A system for room acoustic simulation for one's own voice

Manuj Yadav

A thesis submitted in fulfillment of the requirements for the degree of

Master of Philosophy

Faculty of Architecture, Design and Planning

The University of Sydney

November 2011

Preface

I certify that this thesis has not already been submitted for any degree and is not being submitted as part of candidature for any other degree.

I also certify that the thesis has been written by me and that any help that I have received in preparing this thesis, and all sources used, have been acknowledged in this thesis.

I also certify that approval from The University of Sydney's Human Research Ethics Committee (HREC) was obtained prior to the experiments involving human participants that form part of this thesis, and the experiments were conducted in strict adherence to the HREC guidelines. The following are the details of the HREC approval for the current thesis.

Title: Characterizing Auditory Room Size Perception using Oral Binaural Room Impulse Responses

Protocol No: 13210

First Approval Date: 5/10/2010

Manuj Yadav

Abstract

The real-time simulation of room acoustical environments for one's own voice, using generic software, has been difficult until very recently due to the computational load involved: requiring real-time convolution of a person's voice with a potentially large number of long room impulse responses. This thesis is presenting a room acoustical simulation system with a software-based solution to perform real-time convolutions with headtracking; to simulate the effect of room acoustical environments on the sound of one's own voice, using binaural technology.

In order to gather data to implement headtracking in the system, human head-movements are characterized while reading a text aloud. The rooms that are simulated with the system are actual rooms that are characterized by measuring the room impulse response from the mouth to ears of the same head (oral binaural room impulse response, OBRIR). By repeating this process at 2º increments in the yaw angle on the horizontal plane, the rooms are binaurally scanned around a given position to obtain a collection of OBRIRs, which is then used by the software-based convolution system.

In the rooms that are simulated with the system, a person equipped with a near-mouth microphone and near-ear loudspeakers can speak or sing, and hear their voice as it would sound in the measured rooms, while physically being in an anechoic room. By continually updating the person's head orientation using headtracking, the corresponding OBRIR is chosen for convolution with their voice. The system described in this thesis achieves the low latency that is required to simulate nearby reflections, and it can perform convolution with long room impulse responses. The perceptual validity of the

system is studied with two experiments, involving human participants reading aloud a set-text.

The system presented in this thesis can be used to design experiments that study the various aspects of the auditory perception of the sound of one's own voice in room environments. The system can also be adapted to incorporate a module that enables listening to the sound of one's own voice in commercial applications such as architectural acoustic room simulation software, teleconferencing systems, virtual reality and gaming applications, etc.

Acknowledgements

I wish to thank my supervisor, Dr. Densil Cabrera and my associate supervisor, Assoc. Prof. William L. Martens, for their insights, guidance and support throughout the process entailed in this thesis.

I also wish to thank everyone in the acoustics postgraduate research program and Ken Stewart for their technical assistance and maintaining an entertaining atmosphere in the acoustics laboratory.

Finally, thanks to my family for their long-distance moral support.

TABLE OF CONTENTS

Chapter 1

Introduction

Self-created sounds, whether conscious or otherwise, are an important part of daily activities for humans. These sounds serve many purposes such as self-identification, communication, etc., in a number of scenarios. One such scenario is speaking in a room environment and listening to the sound of one's own voice. Autophonic perception (Lane et al., 1961) in rooms, i.e., perception of the sound of one's own voice, can follow three pathways (Figure 1.1) (Pörschmann, 2000; Lehnert & Giron, 1995; v. Békésy, 1949):

a) Directly conducted from the mouth to the ears of the same head (direct-airborne conduction.

b) Conducted through the internal structures of the human head to the cochlea (bone or body conduction).

c) Room reflected sound, which includes reflections from relevant surfaces in the room's environment (indirect-airborne conduction).

The autophonic perception can vary from room to room, mostly determined by the characteristics of the reflective surfaces as mentioned in the third pathway above (c), while the transmission through the other pathways (a, b) is held constant. In order to systematically study this variation and the effect it has on a talker listening to his/her autophonic output, this talking-listener would have to be tested in many rooms that are different in one (or some) of the room's auditory characteristics, such as the arrival time of the earliest reflection, the absorptive nature of the surfaces, reverberation time, etc.

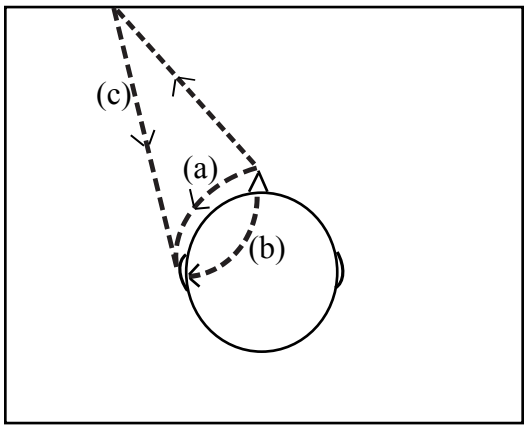


Figure 1.1: The three pathways of autophonic perception in rooms. Here (a) is the direct-airborne conducted sound, (b) is the bone (or body) -conducted sound and (c) is the indirect-airborne conducted sound.

When moving to different rooms, not only are the auditory characteristics of the rooms changing, but also the visual characteristics, which can have an undesirable effect on a study designed to test only auditory characteristics. A way around this drawback is to sufficiently eliminate the impact of visual stimuli through appropriate means, by blindfolding the talking-listeners for instance. But such solutions are generally limited in the improvements they provide by being logistically cumbersome, and they might be completely prohibitive if normal usage of human vision is required.

Another solution is to acoustically recreate room environments (real, or computer generated) while providing no visual information about the rooms. This can be accomplished through acoustically measuring the room environments, using impulse response techniques. A correctly measured impulse response has been shown to contain all the acoustical information about a room, with respect to the position in the room where the measurement was made (Kutruff, 2009). In room acoustical literature, there are many studies that use the impulse response method to render virtual environments (see

Blauert (1993) and Novo (2005) for a review) of different rooms, while the listener to the sounds in these virtual rooms is physically present in anechoic conditions (devoid of any reverberance). The sounds that are used as the stimuli are most commonly measured in anechoic conditions, and then convolved with the impulse response of a room. The resulting stimulus is played back to a listener through a chosen reproduction format (headphones, loudspeakers, etc.) (Novo, 2005). In colloquial terms, this process essentially leads to 'colouring' of a dry (anechoic) sound sample with the acoustical characteristics of a room.

However, most of the studies in the literature above have been limited to presenting convolved sounds that are external, or exocentric, to the listener and only a few studies have addressed the more commonly encountered egocentric sounds: in the form of one's own voice or other self- produced sounds such as the sound of clapping, keyboard clicks, etc. (Pörschmann & Pellegrini, 2010). One of the major reasons for this trend in the literature of room acoustics has been the computational complexity of performing convolution of long impulse responses that correspond to room environments, with real-time voice of a talking-listener (Torger & Farina, 2001). Though hardware solutions for performing real-time convolution exist (such as the Huron platform or other artificial reverberation generators), such solutions are generally expensive and can only perform convolution with respect to one head-position.

Due to the change in the orientation of the head, which changes its distance from the reflective surfaces of a room environment, the auditory scene that is perceived by the two ears generally changes in its characteristics. Such changes in the interaural characteristics are important for tasks that involve deriving auditory information from

these changes, for example the cues for distance from surfaces, size of a room environment, etc. (Blauert, 1997). When the auditory scene does not change with the change in head-position of talking-listener, it can affect the realism within a simulation for tasks that involve normal or exploratory head-movements that are accompanied with human talking. This limits the applicability of a hardware-based or indeed software-based solution that does not incorporate head-movements, within a real-time simulation of a room environment for egocentric sound.

To address the issues outlined above, especially in the last two sentences, this thesis is presenting a software-based solution for the acoustical simulation of one form of egocentric sound, i.e., the sound of one's own voice, with a modified method of impulse response measurements that also incorporates human head-movements. In other words, the thesis is presenting a room acoustical simulation system in which a person (physically present in anechoic conditions) can speak or sing, and listen to his/her voice as it would sound in a particular room environment corresponding to the third pathway (c) in Fig 1. In the system, the other two pathways (a, b) are not simulated as they are naturally present with a person speaking or singing while listening to his/her own voice.

The software-based solution presented in this thesis uses generic software that is relatively inexpensive compared to the hardware-based solutions. As the software components can be more easily customized (compared to hardware components) to incorporate changes accompanied with head-movements for a greater degree of realism, the current system can be applied in a wider range of auditory tasks involving listening to the sound of one's own voice.

There are three major sections in this thesis that are divided into 7 chapters. The first two chapters (including the current) introduce the theoretical issues that relate to hearing the sound of one's own voice in rooms. Chapter 2 refines the introductory discussion in this chapter regarding the sound of one's own voice in rooms, and outlines the three steps involved in the design of the room acoustical simulation system presented in this thesis, which has software-based solution for real-time convolution. Chapter 2 also elaborates on the first step in the implementation of the system, which involves the description of an impulse response measurement method from the mouth to the two ears of the same head.

The next section, which forms the core of the thesis, addresses the next two steps in the implementation of the system, as mentioned in Chapter 2. The concept the human head-movements that is introduced in Chapter 2, is illustrated with a case study in Chapter 3. The results from the study present quantitative data that can be used to design a system that incorporates human head-movements through headtracking.

Chapter 4 builds on the discussion in the preceding chapters to describe the integration of hardware and software components included in the design of a real-time room acoustical simulation system, which is the main deliverable of this thesis. The system is analyzed in terms of its input, output and real-time performance, i.e., the time-lag between the instant a talker produces a vocal transduction and the moment he/she listens to the sound of his/her voice in a simulated room. The system is designed to have an electroacoustic time-lag (latency) of essentially 0 ms.

Chapter 5 & 6 present two case studies, one per chapter, which were designed for perceptual evaluation of the system described in Chapter 4. The first study, which

addresses the headtracking incorporated in the system, involved participants performing a headtracking detection test. One of the motivating factors behind this study was to compare the detectability of headtracking in the present egocentric task, to past research that involved primarily exocentric tasks. Here, the participants use their own voice to detect whether the room reflected sound (room acoustic simulation of their voice in the system) changes in accordance with their head-movements, by testing with and without headtracking. The results show that headtracking is shown to improve the simulation's effectiveness in creating impressions of reverberant rooms – in which head rotation is associated with a significant change in the interaural cross correlation function.

The second study provides an example of how the system can be used to design experiments involving the performance of an everyday task, but with only egocentric auditory stimuli. In this study, human participants rated the aurally perceived size of different rooms that were simulated in real-time within the system. The motivation here came from previous studies involving mainly exocentric auditory (sometimes combined with visual) stimuli, where factors other than the room's actual volume were shown to be more important in judging the auditory room size. It is shown here that human participants were able to judge the sizes of a variety of room based on just egocentric auditory stimuli, when no other information about the rooms was provided. The level of the room reflected sound that reaches the talking-listeners, quantified here as room gain, is shown to be the most significant cue in judging the room's auditory size.

The last section (Chapter 7) comprises of a summary of the thesis, which is concluded with an outline of future research that can be undertaken with the system described.

Appendix A is a companion to Chapter 4, and it describes the software-based implementation of the real-time acoustical simulations system in detail. Appendix B describes some of the physical characteristics of the room conditions that were simulated through the system presented in this thesis. Appendix C comprises of a discussion of a selection of existing real-time room acoustical simulation systems, to elaborate the scope and limitations of the system presented in this thesis. Appendix D lists the publications arising from this thesis.

Chapter 2

Listening to one's own voice in rooms

As mentioned in Chapter 1, there are three pathways from the point of view of autophonic perception in rooms: (a) direct-airborne conduction; (b) bone conduction; and (c) indirect-airborne conduction. In this chapter, each of these pathways will be addressed in a separate section (sections 2.1, 2.2 and 2.3), with a description of how they are incorporated in the system described in this thesis. Section 2.4 describes the steps in the implementation of the system.

2.1 Direct-airborne conducted sound

According to the study conducted by Dunn and Farnsworth (1939), the sound field around a human speaker (measured at 76 positions and 13 frequency bands in their study) is almost entirely caused by the sound radiated from the mouth. They found that the radiation from nearby body surfaces such as throat, chest and the back of the head had a relatively insignificant effect on the measurement, when compared to the mouth radiation. In this thesis, the direct-airborne conducted sound is simply referred to as the direct sound, where it signifies the directly radiated sound from the mouth to the two ears of the same head. The direct sound can change from person to person, depending on many factors such as the shape of the head, the presence of hair, etc., and the system presented in this thesis takes into account these considerations, which will be covered in section 4.2.1 of Chapter 4.

As the direct sound is already present with a person listening to the sound of his/her own voice, this pathway is not being simulated in the system described in thesis. However, in the simulation, the sound radiating from the mouth of a talking-listener is the signal that is used for convolution to simulate the indirect-airborne conducted sound, which will be described in section 2.3.

2.2 Bone-conducted sound

Even though it is generally referred to as just the bone-conducted sound, this pathway includes the sound conducted to inner ear from the tissue and muscles of the human head. As an example of the contribution of bone-conducted sound, it is instructive to listen to the recording of one's own voice. Here, as the bone-conducted pathway is missing, the sound of one's own voice could be strikingly different from what one normally associates with.

However, an in-depth description of this pathway of one's own voice is quite extensive (Tonndorf, 1962) and beyond the scope of this thesis. As with the last pathway, the bone-conducted sound of one's own voice is not simulated, because it is already present with a person listening to the sound of one's own voice.

2.3 Indirect-airborne sound

The indirect-airborne sound, referred to as the room reflected sound in thesis, is the third pathway in the auditory perception of the sound of one's own voice in rooms. Unlike the last two pathways, the sound corresponding to this pathway is simulated in the system described in this thesis. The contribution of this pathway to the overall autophonic

perception can be appreciated by anyone who has had the experience of speaking in a variety of room conditions that differ in acoustical or physical (including atmospheric) qualities. This contribution can be almost none (in anechoic conditions) or overwhelming (in highly reverberant conditions) or anything in between.

A system that simulates this pathway has to characterize how the room behaves on being excited by a sound transduction. Although there are various ways of implementing such a system, our system is based on a binaural head-tracked approach, which has the advantage of short acoustic latency. This is based on the assumption that the signal input microphone is near the mouth and the output loudspeakers are near the ears, allowing small rooms and close reflections to be simulated: a condition met in the current system.

2.4 Steps involved in implementing the system

Figure 2.1 shows the summary of the implementation of the system, which goes through the following steps.

1. Measuring room impulse responses from the mouth (oral) to two ears (binaural) of a Head and Torso Simulator (HATS) in a fixed position, referred to as oral binaural room impulse response (OBRIR).
2. Feeding the appropriately treated set of OBRIRs of a room into the software based real-time convolution system that uses head-tracking to determine which OBRIR to choose to convolve with the user's voice.

3. Convolving the voice of the person using the simulation with the appropriate OBRIRs (selected via head-tracking) and returning the result to their ears in real-time.

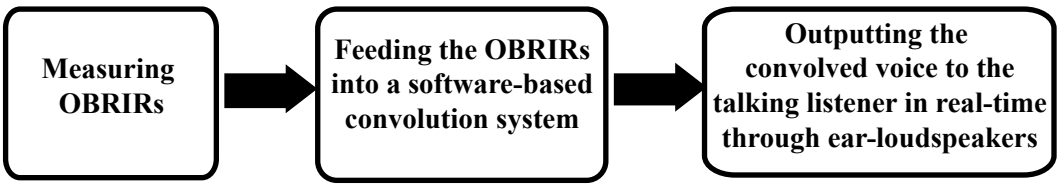


Figure 2.1: Steps in the implementation of the system

In this chapter only the first step will be described, while the remaining two steps are only introduced and are the subject of subsequent chapters. The limitations of the current implementation, in terms of the measurement method, are also addressed.

