplifiers and 48 V phantom power; as well as four balanced gold-plated inputs (5-8) for line-in connection. The interface is bus powered and transmits 24-bit audio samples via FireWire (IEEE 1394) to the personal computer. The available sampling rates are:

- **1x**: 44.1 kHz and 48 kHz
- **2x**: 88.2 kHz and 96 kHz
- **4x**: 176.4 kHz and 192 kHz

In the present study the sampling rate used was $f_s = 44.1$ kHz. The buffer size of the interface was selected as well, which is proportional to acquisition latency: a
4.3. AUDIO DRIVERS FOR PD

higher buffer size increases the processing time, thus slight delays or signal clicks might be perceived. The author chose a buffer size of 512 samples, from the available range of 64 to 1024, as a compromise between perceptual latency and signal resolution. Details on audio latency are mentioned in Chapter 10.

4.3 Audio Drivers for Pd

In order to work with the MOTU interface channels with no perceivable latency in Pd, a cross-platform driver (ASIO4ALL v2.10 [20]) for Audio Stream Input/Output (ASIO) compatibility was installed in PC2.

However, it was not trivial to make ASIO drivers work with the extended version of Pd (Pd-extended). ASIO drivers are compatible with versions of Pd Vanilla, whereas with Pd-extended multiple channels cannot be read from an audio interface. Only two input and two output channels can be used at a time with Pd-extended, e.g. MOTU Analog-In 1-2 and MOTU Analog-Out 3-4.

When choosing the standard Memory-Mapped Input/Output (MMIO) settings in Pd, which makes use of the internal sound card of the PC, the computing capabilities of the core(s) are distributed to many other things rather than just Pd. The latter means there is no way to reach from the sound card to Pd and vice versa in a direct way without interruptions from other processes such as system sounds, audio from other applications, etc. Thus noticeable latency was found when using MMIO, attributed to the fact that Pd was not running as a dedicated audio process.

On the other hand, Pd Vanilla could not be used for the present study because externals are not supported, thus Doukhan’s binaural external cw_binaural~ [11] could not be implemented. However, an experimental beta release of Pd-extended was found in the Pd website [41], which was tested by the author with the ASIO drivers. Fortunately the drivers worked well, all the MOTU channels could be used and very low latency was achieved.
Chapter 5

Motion Capture Scheme

The motion capture architecture is shown in Figure 5.1. $2M$ (4) rigid bodies are tracked, where $M$ corresponds to the number of musicians (2). By means of a UDP communication network, an OSC server sends six degrees of freedom for each rigid body (6DOF/RB) to an OSC client, which streams the data to a receiver. The following sections will overview the relevant aspects of the motion capture system (MCS) used in the study.

![Figure 5.1. Block diagram of motion capture scheme.](image)

5.1 OptiTrack Motion Capture System

The MCS used consists of 8 infrared cameras from Natural Point OptiTrack [40], one of them depicted in Appendix C.4. Specifications, protocols and features of the system, such as latency, spatial resolution, frame rate, among others, are overviewed in the next sections; thus pointing out the requirements for adequate dynamic tracking and auralization presented in the theory Section 2.2.1.
CHAPTER 5. MOTION CAPTURE SCHEME

5.1.1 Latency and Frame Rate

Of the eight infrared cameras used, four of them are model Flex 13 and the other four are model V100:R2. Latencies are respectively of 8.33 ms and 10 ms, which correspond to an effective frame rate of 120 FPS and 100 FPS. As presented in the theory, the minimum update rate required for optimum dynamic tracking is of 60 FPS, thus a security factor of around 1.67 and 2 is used in the present study.

5.1.2 Spatial Resolution

Performing a precise and adequate calibration, the minimum spatial measurement the cameras can perform is of around 1 mm. This measure, along with the fine spatial resolution of the Pd external \texttt{cw\_binaural}~\cite{11}, provided a total resolution that was not measured in this study. However full and smooth spatial exploration was achieved within the angular boundaries of the HRTF databases used.

5.1.3 Calibration

The process of calibration is divided into two sub-processes. At first, a task called 'wanding' has to be performed, which consists of moving a rigid-body wand (with three infrared markers) within the visual range of the cameras.

After several samples of the wand are captured by the cameras, which was considered approximately around ten thousand as enough, the Calibration Wizard tool in ARENA performs an optimization of the tracking volume and creates a calibration file (.cal).

5.1.4 Tracking Volume

The tracking volume $V$ obtained in the present study, under well-calibrated conditions was approximately 10 m$^3$. The area of the room where the motion capture system was placed was relatively small (around 20 m$^2$), thus it was considered by the author as an additional restriction for the number of musicians using the system.

5.2 Rigid Bodies

The construction of 4 rigid bodies was done with four infrared markers each placed asymmetrically. In the present study, two rigid bodies were constructed respectively for two headphones, and two more respectively for two musical instruments. The Rigid Body Wizard tool in ARENA was used to construct the rigid bodies and create the rigid body files (.rb).

5.2.1 Headphones

In order to track the translation and rotational movements of the musician’s head, a rigid body was constructed with the headphones used for playback. Appendix
5.3. THE OSC THREAD

C.5 depicts the RB in one of the Sennheiser wireless headphones.

5.2.2 Instruments

A rigid body was constructed and placed on every musical instrument, in order to track its orientation and movements. Appendixes C.6 and C.7 respectively depict the RBs mounted in the violin and in the bassoon.

5.3 The OSC Thread

A UDP Network was designed with the purpose of communicating ARENA with Pd. In this way, a standardized OSC thread was established for the network, in order to easily route out the data for each rigid body in Pd. Appendix B.2 depicts the construction of the thread with Pd. The OSC thread is shown in Figure 5.2

| OSC | /rb | M | xyz | /rb | M | ypr |

Figure 5.2. OSC thread for rigid body data used in the UDP network.

where $M$ is the number corresponding to the order in which the rigid body file (.rb) was loaded in ARENA. The string $xyz$ corresponds to the three DOF for translation motion and $ypr$ corresponds to the three DOF for rotational motion. The order in which the RBs were loaded in ARENA was:

- **RB 1**: Pair of headphones 1
- **RB 2**: Pair of headphones 2
- **RB 3**: Violin
- **RB 4**: Oboe/bassoon
Chapter 6

Towards Directivity of Acoustic Instruments with Pd

In the present chapter, an algorithm for implementing real-time directional radiation of acoustic instruments with Pure Data is devised. The definition of directional transfer functions is presented, as well as the use of a directivity database and the algorithm alternatives to compute the model in real-time. In the experimental tests the author was not able to implement these algorithms because of severe computational cost and latency when used together with the rest of the signal processing chain. However the algorithm was tested separately with audio files read from disk (contrabass and violin loops) and changing the rotation of the instrument.

**Figure 6.1.** Block diagram of directivity radiation scheme.

### 6.1 Radiation Orientation

Sound radiation of acoustic instruments was modeled as a function of frequency, and azimuth and elevation angles $H_{rad}(f, \theta, \psi)$, which from now on we will refer to as Directional Function (DF). For a given sound heard in a certain listening point $P$ in space a directional transfer function can be computed with FFT algorithms and
a database of directional radiation of acoustic instruments. In the present study, the author used Pätynen’s directivity database of orchestral instruments [38].

The goal of the following algorithm is to measure the relative orientation between an acoustic instrument and a listening point, which corresponds to the angles \( \theta_{dir}, \psi_{dir} \). Such orientation was modeled with 3-D vector analysis and spherical coordinates, taking as input the six DOF obtained with the motion capture system. In the following sections, the vector analysis is divided into two planes (x-z and x-y) corresponding to radiation azimuth and elevation respectively.

![Image](image.png)

**Figure 6.2.** A certain orientation \( (\theta_{dir}, \psi_{dir}) \) between the head of a musician \( M \) relative to an instrument \( \beta \). The blue arrow and the yellow circle are the frontal axis and the median plane of the instrument, respectively.