2.4.1 First step

Impulse response measurements, introduced in Chapter 1, normally involve a sound source and receiver that are physically separate from each other. In this thesis, a slightly modified method of impulse response measurement is used, which was first described by Cabrera, Sato, Martens and Lee (2009). Here the impulse response is measured from the mouth (sound source; oral) to the two ears (sound receivers; binaural) of the same head, with respect to a fixed position and orientation of the head. Such oral-binaural room impulse responses (OBRIRs) can be measured for a range of positions and orientation (for example, by rotating the measurement instrument) in a process referred to as binaural room scanning (BRS). Incorporating BRS in a room simulation leads to a

more realistic simulation, as the simulation tracks the head-position and orientation of a talking-listener, and accordingly changes the auditory scene by changing the OBRIR used for convolution (Lehnert and Blauert, 1992).

The position and orientation of a head can be measured in the six degrees of freedom, as seen in Figure 2.2.

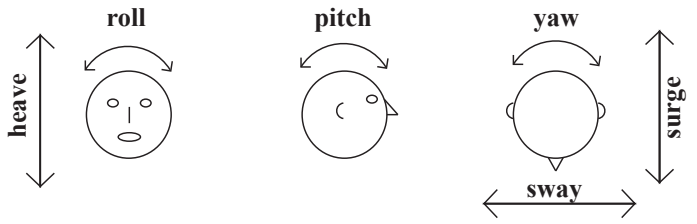


Figure 2.2: The three angular variations (roll, pitch and yaw angles) of three linear translational (sway, surge and heave) degrees of freedom of human head movements.

In the present thesis, head movements in only the horizontal plane, which corresponds to yaw angles, are tracked and incorporated in the system; the reasons for which will be discussed in Chapters 3, 4 and 5. The method for measuring OBRIRs for these angles is dealt in detail by Cabrera *et al.* (2009) and can be seen as two distinct processes of collection and processing. Briefly, the collection stage involves using a HATS. A swept sinusoid (or any other appropriate measurement signal) is emitted from the HATS mouth simulator, and is recorded by three microphones – one at the mouth reference point, and one in each ear (at the entrance of the ear canals) of the same HATS (B&K 4128C) in an arrangement as displayed in Figure 2.3. By using a long sweep duration (15 s), this method is also suitable for exciting large rooms, such as the one seen

in Figure 2.4 (b) (7650 m^3), where sometimes the sound level from the HATS mouth loudspeaker could be insufficient for sufficient signal-to-noise ratio.

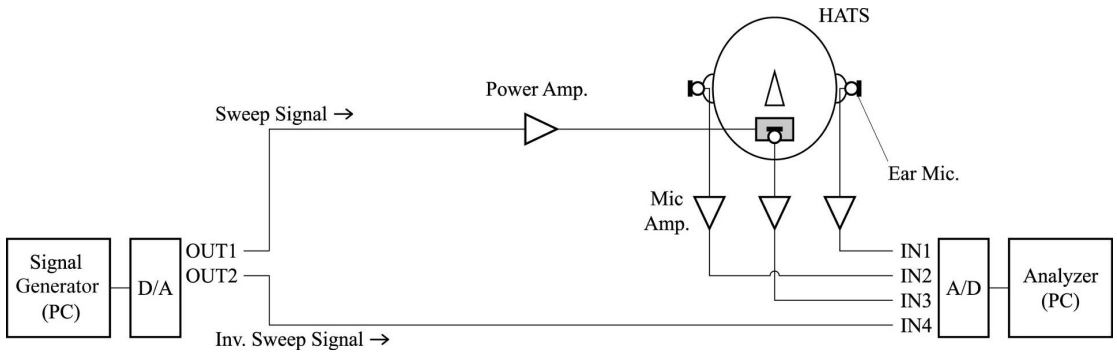


Figure 2.3: Block diagram for OBRIR measurement with a HATS

The impulse responses are derived from the transfer functions between the mouth reference microphone and the ear microphones. This method is repeated for each position that contributes to binaural room scanning (Cabrera *et al*. (2009) covered 121 yaw angles (Figure 2.2) with a range of -60 to 60 degrees and 2º resolution, using the HATS mounted on a turntable as in Figure 2.4 (a)). The result is a collection of OBRIRs for the left and right ears, indexed by the angle of horizontal rotation of the HATS.

The processing stage, that uses a MATLAB routine, involves filtering in the frequency domain to band-limit the OBRIRs between 100 Hz and 10 kHz. This filters out the high- and low- frequency noise acquired during the measurement process (because it is outside the frequency limits of the mouth simulator). Following this first stage of filtering, further improvement of the reverberant tail is done by fading out any noise floor such that it acts as an extrapolation of the measured reverberant tail. This extrapolation process is done in octave bands centered on 125 Hz – 8 kHz (except the lowest and highest bands, which are implemented as low and high pass filters respectively). Zero phase filtering is used to maintain synchrony between the frequency bands. The routine

estimates for each band the point at which noise overwhelms the impulse response and applies the smoothing to match the decay rate of the reverberation time. The processed bands are recombined, yielding impulse responses with no apparent noise floor. This process, first implemented for individual binaural room impulse responses by Lee, Cabrera, and Martens (2009) has been updated to derive multiple OBRIRs to be used for room simulation.

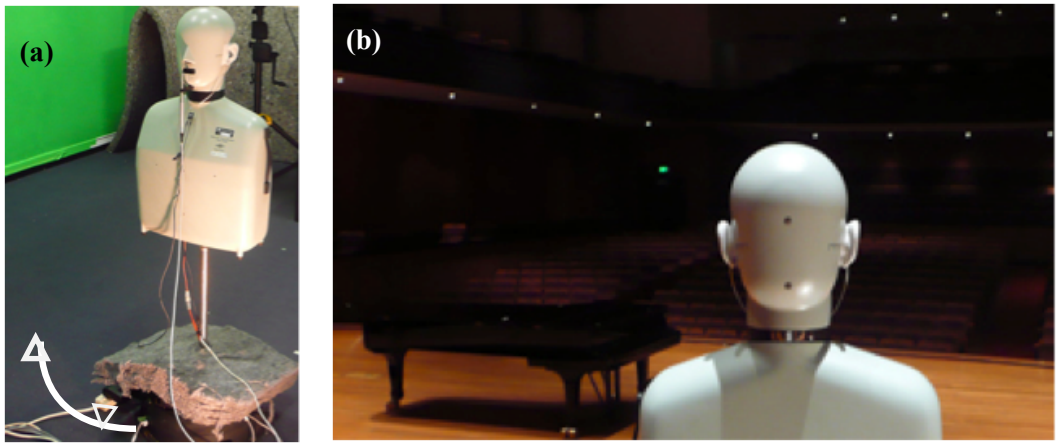


Figure 2.4: (a) The figure shows the HATS mounted on a turntable that can rotate in 2° (can be variable) increments to binaurally scan a room. Also seen are the mouth and ear microphones. (b) The view from behind the HATS in a large room.

As there is no need to simulate the direct sound in the system and the measurement process described above can result in OBRIRs with a longer duration than what can be expected from the reverberation time of the room, the OBRIRs are truncated at two points, (1) from the beginning of the OBRIR to number of samples corresponding to the current system's throughput latency and (2) from the point where the samples are all 0's in the OBRIR. Floor reflections are also not being simulated as they are obtained

by carpeting the floor around a user (on which the chair they sit on is placed) in an anechoic room.

2.4.2 Second and third step

The second and third steps, as seen in Figure 2.1, are described in chapter 4.

2.4.3 Limitations of the measurement method

Measuring the OBRIRs using a HATS has many advantages, such as the time invariance of the signal used, the ease in the repeatability of the process, use of generic head related transfer function, generic long term voice directivity of the HATS etc., but also has a major limitation, at least in the current implementation of the measurement method (Cabrera *et al.*, 2009). As rotating the HATS for BRS measurements rotates the whole HATS, i.e., both the head and torso at the same time, when the rooms are simulated using the measured OBRIRs, the talking-listener is also required to move both head and torso together when executing head-movements in the horizontal plane. This limitation could be addressed in future refinements of the measurement method, where only the head of HATS from the shoulders up would be rotated in any given degree of freedom. This refinement currently poses technical problems, as it would require motorizing the HATS' neck to move the head in desired increments.

Also, as will be discussed in more detail in Chapter 3, currently the BRS is only implemented in the horizontal plane with 2° yaw increments. For an even finer resolution, the BRS could incorporate interpolation between the yaw angles in the future.

Chapter 3

Head movements while reading aloud horizontally arranged English texts

This chapter is presented in the form of a case study, which is motivated by the characteristics of head movements within room acoustical simulations, in particular the simulation system to be described in the next chapter. As the room acoustical simulation system presented in this thesis is described in detail in the next chapter, the following is only a brief overview of how room reflections are simulated and how head movements are an important part of such a design, to facilitate a discussion related to such simulations later. Incorporating head movements in a binaural room simulation system requires various head positions in a room to be individually measured in a process referred to as binaural room scanning, in the desired degrees of freedom with a certain spatial resolution (Mackensen, Fruhmann, Thanner, & Theile, 2000).

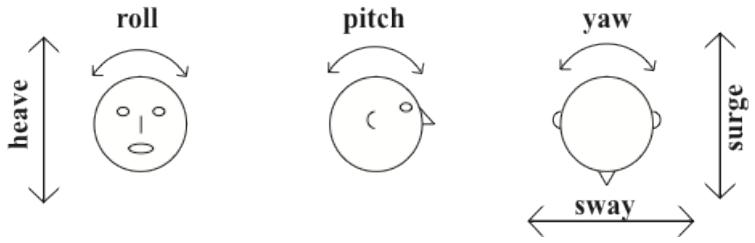


Figure 3.1: (same as figure 2.2) 6 degrees of freedom; 3 angular variations (yaw, pitch, roll) and 3 linear translations (sway, surge, heave)

Out of the 6 degrees of freedom shown in Figure 3.1 for example, a room could be simulated with 2° resolution for a yaw range of 81° (-40° to +40°). In the simulation, when a talking-listener projects his/her voice from a certain position within this range, his/her position is tracked in real-time and his/her voice is convolved with the room's binaural impulse response (BIR), with respect to that position of the room and played back to the talking-listener with a chosen reproduction format in real-time. However, the convolution of real-time voice with the directional BIR of a certain position is a computationally intensive task (Torger & Farina, 2001) with only a handful of BIR and other real-time convolution systems that are currently implementing this with an acceptable acoustical accuracy (Cabrera, Sato, Martens & Lee, 2009; Pörschmann & Pellegrini, 2010; Wefers and Vorländer, 2010; Ueno & Tachibana, 2010). To incorporate many such positions and directions required to create a realistic simulation thus becomes a non-trivial task.

One way to limit the number of such positions is to quantitatively determine the characteristics of head movements of people when they are reading a text aloud. This chapter is presenting the results of such an investigation. Though such information is directed towards optimizing the number of head positions simulated in a real-time system, which can potentially reduce the overall computational load, the results could also be used in other cases where natural head movements while reading a text aloud are considered.

3.1 Reading silently vs. reading aloud

Reading silently has been considered to be a complex process with a module for perceptual gathering of information and a module for cognitive processing of that information (McConkie, Reddix, & Zola, 1992). According to recent research however, the information-gathering module shows coordination between eye and head movements over time (Lee, 1999; Proudlock, Shekhar, & Gottlob, 2003). This indicates a differential level of cognitive control influencing information-gathering also, with a heterogeneous interplay between perceptual and cognitive processes overall (Lee, 1999). These studies have mainly examined the coupling of head and eye movements and their contributions to coordinate gaze shifts while reading horizontal or vertically arranged texts silently to one's own self under experimental conditions. Although these findings have practical applications in vision research and ophthalmology, head movements while reading a text aloud has not been that well studied, that has applications in the present thesis.

In this chapter, reading aloud is meant to indicate the act of enunciating the text at a level one is most comfortable with. No other parameters are imposed, which implies that the participants are free to hold the text at a distance and angle that they find most comfortable while reading, unlike previous studies where these parameters are fixed for all the participants (Hutchings, Irving, Jung, Dowling, & Wells, 2007; Lee, 1999; Proudlock et al., 2003). The rationale for such a choice is that the room acoustical simulation system, which is to be described in the next chapter, is used in studies that have participants reading a set-text (common in studies that require all the participants to be tested under the same experimental condition) in their natural reading state (case studies in Chapters 5 and 6).

It is known from room acoustical research that unrestricted head movements while speaking aloud, with or without moving the lower parts of the body in a sitting position, causes human voice to be projected to different parts of the room (potentially differing in physical aspects such as distance from the talking-listener and surface properties of the walls, hangings etc.), and a corresponding change in the arriving room reflections to the ears plays a major role in determining human behaviour in such environments (McGrath, Thomas, & Fernstrom, 1999).

3.2 Experimental design

3.2.1 Choice of text and its presentation

An important issue associated with reading aloud studies is choosing a text that suits the purpose and design of the study. To avoid any issues regarding the choice of the text being read, in terms of the familiarity with the words and ease of diction, a simple approach was adopted. This involved different groups of participants reading two different texts, having a marked variation in degree of difficulty encountered while reading aloud. One text was an excerpt from James Joyce's *Ulysses* (classified as a *difficult* text) and the other an excerpt from Mark Twain's autobiography (classified as an *easy* text). Both these texts can be accessed freely from websites such as Project Gutenberg, without violating any copyrights.

These two texts were printed out on 10 one-sided A4 sheets each with double line spacing (Font -Times new roman; size- 12) with no footnotes. No further manipulation of the text was done. The horizontal and vertical margins were 0.032 m and 0.027 m respectively. The mean distance from the middle of the text to the mid-point of the line

joining the eyes (*d* in Figure 3.2) for the 10 participants in reading II was 0.44 m. The mean angle the text was held at from the horizontal plane orthogonal to the participant was 37º (β in Figure 2). Scanning the horizontal (α in Figure 2) and vertical extremes of the text subtended a mean angle of 135º and 126º respectively at the mid-point between the eyes.

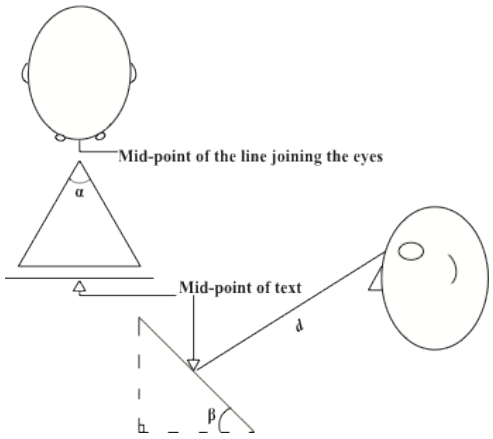


Figure 3.2: α is the angle that the horizontal extremes of the text subtend at mid-point of the line joining the eyes. The angle subtended by the vertical extremes can be calculated similarly. β and *d* are the angle and distance respectively, that the text is held at.

3.2.2 Participants

Ten participants (50% male and female) took part in each of the two reading task. 4 participants were common to both groups, giving a total of 16 unique participants. All participants were either postgraduate students or administrative staff members at the Faculty of Architecture, Design and Planning at the University of Sydney. Their ages ranged from 20 to 60 years and 11 of the 16 participants were native English speakers. The participants were explained the reading aloud task and their consent was received

before the task. One participant had read *Ulysses* previously but did not realize it till the source of the text was revealed after the task was completed. So unfamiliarity with both the texts is assumed for all the participants.

3.2.3 Apparatus

Apart from the texts discussed above, a Polhemus Fastrak® unit was used as the head-tracking device. A Max/MSP (Cycling '74, 2011) patch (Appendix B) interacted with the Fastrak® through the serial bus to retrieve and store the data on a Windows machine at a sampling frequency of 5Hz, for offline statistical analysis using MATLAB. The Fastrak® has an electromagnetic field transmitter that was placed on a wooden stand behind the participants at a distance of 0.3m. A receiver that transmits the head-tracking information was placed on a pair of ear-loudspeakers that the participants wore as in Figure 3B. There was no sound being played back through the ear-loudspeakers. As this set-up is being used in a real-time room acoustic simulation designed by the authors that incorporates headtracking, the same set-up was used in this study to maintain consistency.

3.2.4 Reading aloud tasks

For both tasks, the participants were seated on a wooden chair in the Recording studio (Figure 3.3 A) at the Faculty of Architecture, Design and Planning, the University of Sydney. The participants were asked to read the text with natural movements of the head and limiting any torso movement. The experimenter was in the adjacent control room, unable to hear what the participants were reading. This was done to provide the participants a sense of privacy though the author could visually inspect the participants at

all times, but not *vice-versa*. The participants were allowed to hold the text at a distance most comfortable to them. This is in contrast to other studies that have examined head movements, where the distance and orientation of the text was fixed.

Once the participant's head position and orientation were calibrated to be the reference point in all degrees of freedom, their head movements were recorded in 6 degrees of freedom (Figure 3.1), while they read as much of the text they could in 10 minutes at their normal reading speed. The duration of the reading allowed both short term and long-term phenomena to be studied.

Figure 3.3: The set-up for the readings showing the text being held at a comfortable level (A) and the head-tracking receiver on the ear loudspeakers and the transmitter (B) behind a seated user (not a participant).

3.3 Data Analysis

For the long-term characteristics, the first point of analysis involved deriving the statistical summary of the data by concatenating the results for all participants per degree of freedom. Additionally, for the linear translations, the values in metres were converted to degrees to compare these values with the angular degrees of freedom. This was done

by approximating the radius of human head as 0.09 m (p) (Blauert, 1997) and deriving the angle (γ) subtended at the centre of the head by the linear distance (m) covered, as follows

$$\gamma = tan^{-1} (p / m)$$

Though there are better approximations, the value in angles derived in this manner provides a simple way to compare the linear and angular degrees of freedom. A paired sample *t*-test was conducted on the range, RMS velocity and displacement of movement in all degrees of freedom, to determine if the head movements changed from the *difficult* to the *easy* reading task. This was only done for the four participants who read both the texts. For a visual comparison of the average measure of head movements over time and space, the root mean squared (RMS) velocities were plotted alongside the RMS displacements for each participant.

The short-term characteristics were studied by plotting the Hilbert transform of each participant's head movement data in each degree of freedom, against time. The Hilbert transform has been employed in head movement studies previously (Seo & Lee, 2002), as it is a useful measure of the instantaneous attributes of a time series such as the instantaneous head movements in the present case.

As the participants were free to hold the text at the position they were most comfortable with, there was variability in the distance d (Figure 3.2) and angle β (Figure 3.2) across all participants. To understand whether there was any interaction between these two parameters and the RMS displacements and velocities calculated above, their correlation coefficients were compared for significance.

3.4 Results

3.4.1 Range of movements

Overall for the two different reading tasks, the range of angular variations was largest for the yaw angles and least for the roll angles. The values of sway and surge, when converted to degree from metre are larger than all the angular degrees of freedom, though the conversion is only a first-order approximation. More accurate conversion is likely to shrink the converted values. The values are as shown in Table 3.1.

The *difficult* and *easy* reading tasks did not make a significant difference in terms of the range of head movements executed by the 4 participants who read both the texts. A paired-sample *t*-test to compare the means of the scores of participants reading texts I and II did not reject the null hypothesis ($p > 0.05$ in all the cases). The *r* (effect size) is also less than 0.5 (a common criterion for concluding a large effect) for all the degrees of freedom. The null hypothesis in this case is that a change in the text being read aloud (*difficult* vs. *easy* text) has no effect on the participants. The statistics for the *t*-test are elaborated in Table 3.2.

Degree of freedom	Standard Deviation		Range		Interquartile Range	
	I	II	I	II	I	II
yaw	5.96	6.03	54	51	8	8
pitch	3.54	5.63	40	55	4	7
roll	4.07	3.85	46	31	4	4
x	0.02(12.53)	0.03(18.43)	0.31(73.81)	0.19(64.65)	0.02(12.53)	0.05(29.05)
y	0.02(12.53)	0.04(23.96)	0.33(74.74)	0.38(76.67)	0.03(23.96)	0.02(12.53)
z	0	0.01(6.34)	0.03(18.43)	0.09(45)	0	0.02(12.53)

Table 3.1: The standard deviation, range and interquartile range of the data for all participants for all the degrees of freedom. Reading I had *Ulysses* as the text and II had Mark Twain's *Autobiography.* Yaw, pitch & roll are expressed in degrees while sway, surge & heave are expressed in metres, with degree values in brackets derived from the approximation described in the data analysis section.

3.4.2 RMS velocity and displacement

When the results for readings I and II are combined- for the angular degrees of freedom, RMS velocity and displacement for yaw with peak values of 14.06º/s and 9.54º respectively was more than the ones for pitch and roll. For the linear degrees of freedom, RMS velocity and displacement for surge with peak values of 0.15m/s and 0.1 m respectively was more than the ones for sway and heave, as seen from Table 3.3. For the 4 participants who read aloud both the texts, a paired-sample *t*-test to compare the means of the RMS velocity in readings I and II did not reject the null hypothesis ($p > 0.05$ in all the cases). The *r* (effect size) is also less than 0.5 (a common criterion for concluding a large effect) for all the degrees of freedom. The null hypothesis in this case is that a change in the text being read aloud (*difficult* vs. *easy* text) has no effect on RMS velocity of the head movement of participants. The statistics for the *t*-test are elaborated in Table 3.2. A similar *t*-test for RMS displacement (not shown) also did not reject the null hypothesis.

Kendall's correlation coefficient (τ) was derived to analyze the interaction of the distance and angle that the text was held at, with the RMS velocity and displacement for each degree of freedom. The distance the text was held at had a significant relationship with RMS pitch displacement, RMS pitch velocity and RMS roll velocity with $\tau = .53$, .45, and .45 respectively, p (one tailed) < 0.05 in all cases.

Degree of freedom	Mean, SE				
	I	II	*t*	*p* (2-tailed)	*r*
yaw	-2.82, 2.76	-2.21, 3.02	-0.16	0.88	0.45
pitch	0.09, 1.98	-0.07, 2.63	0.07	0.95	0.48
roll	-0.31, 0.79	-0.97, 2.38	0.31	0.77	0.49
x	0.03, 0.02	0.42, 0.62	-0.66	0.56	0.31
y	0.02, 0.02	-0.33, 0.41	0.85	0.45	0.25
z	0.01, 0.01	0.33, 0.27	-1.18	0.32	0.18

Table 3.2: Comparing means from readings I (*Ulysses*, *difficult* text) and II (Mark Twain's *Autobiography*, *easy* text). The paired-sample mean and standard error (SE) of the differences between participants' scores (who read both the texts) for readings I and II are tabulated in the 2nd and 3rd column respectively followed by the *t*-statistic (I – II) for 3 degrees of freedom (statistical), 2-tailed significance (*p*) with a confidence interval of 95% and effect size *r*.

Degree of freedom	RMS velocity		RMS displacement	
	Max	Mean, SD	Max	Mean, SD
yaw	14.06	4.96, 1.45	9.54	0.88
pitch	8.65	3.84, 1.23	0.07	0.95
roll	5.72	3.55, 0.62	0.31	0.77
x	0.05	0.03, 0.01	0.07	0.03, 0.01
y	0.15	0.05, 0.02	0.1	0.03, 0.02
z	0.05	0.02, 0.01	0.04	0.02, 0

Table 3.3: The maximum, mean and standard deviation of RMS velocity and displacement values for all the degrees of freedom. The units for the RMS velocity and displacement columns for the first three rows are º/s and º respectively. The same for the last three rows are m/s and m respectively.

Degree of freedom	Mean, SE				
	I	II	*t* (Wand, et al.)	*p* (2-tailed)	*r*
yaw	3.67, 0.62	4.79, 1.3	-1.34	0.27	0.15
pitch	2.37, 0.33	3.95, 0.86	-2.71	0.07	0.04
roll	2.87, 0.15	3.15, 0.31	-0.77	0.49	0.27
x	0.03, 0	0.03, 0	-0.03	0.98	0.49
y	0.04, 0.01	0.03, 0	1.8	0.17	0.09
z	0.03, 0.01	0.02, 0	0.36	0.74	0.39

Table 3.4. Comparing means of RMS velocity from readings I (*Ulysses*, *difficult* text) and II (Mark Twain's *Autobiography*, *easy* text). The paired-sample mean and standard error (SE) of the differences between participants' scores (who read both the texts) from readings I and II are tabulated in the 2nd and 3rd column respectively followed by the *t*-statistic (I – II) for 3 degrees of freedom (statistical), 2-tailed significance (*p*) with a confidence interval of 95% and effect size *r*.

3.5 Visual Comparison

3.5.1 RMS velocity and displacement

The values of RMS velocity and displacement for both the readings are charted in Figure 3.4 to visually inspect the average variation amongst the participants One obvious feature for most of the participants is that a large value of RMS velocity corresponds to a large RMS displacement and *vice-versa*. More specifically, participants 1, 5, 8 and 10 are the ones that aloud read both the texts and their results in reading II can be seen to be a scaled version of experiment I. Most participants confirm to the finding above that the peak RMS values are seen for yaw and surge, though some participants show individual peak values for other degrees of freedom. Participants 1, 3 and 7 for example, show their peak RMS velocity for roll (upper-left quadrant of Figure 4A) and participants 3, 4, 5 and 7 show their peak RMS displacement for roll (upper-right quadrant of Figure 4A).

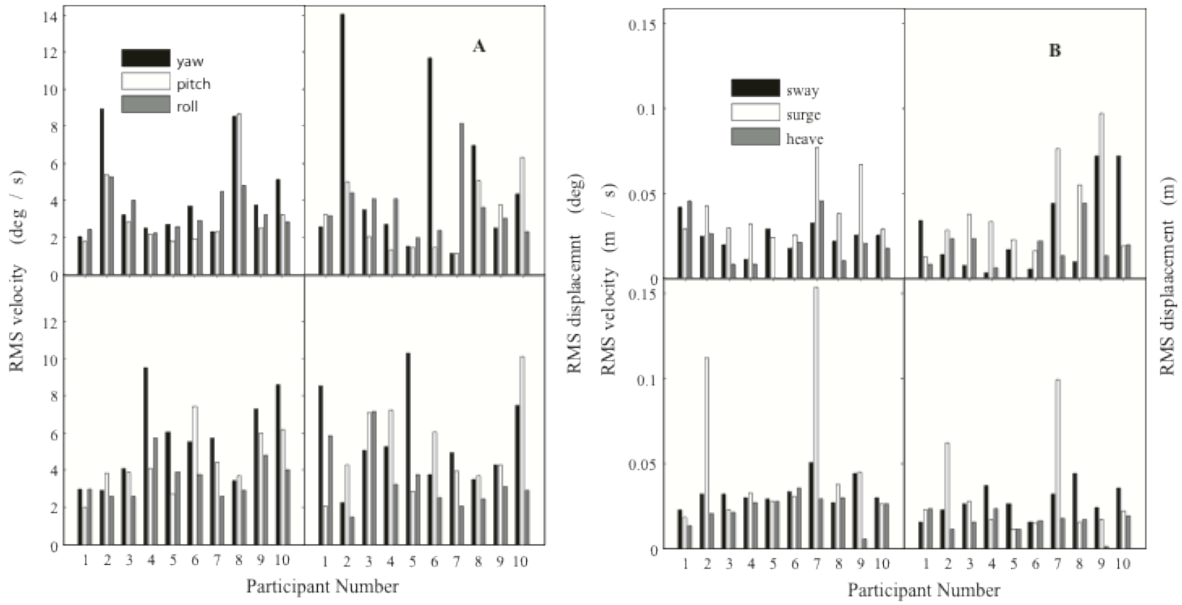


Figure 3.4: RMS velocities and displacements for the angular (A) and linear (B) degrees of freedom. The upper quadrant in both A and B show the results from reading I and the lower quadrant shows the same for reading II. For A, left quadrant shows RMS velocities in °/second and right quadrant shows RMS displacement in °. For B, left quadrant shows RMS velocities in metre/second and right quadrant shows RMS displacement in metre.

3.5.2 Instantaneous attributes

Another aspect of head movements can be inspected from Figure 3.4, where the instantaneous amplitudes for all degrees of freedom of 3 participants (1, 6, 9 - randomly chosen) from reading II are plotted over the experimental duration of 10 minutes. For all the degrees of freedom except yaw, a staircase-like pattern, as seen in other studies (Lee, 1999; Proudlock et al., 2003), can also be seen here in almost all the subfigures, where the staircase can be described as a gradual rise and then fall (or *vice-versa*) in amplitude with a jump in between. For pitch, the jump happens when the page is read completely in

the vertical direction, indicating the minimum amplitude, and the next page starts from the top. A similar line of reasoning applies for all degrees of freedom. As it is a horizontal reading task, staircase pattern can also be noticed for yaw if the resolution of the graph is lowered (not shown). The staircase in this case indicates reading a line from left-right, then shifting to the new line and repeating the process. The data presented in the Figure 3.5 looks noisy for most part for yaw as the participants are reading many lines per page, in effect confounding the jump which is seen more clearly in other degrees of freedom (except perhaps for sway, which can be explained similarly as it is on the same plane as yaw).

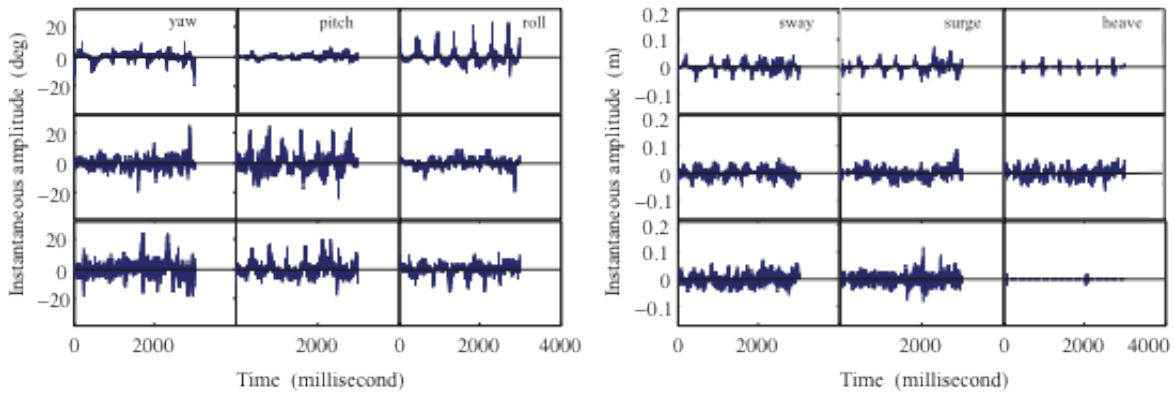


Figure 3.5: The Hilbert transform showing the instantaneous amplitude in ° (for yaw, pitch, roll) and metre (for sway, surge, heave) for 3 participants (1, 6, 9) from reading II.

3.6 Discussion

As the main aim of the case study presented in this chapter is characterizing the head movements for room acoustic simulations and similar areas, a conservative approach is adopted while comparing the results to studies that have

analyzed coordinated eye-head movements, largely due to difference in scope. So instead of focusing on comparison with the finer details of other studies' results, the focus here is on deriving a meaningful analysis of head movements while reading a text aloud.

3.6.1 Influence of text on head movements

As seen from the results of Tables 3.2 and 3.4, the choice of text did not influence the head movements of participants who read both the texts aloud. Although only two texts were compared, it can be reasoned that since the participants are scanning the text from one (horizontal and vertical) extreme to another, they are not likely to go beyond the range that corresponds to these extremes (there can be exceptions here due to individual scanning styles), where out of all the parameters, the age of a participant has been previously shown to have no influence on the head movements (Proudlock et al., 2003) while other parameters are yet to be fully researched. Familiarity with the text was not directly measured for any effects, but is unlikely to cause an effect based on the reasoning above. It was expected however, that the head movement velocity might differ from reading a *difficult* to an *easy* text, though the statistical analysis of RMS velocity and displacement shows that this was not the case for the texts used. So while the choice of text could be task-constrained, these results indicate that researchers can exercise a wide range of freedom while choosing the difficulty of a horizontally arranged English text printed on A4 sheets.

3.6.2 Optimizing head movement range within simulations

For a task that is similar to the reading aloud task described here, the results of the head movement range from Table 3.1 can be used to optimize the performance of a

simulation while incorporating head-tracking. Also, the relationship between the distance *d* (Figure 3.2) and RMS displacement indicates that displacement along a degree of freedom can be reduced by reducing this distance, perhaps to the extent where it is the closest without being restrictive in any other sense. This would lead to a reduction in the range that is simulated, and a standard distance value could be derived from studying this relationship specifically.