### 6.1.1 Radiation Azimuth

In order to calculate the radiation azimuth angle \( \theta_{dir} \), we need to present vector equations regarding the musician’s head \( M \) and the instrument \( \beta \). We know the following variables from the MCS:

- \((x_M, y_M, z_M)\): translation motion of the head of musician \( M \)
- \((x_\beta, y_\beta, z_\beta)\): translation motion of instrument \( \beta \)
- \(\theta_\beta\): azimuth angle relative to frontal axis of instrument \( \beta \)

Therefore we know the vector that represents the head of the musician \( M \) relative to the reference coordinate system \((x_0, y_0, z_0)\), referred to as \( \vec{r}_{mxyz} \).
6.2. COMPUTATIONAL ALGORITHMS FOR DFS

\[ \vec{r}_{m_{xz}} = x_M \hat{x} + z_M \hat{z} \]  
\hfill (6.1)

and the vector that represents instrument \( \beta \) as well

\[ \vec{r}_{\beta_{xz}} = x_\beta \hat{x} + z_\beta \hat{z} \]  
\hfill (6.2)

Thus using equations 6.1 and 6.2 we can calculate the relative vector between musician \( M \) and instrument \( \beta \)

\[ \vec{r}_{rel_{xz}} = \vec{r}_{m_{xz}} - \vec{r}_{\beta_{xz}} \]  
\hfill (6.3)

The frontal axis of the rigid body on the instrument \( \beta \) coincides with zero-degrees azimuth. Thus we also know the unity vector, \( \hat{I}_{\beta_{xz}} \), that characterizes the orientation of instrument \( \beta \) relative to its frontal axis

\[ \hat{I}_{\beta_{xz}} = \cos \theta_\beta \hat{x} + \sin \theta_\beta \hat{z} \]  
\hfill (6.4)

Finally we can compute the desired radiation angle, \( \theta_{dir} \), via vectorial dot product between \( \vec{r}_{rel_{xz}} \) and \( \hat{I}_{\beta_{xz}} \) following equation 6.5

\[ \theta_{dir} = \arccos \left[ \frac{\vec{r}_{rel_{xz}} \cdot \hat{I}_{\beta_{xz}}}{|\vec{r}_{rel_{xz}}|} \right] \]  
\hfill (6.5)

6.1.2 Radiation Elevation

The procedure to calculate the radiation elevation angle \( \psi_{dir} \) is very similar to that of the azimuth angle. The only difference is that instead of using \( \theta_\beta \) of instrument \( \beta \), we need \( \psi_\beta \). In this way, we will only present the formula to compute radiation elevation according to the X-Y plane variables between the head of musician \( M \) and instrument \( \beta \).

\[ \psi_{dir} = \arccos \left[ \frac{\vec{r}_{rel_{xy}} \cdot \hat{I}_{\beta_{xy}}}{|\vec{r}_{rel_{xy}}|} \right] \]  
\hfill (6.6)

6.2 Computational Algorithms for DFs

The author developed a real-time computational algorithm with Pd to find the SPL needed for a certain frequency component \( f_i \) of the musical signal of instrument \( \beta \), with position \((\theta_{dir}, \psi_{dir})\) relative to the head of a musician \( M \). This method was done following a trilinear interpolation with Päätynen’s database [38], although an additional and more elegant solution is described.
6.2.1 Directivity Database

An overview to the database [38] has to be presented prior to describe the computational algorithms, in order to understand the variables that are handled.

Due to the positioning of the microphones realized by Pätynen, there are two different array configurations that depend on the elevation angle. Equation 6.7 depicts the azimuth angles of the microphones for the two array configurations [38].

\[ \theta = \begin{cases} 
0^\circ, 72^\circ, 144^\circ, 216^\circ \text{ or } 288^\circ, & \text{for } \psi = 11^\circ \text{ and } 53^\circ \\
36^\circ, 108^\circ, 180^\circ, 252^\circ \text{ or } 324^\circ, & \text{for } \psi = -53^\circ \text{ and } -11^\circ
\end{cases} \] (6.7)

However, as an elevation angle can be found in the interval \([-11^\circ, 11^\circ]\), two pairs of two boundaries are found for the azimuth angle. Equation 6.8 generalizes \(\theta_{\text{low}}\) and \(\theta_{\text{high}}\) as a piecewise function of the elevation interval. We will refer to low and high boundaries as the theory on interpolation algorithms refers to the normalized lattices: "0" with low and "1" with high (see Section 2.1.7).

\[ \theta_{\text{low}} = \begin{cases} 
36^\circ, 108^\circ, 180^\circ, 252^\circ \text{ or } 324^\circ, & \text{for } \psi_{\text{low}} \in [-53^\circ, -11^\circ] \\
0^\circ, 72^\circ, 144^\circ, 216^\circ \text{ or } 288^\circ, & \text{for } \psi_{\text{low}} = 11^\circ
\end{cases} \]  (6.8)

\[ \theta_{\text{high}} = \begin{cases} 
36^\circ, 108^\circ, 180^\circ, 252^\circ \text{ or } 324^\circ, & \text{for } \psi_{\text{high}} = -11^\circ \\
0^\circ, 72^\circ, 144^\circ, 216^\circ \text{ or } 288^\circ, & \text{for } \psi_{\text{high}} \in [11^\circ, 53^\circ]
\end{cases} \]

thus, both \(\theta_{\text{low}}\) and \(\theta_{\text{high}}\) will have individual low- and high-edge boundaries and we define \(\theta_{\text{dir}}\) with equation 6.9

\[ \theta_{\text{dir}}(\psi_{\text{dir}_{\text{low}}}, \psi_{\text{dir}_{\text{high}}}) = \begin{cases} 
\theta_{\text{dir}_{\text{low}}}, & \text{when using } \psi_{\text{dir}_{\text{low}}} \text{ boundaries} \\
\theta_{\text{dir}_{\text{high}}}, & \text{when using } \psi_{\text{dir}_{\text{high}}} \text{ boundaries}
\end{cases} \]  (6.9)

6.2.2 Trilinear Interpolation of SPL

A first approach developed by the author was to perform a trilinear interpolation of sound pressure levels using the database [38]. Thus following the previous discussion on the variables of Pätynen’s database, the algorithm is done in Pd with the construction of a radiation database, consisting of 22 x 28 tables for each acoustic instrument in the study [38], corresponding to the 22 microphone positions and the 28 one-third octave frequency bands. An example of the database is depicted in Figure 6.3. At the output, the corresponding SPL for a given orientation of instrument \(\beta\) is obtained. Further details on real-time FFT filtering are presented in Sections 6.3 and 6.4.

We need to overview the extraction of frequency partials according to digital signal processing algorithms and psychoacoustic criteria presented in the theory. The PD object fiddle~ is used for the frequency analysis. Five frequency components with their corresponding SPL are found. From now these frequencies will be
6.2. COMPUTATIONAL ALGORITHMS FOR DFS

Figure 6.3. Cello radiation table in Pd for 28 FFT bins and a microphone positioned at (108°, -11°).

referred to as FFT bins. A window of 2048 elements is used, making an analysis every 1024 samples. Equation 6.10 depicts the processing time for every analysis

\[ t_{\text{fiddle}} = \frac{\text{Analysis size}}{f_s} = \frac{1024}{44100 \text{Hz}} \simeq 23.2 \text{ms} \quad (6.10) \]

The lowest frequency that \textit{fiddle}~ will detect is found in equation 6.11, corresponding to 2.5 cycles per analysis window size

\[ f_{\text{Low}} = \frac{f_{\text{Low, Nyquist}}}{2} = \frac{2.5}{2t_{\text{fiddle}}} \simeq 53.8 \text{Hz} \quad (6.11) \]

By simultaneously following the criterions in equations 2.23 and 2.25, a discrimination of the five partials is performed in order to choose the most noticeable \( \mu \) frequency components that are going to be processed and used for the construction of the DTF. In the case that condition 2.23 is satisfied but 2.25 is not, the loudest partial is selected and the other is left out. We can observe that a set of \( \mu + 1 \) functions will be found with the frequency analysis (see equation 6.12 and Figure 6.4).

\[
H_{\text{rad}} = \begin{cases} 
H_0[\text{bin}, \theta_{\text{dir}}, \psi_{\text{dir}}], & \text{for } \text{bin} = \text{bin}_0 \\
H_1[\text{bin}, \theta_{\text{dir}}, \psi_{\text{dir}}], & \text{for } \text{bin} = \text{bin}_1 \\
& \vdots \\
H_{\mu}[\text{bin}, \theta_{\text{dir}}, \psi_{\text{dir}}], & \text{for } \text{bin} = \text{bin}_\mu 
\end{cases} \quad (6.12)
\]
where \( \text{bin}_0 \) corresponds to fundamental frequency and \( \text{bin}_1, ..., \text{bin}_\mu \) correspond to 1st, ..., \( \mu \)-th partials.

In this study the author decided to choose the value of \( \mu = 2 \) in order to reduce computational cost but still affect the directivity of, what usually are, the most prominent spectral components.

Using the triplet of variables \((\text{bin}_i, \theta_{dir}, \psi_{dir})\), with \( i = 0, 1, 2 \), the SPL at a certain discrete FFT \( \text{bin}_i \) can be calculated via trilinear interpolation with the lattices of the cube in Figure 2.4. We change \( F[x, y, z] \) for \( H_{rad}[\text{bin}_i, \theta_{dir}, \psi_{dir}] \). However, as the values for radiation azimuth angles \( \theta_{dir} \) depend inherently on the values for radiation elevation angles \( \psi_{dir} \), the interpolation is not entirely commutative. This means that first we need to compute the interval in which \( \psi_{dir} \) is found and then search for the intervals in which \( \theta_{dir} \) is located. The order in which the frequency variable \( f \) is interpolated is commutative.

Once the lattices are found at the database [38], a normalization of the intervals from \([\text{Low} - \text{edge}, \text{High} - \text{edge}]\) to \([0, 1]\) is computed. Thus variables \( \text{bin}_i, \theta_{dir} \) and \( \psi_{dir} \) of the individual components of \( H_{rad} \) are normalized within the new intervals via equation 6.13

\[
\begin{align*}
0 < \text{bin}_i < 1, & \quad \text{for } i = 0, 1, 2 \\
0 < \theta_{dir_{low}} < 1, & \quad \text{for } \psi_{low} \text{ boundaries} \\
0 < \theta_{dir_{high}} < 1, & \quad \text{for } \psi_{high} \text{ boundaries} \\
0 < \psi_{dir} < 1
\end{align*}
\] (6.13)

Therefore trilinear interpolation of the discrete FFT \( \text{bin}_i \) can be done via equation 2.14 by substituting the variables \([x, y, z]\) for \([\text{bin}_i, \theta_{dir}, \psi_{dir}]\). This algorithm is running in parallel \( \mu + 1 \) times (\( \mu + 1 \) frequency components) for each audio channel.
6.2. COMPUTATIONAL ALGORITHMS FOR DFS

\[ H_{ij}[\text{bin}_i, \theta_{dir}, \psi_{dir}] = H[0, 0, 0][1 - \text{bin}_i][1 - \theta_{dir_{low}}][1 - \psi_{dir_{low}}] + \\
H[1, 0, 0]\text{bin}_i[1 - \theta_{dir_{low}}][1 - \psi_{dir_{low}}] + \\
H[0, 1, 0][1 - \text{bin}_i][1 - \theta_{dir_{low}}][1 - \psi_{dir_{low}}] + \\
H[0, 0, 1][1 - \text{bin}_i][1 - \theta_{dir_{high}}]\psi_{dir_{high}} + \\
H[1, 0, 1]\text{bin}_i[1 - \theta_{dir_{high}}]\psi_{dir_{high}} + \\
H[0, 1, 1][1 - \text{bin}_i]\theta_{dir_{high}}\psi_{dir_{high}} + \\
H[1, 1, 0]\text{bin}_i\theta_{dir_{low}}[1 - \psi_{dir_{low}}] + \\
H[1, 1, 1]\text{bin}_i\theta_{dir_{low}}\psi_{dir_{low}} \tag{6.14} \]

6.2.3 Bilinear Interpolation of DFs

A more elegant solution to directional functions with Päätynen’s database is bilinear interpolation of DFs. Such method will be presented now, however the author did not manage to implement nor compare its performance with the trilinear interpolation algorithm.

We can forget about extraction of frequency partials and psychoacoustic criterions presented in the previous sections, and think about directional functions as transfer functions depending on frequency. The author introduces now the concept of Directional Transfer Function (DTF), as the frequency response of an instrument presenting a certain orientation from its frontal axis relative to the head of a listener.

By means of Päätynen’s database, we currently know discrete versions of the DTFs at 22 angular combinations (see Figure 6.5). Hence the DTF \( H_{ij}[k, \theta_{dir}, \psi_{dir}] \) is defined via equation 6.15

\[ H_{ij}[k, \theta_{dir}, \psi_{dir}] = \begin{bmatrix}
H_{ij}[0, \theta_{dir}, \psi_{dir}] \\
H_{ij}[1, \theta_{dir}, \psi_{dir}] \\
\vdots \\
H_{ij}[N - 1, \theta_{dir}, \psi_{dir}] 
\end{bmatrix} \tag{6.15} \]

where \( k = 0, 1, ..., N - 1, N = 28 \) FFT bins, and \( H_{ij}[k, \theta_{dir}, \psi_{dir}] \) corresponds to the directional radiation of instrument \( j \) relative to musician \( i \).

Therefore it is necessary to interpolate the radiation azimuth and elevation angles. Again, as the radiation azimuth depends on the radiation elevation, we first interpolate the elevation variable. Then, azimuth interpolation is executed and we have a transfer function that has 28 FFT bins. However the problem is not as simple as it looks because an interpolation of 28 elements times four lattices (see Figure 2.3) has to be carried out. Equation 6.16 describes the computation of bilinear interpolation to the orientation of four known 28-bin DTFs, and conversion of frequency resolution in the FFT algorithm (\( N' = 8192 \))

43
$D(i,j) = H_j[k', \theta_i, \psi_i](\theta_h - \theta_d_{ij})(\psi_h - \psi_d_{ij}) + H_j[k', \theta_i, \psi_i]\theta_h(\psi_h - \psi_d_{ij}) + H_j[k', \psi_h]\theta_d_{ij}(\psi_h - \psi_d_{ij}) + H_j[k', \theta_h, \psi_h]\theta_d_{ij}\psi_d_{ij}$

where $k' = 0, 1, ..., 8191$, evaluating the bins $k = 0, 1, ..., 27$ at the corresponding values in $k'$. Since the frequency values correspond to one-third octave bands from 25 Hz to 12.5 kHz, the FFT bins altered are found via equation 6.17, and the magnitude of the remaining $k' - k$ bins is set to unity.

$k' = \min \left\{ \| \frac{f_k \cdot 2N'}{f_s} \| \cdot \frac{f_s}{2N'} - f_k \| , \| \frac{f_k \cdot 2N'}{f_s} \| \cdot \frac{f_s}{2N'} - f_k \| \right\}$

where $f_k$ corresponds to the $k$ nominal values of the one-third octave bands, $N'$ is the new FFT resolution (8192), $f_s$ the sampling rate of 44.1 kHz. The filter that convolves the audio signals is very similar to that used in trilinear interpolation (see Section 6.4). Still, the author does not know to what extent bilinear interpolation of DFs will be more efficient than the trilinear interpolation algorithm.
6.3 Hann Windowing

The code to generate four-overlapping Hann window functions with \( N = 8192 \) samples is depicted in Figure 6.6. The block~ object is used to change the amount of FFT samples, as well as oversampling/downsampling or overlapping windows.

![Figure 6.6. Pd code to generate four-overlapping Hann window functions.](image)

The Hann windows multiply the signal of instrument \( \beta \), \( x_\beta[n] \), acquired with the MOTU before converting it via FFT, and after performing IFFT. The same procedure is repeated for the remaining \( \beta - 1 \) instruments. Now the missing process is the convolution in the frequency domain, which is carried out with an Adaptive Gain Filter (AGF), presented in the next section.

6.4 Adaptive Gain Filter

A set of \( \mu_\beta + 1 \) SPLs are obtained for instrument \( \beta \) throughout the trilinear interpolation algorithm, corresponding to the attenuation of the \( \mu_\beta + 1 \) partials found. Therefore a conversion of decibels to root-mean square (rms) pressure is done via equation 6.18

\[
p_{rms,i_\beta} = 10^{\frac{dB_{i_\beta}}{20}}
\]  

(6.18)

where \( p_{rms,i_\beta} \) is the rms value of a given partial \( f_{i_\beta} \) with an attenuation of \( dB_{i_\beta} \) decibels found via trilinear interpolation.