3.6.3 Inclusion of roll within room acoustic simulations

The discussion in this section addresses one of the limitations of the system to be described in the next chapter, which is common with most room acoustic simulations which headtrack yaw (Moldrzyk, Ahnert, Feistel, Lentz, & Weinzierl, 2004; Welti and Zhang, 2010), and while some have incorporated pitch (Lindau, Hohn, & Weinzierl, 2007), not many incorporate roll. Firstly, it has to be acknowledged that incorporating any angular degree of freedom other than yaw poses technical problems; such as the dummy head that is used to measure the binaural impulse responses can generally only be moved along the yaw plane without extensive manipulation (Lindau et al., 2007). Then there is the issue of optimizing computational performance while maintaining accurate sound localization in the yaw plane. But as the computational performance related to simulation tasks is increasing, a case can be made from the results of the current study towards the inclusion of other degrees of freedom, especially roll (either measured or approximated), in room acoustic simulations that have users speaking or singing while being seated. An inspection of the results of Table 3.1 shows that there is substantial movement for roll. Also Figure 3.4A show some participants showing more roll movement than yaw or pitch. It can be argued that these are task-dependent here. But the

fact that there is substantial roll movement warrants a closer inspection of the issue of its inclusion in room simulations, given that it has a major role in upper-lower hemisphere (along the median plane) sound localization task (Blauert, 1997). The second point, arising from the first, indicates that a roll-centred sound localization study is likely to get results in favour of the overall contribution of roll movement. Moreover, since the aim of a simulation is to be as accurate as possible, these two points cannot be ignored from the point of view of optimizing the degree of 'presence' (Witmer & Singer, 1998) within simulations.

3.6.4 Translations along linear degrees of freedom

Even though the text was read aloud while being seated, there was not only linear swaying and surging but also (surprisingly) heaving, although the last was minor in comparison. The main reason behind this finding can be understood in the context of the headtrack measurements, where the linear translations include the corresponding angular variations on the same plane. For example, in the horizontal plane, a change in yaw corresponds to a change, albeit smaller, in sway and *vice-versa*. The same reasoning applies to the other planes.

The influence of these linear translations to the sound within a room acoustic simulation and whether they can be approximated is yet to be directly studied.

3.6.5 Evaluation of presence

The texts in the in the current study were read aloud in a *real* room and the results can be applied to incorporate headtracking in simulating the same *real* room. Apart from this direct application, it would be interesting to investigate whether the presence in the

simulated version (with the head movement information from the present research) of the *real* room can be used to evaluate and refine the current results qualitatively. In this thesis, Chapter 5 and 6 address some of the criteria of presence in a simulation.

3.7 Conclusions

The current study investigated the head movements in angular (yaw, pitch, roll) and linear (sway, surge, heave) degrees of freedom while reading aloud two (*difficult* and *easy*) horizontally arranged English texts. The results show maximum range and peak amplitude for yaw (angular) and surge (linear) though there are substantial movements in other, especially angular, degrees of freedom. The choice of text did not influence the head movements of participants but the distance the text was held at influenced RMS pitch displacement, and RMS pitch and roll velocity. The results of the research can be most directly applied to optimize the headtracking range of room acoustic simulations (the system in the next chapter, for example) for reading aloud tasks, though the results can also be useful in tasks where a text is read aloud (within the constraints of the current study) to one's own self such as classrooms, meetings etc. In future, a thorough examination of the recorded voice parameters such as voice projection etc. that are accompanied with head movement (with or without eye movement) while reading a text aloud might lead to findings that demonstrate coordination between the cognitive processes underlying voice tasks and head movement, similar to eye and head coordination. These findings could have applications in areas such as speech pathology, professional speaking and singing, gaming, virtual worlds etc.

Chapter 4

A room acoustical simulation system with real-time convolution of one's own voice

As was discussed in chapter 2, the real-time simulation of room acoustical environments for one's own voice using generic software has been difficult until very recently due to the computational load involved (Torger & Farina, 2001; Schröder *et al.*, 2010). The main focus of this chapter is to describe a software-based solution for real-time convolution with headtracking to simulate the effect of room acoustical environments on the sound of one's own voice, using binaural technology. In doing so, this chapter consolidates the theoretical framework for real-time room acoustical simulation for one's own voice described in chapter 2, and the characteristics of human head movements that are illustrated in chapter 3.

In the simulation, a person equipped with a near-mouth microphone and near-ear loudspeakers can speak or sing, and hear their voice as it would sound in the recorded rooms, while physically being in an anechoic room. By continually updating the person's head orientation using headtracking, the corresponding OBRIR is chosen for convolution with their voice. The system described in this chapter achieves the low latency that is required to simulate nearby reflections, and it can perform convolution with long room impulse responses. For a binaural system sensitive to head position, this number of room impulse responses depends on how many head positions are incorporated within the

simulation and their duration is guided by the reverberation time of the room and perhaps the background noise level of the simulation environment (Mershon & Bowers, 1979).

The system can be divided into its software and hardware modules, which are discussed in sections 4.1 and 4.2, respectively. Each module is elaborated with respect to its components. The software module includes an overview of the software components used for implementing real-time convolution and headtracking, which are explained in-depth in Appendix A. Similarly, the hardware module describes the various hardware components that are part of the system in sections 4.2.1, 4.2.2 and 4.2.3. The characteristics of the complete system as an arrangement of these two components is the subject of section 4.3, where the issue the system latency is addressed in section 4.3.1, gain calibration is discussed in section 4.3.2, effect of loop gain and cross-talk is discussed in section 4.3.3, and the categorization of the system as a mixed-reality environment for a talking-listener is elaborated in section 4.3.4. This is followed by a summary of the whole chapter.

4.1 Software module

The software component of the system is implemented by hosting the commercially available VST plugin SIR2 (Appendix A) for real-time convolution (Knufinke, 2010) in Max/MSP (Cycling'74). The signal processing necessary for implementing headtracking is done in Max/MSP.

4.1.1 Introduction to Max/MSP

Max/MSP is a graphical music/signal processing environment, where instead of constructing a program or application in a text format (like C++, Java etc.), an application

is designed by using Max/MSP "objects". In Max/MSP, each object is a program in itself, whose implementation in terms of the coding is generally hidden from the user. Each object is referred to by a name and a user can include many objects in a Max/MSP application that can be connected to other objects by cords, much like the connection cords in an analog synthesizer, constrained by the number of inputs and outputs allowed for the respective objects.

As an analogy, an object could be understood as an alphabet and a user can construct a Max/MSP application that contains words and then sentences made from the alphabet. Though Max/MSP comes with a vast collection of objects, there is also a provision for users to program their own objects (generally in programming languages like C, Java, etc.). Information could also be transmitted through messages to objects, where a message to an object is determined by the functionality of the object. Figure 4.1 provides an example of a simple Max/MSP patch.

4.1.2 Real-time convolution implemented in Max/MSP, with headtracking

As mentioned above, a detailed description of the real-time convolver (SIR2) and how it is integrated in the Max/MSP patch is the subject of Appendix A. This section describes the implementation of the software module of the complete system in Max/MSP.

Based on the findings of chapter 3, headtracking has been enabled for a yaw range of -40º to 40º with a 2º resolution in the system, which could account for incidental head movements encountered during reading an English text aloud and perhaps even normal conversational speech. This degree range can be expanded to include more angles or

other head positions (also discussed in chapter 3), which would require measuring the OBRIRs for the extended range.

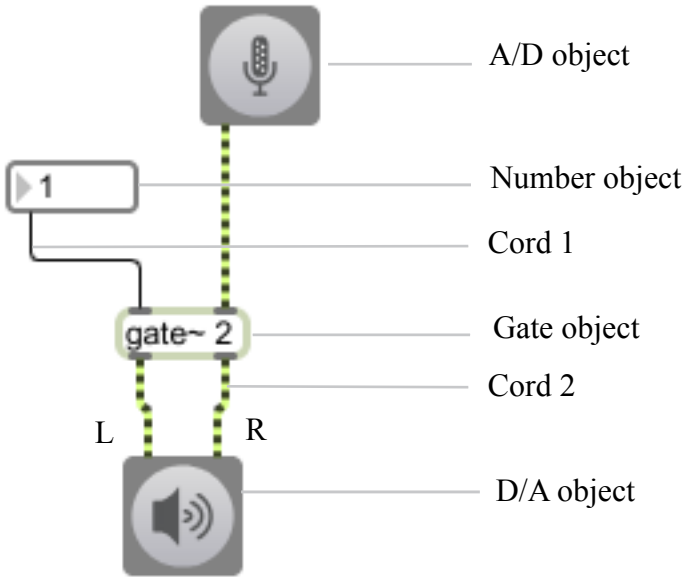


Figure 4.1: An example of a Max/MSP patch. There are 4 objects in the patch. The gate~ object has two inputs, one from the number object and the other as a signal streaming from an analog to digital (A/D) conversion object. The gate~ object is routing an incoming signal from the A/D object either from its left output or from its right output to the corresponding inputs of the D/A object. The switch to change the channel is the number box where 1 corresponds to the left channel and 2 to the right channel. Patch cords 1 and 2 are of different types where type 1 can transmits integers or floats and type 2 transmits signals.

A 2º resolution within the 81º range leads to 41 spatial *zones*, each *zone* implemented within Max/MSP as a 2-channel convolver (using SIR2 plugin). Figure 4.2 provides the flow chart of the Max/MSP implementation. In the simulation system, the opening of the spatial *zone*(s) is determined by the yaw of the user's (a talking-listener)

head, with data provided by the headtracker (a hardware component described in section 4.2). The *zone* that is currently open performs real-time convolution of the user's voice with the loaded OBRIR for that angle. There are two possibilities here, as seen in step 6 of the flow chart and illustrated in Figure 4.3,

(a) the user stays in the current position (pos. A, the NO branch of step 6) and,

(b) the user rotates his/her head position (pos. B, the YES branch of step 6).

In case (a) only two channels of convolved audio are *output*. In case (b), as the user moves his/her head from position A to B, the audio *input* to position A's *zone* is cut as soon as the user moves his/her head from this position, but the convolved audio *output* continues for the duration of the loaded impulse response, after which position A's *output* is also cut. And, as the user could be moving his/her head much faster than the length of the current impulse response (some of which can extend beyond 4 seconds), the process entailed in case (b) continues in an iterative manner, leading to a potentially large number of *zones* that are concurrently open, all outputting audio streams for that angle range. This provides for a highly realistic simulation of the interaction between a person's voice and the room since the audio *output* from a particular angle *zone* continues, as in the real world, even though the person is in a different *zone* (please refer to Appendix A, section A.5.1 for more information regarding this feature). It must be noted however, that the system follows a 1*in*/2*out* format (one mouth, two ears). So no matter how many angle *zones* are open allowing for multichannel convolution, the *output* is still two-channel binaural.

A crossfade of 10 ms is applied in the time domain after convolution to implement simple interpolation between the *outputs* from the closing angle *zone* to the *input* of the opening angle *zone*.

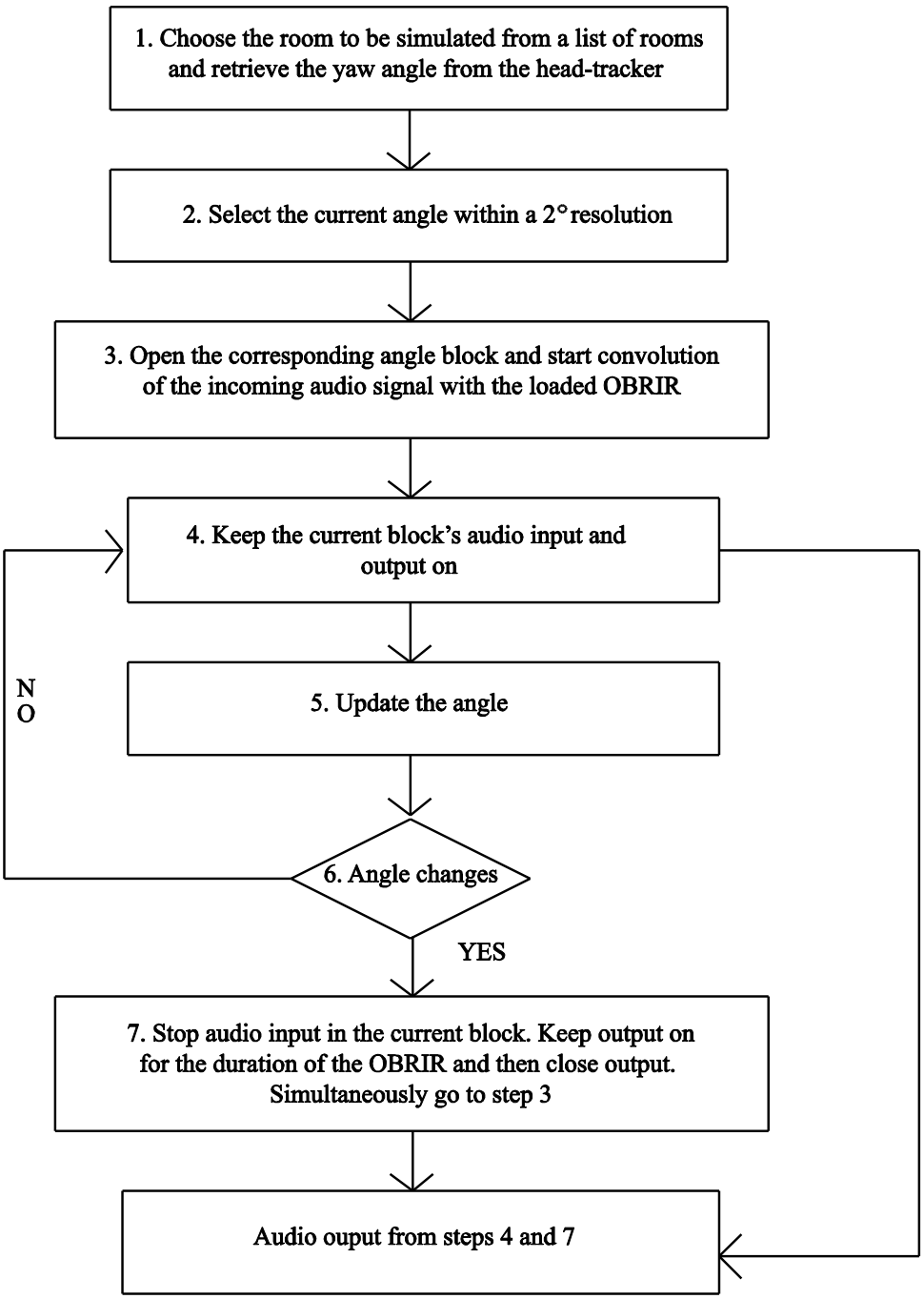


Figure 4.2: Schematic flow chart of the Max/MSP real-time convolution (using SIR2 plugin) system with convolution in steps 4, 8 and 9.

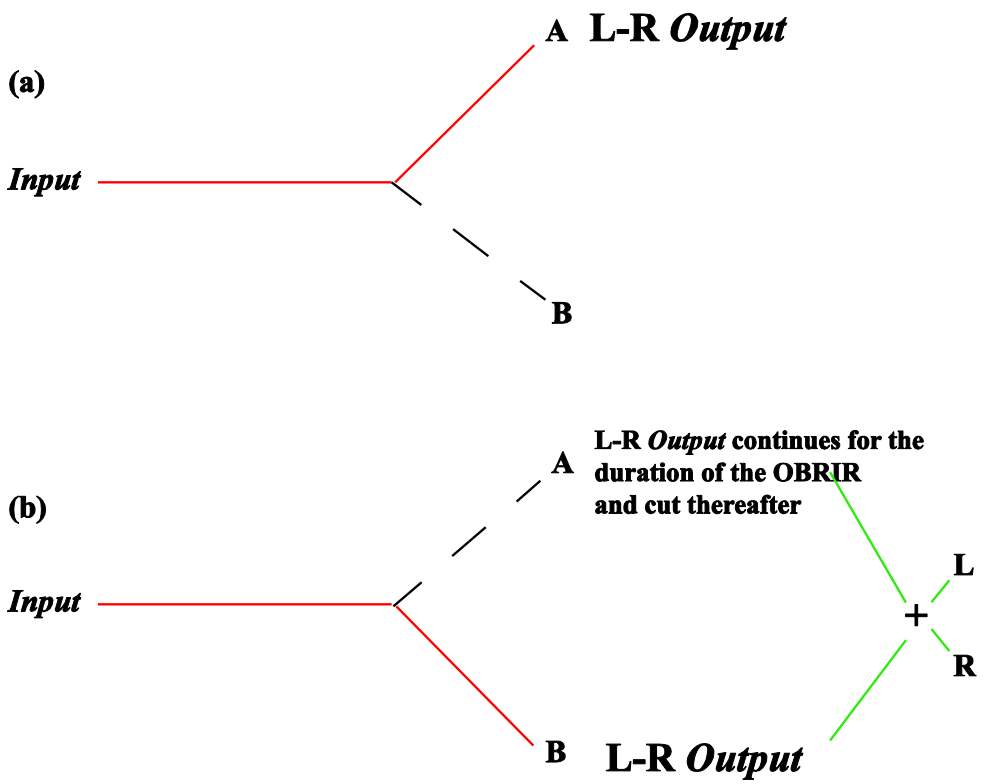


Figure 4.3: The two scenarios (a) and (b). Step (b) continues in an iterative manner. Output is always binaural (L-R).