As we are not using the frequency scale but the bin scale, a frequency-to-bin conversion for \( N = 8192 \) bins and \( f_s = 44.1 \) kHz is done via equation 6.19

\[
bin_{i_\beta} = \frac{N \cdot f_{i_\beta}}{f_s} \approx \frac{f_{i_\beta}}{5.4}
\]  

(6.19)
where $bin_{i,\beta}$ is a decimal number corresponding to scaling $f_{i,\beta}$ to bins; thus a rounding algorithm to the nearest integer value is done with equation 6.20 to obtain $binr_{i,\beta}$

$$binr_{i,\beta} = \left\lfloor \left\{ \frac{f_{i,\beta}}{5.4} \right\} + 0.5 \right\rfloor \tag{6.20}$$

where $\{f_{i,\beta}/5.4\}$ is the decimal part of $bin_{i,\beta}$. Thus, the frequency resolution is about 5.4Hz and, according to equation 2.7, the frequency deviation $|\Delta f_{i,\beta}| \leq 2.7$ Hz. The AGF is then modified at $binr_{i,\beta}$ with the corresponding $rms$ pressures obtained in 6.18.

The $rms$ values are raised to four and divided by a normalization factor via equations 2.10 and 6.21 according to the FFT block size ($N = 8192$) and the amount of overlapping windows (4)

$$G_{rms_{i,\beta}} = \frac{p_{rms_{i,\beta}}^4}{\tau} = \frac{p_{rms_{i,\beta}}^4}{1.5 \cdot N} = \frac{p_{rms_{i,\beta}}^4}{12288} \tag{6.21}$$

where $G_{rms_{i,\beta}}$ are the normalized values of the FFT coefficients used in the convolution. Multiplication in the frequency domain of $G_{rms_{i,\beta}}$ with real and imaginary parts of the FFT of the audio signal is computed. This procedure is done for every channel acquired with the audio interface, providing the signal $x_{dir_{M,\beta}}[n]$ for the input of the RIR model corresponding to musician $M$ relative to instrument $\beta$. 
Chapter 7

Room Acoustic Model with Pd

The current chapter devises the real-time FFT algorithm that adds reverberation to the incoming audio signal (after the directivity model) \( x_{\text{dir},M,\beta}[n] \) by means of a room impulse response and compensation of direct-to-reverberant sound ratio, depicted in Figure 7.1. The same notation is used regarding the acoustic signal of instrument \( \beta \) for the playback channel of musician \( M \).

![Figure 7.1. Block diagram of the room acoustic model.](image)

7.1 The Room Impulse Response

An RIR was taken from the OpenAIR Library [48]. The response corresponds to the central part of a cathedral. It is worth pointing out that depending on the size of the response selected, a certain number of RIRs will exist, thus transitions from one to another have to be done with a switch~ object. However the chosen RIR simulates the size of a response that matches with the size of the tracking volume \( V \). Thus re-scaling of the IR was not required in the present study.
CHAPTER 7. ROOM ACOUSTIC MODEL WITH PD

The convolution process was done with an FFT algorithm of 2048 samples. The Pd external `partconv` was used to implement the mathematical operation. Equation 7.1 describes the convolution of the RIR coefficients, $h_{RIR}[k]$, with the FFT of the directional signal, $x_{dir_{M,\beta}}[n]$, of instrument $\beta$ relative to the head of musician $M$. The resulting signal, $x_{RIR_{M,\beta}}[n]$, is the input for the binaural synthesis which is presented in Chapter 8.

$$x_{RIR_{M,\beta}}[n] = \frac{1}{1.5N}\text{IFFT}\{X_{dir_{M,\beta}}[k] \cdot h_{RIR}[k]\}$$ (7.1)

7.2 Direct-to-Reverberant Sound Ratio

One of the problems encountered with the FFT algorithm used is that the direct-to-reverberant ratio (DRR) was not found perceptually correct. When the listener was closer to the sound source, the early reflections of the IR were more prominent than the direct sound. Thus a compensation for the distance between sources and listeners (instruments and musicians) was developed by the author taking as a reference a Pd external `openair` [48].

The goal of this algorithm was to reproduce a more realistic impression of the balance between direct sound and reverberation, whose perceptual cues depend largely on the distance between the source and the listener $d_{M,\beta}$. Thus the distance between the head of musician $M$ and instrument $\beta$ was then computed via equation 7.2. Inverse of $d_{M,\beta}$ was calculated as well.

$$d_{M,\beta} = \sqrt{(x_M - x_\beta)^2 + (y_M - y_\beta)^2 + (z_M - z_\beta)^2}$$ (7.2)

It is worth to mention that the maximum separation from one extreme to the other of the tracking volume is about 2 m. Thus the highest $d_{M,\beta}$ and lowest $d_{M,\beta}^{-1}$ are approximately 2 m and 0.5 m$^{-1}$ respectively. A `maxlib/scale` Pd object is used for scaling both the distance $d_{M,\beta}$ and inverted distance $d_{M,\beta}^{-1}$. The ranges and values selected by the author are as follows:

- $d_{M,\beta}$: Input range of 0.3 to 2 ⇒ Output range of 0.15 to 0.4
- $d_{M,\beta}^{-1}$: Input range of 0.4 to 2 ⇒ Output range of 0.5 to 0.6

The reverberant signal and the direct signal were multiplied respectively by the values obtained after scaling $d_{M,\beta}$ and $d_{M,\beta}^{-1}$, and mixed to obtain the resulting signal $x_{RIR_{M,\beta}}[n]$ (see Appendix B.5).
Chapter 8

Binaural Sound with Pd

In the present chapter binaural synthesis with Pd is described. The sequence obtained from the RIR model, \( x_{RIR_{M,\beta}}[n] \), is transformed into Fourier and convolved with an HRTF database. The output of this stage will be referred to as \( x_{LM,\beta}[n] \) and \( x_{RM,\beta}[n] \), corresponding respectively to left- and right-channel signals for musician \( M \) relative to instrument \( \beta \). Compensation for the motion of the musicians and instruments was calculated in order to maintain the sound image independent of the position and relative orientation between them. The following sections present the binaural rendering algorithm, HRTF databases used as well as some insights about generalized and individualized responses.

8.1 Angular Compensation

A twofold angular compensation is done for azimuth and elevation angles of the frontal axis of the head of the musician \( M \) relative to the instrument \( \beta \). In this sense, a similar algorithm as presented in Section 6.1 is implemented in order to compute binaural azimuth and elevation angles \( (\theta_b, \psi_b) \).

8.1.1 Binaural Azimuth

Resolving the binaural azimuth angle of the head of a musician \( M \) relative to an instrument \( \beta \) requires the use of the coordinate pairs \((x_M, z_M)\) and \((x_\beta, z_\beta)\). We can think of the problem as a Pythagoras triangle, as depicted in Figure 8.1, where the blue arrow is the frontal axis of the head of the musician.

Thus \( \theta_b \) can be computed via the following equation:

\[
\theta_b = \arctan \left( \frac{(x_M - x_\beta)}{(z_M - z_\beta)} \right)
\]  

8.1.2 Binaural Elevation

In the same way, but looking at a Pythagoras triangle in the plane X-Y, we have calculated the distance \( d_{xz_{M,\beta}} \) in the RIR stage, and we calculate the height dif-
ference in the vertical axis $Y$. The procedure for computing $\psi_b$ is similar to the presented for $\theta_b$. The equation to calculate $\psi_b$ is as follows

$$\psi_b = \arctan 2\left(\frac{(y_M - y_{\beta}), \sqrt{(x_M - x_{\beta})^2 + (z_M - z_{\beta})^2}}{\sum} \right)$$ (8.2)

A difference regarding the interval boundaries of $\theta_b$ and $\psi_b$ is found as regards to the limit points sampled in the HRTF database. For binaural azimuth, no restriction is found since a full $360^\circ$-circumference is defined in the dataset. However, binaural elevation is restricted to the measured interval $I_e = [-45^\circ, 90^\circ]$ in [2]. This is solved with the use of a clip~ object, limiting $\psi_b$ to the interval $I_e$.

### 8.2 Individualized vs. Generalized HRTFs

Given the resources for binaural rendering with Pd, the HRTF databases available to use with cw_binaural~ [11] are the Listen [2] and the CIPIC datasets [15]. Such databases correspond to HRTF measurements of human subjects taking into consideration their anthropomorphic data.
As a matter of clarification, normally when referring to individualized HRTFs we consider the response of the individual who is currently under scrutiny. Therefore, it is not entirely correct to say that in the present study individualized HRTFs were used, because they were not measured from the musicians used in the experiments. However, they are not generalized either because none were done with a dummy head. Therefore, the author will refer to such databases as individualized HRTF measurements from human subjects.

It is worth to remind the fact that every individual will have a unique HRIR. In this way, the author was intending to explore an algorithm for anthropomorphic measurements in order to approximate the response $HRIR_\alpha$ of a given subject $\alpha$ to the closest HRIR in the database. Such algorithm is overviewed in the next section. However an HRTF dataset measured with a human subject in [15] was chosen for all the playback channels.

### 8.3 An Overview to the SVM Algorithm

Regarding the previous discussion, closest HRIR means a twofold condition:

- The differences between the features, $\Delta \alpha_i$, of a subject $\alpha$ and a subject in the individualized HRIR database are minimum
- There is a maximum number of features, $F_{\alpha_{\text{max}}}$, whose differences are minimum

In this way, computing an SVM classification algorithm of multiple features might provide the solution to such condition, thus yielding the most approximate HRIR respect to anthropomorphic data of a given individual $\alpha$. The author started thinking of such algorithm at the beginning of the research, given the uniqueness and perceivable effects of HRTFs on human subjects. A database with eleven (11) features $F_\alpha$ was coded in Pd with common anthropomorphic measurements from individualized HRTF databases [15, 2]. The features used were:

- $\alpha_0$: Sex gender
- $\alpha_1$: Head width
- $\alpha_2$: Head height
- $\alpha_3$: Head depth
- $\alpha_4$: Neck width
- $\alpha_5$: Neck height
- $\alpha_6$: Neck depth
- $\alpha_7$: Torso top width
• $\alpha_8$: Torso top depth
• $\alpha_9$: Shoulder width
• $\alpha_{10}$: Height

The intention of the author was to obtain the eleven features of the individuals used for the experiments, thus send a list as input to the algorithm and retrieve the most approximate HRIR found in the databases. However due to matter of time this algorithm was never tested at the moment of the experiments. Still, its development becomes of great interest in future work.
Chapter 9

Spatial Audio Playback

The audio playback is described in the present chapter. It will be devised as a multi-channel algorithm, represented with matrices mixing the signals $x_{LM,\beta}[n]$ and $x_{RM,\beta}[n]$ for $\beta = 1, 2, ..., M$. The author presents the criteria for selecting the headphones, as well as the definition of the playback matrix $Y$.

9.1 Headphone Selection

Considering that musicians might want to move freely, wired technologies may not be the best alternative to take. Adding more wires may put the musicians under undesired conditions of restricted motion. In this sense, the motivation of using wireless headphones seems to be well justified.

However it is not trivial to use any model of wireless headphones. As every channel has an unique spatial rendering of audio in the VAE, the headphones have to receive different $M$ signals corresponding to the $M$ musicians. Traditionally wireless headphones were used for common playback between a single transmitter and multiple receivers, which is not the case in the present study. Thus the need of using multi-channel wireless headphones was encountered.

Furthermore, the fact that the real sound from the instruments can be still perceived when using headphones, another restriction arises. The headphones have to be closed (or semi-closed) in order to significantly attenuate outside sounds and let the musician perceive mostly the rendered playback of the VAE.

In this way, two semi-closed wireless headphones Sennheiser HDR-220 and two digital transmitters TR-220 are used in the present study [8]. The transmitters are connected via RCA to the MOTU analog outputs (1, 2) and (3, 4), providing the spatial playback to musicians 1 and 2 respectively. The following section presents the construction of the generalized playback matrix $Y$ for all audio channels involved.
9.2 The Spatial Mix

Let us consider a \(2 \times M\) matrix with the contributions of the \(2M\) stereo audio channels corresponding to the \(M\) musicians respect to every instrument \(\beta\), where \(\beta = 1, 2, ..., M\)

\[
Y = \begin{bmatrix}
x_{L_1}(t) & x_{R_1}(t)
x_{L_2}(t) & x_{R_2}(t)
\vdots & \vdots
x_{L_M}(t) & x_{R_M}(t)
\end{bmatrix}
\]  

(9.1)

Thus following the notation of a stereo playback channel \((x_{L,\beta}[n], x_{R,\beta}[n])\) for musician \(M\) respect to instruments \(\beta = 1, 2, ..., M\), we define an \(M \times M\) full-playback matrix \(L_c\) for left channels

\[
L_c[M,\beta] = \begin{bmatrix}
x_{L_{1,1}}[n] & x_{L_{1,2}}[n] & \cdots & x_{L_{1,M}}[n]
x_{L_{2,1}}[n] & x_{L_{2,2}}[n] & \cdots & x_{L_{2,M}}[n]
\vdots & \vdots & \ddots & \vdots
x_{L_{M,1}}[n] & x_{L_{M,2}}[n] & \cdots & x_{L_{M,M}}[n]
\end{bmatrix}
\]  

(9.2)

and an \(M \times M\) matrix \(R_c\) for the right channels

\[
R_c[M,\beta] = \begin{bmatrix}
x_{R_{1,1}}[n] & x_{R_{1,2}}[n] & \cdots & x_{R_{1,M}}[n]
x_{R_{2,1}}[n] & x_{R_{2,2}}[n] & \cdots & x_{R_{2,M}}[n]
\vdots & \vdots & \ddots & \vdots
x_{R_{M,1}}[n] & x_{R_{M,2}}[n] & \cdots & x_{R_{M,M}}[n]
\end{bmatrix}
\]  

(9.3)

The rows of the matrices in equations 9.2 and 9.3 correspond to the contributions of all the instruments in the playback of a given musician. The columns correspond to every instrument. In this way, following equation 9.1, the spatial mix for musician \(M\) respect to all the instruments \(\beta = 1, 2, ..., M\) can be found via equation 9.4

\[
Y = \begin{bmatrix}
\sum_{\beta=1}^{M} L_c[1,\beta] & \sum_{\beta=1}^{M} R_c[1,\beta] \\
\sum_{\beta=1}^{M} L_c[2,\beta] & \sum_{\beta=1}^{M} R_c[2,\beta] \\
\vdots & \vdots \\
\sum_{\beta=1}^{M} L_c[M,\beta] & \sum_{\beta=1}^{M} R_c[M,\beta]
\end{bmatrix}
\]  

(9.4)

Finally each stereo pair of signals in \(Y\) are sent back to the audio interface (D/A converter), assigning unique stereo channels for every musician’s playback.
Chapter 10

Analysis of the Method

The author evaluated the presented method by means of latency measurements, as well as experimental tests with musicians using conventional and binaural monitoring systems. The author also carried out listening tests with professional musicians using two stereophonic recordings of a short musical composition: one with conventional monitoring systems and one with the presented method. The goal was to make a comparative study to assess the musicians’ response to both technologies, and to identify crucial features in the live music-making process and sound reinforcement technology relevant for this study.

10.1 Audio Delay and Acquisition Latency

Two processing times were measured: Pure Data’s audio delay and acquisition latency of the MOTU interface. Together they provide the delay in time from a note recorded by the microphones to the playback on the headphones.

Figure 10.1 depicts the audio settings in Pd. In this window the user can specify the sampling frequency in kHz and the audio delay in milliseconds, as well as select the I/O devices for acquisition/playback.

![Figure 10.1. Audio settings window in Pure Data.](image)

The amount of audio delay can be selected freely by the user, although too much or too little will affect the quality of the playback, particularly in situations that demand real-time implementation such as we find in the present study. Too short

55
a delay may cause some clicks in the playback, whereas too much delay will make real-time implementation unfeasible. Thus for the current study, an audio delay of 14 ms was chosen as the reasonable compromise between real-time requirements and playback quality.

On the other hand, when choosing these 14 ms for audio delay, the author used a Pd patch, depicted in Figure 10.2, to measure the acquisition latency of the MOTU audio interface. The procedure consisted of plugging the analog output 1 of the interface to the analog input 1. The acquisition delay was then found to be approximately between 4 and 6 ms. Even though the real latency may in theory be considered high (18-20 ms), it did not affect the listeners’ perception.

Figure 10.2. Measurement of acquisition latency with Pure Data.