4.2 Hardware module

The audio hardware dedicated to perform the analog-digital and digital-analog conversion is a RME ADI-8 QS AD/DA converter, interfacing with a RME HDSPe AES pci-express card on-board an Intel Xeon machine on a Windows platform. Audio streaming is done through Steinberg's ASIO audio driver interface, performed at a sampling rate of 48 kHz, 32-bit quantization with a variable buffer size (128 – 256; to be

elaborated in the latency discussion below) samples. The headtracker used is a Polhemus® Fastrak® unit.

A headset microphone (DPA 4066) is used as the input microphone and AKG K1000 as the ear-loudspeaker (these are loudspeakers near the ears, without any circumaural cushion or contact with the ears), which also holds the receiver as in Figure 4.4. The user is seated on a wooden chair, just in front of the transmitter of the headtracker, which is mounted on a wooden stand. The presence of metallic objects in the vicinity of the head-tracking apparatus is minimized, because they could interfere with the electromagnetic field that is generated by the transmitter and used to communicate with the receiver.

For this module, the positioning of the microphone and the effect of the ear-loudspeakers are considered with the help of a case study in section 4.2.1. The results of the case study in terms of microphone positioning on a talking-listener and the effect of the presence of ear-loudspeakers are discussed in 4.2.2 and 4.2.3, respectively.

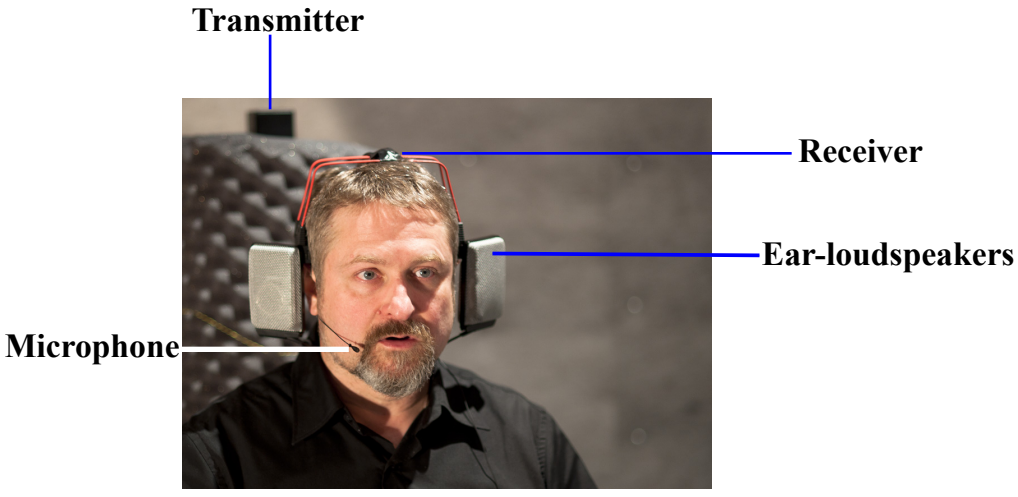


Figure 4.4: The positioning of the headtracker transmitter and receiver, microphone and ear-loudspeakers on a user of the system, in an anechoic room.

4.2.1 A case study to consider the microphone positioning and effect of ear-loudspeakers

This case study addresses two issues. The first is determining the position of the headset microphone for participants of different head shapes, and how this might affect the gain calibration of the system. The main cause for determining a microphone position arises from the consideration of placing the microphone away from the direct airstream out of the mouth, which has the advantage of eliminating to a large extent the detrimental effects of air turbulence from plosive and fricative sounds.

The second issue is the effect of the presence of the ear-loudspeakers, which are placed on a talking-listener as seen in Figure 4.4. Each ear-loudspeaker can be swivelled on the horizontal plane, ranging from the configuration shown in Figure 4.5 (a) to 4.5 (b). Here the configuration seen in Figure 4.5 (b) is the same as the one seen in Figure 4.4. The configuration seen in Figure 4.5 (b) is chosen as the preferred configuration for the current system as it minimizes the reflections of the direct sound from the surface of the ear-loudspeaker to the ears while maximizing the externalization of the convolved sound from the last step of Figure 4.2, which corresponds to the room reflected sound from autophonic stimuli.

The case study examined the two issues with 5 participants who read aloud a set text into a DPA 4066 headset microphone positioned at a distance of 7 cm from the centre of the lips to determine the effects of microphone placement on the gains of both ears, in the presence or absence of the ear loudspeakers.

Figure 4.5 (a) The close-to-ear and (b) the maximum angle away-from-ear configurations possible with the AKG K1000 ear-loudspeakers. Here (b) is the preferred position for the current system.

The microphone distance of 7cm from the centre of the lips is similar to a position used in several studies of singing acoustics (Cabrera, Davis, Barnes, Jacobs, & Bell, 2002). The same microphone and a similar distance are also being used in another real-time convolution set-up (Pelegrín-García, Fuentes-Mendizábal, Brunskog, & Jeong, 2011a), which is closest in scope to the current system in terms of its convolution implementation. The participants also wore microphones at the entrance of ear canal of each ear (Brüel & Kjær 4101 Binaural Microphone) and a microphone was positioned 1 m in front of the participant (Brüel & Kjær 4190). Hence there were a total of 4 signals being recorded per participant (1 mouth and 2 ear, and 1 m microphones). The same microphone arrangement was used on a HATS, for which a pink noise signal was emitted from the HATS mouth simulator and recorded at the microphones. Measurements were made with and without the ear-loudspeakers present. In order to gauge how consistently the headset microphone position represents the far-field radiated sound, transfer functions from the headset microphone to the 1 m microphone were derived for HATS and human participants using the cross-spectrum method (Knapp & Carter, 1976) from which the

average gain (in dB) over 7 octave bands (125 Hz to 8000 Hz) was determined. The transfer function from the headset microphone to the right ear microphone (on the same side of the face as the headset microphone was) were determined without and with the participant (and HATS) wearing the ear loudspeakers. This gives an indication of the effect of the ear loudspeakers on the direct sound.

4.2.2 Discussion of microphone positioning

As can be seen from Figure 4.6, the octave band gains for the HATS are largely within the range of gains seen for the human participants. The un-weighted power average of these octave band gains yields a gain difference of 1 dB between HATS and the human measurements, and the A-weighted gain difference between the HATS and humans is 0.2 dB. However, clearly there is substantial variation between individual participant gains across the frequency range, with a 5-7 dB range of values for each of the bands from 500 Hz and above, and a 12 dB range in each of the two lower bands. As the individual participant gains tend to fluctuate across the frequency range, the variation between participants is reduced when the octave band gains are power-averaged: the participants' A-weighted gain spans 3.3 dB, and their unweighted gains span 2.9 dB.

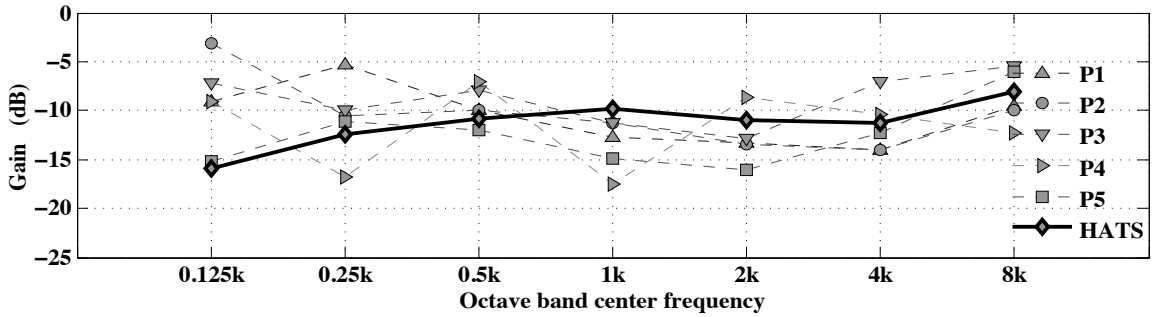


Figure 4.6: The gain (in dB) calculated from the transfer function of the headset microphone to the 1 m microphone for the left ear for 5 human participants and HATS.

4.2.3 Discussion of the effect of ear-loudspeakers

Table 4.1 and Figure 4.7 show that the presence of ear-loudspeakers has scarcely any effect on the octave band gains for both the HATS and human participants.

Octave band centre frequency	**125**	**250**	**500**	**1000**	**2000**	**4000**	**8000**
HATS with (-) without	-0.12	-0.19	-0.25	-0.19	-0.16	-0.11	0.14
Humans with (-) without	0.20	0.46	0.17	-0.08	-0.26	-0.06	0.32

Table 4.1: The difference between the octave band gain values for the HATS and Humans with and without the ear-loudspeakers.

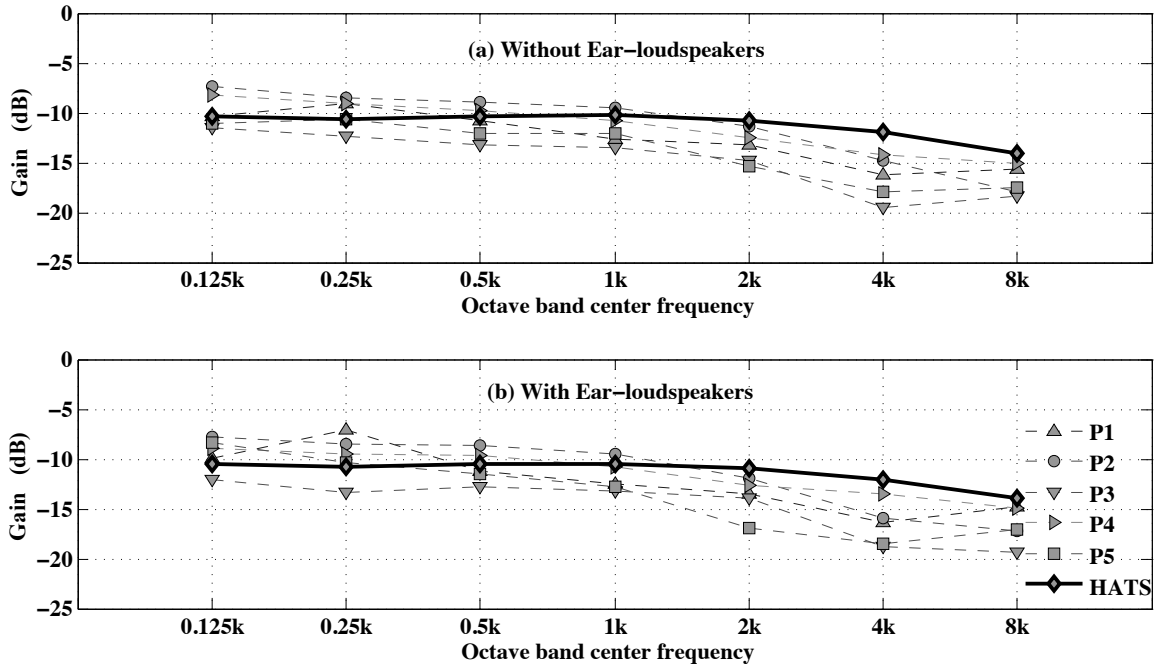


Figure 4.7: The transfer function from the headset microphone to the right ear microphone (on the same side of the face as the headset microphone was), without (a) and with (b) the 5 participants (and HATS) wearing the ear loudspeakers.

4.3 Characteristics of the complete system

Figure 4.8 provides a simplified view of how the system is actually set-up in the Acoustics Laboratory at the Faculty of Architecture, Design and Planning, The University of Sydney.

This section describes four important characteristics of the complete real-time room acoustic simulation system, which uses the software and hardware components listed in the previous two sections.

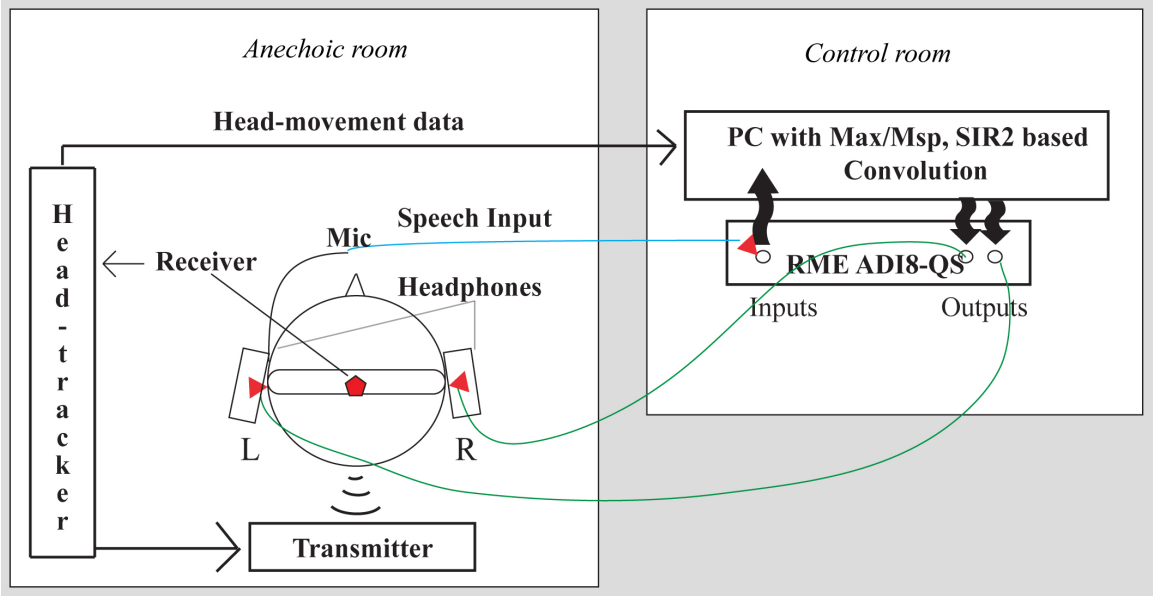


Figure 4.8: The complete set-up in the Acoustics Laboratory at the Faculty of Architecture, Design and Planning, The University of Sydney.

4.3.1 System latency

There are two stages within the current system where latency is introduced. The throughput (electroacoustic) latency is 7.6 ms and 11 ms with a buffer size of 128 and 256 samples respectively, at a sampling rate of 48 kHz. This latency is the round-trip time taken from the instant a person's voice is recorded at the headset microphone to the time that the convolved speech is emitted from the ear-loudspeakers, which the person can hear. For rooms of reverberation times of upto 2.5 s, a buffer size of 128 samples is used and 256 samples for reverberation times above 2.5 s. This variable buffer size implies that the earliest reflections that can be simulated by the system (excluding the floor reflection) are of the order of 2.04 and 3.2 m distances with 128 and 256 buffer size respectively. It is worth noting that this latency is compensated for within the system by truncating the initial part of the OBRIRs, as described in Chapter 2, section 2.4.1) by the number of samples represented by 7.6 or 11 ms of audio signal (depending on the buffer

size), so the latency of the room reflections reaching the talker's ear through the ear loudspeakers is essentially 0.