### 10.2 Experimental Tests

The experimental tests involved four musicians, grouped in pairs, performing with two monitoring techniques referred to as systems A and B. The former system fed the signal from the RIR model as the playback, whereas the latter sent the signal obtained from the entire processing chain. The tests were mainly intended to be a pilot for future experiments, given the few contributions provided by only four non-professional musicians. The following sections describe in detail the subjects used, the tasks performed and the survey.
10.2. EXPERIMENTAL TESTS

10.2.1 The Subjects

The subjects used were four musicians (3M, 1F) with ages between 20 and 47 years-old, all of them amateur (although one of them is currently carrying studies in music), and with a musical experience ranging from 13 to 40 years. Two of them played the violin, one the oboe and one the bassoon. All of them have performed with monitoring systems before.

10.2.2 The Score

The score is a trio for oboe, violin and violoncello composed by Luis Zea. However it was suitable for playing as a duo as well. It is a short, tonal piece in polyphonic style, easy to read and play at first sight. The entire score is shown in Appendixes A.1 and A.2.

10.2.3 The Tasks

The musicians were asked to perform the musical score under two conditions for each of the monitoring systems A and B. The tasks are described as follows:

- **Task 1**: Perform the score while sitting
- **Task 2**: Perform the score while walking slowly around the sitting position in Task 1

The musicians were told to focus their attention on the sounds coming from the headphones, rather than outside sounds, given that the Sennheiser HDR-220 are semi-closed. The second task was originally intended to be walking clockwise. However due to the wires of the microphones and the tracking limitations of the motion capture system the musicians were asked to move just a little within their original sitting position.

10.2.4 The Survey

After the musicians finished the first task, they were asked to answer two questions:

- Describe briefly the differences that you notice between both systems(*)

1. To what extent each system reproduces the spatial localization of the instruments through your headphones?

   **System A**: Very little 1 2 3 4 5 6 7 Very much
   **System B**: Very little 1 2 3 4 5 6 7 Very much

   After the second task, the question marked with (*) was asked again in order to identify any training effects of the system in the musicians. The survey also included the following:
• Describe briefly the differences that you notice between both systems

2. To what extent each system reproduces your motion, that of the other musician, and the relative position between the two of you through your headphones?

   **System A:** Poor 1 2 3 4 5 6 7 Excellent
   **System B:** Poor 1 2 3 4 5 6 7 Excellent

3. Rate how much attention did you pay to the monitoring technique being used as opposed to your musical interpretation while playing

   **System A:** Monitoring technique 1 2 3 4 5 6 7 Interpretation
   **System B:** Monitoring technique 1 2 3 4 5 6 7 Interpretation

4. To what extent did each system enhance your musical experience while playing

   **System A:** Very little 1 2 3 4 5 6 7 Very much
   **System B:** Very little 1 2 3 4 5 6 7 Very much

5. If you needed to monitor your musical performance, rate how much you would prefer to use systems A and B

   **System A:** Less likely 1 2 3 4 5 6 7 Most likely
   **System B:** Less likely 1 2 3 4 5 6 7 Most likely

The original document of the survey for experimental tests can be seen in Appendices A.3, A.4 and A.5.

### 10.3 Listening Tests

While the experimental tests were carried out, two recordings (referred to as Version A and Version B) were done, corresponding respectively to performances with the monitoring Systems A and B. The playback for the violin was used for both recordings. In this way, an additional survey was designed and sent out to professional musicians. A file was made with the two recordings, published in Vimeo [58] and distributed to the participants.

#### 10.3.1 The Subjects

Twenty three (23) professional musicians (15M, 8F), with an average of 29 years of musical experience and age of 46, were asked to participate in the listening test and to answer to a questionnaire. All of them have frequently used monitoring systems when performing live music.
10.3. LISTENING TESTS

10.3.2 The Recordings

A violinist and a bassoonist were being recording while performing both versions of the musical score. The recordings were done during a repetition of the second task, thus the musicians were meant to move around. However, in this case only the violinist was moving while the bassoonist remained seated. The spatial playback sent to the violinist was the recorded channel. The audio files, referred to as Version A and Version B, were put together in a file [58] whose link was forwarded to the professional musicians.

10.3.3 The Survey

The survey for the listening tests was similar to the survey for experimental tests, however in this case the questions were geared towards an assessment of the initial reaction of professional musicians to the monitoring technology presented in this study. After listening the audio files in [58], the musicians were asked to complete the following survey:

- Briefly describe the differences between both versions of the musical segment that most strike your ears

7. Rate to what extent the monitoring system in each version reproduces (through your headphones) the location of the instruments in space
   - **Version 1**: Very little 1 2 3 4 5 6 7 Very much
   - **Version 2**: Very little 1 2 3 4 5 6 7 Very much

8. Rate to what extent the monitoring system in each version reproduces (through your headphones) the movement of the instruments in space
   - **Version 1**: Very little 1 2 3 4 5 6 7 Very much
   - **Version 2**: Very little 1 2 3 4 5 6 7 Very much

9. Rate to what extent did the monitoring system in each version enhance your musical experience while listening?
   - **Version 1**: Very little 1 2 3 4 5 6 7 Very much
   - **Version 2**: Very little 1 2 3 4 5 6 7 Very much

10. If you needed to use amplification for your musical performance, rate which monitor system you would probably prefer?
    - **Version 1**: Less likely 1 2 3 4 5 6 7 Most likely
    - **Version 2**: Less likely 1 2 3 4 5 6 7 Most likely

The original document of the survey for the listening tests is found in Appendixes A.6 and A.6.
Chapter 11

Discussion of Results

The results of the experimental and listening tests are shown in the present chapter. An average for the perceived quality obtained in each question of the surveys is shown, thus pointing out the most relevant findings yielded with the evaluation method.

11.1 Experimental Tests

As previously mentioned, the experimental tests were mainly done as a pilot to identify and eliminate noticeable errors or disadvantages of the method. Indeed, the author was interested in validating the integration of the parts of the system, something which had not been attempted with musical instruments before. Statistical data shown in Figure 11.1 was calculated with the musicians’ feedback.

As we may see from Figure 11.1, none of the questions except number 5, provides a significant result that could suffice at least as an indicator. This question was about which monitoring system would the musicians prefer to use. We can clearly see that the musicians would rather use System A. The musicians referred to System B as having terrible clicks and glitches while performing the second task (moving while playing). One of them asserted that System B was very bad in every aspect and some others found it very hard to perform with. System A was generally described as providing a more realistic playback than System B. One performer felt that the playback was a failed attempt to put their sounds inside of a church. However, the author attributes the comments regarding glitches to the size of the tracking volume. When the rigid bodies were not tracked, the 3-D feedback jumped from the musician’s position to the origin of the tracking volume. Hence, the sound presented these glitches whenever the rigid body was outside the tracking volume.

The survey presented questions in between the two tasks in order to explore any training effect of the system on the individuals. It is interesting to point out that, in fact, the perceived quality of System B increased in about 14% from Question 1 to Question 2; and that of System A decreased in about 10%. In the author’s opinion, this result indicates that musicians are not aware of tridimensional sound.
One musician confused spatial rendering with panning of the signal. The author considers this to be one of the biggest challenges of using 3-D audio: people are not used to it.

It is reasonable to assume that musicians would not fully understand what the playback of System B was all about unless further information was provided to them. One a hint is given (such as related to the spatial localization of the instruments after the first task), it is likely that musicians would have a better chance to appreciate the usefulness or applicability of the system, and be able to focus on it more carefully while doing the tests.

Unfortunately, due to lack of time the pool of participants for experimental tests was not as big as the author originally intended. Thus, listening tests with a bigger number of musicians were carried out to complement the experiments, and this led towards curious results and interesting ideas.

### 11.2 Listening Tests

One of the primary differences between the experimental and listening tests was, evidently, related to the number of participants. Also, the fact that the musicians could listen to the audio files as many times as they wanted ruled out the idea of considering the influence of training effects on the subjects. As previously mentioned, the listening test was carried out by means of recording and distributing a
11.2. LISTENING TESTS

file in Vimeo [58] and a fairly short survey. The participants were all professional musicians, which provided a higher degree of reliability on the results obtained. Figure 11.2 shows the statistical results of perceived quality yielded in questions 7, 8, 9 and 10 of the survey.

![Figure 11.2. Results for perceived quality obtained in the questions of the survey for listening tests.](image)

At first sight, we might say that the survey was not correctly designed. The author believes that providing more information about 3-D audio could have prevented the confusion in the participants when answering the survey. The higher standard deviation in most of the questions suggests to the author that unfortunately the musicians probably misunderstood the experiment. And indeed, some of the additional comments in the surveys seem to confirm this conjecture. Some musicians commented that Version B was annoying as a monitoring technique. Others commented that Version B somehow produced an interesting effect. Others mistakenly referred to both versions as a mono and stereo duality, and confused panning with spatial rendering. Such comments strengthen the author’s point that musicians do not know about 3-D sound. Therefore, it is difficult to answer questions concerning the potential of this new technology unless we have a minimum of knowledge about it.

An interesting point discussed with Prof. Ternström, the project examiner, was the usefulness of this monitoring technology for singing choirs. Two of the musicians that participated in the test are singers, and they suggested indicators that point towards the use of spatial rendering of voices under monitored conditions.
CHAPTER 11. DISCUSSION OF RESULTS

One of them mentioned that musical creativity could be stimulated with the use of spatialized sound. On the other hand, Prof. Ternström suggested that the present monitoring technique could be explored in live performances of choirs, when one singer is out of tune; hence he/she could identify his/her sound and that of the others.

In addition, we can recognize that movements were perceived about 70% more in Version B compared to Version A when looking to statistics of Question 8. This is a good feedback for the author because it means that the method may have achieved one of its goals: real-time spatial rendering. Nevertheless, these results by themselves do not take us very far and further tests are necessary. In future work, some ideas will be discussed regarding the listening tests and the collected data.
Chapter 12

Future Work

Given the lack of time, as well as some inconveniences with software and hardware during the development of the present study, the author would like to provide additional comments and suggest further research lines that might be of interest in the field.

A test with the directivity of instruments along with the entire processing chain of the presented method was not implemented in the pilot experiments. The main reason was significant amount of delay due to higher computational needs in the signal processing. Therefore, code optimization of the directional radiation functions and/or performing the experiments with more powerful personal computers is highly encouraged (e.g. quadratic interpolation).

The next step in research is the computation of a numeric threshold that will define the use of either static or dynamic sound mixing. The algorithm would track the energy of the musicians according to their movements and establish the need of compensating the spatial presentation, which means, the requirement of computing continuous interpolation and adaptive convolution with the DTFs and the HRTFs. The latter implies that a single (static) DTF and HRTF will be used for the convolutions. Hence, if a musician is not moving too much, the energy is below a certain threshold and an SSM scheme is implemented for his/her particular feedback, thus reducing computational cost on the code. Whilst in the case of energy above the threshold, interpolation and filtering with dynamic DTFs and HRTFs are performed, leading to a DSM scheme for that audio feedback channel.

Regarding the data collected in the listening tests, it could be interesting to identify the possible correlations between the extent of the musicians’ experience and the answers provided in the survey. It could also be of great interest to find out what professional musicians think about tridimensional sound, and in case they do not know about it, to show them the potential of this technology and assess their answers. This important element was missing in the present study and it should have given a more complete evaluation of the method.

In the same way more experimental tests with musicians and a larger tracking volume are highly encouraged for further research in the field. The author sug-
CHAPTER 12. FUTURE WORK

gest the inclusion of more infrared cameras and the use of accelerometers and/or gyroscopes.

Regarding the individualized HRTFs, the author already started developing an algorithm with SVM classification in order to find the closest response within a database of subjects with anthropomorphic features. In this way, sound localization errors can be minimized and aural perception may be enhanced. Therefore further development of this algorithm is well justified for improving binaural synthesis.

In order to present a simplified model and focus on the design of a standardized method, the case of sound diffraction has been excluded in the present study. Such is the case, expected in real-life situations, when a musician is listening to a certain instrument within a direct wavefront, while another musician shows up in between. Careful future studies of such cases may enhance the realism of the acoustical events. However this may not be a compulsory consideration for the framework to effectively work.

Once a binaural monitoring system may become standardized, we can think about transaural audio with crosstalk cancellation techniques. In such a case, the room acoustic model differs from the one presented in the report, due to the fact that the wedge-floor monitors have a directivity response, and complex reflection patterns are likely to occur with the surfaces of the room where the performance is taking place.
A computational model for a binaural monitoring system has been presented in this report. The framework was taken from a previous research carried out by Savioja and co-workers [44]. Hence four considerations were taken into account: motion capture of the musician’s head and instrument, directivity responses of the acoustic instruments, room acoustics, and binaural audio. The motivation for the study were two aspects identified by the author as missing in traditional monitoring systems: directional radiation, and spatialization of the sounds on stage.

The research was based on the hypothesis that the inclusion of these features into the conventional monitoring schemes will enhance the music-making process, from the musician’s as well as the listener’s point of view, hence the artistic value of the musical experience may be enriched. The limitations of the study were identified, pointing out the acoustical conditions of the room and the needs of the musicians. Thus when the acoustics of the room are satisfactory, monitoring (amplification) of the sounds on stage is not normally required, and musicians have a tendency to see such amplification as undesirable. However, when room acoustics are poor, amplification may become necessary and the monitoring system presented in this thesis may enhance the musical experience.

A detailed description of the computational model with a UDP network between Pure Data (Pd) and the motion capture software (ARENA) through Open-Sound Control messages was presented (see Section 3.2.3 and Figure 3.2). The most relevant algorithms and the resources used in the study were devised, including Pd patches as well as references to directivity measurements [38], room acoustics [48], and binaural audio [15, 2, 11]. Also, an evaluation of the monitoring system was done via three methods: (i) measurements of delay in the processing chain, (ii) a pilot experiment with four amateur musicians performing a duo for violin and oboe in a small room, and (iii) a listening test with twenty-three professional musicians evaluating two audio recordings.

The latency of the system was measured with Pure Data, and was found in the order of 20 ms (see Section 10.1). Although it was not considered perceptually disadvantageous by the musicians, more improvements can be done in order to
reduce the latency, such as more powerful personal computers or audio interfaces, and code optimization.

The two monitoring systems evaluated were:

- System A: The output of the RIR model in the processing chain
- System B: The output of the binaural monitoring system

Two recordings were made with the playback channel of the violinist player, during the second task in the pilot experiments (refer to Section 10.3):

- Version A: recording using System A as a monitoring technique
- Version B: recording using System B as a monitoring technique

As regards the pilot experiments, although a very small pool of participants was used, the survey provided interesting results. Two tasks were performed with four musicians (none of them professional), grouped in pairs, using two monitoring systems. The tracking volume was one of the main issues in the method, given the sound glitches the musicians identified during the pilot tests (refer to Section 11.1). An important result was found regarding training effects from Question 1 to Question 2 in the survey (refer to Appendix A.4); which addressed the perception of spatialization, and of motion of the sounds, respectively. Two tasks were asked to the participants: (i) perform while being seated, and (ii) perform while slowly moving around the sitting position in task (i). Once the first task was performed, they were asked to describe the differences between each system, followed by rating the perception of spatial localization of the sounds. After performing the second task, the differences between A and B were asked again, followed by rating the perception of motion of the sounds, in which a training effect of increasing 14% for System A and decreasing 10% for System B were identified (see Section 11.1). The author considered that presenting the survey after each of the two tasks worked as a hint which made the performers focus their attention more on the spatial rendering of the instruments, thus justifying the training effect found. This is one of the most important conclusions of the research and is related to the musicians’ degree of awareness of tridimensional sound. The author believes that if the musicians had been familiar with 3-D audio beforehand, the results and the evaluation of the method may had been more successful.

The listening test was carried out with 23 professional musicians. As mentioned above, two recordings were done during the second task of the pilot experiments with the violinist’s playback channel (Version A and Version B, respectively). A file with the two versions was uploaded in Vimeo [58], and the website link was sent out with a survey to the professional musicians in order to evaluate the listening experience with the two monitoring systems. Training effects were disregarded in this test. The results and comments of the survey of the listening tests suggest that Questions 7 and 8 were confusing for some of the musicians: i.e. spatial rendering
with panning (see Section 11.2). Questions 7 and 8 addressed the perception of spatial localization and of motion of the sounds, respectively (refer to Appendix A.6). Thus the first conclusion is that the design of the survey was not entirely successful. On the other hand, high standard deviation was found in most of the answers and none of the musicians identified the use of binaural audio in Version B. Thus, the musicians’ lack of familiarity with 3-D audio was found once again to be a problem for them in order to properly evaluate the potential of binaural monitoring. In addition, three of the musicians pointed out possible applications of the System B in singing choirs, referring to ideas such as enhancement of musical expression and creativity on stage. Furthermore, the perceived motion of the instruments (Question 8 of the survey in Appendix A.6) was identified about 70% clearer in Version B rather than in Version A, pointing out the fact that binaural technology does render sound spatialization successfully.