The other source of latency is the head-tracker update rate. Although the current system can withstand a head-tracker update rate of 100 ms (or 10 Hz) without any degradation in the output audio, any value less than this leads to clipping, which is environment (Max/MSP) dependent. Specifically, a fast update rate (less than 10Hz) causes the angle *zones* to be closed and opened at a rate faster than the signal processing in the environment (Max/MSP *poly~* object) allows.

The total system latency (TSL) in a simulated environment is defined as the time delay between the onset of an event, such as audio input or head movement, and the response of this event (Wenzel, Arruda, Kistler, & Wightman, 1993). TSL is therefore determined by the slowest component of the system. Going by this definition, the latency of the convolution system with head-tracking works out to be approximately 100 ms (head-tracker latency). However, TSL has typically been used in studies focusing on sound localization (Wenzel & Foster, 1993) where TSL of 250 ms is just noticeable (Wenzel, 2001) for long stimuli. Other studies have contested this value and have suggested a TSL threshold of 75 ms where values more than this threshold are detectable by the listener and are reported as causing incongruence in the virtual environment for the listener. The performance in sound localization did not improve with TSL more than 75 ms (Yairi, Iwaya, & Suzuki, 2008) in that study.

As the current system does not simulate the direct sound, only the changes in direction (relative to the head) of the room reflections are delayed by 100 ms. It is suggested here that the auditory image of the room reflections does not suffer much

degradation by a head-tracking response delay of 100 ms, although there are no studies that have directly addressed this issue.

4.3.2 Gain calibration

In order to calibrate the system gain a set of OBRIR are loaded into the convolution software, and a HATS is put in the simulation system instead of the human user. The system gain is adjusted so that the OBRIR measured through the simulation system (which includes the direct sound from mouth to ears) matches the original OBRIR that was loaded into the simulation system. Individualized gain correction for an experiment participant can then be efficiently derived once they are wearing the headset microphone by measuring the level difference between the headset microphone and 1 m microphone, compared to the corresponding level difference for the HATS (-10.8 dB unweighted and -11.7 dB A-weighted). It must be noted here that the gain changes are implemented by adjusting the convolver gain, with the headset microphone preamp and AD/DA gains being kept constant.

4.3.3 Loop gain and cross-talk

After gain calibration, there was a negligible effect (< -16 dB) due to loop gain (feedback from the ear-loudspeakers to the headset microphone) and cross-talk caused by the ear-loudspeakers for a frequency range of 100 Hz – 10 KHz, which covers the range of human autophonic output.

Here, the method for determining loop gain involved loading the convolver with an impulse that was gain matched with the loudest component of the OBRIRs used in the thesis. The system was then excited with a logarithmic sweep from the HATS' mouth

loudspeakers that was picked-up by the headset microphone. The loop gain was calculated as the average gain difference between direct and reflected sound components of the IR recorded from the sound (appropriately delayed to follow the sweep) from the ear-loudspeakers to the headset microphone.

Cross-talk was determined as the sound from each ear-loudspeaker to the microphone placed at the entrance of the corresponding contralateral ear-canal microphone (Brüel & Kjær 4101 Binaural Microphone).

4.3.4 Categorization as a mixed-reality environment

A distinction in terminology that is introduced in this chapter comes with the observation that the current system can be better classified as an auditory mixed-reality (MR) environment, a term consistent with the framework suggested by Milgram & Colquhoun (1999) for visual MR. The basis for this observation comes from the fact that the sound that the talking-listener hears is the sound of his/her own voice (*real* component), and the truncated (7.6 ms or 11 ms, depending on the buffer size) room acoustical response of a simulated room (*virtual* component). In other words, the real and virtual acoustic stimuli that are referred to here support the experience of a perceptual space that can be explored by humans and computers in a mixed sense of reality. In these environments, the distinction between real and virtual (computer modeled) can be sometimes hard to define but can be described as a continuum (Figure 4.9) that spans these two extremes (Milgram & Colquhoun, 1999). 'Presence' within such a continuum can be characterized by how immersive it is and the level of interaction that it permits (Witmer & Singer, 1998)

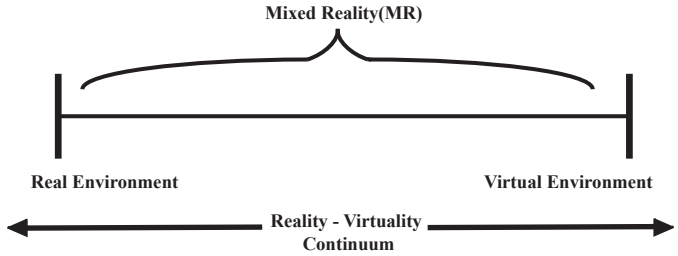


Figure 4.9: The concept of mixed reality and the reality-virtuality continuum adapted from Milgram and Colquhuon (1999)

4.4 Summary

This chapter is presenting a software-based real-time convolution solution to simulate one's own voice in binaurally measured acoustic environments, while one is physically present in an anechoic environment. The system described in this chapter can be used to design experiments that focus on studying the effect of room acoustics on the sound of one's own voice in simulated environments. The ability to switch between these environments with no latency enables the exploration of environments with various acoustical properties and physical volumes. The system can simulate real rooms that have been measured using a HATS, but could also be used with computer modeled rooms with the possibility of a much larger range of positions and orientations available for head-tracking control.

Although there are systems that simulate acoustic environments (Pörschmann & Pellegrini, 2010; Favrot & Buchholz, 2010; Schröder et al., 2010; Ueno, Kato, & Kawai, 2010; Silzle, Novo, & Strauss, 2004), some even simulating self-produced sound (Pelegrín-García *et al*., 2011a; Pörschmann & Pellegrini, 2010; Schröder et al., 2010; Ueno, Kato, & Kawai, 2010, the current system can be used to 'accurately' (Novo, 2005)

simulate one's own voice using minimal hardware and software compared to other systems.

As the system is based on binaural technology, it is worth discussing some key features of this approach of room simulation. Despite the short acoustic latency inherent to the binaural systems, they have been known to suffer from two limitations (Favrot & Buchholz, 2010). The first limitation is the argument that, in order to achieve the authenticity required for psychoacoustic experiments, the binaural system should use individualized, listener-specific head-related transfer functions (HRTFs). However, findings from a recent study (Zahorik, 2009) show little benefit in using individualized HRTFs over non-individualized HRTFs for room simulation involving fixed source direction, which is somewhat similar to the current simulation. The other limitation is the inside-the-head localization sometimes experienced in binaural systems. This should be greatly reduced or solved by head-tracking and using the ear-loudspeakers, and furthermore the more relevant question is whether or not the auditory impression of a room excited by one's own voice is externalized (as opposed to the auditory image of a sound source).

Chapter 5

Detection of headtracking in room acoustic simulations for one's own voice

This chapter contains the first of the two case studies that are presented as experiments involving human participants, to validate the simulation system presented in this thesis. The case study in this chapter addresses the headtracking implemented in the system. It begins with an overview of the concept of headtracking in section 5.1 that has been presented in more detail in the preceding chapters. The experimental set-up is described in section 5.2. The results of the experiment are discussed in section 5.3 and section 5.4 summarizes the results of the experiment in the context of the thesis.

5.1 Headtracking in room simulation systems

The impression of the sound of one's own voice in a room can go on from being unnoticed to remarkably striking depending on the interplay of many factors such as the level of directed concentration, context, situation, etc. but more importantly, on the acoustical characteristics of the room in which the talking-listener is present. People visiting an anechoic room for the first time can sometimes be overwhelmed with how 'dead' their voice sounds to themselves. On the other extreme, listening to one's own voice in a highly reverberant room could be accompanied with a sense of grandeur due to the minimal vocal effort required to produce very high levels. Voice projection in more

'everyday' rooms is what most people are familiar with, and here the room reflected sound of one's own voice can be a rich source of information to determine the room's characteristics when visual (McGrath *et al*., 1999) and other sensory stimuli are not present or augment other sensory inputs when they are present. This case study is limited to the first case where only auditory stimulus is present in a real-time simulation of room reflected sound of one's own voice.

In a room of fixed volume, the sound that reaches the two ears is determined not only by the various design features of the room such as the building material used, furnishings, etc., but also by the head position of the talking-listener. By being closer to one wall than others can result in room reflections that are distinctive in qualitative and quantitative features that can change when the head position changes. Hence, in the room acoustic simulation for one's own voice that is presented in this thesis, while the talking-listener is speaking or singing, his/her head-position is continually tracked and his/her voice is convolved with the appropriate OBRIR and reproduced on a pair of ear-loudspeakers that the talker-listener wears. Using the appropriate OBRIR corresponding to the talking-listener's head position leads to reflected sound images received at the ears that are stationary in external space even when the talking-listener moves his/head, closer to what happens in a *real* room environment – and this should reduce the incidence of inside-the-head localization. This is more likely to create a higher degree of 'presence' (Witmer & Singer, 1998) within the simulation, implying greater task based performance.

There is however, an aspect of the real-time simulation for one's voice that, due to the nascent nature of research in this field, has not been fully studied. This issue addresses the detectability of headtracking by the talking-listeners using their own voice

as the stimulus. In other words, it has not previously been shown whether incorporating various head-positions within simulations leads to a change in the perception of auditory scenes as the talking-listeners move their heads: given the measured OBRIRs have quantifiable differences over the head-movement range in question. Some of the rooms used in the present study have been shown to have a large variation in interaural features (such as early *IACC* range) amongst other acoustical parameters, over a BRS range (Cabrera *et al*., 2011) of -60° to +60° yaw angles. Such changes are likely to cause a change in the auditory scene as the head is moved over the BRS range and consequently change the room reflections associated with the talking-listener's vocal transduction.

Previous studies that have addressed headtracking in simulations have focused primarily on exocentric sound sources (Yairi, Iwaya & Suzuki, 2008; Welti & Zhang, 2010) and the applicability of those studies to egocentric sound sources may be limited. The current study addresses the detectability of headtracking within real-time simulation of one's own voice (egocentric) by conducting an ABX detection test within six simulated rooms. Here the rooms used are physically measured *real* rooms, not computer generated rooms.

Apart from serving as a validating experiment for the system presented in this thesis, the findings of this case study are also likely to influence future studies that are focusing on real-time simulation of egocentric sounds.

5.2 Experimental set-up

5.2.1 Participant information

Five participants (all male) took part in the detection test. They were selected to provide a reasonable variation in listening capabilities within the limited sample size. The participants ranged from being expert listeners (2), architecture postgraduate students architecture with no formal musical training (2) and one postgraduate student in acoustics who reported slight hearing loss (not quantified here). The experiment was conducted in an anechoic room in the Acoustics Laboratory, Faculty of Architecture Design and Planning, The University of Sydney.

5.2.2 Real-time room acoustic simulation system

For a description of the system used in this case study, please refer to chapters 2 and 4.

5.2.3 Rooms used

The rooms simulated in the experiment were real rooms in the Faculty of Architecture Design and Planning, The University of Sydney, ranging in volume from 125 m^3 to 7650 m^3. The characteristics of the rooms measured are described in detail in a paper by Cabrera *et al*. (2011) and in Appendix B (with a permission by the authors), so this section is only presenting their volumes and mid-frequency reverberation times in Table 5.1.

Room no.	Volume (m^3)	Reverberation time (s)
1 (3)	125	0.6
2 (6)	152	0.35
3 (7)	170	0.4
4 (8)	188	0.9
5 (10)	610	0.6
6 (11)	7650	1.7

Table 5.1: The rooms used in the experiment, indexed from 1-6 with a number in bracket showing their index in the study by Cabrera *et al.* (2011), followed by their volume and mid-frequency reverberation times.

5.2.4 ABX headtracking detection test

In an ABX test, each trial consists of three stimuli (A, B and X). The stimuli A and B are different, but one of them is identical to X. The participant's task is to determine which one (of A and B) is the same as X. Before the experiment, it was explained to the participants that the presence/absence of headtracking was being tested in the ABX test. The participants were seated on a wooden chair placed on a carpeted portion of the anechoic room (with a large wooden board underneath the carpet). They were given a few sheets of printed text with the choice that they were free to either read from the text or make any other kind of vocalization that would enable them to make a match between either A-X or B-X. The A, B and X stimuli were the same simulated room but were randomized to have headtracking state either *on* or *off* with A & B having the opposite state per trial and X having an *off* or *on* state. Hence a correct answer required matching an *on-on* or *off-off* pair. The experimenter was controlling the state of the

headtracker switch (randomly generated) in the Max/MSP patch from a control room, while being able to communicate with the participants throughout the experiment over a two-way monitoring audio channel. This was done to avoid any issues that are introduced by putting computers screens (source of reflection) or projectors with fan noise, in an anechoic room.

Each of the rooms was tested twice (in randomized order) giving a total of 12 trials (*N*) with each of the two possible X states tested (*on* and *off*) per simulated room. All the participants performed the experiments without any break and there was no limitation on how long and in what order they wanted to listen to any of the three stimuli (A, B and X).

5.3 Results and discussion

5.3.1 Analysis per participant

In order to perform a statistical analysis on ABX test data it is necessary to decide on the values of *r*, *p*, Type 1 error, Type 2 error and a fairness-coefficient (FC_P). Here *r* is the threshold for the number of correct detections and *p* is the percentage of correct detections required in an independent *Bernoulli trial*, which determines what is considered detectable. For N=12 independent trials, a value of r=7 implies that the listeners are correctly identifying more than 50% (guesswork) of the stimuli. As the listening skill of the participants in the current experiment was assumed to vary over a large range, *p* was taken to be just above chance, i.e. 0.6. In other words the probability (*p*) of the participants getting more than 50% correct detection was taken as 0.6. Type 1 and Type 2 can be calculated from binomial distribution for a value of *p.* They arise from results that indicate that different stimuli are identical (Type 1 error) and that identical

stimuli are different (Type 2 error). It must be noted here that as the system required the participants to speak for them to hear the simulated room reflections, the sound of no two stimuli were actually identical. FC_p (Leventhal, 1986) has been used as a measure of the degree to which the two error risks have been equalized for a given p, and can be calculated as

$$FC_p = \text{smaller probability / larger probability} \quad (1)$$

Table 5.2 shows that the maximum FC_p value of 0.865 is attained with Type 1 and Type 2 errors as 0.387 and 0.334 respectively. So 86.5% of the time, there is a 38.7% chance of different headtracking states being heard as identical and 33.4% chance of no change in headtracking state being heard as different when it is identical.

By inspecting Figure 5, it can be seen that all participants performed r=7 or better for correct detections. Participants 1 and 2 were classified as expert listeners at the beginning of the test and they had more correct detections than the others. Another interesting finding was the fact that all listeners correctly identified the stimulus pair for the room with the highest early *IACC* range as the head turned (median values; for OBRIRs ranging from -60° to +60° in Cabrera *et al.*, 2011).

r	Type 1 error p=0.5	Type 2 error p=0.55	p=0.6	p=0.65
11	0.0032	0.9917	0.9804	0.9576
10	0.0193	0.9579	0.9166	0.8487
9	0.0730	0.8655	0.7747	0.6533
8	0.1938	0.6956	0.5618	0.4167
7	**0.3872**	0.4731	**0.3348**	0.2127
6	0.6128	0.2607	0.1582	0.0846
5	0.8062	0.1117	0.0573	0.0255

Table 5.2: The calculated values (bold face) of Type 1 and Type 2 errors for different number r correct detections. Number of binomial trials, N = 12

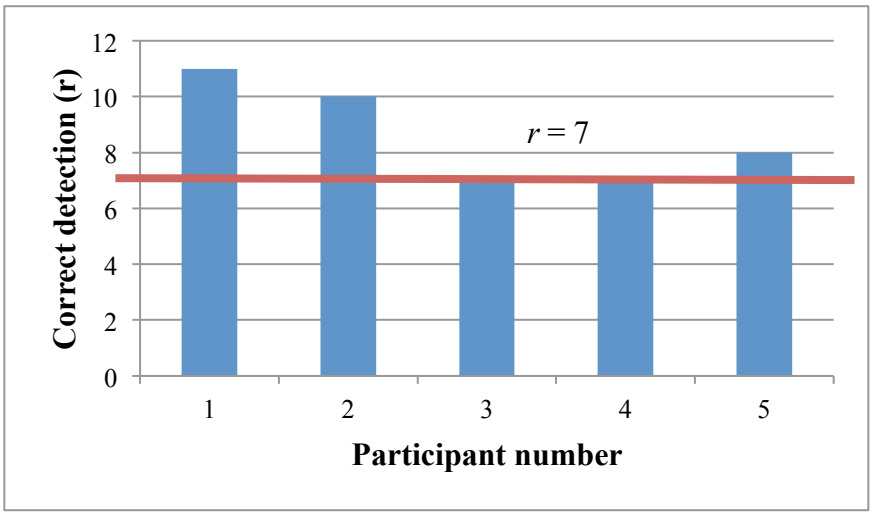


Figure 5.1: Number of correct detections (r) in the ABX test plotted for each participant. r=7, the detection threshold is represented as the thick red line on the plot

It must be emphasized here that the significance level of 0.39 and 0.33 is much higher than 0.05 or 0.01, which is typically used in statistical analysis (Type 1 error). But the whole rationale behind introducing the measure of fairness coefficient is to reduce Type 2 errors so that detectable differences are not concluded to be undetectable for a relatively difficult test such as the current one, where participants are tested under unfamiliar circumstances while also being protected from becoming exhausted from doing too many trials (increasing N). Leventhal (1986) suggests that in order to equalize

Type 1 and Type 2 errors for a given N and p, it is better to have leniency in allowing a higher value of Type 1 and Type 2 errors that are still comparable to each other, as it increases the overall statistical power of the analysis by reducing the probability of overlooking Type 2 errors. But it also has to be acknowledged that a study with more trials could be more conclusive and with smaller values of Type 1 and Type 2 errors.

5.3.2 Analysis for participants' concatenated results

On the other hand, as each trial in the current experiment is an independent Bernoulli trial, the results from all the participants can be concatenated, which gives us N=60. For this value of N, r can be set as 37, for relatively higher probability of p=0.7 leading to corresponding Type 1, Type 2 errors and FC_p values of 0.046, 0.063 and 0.73. The number of current detections in the concatenated case is 43, which are greater than the threshold r. So here, 73% of the time, there is a 4.6% chance of different headtracking states being heard as identical and 6.3% chance of headtracking states being heard as different when it is identical. However, the previous line of analysis is preferred here, as it provides a clearer picture of detection performance across the participants.

5.3.3 Detection for each room

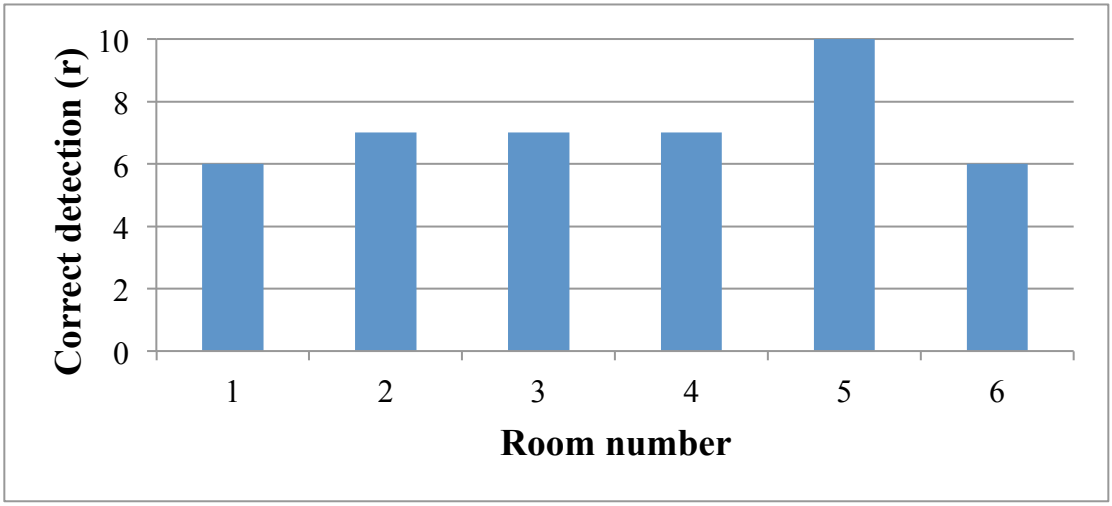


Figure 5.2: Number of correct detections (*r*) out 10 in the ABX test for the concatenated participant results for the six rooms that were simulated arranged in ascending order of physical volume in m^3

Figure 6 shows the number of correct detections per room where the rooms are ordered in ascending order of room volume as listed in Table 1, where 100% correct detection is seen for room 5 and more than 50% detection seen for all the rooms. Even though not undertaken in the current study, there is scope for further research into correlating the findings of Figure 5 with the room characteristics detailed in Cabrera *et al.* (2011).

5.3.4 Presence in simulated rooms

All the participants reported the phenomena of being in a different room from the one they were physically in, just by hearing the convolved sound of their own voice

coming from the ear-loudspeakers, for both the headtracked and non-headtracked stimuli. The detection of headtracking implies that the participants experienced a higher degree of presence (Witmer and Singer, 1998) in the rooms when the room reflections changed in accordance with their head-movements. In the future, a presence questionnaire could be used to determine the degree of presence more explicitly.

5.4 Summary

Although the validity of the equalizing Type 1 and Type 2 errors as suggested by Leventhal (1986) can be questioned from an engineering perspective (Shanefield, Clark, Nousaine & Leventhal, 1987), where sometimes a significance level of 0.05 or lower may be essential, such equalization nevertheless increases the statistical power of the analysis of ABX test data for limited *N* studies. Increasing the statistical power is a more reasonable approach in cases such as the current, where the difference between the stimuli being tested can be small to moderate.

Another aspect related to choosing a significance level is that it has not been studied whether the talking-listeners executing similar head-movements in a *real* room would in fact detect a change in the sound field, which is *simulated* with headtracking in the current system. In the current study 100% detection was noticed for the room with the largest *IACC* range (Cabrera *et. al*, 2011) over the range of yaw angles (Figure 5.2), with lesser detection rates for other rooms (however more than 50% in all cases). This suggests that detecting a change in the sound field might be intrinsically difficult for some rooms for a set of talking-listeners with a variable range of listening skills. However, this can only be tested in an experiment with human participants where they

directly compare the sound field in the *real* vs. *simulated* conditions of the same room environment, with similar head-movements. This can be seen as a further step in the perceptual validity of the system. Also, by controlling relevant acoustical parameters (such as the *IACC*), the results of such a study could help the experimenters make decisions about choosing an appropriate significance level and detection threshold for rooms with certain acoustical characteristics.

Following the above discussion and the discussion in section 5.3, it can be stated after Type1 and Type 2 error equalization as described by Leventhal (1986), headtracking was shown to be detectable by five participants performing an ABX detection test for an egocentric sound source (one's own voice). The participants were tested in six measured real rooms that were simulated using the real-time room acoustic simulation system that is described in chapters 2 and 4 in this thesis. By concatenating the results for all the participants, each being an independent trial (N=60), higher probability of detection with considerably lower values of Type 1 and Type 2 errors is noticed. Future studies could be organized to include more participants, include more rooms with a larger variation in acoustical parameters (especially *IACC* and reverberation time), and modify existing experimental set-up to permit more trials per participant while avoiding participant exhaustion. As the current simulation is only incorporating head-movements along the horizontal plane (yaw) over 81º range, it would be interesting to incorporate at least yaw and pitch, all with a larger range; and also to detect the threshold of detectability by constraining the head-movements. Finally due to the technical issues that may be involved with measuring real rooms, computer simulated room could be used to generate OBRIRs over the degrees of freedom.

Chapter 6

Auditory room size perceived from a room acoustic simulation with autophonic stimuli

This chapter describes the second of the two case studies that are presented as experiments involving human participants, to validate the simulation system presented in this thesis. This case study was conducted to determine the acoustical parameters that influence the judgement of a room's size, based on primarily auditory stimuli (one's own voice). As in the previous case study, no visual information about the rooms was presented. Human participants performed talking tasks in rooms that were acoustically simulated in real-time, and rated the aurally perceived size of each room.

Section 6.1 contains a review of literature pertaining to auditory room size perception. Section 6.2 describes the experimental set-up. Section 6.3 provides an analysis and discussion of the results of the experiment. The study is summarized in section 6.4.

6.1 Auditory room size perception

The size of a room is one of its most basic attributes, and this study examines the perception of room size using sound alone. Although it can be argued that the most reliable judgement of room size can be arrived at from visual inspection, it is also possible to judge the size of a room using auditory stimuli, without accompanying visual

stimulus ((Tajadura-Jiménez, Larsson, Väljamäe, Västfjäll, & Kleiner, 2010; Pop & Cabrera, 2005; Hameed, Pakarinen, Valde, & Pulkki, 2004; McGrath et al., 1999; Mershon & King, 1975). This involves exciting the room with an appropriate sound source and hearing the characteristics of the acoustic reflections from the walls, furnishings etc. Experimental studies eliciting auditory room size judgments can provide insight to space perception processes of people with a significant visual impairment (McGrath *et al*., 1999); contribute to the understanding of reverberance in concert halls (Lokki, Pätynen, Kuusinen, Vertanen, & Tervo, 2011); and extend the understanding of psychoacoustics relating to autophonic output (one's own voice) (Lane, Catania & Stevens, 1961) in rooms (Pelegrín-García *et al.*, 2011a; Cabrera, Jeong, Kwak, & Kim, 2005).

In listening to the sound of a room, the sound source can be the listener him/her-self (egocentric stimulus) or there can be a sound source physically distinct from the listener (exocentric stimulus). The scenarios arising from these exocentric and egocentric stimuli constitute exocentric and egocentric tasks, respectively. Previously, in mostly exocentric tasks, auditory room size perception has been shown to be more strongly affected by acoustical parameters (specifically the room's reverberation time, source-receiver distance and clarity index) than the room's physical volume (Cabrera, 2007). Table 6.1 provides an overview of the past research in auditory room size perception using exocentric stimuli.

This study however, investigates auditory room size perception in an egocentric task; based on an auditory mixed-reality (MR) environment, a term explained in chapter 4, section 4.3.4.

Author	Parameter(s) associated with auditory room size judgements
Sandvad (1999)	Reverberation time in a positive relationship, direct to reverberant energy ratio, energy measures.
Hameed *et al.* (2004)	Reverberation time in a positive relationship.
Cabrera *et al.* (2005)	Clarity index (C_{80}) in a negative relationship, where the actual room size is held constant, and reverberation time and source-receiver distance are varied.
Mershon *et al*. (1989)	Reverberation time, source-receiver distance and background noise level, in a positive relationship.
Pop *et al.* (2005)	Clarity index (C_{80}) and stimulus sound level, in a negative relationship.

Table 6.1: A summary of past research in auditory room size perception with human participants, using exocentric stimuli. The last two shaded rows indicate that the auditory stimuli were presented in real rooms, whereas the first three rows indicate that auditory stimuli from computer modelled rooms were used.

6.2 Experimental set-up

6.2.1 Participant information

Room size judgements were made by 8 participants (ages 23-45; 7 male, 1 female; 4 acoustically knowledgeable and 4 acoustically naïve university students), who were seated on a wooden chair placed on a carpeted floor in an anechoic chamber (with a large wooden board underneath the carpet, as described in Chapter 4, section 4.3). They were given a few sheets of printed text with the choice that they were free to either read from the text or to use any other speech or vocalization that would enable them to judge the size of the simulated room, with typical or more exploratory head movements. The participants were tested in the seven room simulations according to a random order, with

two trials per room. They gave a room size rating for each trial using a numerical scale ranging from 1 (the size of the anechoic room in which the talking-listeners were physically present) to10.

6.2.2 Real-time room acoustic simulation system

For a description of the system used in this case study, please refer to chapters 2 and 4.

6.2.3 Rooms used and the acoustical parameters tested against the room size judgements

The rooms simulated in the experiment were real rooms in the Faculty of Architecture Design and Planning, The University of Sydney, ranging in volume from 125 m^3 to 7650 m^3. The characteristics of the rooms measured are described in detail in a paper by Cabrera *et al*. (2011) and in Appendix B (with a permission by the authors). As a control, one of the rooms was measured in two conditions, differing only by the presence of a small curtain near the measurement position, leading to a slight change in the acoustical parameters. These two conditions of the same room were included in the current experiment to test the variation in the room size judgements of essentially the same room, leading to a total of seven simulated room conditions.

Based on the findings of previous studies, room size judgments were examined in relation to measures of physical room size (volume, *V*) and to acoustical parameters derived from the OBRIRs. The acoustic parameters include the following: mid-frequency (500 Hz) reverberation time (*RT*) with an evaluation range from -10 dB to -30 dB (amended from the more commonly used -5 dB to -25 dB range, to account for the higher

gain of the direct sound); room gain ($G_{RGearly}$) derived from the amended procedure outlined by Pelegrín-García (2011), which was first proposed by Brunskog, Gade, Bellester, & Calbo(2009) as a measure of the energy of the room-reflected sound that the talking-listener hears (power average of the two ears, expressed in dB); clarity index (C_{50}) (Pop & Cabrera, 2005; Cabrera *et al.*, 2005; Cabrera, 2007) and interaural cross-correlation ($IACC_{early}$) (Cabrera, 2007), using 80 ms as the boundary between early and late. One distinction in the calculation of the room gain values here from the procedure described by Pelegrín-García (2011) was the duration of direct sound, which in the current study was taken as 7.6 ms and corresponded to the duration of the direct sound and first floor reflection of the OBRIR (Chapter 4, section 4.3.1). In the case of the room gain, the values presented corresponded to the 0° OBRIR while the *IACC* values were the octave band mean values over the entire headtracking range of -40° to +40° yaw as described in Cabrera *et al.* (2011) (please refer to Appendix B for a review of the parameters). Early decay time was not calculated because it is not well-defined for a source very close to a receiver.

V_{est} is a quasi-acoustical parameter calculated from an empirical function relating room volume to reverberation time ($RT \approx 0.26 \ln(V) - 0.75$) that was derived by Shabtai *et al.* (2010).

6.2.4 Results of the experiment

The room size judgements of each participant were centered (by dividing each rating by their mean rating) so that the participants would have a more equal weight in the analysis of combined results. The subjective room size ratings and physical parameters are shown in Table 6.2.

Room	Rated Size	V(m^3)	RT(s)	G_{RG}(dB)	C_{50}(dB)	$IACC_{early}$	V_{est}(m^3)
1 (3)	0.91	125	0.60	1.05	11.8	0.25	179
2 (6)	0.76	152	0.35	0.81	18.3	0.26	68
3 (7)	0.70	170	0.40	0.83	20.7	0.21	83
4 (8)	1.25	188	0.90	1.59	11.6	0.21	570
5	1.27	188	0.90	1.54	12.5	0.23	570
6 (10)	0.63	310	0.50	0.68	20.5	0.54	122
7 (11)	1.48	7650	1.70	0.29	31.6	0.54	12370

Table 6.2: The data used for the statistical analysis. The rooms are numbered from 1-7 with a bracketed number showing their index in the paper by Cabrera *et al.* (2011), which characterized the rooms used in this study in detail (see Appendix B). The next columns consecutively show the mean rated room sizes; volumes; mid-frequency reverberation times; early room gains; clarity index; early *IACC* values; estimated volumes from the linear regression model described by Shabtai *et al*. (2010) Rooms 4 & 5 (a control) were the same room measured in two slightly different conditions, but only room 4 is characterized in Cabrera *et al.* (2011)

6.3 Analysis and discussion

To determine whether there was a variation in the rated values of room size with different room conditions, a one-way ANOVA was conducted, and the result indicates a significant effect ($F(6, 49) = 24.63$, $p < 0.01$). As the near-identical stimuli (rooms 4 and 5) received very similar size ratings (Table 1), this suggests that the participants were consistent in judging the size of the same room in two slightly different conditions. Condition 5 was excluded from further statistical analysis, as its subjective ratings and objective parameters were so similar to condition 4. Also, condition 7, which represented the autophonic perception in a large reverberant room environment (a recital hall) was identified as an outlier and consequently not included in further statistical analysis.