Further development of the current work was presented, leading to code optimization of directivity functions (e.g. bilinear interpolation), as well as computation of correlation and statistics with the survey for the listening tests. Larger pool of participants for experimental tests and improvements of the tracking volume of the motion capture system are encouraged. The author left some thoughts for future research on individualized HRTF responses, enhancements on the perception of acoustic events such as sound diffraction within the body of the musicians, and transaural playback with floor-wedge monitors. The author stated the next step on the research, pointing out the importance of a numerical threshold of energy according to the movements of the musician, in order to choose whether his/her feedback will be compensated or not (DSM vs. SSM), hence optimize the computational cost of the algorithm (refer to Chapter 12 for more details).

As pointed out before, the author identified the tracking volume of the motion capture technique not sufficient. And that is why nowadays it might be very difficult to implement such monitoring technique in real-life scenarios. Still it may be useful to remember the discussion of static vs. dynamic sound mixing addressed in Section 2.6. In cases where the musicians are moving very much (e.g. rock/pop concerts), the spatialization with binaural audio might be dynamic; otherwise the sense of spatial rendering is lost. However, in cases where musicians are mostly sitting or standing on the same spot (e.g. academic music and jazz), motion capture techniques are not as necessary as in the dynamic case; thus spatial rendering can be performed statically. Therefore, for the time being, we might ask ourselves to what extent such monitoring technique with motion capture becomes useful given the difficulties found in the present study. Nowadays with the exponentially rising technology, in the next years such problems as latency and tracking techniques will be overcome. The author still considers the initial hypothesis of enhancing the artistic value of musical performances worth being supported, and encourages further research in the area.
Bibliography


BIBLIOGRAPHY


Appendix A

Musical Score and Surveys

Figure A.1. Musical score used in the experimental tests (1/2).
Figure A.2. Musical score used in the experimental tests (2/2).
QUESTIONNAIRE FOR EXPERIMENTAL TESTS

INITIAL QUESTIONS: Please answer the following questions before you proceed to the next section

• Sex: M ___  F ___
• Age: _______
• Instrument(s): _________________________________________
• Are you a professional musician: YES ___ NO ___
• Years of musical experience: _______
• Have you performed live music with other musicians using a monitoring system (that is microphones and amplification)? YES ___ NO ___

BEFORE THE TEST

The test involves a musical performance (of a short piece) and has two parts: (i) performing while sitting down and (ii) performing while walking slowly in clockwise direction around the marked area on the floor. For each part of the test you will experience TWO different monitoring techniques through your headphones. The first one will be referred to as system A and the second as system B. Please focus your attention on the sounds coming from your headphones and ignore any other sounds.

Figure A.3. Survey for experimental tests (1/3).
AFTER THE 1st PART:

- Describe briefly the differences that you notice between both systems

1) To what extent each system reproduces the SPATIAL LOCALIZATION of the instruments through your headphones?

   **System A:** Poor 1 2 3 4 5 6 7 **Excellent**
   **System B:** Poor 1 2 3 4 5 6 7 **Excellent**

AFTER THE 2nd PART:

- Describe briefly the differences that you notice between both systems

2) To what extent each system reproduces your MOTION, that of the other musician, and the relative POSITION between the two of you through your headphones?

   **System A:** Poor 1 2 3 4 5 6 7 **Excellent**
   **System B:** Poor 1 2 3 4 5 6 7 **Excellent**

3) Rate how much attention did you pay to your musical interpretation while playing

   **System A:** Very little 1 2 3 4 5 6 7 **Very much**
   **System B:** Very little 1 2 3 4 5 6 7 **Very much**

4) To what extent did each system enhance your musical experience while playing

   **System A:** Very little 1 2 3 4 5 6 7 **Very much**
   **System B:** Very little 1 2 3 4 5 6 7 **Very much**

Figure A.4. Survey for experimental tests (2/3).
5) If you needed to monitor your musical performance, rate how much you would prefer to use systems A and B

<table>
<thead>
<tr>
<th>System</th>
<th>Less likely</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>Most likely</th>
</tr>
</thead>
<tbody>
<tr>
<td>System A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

VI) Please add any additional comments you find relevant regarding the test.

Additional comments:
____________________________________________________________________________
____________________________________________________________________________
____________________________________________________________________________
____________________________________________________________________________

THANK YOU!

Figure A.5. Survey for experimental tests (3/3).
QUESTIONNAIRE FOR LISTENING TESTS

PRELIMINARY INFORMATION: (in multiple choice questions please highlight your answer by changing the text font to **red and bold**)

1) Sex: M  F

2) Age: ________

3) Specify your main area of work:
   Instrument(s): ____________________ Other_______________________

4) Years of musical experience: ________

5) Have you frequently used a monitoring system (microphones and amplification) when performing live music (either on your own or with other musicians)?
   YES - NEVER - FEW TIMES

AFTER LISTENING TO THE AUDIO in [https://vimeo.com/43886864](https://vimeo.com/43886864)

6) Briefly describe the differences between both versions of the musical segment that most strike your ears:
   __________________________________________________________________________
   __________________________________________________________________________
   __________________________________________________________________________

7) Rate to what extent the monitoring system in each version reproduces (through your headphones) the LOCATION of the instruments in space:

   Version 1: Very little 1 2 3 4 5 6 7 Very much
   Version 2: Very little 1 2 3 4 5 6 7 Very much

8) Rate to what extent the monitoring system in each version reproduces (through your headphones) the MOVEMENT of the instruments in space:

   Version 1: Very little 1 2 3 4 5 6 7 Very much
   Version 2: Very little 1 2 3 4 5 6 7 Very much

9) Rate to what extent did the monitoring system in each version enhance your musical experience while listening?

   Version 1: Very little 1 2 3 4 5 6 7 Very much
   Version 2: Very little 1 2 3 4 5 6 7 Very much

**Figure A.6.** Survey for listening tests (1/2).
10) If you needed to use amplification for your musical performance, rate which monitor system you would probably prefer?

Version 1:  Least likely 1 2 3 4 5 6 7 Most likely
Version 2:  Least likely 1 2 3 4 5 6 7 Most likely

11) Additional comments:

________________________________________________________________________

________________________________________________________________________

THANK YOU!

---

**Figure A.7.** Survey for listening tests (2/2).
Appendix B

Pure Data patches

Figure B.1. Network patch in Pure Data showing the UDP connection. PC1 corresponds to "Studio 2" and PC2 corresponds to "Studio 1"
Figure B.2. MoCap-to-IP patch in Pure Data showing the code for the OSC thread used in the network.
Figure B.3. Rigid-body patch in Pure Data showing the OSC messages routed into the six degrees of freedom received from the network.
Figure B.4. Directivity-Instruments patch in Pure Data showing the code blocks for the radiation database, the adaptive gain filter, FFT analysis and trilinear interpolation.
Figure B.5. RIR-model patch in Pure Data showing the Room Impulse Response, and the code blocks of the external `partconv~` and the direct-to-reverberant sound ratio.
Figure B.6. Binaural-model patch in Pure Data showing the code blocks for angular compensation, the external `cw_binaural~` and the HRTF dataset used (CIPIC, subject 003).
Figure B.7. Interface for experimental tests with Pd. Directivity of instruments is excluded.
Figure B.8. Interface for binaural monitoring with Pd. Directivity of instruments is included.
Appendix C

Additional Pictures

Figure C.1. Omnidirectional microphone Behringer ECM 8000.
Figure C.2. Pickup microphone Brüel and Kjær 4021 mounted in the bridge of the violin.
Figure C.3. MOTU Traveler mk3 multi-channel audio interface.
APPENDIX C. ADDITIONAL PICTURES

Figure C.4. A Natural Point OptiTrack infrared camera of the ARENA motion capture system.

Figure C.5. The rigid body constructed and mounted in one of the headphones Sennheiser HDR-220.
Figure C.6. The rigid body constructed and mounted in the violin.

Figure C.7. The rigid body constructed and mounted in the bassoon.
Figure C.8. Perspective of the room where the experiments were carried out (1/2).
Figure C.9. Perspective of the room where the experiments were carried out (2/2).