Following these two deletions in the data set, analysis showed that none of the parameters were significantly correlated with the physical room volume ($p < 0.05$). However, considering these non-significant correlations for their polarity, a negative sign of the correlation coefficient (r) indicates that as the room volume increases, G_{RG} decreases ($R = -0.31$, $p = 0.30$), and *vice-versa*; whereas a positive sign indicates that as the room volume increases, C_{50} ($R = 0.49$, $p = 0.20$) and $IACC$ increase ($R = 0.59$, $p = 0.14$), and *vice-versa.* These signs are at least partly consistent with expectations from room acoustics theory in that a greater diffuse field strength is expected in smaller rooms (leading to increased G_{RG}, and reduced $IACC$); and the expected relationship between C_{50} and room size is more subtle (see [18]). As an important design feature in this study, it is noteworthy that there was no correlation between reverberation time and room volume for the selection of rooms ($R = 0.01$), although a more comprehensive sampling might show a positive correlation between these two parameters (as represented by V_{est}, following Shabtai *et. al*, 2010)

On the other hand, the room size judgements are significantly correlated with all the parameters that are listed in Table 1, except the room's physical volume and $IACC_{early}$. Figure 1 shows the linear regression model (R^2=0.99, F=220.6, p<0.001) that was yielded by room gain as the independent (predictor) variable, which can be expressed as,

$$\text{Predicted room size} = 0.17 + 0.68 \times G_{RG} \quad (1)$$

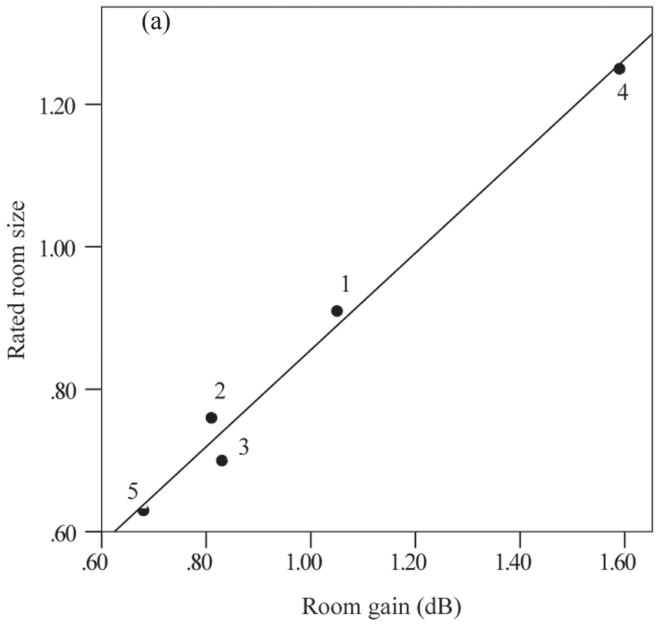
In recent research, higher room gain values have been shown to be important in providing greater vocal comfort and lesser vocal effort for talking-listeners, and vice-versa (Brunskog *et al*., 2009; Pelegrín-García *et al*., 2011b) The results of the present

study are consistent with these findings, with respect to a negative correlation of physical room volume with room gain, as the strength of the reverberant field in a smaller room is generally higher than bigger rooms. Hence, from an objective perspective, room gain values could serve as an important component in the prediction of the room's size. However, the *positive* correlation of the subjective room size responses with the room gain values, modelled in equation (1) is interesting, as it points towards a conjecture that the strength of the reverberant field in the current study was used as an indicator of its reverberance (and that greater reverberance was interpreted as an indicator of greater room size). This conjecture is partly based on the post-experiment interview with the participants, who reported using the reverberation of the rooms as an indicator of their size. Note that the effectiveness of room gain as a predictor in the present study is supported here most strongly due to the zero correlation between reverberation time and room volume.

Future research should focus on studying the interaction between the strength and temporal aspects of reverberant sound fields with respect to auditory room size judgements, where these two parameters are manipulated within rooms of fixed volumes. Similar to the present study, where the reverberation times of the rooms were uncorrelated with their volumes, various levels of correlation between these parameters may be included as a design feature.

There is also scope for improving the experimental design of the current study, by including simulated room conditions with a more uniform scale in terms of their physical size and variety in terms of their purpose (e.g., residential rooms). A method more robust

than magnitude estimation (e.g., paired-comparison, or photograph-matching; Sandvad, 1999) could be employed to validate the findings of this study.



(b)	*B*	*SE B*	*β*
Constant	0.17	0.05	
Room gain	0.68	0.05	0.99

Figure 6.1: (a) The room size judgements by the participants, as a function of reverberation time of the rooms in Table 1, where the rooms are numbered 1-7 in an ascending order of their physical volume. (b) The regression model for predicting perceived room size from reverberation time (*RT*). *B* and *SE B* represent the unstandardized coefficients and their standard error, respectively. β represents the standardized coefficient which gives the number of standard deviations the outcome (predicted room size) will change as a result of one standard deviation in the predictor (*RT*).

6.4 Summary

It can be argued here that the design of the current experiment can be elaborated to include more participants, more simulated rooms and dynamically changing room acoustic parameters. Nevertheless, the main aim of the study was to test human

performance in the system, for a task that has previously been mainly studied with exocentric stimuli. The results of the experiment show that a complex task such as auditory room size perception, which requires a high degree of presence within the simulated environments, can be easily performed by the participants. Here, once the real rooms are measured (which can be from different buildings), they can be simulated and used in an experiment with relative ease, when compared to using the same real rooms in an in-situ experiment. The findings of the experiment in this chapter point to a possible difference between the perception of room size and physical acoustic correlates of room volume, which raises questions for future study. Also, as the summary of the last chapter indicates, there is scope for a systematical comparison of the in-situ and other more conventional exocentric studies (Appendix C) with the system described in this thesis, in terms of their applicability to psychoacoustic research. More specifically, it would be interesting to compare the results of the study conducted in this chapter to a study where the current room acoustical simulation system is used for auditory room size perception (or a similar autophonic stimuli based task such as the one described by Pelegrín-García *et. al*, 2011a) without headtracking.

Chapter 7

Conclusions

Even though the perception of one's own voice in room environments is one of the most common auditory stimuli in everyday activities for people with normal hearing (or people with hearing loss, to an extent), the amount of literature for such autophonic perception in the field of room acoustic is very limited. The handful of studies that have addressed autophonic perception have shown that the sound that is reflected from the acoustically relevant surfaces in rooms, when the room environment is excited by one's own voice, is an important component of human hearing in these environments, which can sometimes change dramatically from room to room (Brunskog *et al*, 2009).

The room reflected sound from one's own voice is referred to as the sound received from the indirect-airborne conducted pathway to the human hearing apparatus. The other two pathways are the direct-airborne conducted sound (or simply direct sound) and bone (or body) conducted sound, as seen in Figure 1.1. As the sound from the direct (within comparable atmospheric conditions and without interference) and bone conducted (when relatively unimpeded) pathways do not differ much from room to room, it is the room reflected sound that primarily affects voice projection (Brunskog *et al*., 2009) and makes some room environments easy to speak in and conversely makes some room environments difficult to speak in (Sato, Bradley, & Morimoto, 2005). The rooms that are difficult to speak in generally offer insufficient room reflections or 'support' for the

talking-listeners, which is a major cause of vocal disorders (vocal strain related) of teachers and other professions requiring public speaking (Brunskog *et al*., 2009), who have to raise their voice beyond a comfortable level in order to be certain that they would be heard by the audience.

Overall, from a psychoacoustic perspective, hearing the sound of the room reflected sound from one's own voice conveys a wealth of information about the room environment to the talking-listener (McGrath *et al*., 1999) most of which has not been studied yet. For another perspective, one can focus on more 'virtual' worlds, where despite the advances in 3-D virtual environment technology, which can be largely attributed to incorporating real-time rendering and immersion into visual spaces, the corresponding virtual acoustic immersion is far from being on par (Schröder *et al*., 2010). Such immersion would allow users to hear their voice as they speak, as it would sound in particular room environments, where the environment can have a variable component of reality depending on the scope of the particular application. If an appropriate representation of one's own voice is incorporated in these environments, the users would hear the room reflected sound change dynamically as they change their head orientation and this would create a much higher level of interaction with the virtual environment, a crucial aspect in creating a sense of 'presence' (Witmer and Singer, 1998) within these environments. It must be noted here that in both these scenarios, the representation and use of externally created (or exocentric) sounds is much higher, both in terms of research and applications, when compared to the representation of self-created (or egocentric) sounds in similar contexts.

A major reason for the disparity between the visual and acoustic aspects, and the exocentric and egocentric aspects as mentioned above has been the computational complexity in rendering a perceptually accurate presentation of the room reflected sound with one's own voice as the stimulus (Torger and Farina, 2001). This thesis has addressed this issue by presenting a room acoustical simulation system for one's own voice, using generic software for real-time convolution and fast hardware for AD/DA conversion on a Windows platform. The system also incorporates headtracking to continually update the auditory scene that is presented to a taking-listener, in accordance with his/her head movements to create an interactive acoustic space, which is comparable to real rooms.

The system uses a modified method of measuring room impulse responses, which are measured from the sound emitted from the mouth (oral sound) and received at the two ears (binaural sound from the oral transduction) of the same head, which is consequently named Oral-Binaural Room Impulse Response (OBRIR) measurement method. These OBRIRs have been measured for a range of yaw angles in the horizontal plane for the current implementation, which can be extended to other degrees of freedom in the future. The OBRIRs can be measured for real or computer generated rooms, which when loaded in the software-based module of the complete system, can be simulated in real-time with instant switching between room environments. Within the system, a talking-listener speaks or sings into a headset microphone in an anechoic room and hears the simulated room reflected sound of a room in which he/she is not physically present, without any time-lag between the onset of the vocal transduction and the arrival of the simulated sound to the ear-loudspeakers that the talking-listener wears.

The system has evolved from room acoustical research in the past, which has been incorporated in the design of other real-time acoustical simulation systems, some of which being comparable to the current system, as described in Appendix C. However, it is the opinion of the present author that even with its present limitations, the current system offers a more 'accurate' representation of the sound of one's own voice in rooms. The system presented in this thesis also has more scope in terms of the relative ease it offers in terms of customization, with or without headtracking, to suit a wider range of research and commercial applications with minimal latency.

Though the applicability of the system is focused towards room acoustic studies that involve the perception of the sound of one's own voice, illustrated by the case studies in Chapters 5 and 6 of this thesis, it can also be used in commercial applications such as teleconferencing, virtual worlds, gaming etc., with appropriate modifications to augment the mixed-reality within these environments.

Appendix A

Software based real-time convolution

This appendix describes the software module of the room acoustical simulation system presented in this thesis. The module is implemented as a Max/MSP application, by hosting a VST plugin, SIR2, to perform the real-time convolutions. In the brief introduction to Max/MSP in the section 4.1.1 of chapter 4, it was mentioned that a Max/MSP application is organized in terms of patcher. A patcher can be likened to a piece of machinery in an automobile, where a patcher can be the whole automobile, or there can be a number of patchers that are arranged according to the functionality of the automobile. Similarly, a Max/MSP application can have just one patcher or an assortment of patchers (and sub-patchers) that exchange information, which is determined by the logic of the application. In the present case, the application is easier to implement and maintain with a number of patchers (and subpatchers) that interact with each other.

In what follows, the Max/MSP application will be described in a format that follows a top-down flow of information. Starting from the front-end, the application will be 'unfolded' at various levels. Figure 1 provides an overview of the how the sections describing the application are organized. Each section is compounded by figures of the corresponding parts of the Max/MSP application. The figures and the text in each section are meant to be referred to as a whole and not separately, as the figure captions in this appendix carry limited information about the various parts in the figures, and most of the

information about the visual map illustrated in the figures is covered in the text of that section.

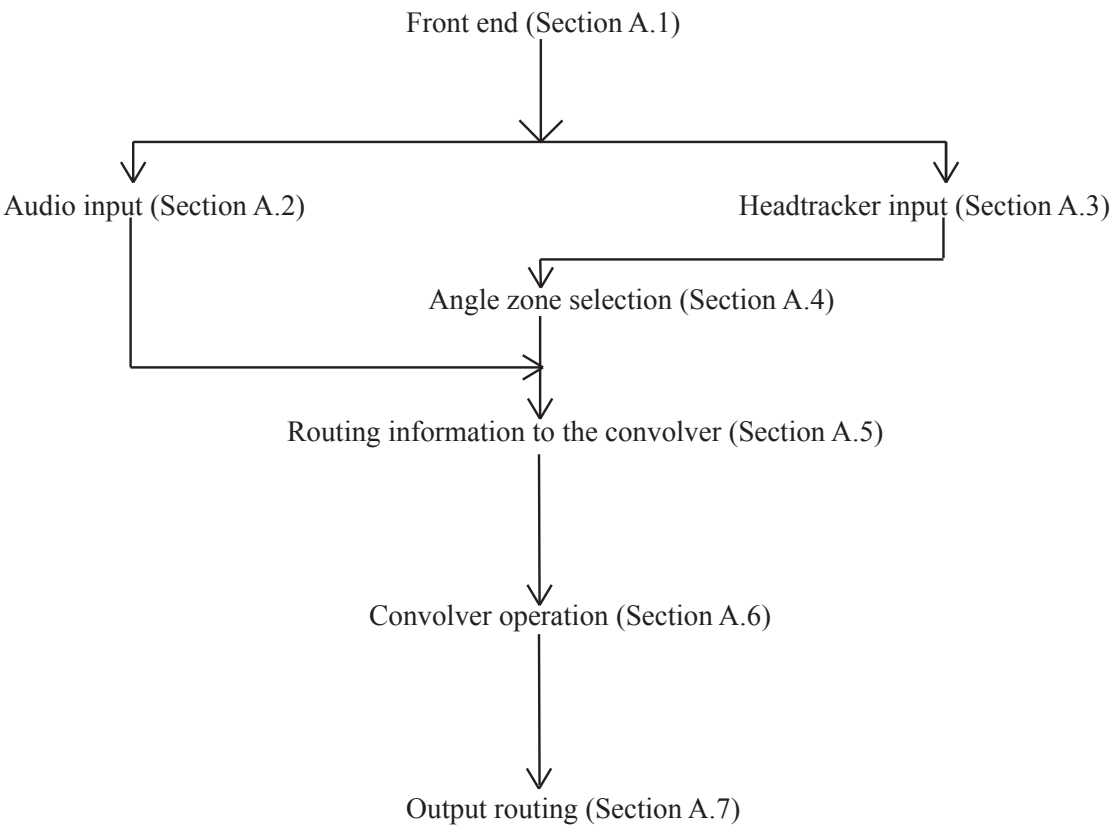


Figure A.1: The organization of the description of the Max/MSP application that is presented in this thesis.

A.1 Front-end

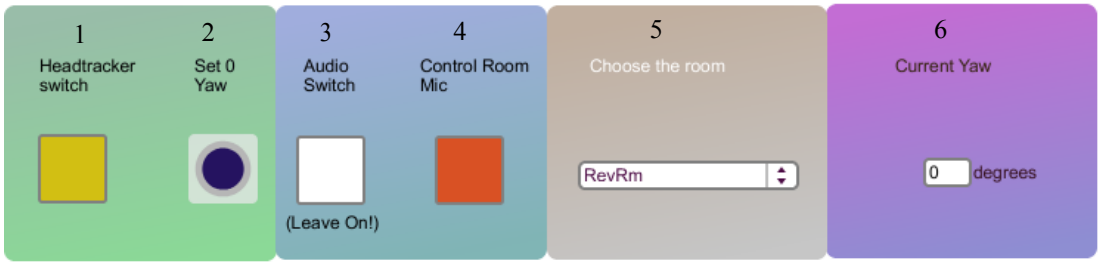


Figure A.2: The front-end of the Max/MSP application. Here, the numbering in the figure is used to illustrate the various parts, which are described in this section. Each of the four differently colored boxes has a separate function, also to be described in this section.

When a user (an experimenter) first opens the application, the front-end as seen in Figure A.2 is all that is visible. Max/MSP has two modes of viewing an application: patching mode and presentation mode. The view in Figure A.1 is in presentation mode, which is generally organized to present front-end graphical user interfaces (GUIs). Each part of the figure is labeled according to its function (the numbering is not part of the front-end), and is described as follows.

A.1.1 Headtracker switch

This is a toggle switch, which can be used to change the on/off state of the headtracker. The output of the switch is routed to a subpatcher (not visible to the user in the front-end) that interacts with the headtracking hardware through the computer's serial port, which will be described in the section A.3.

A.1.2 Set 0 yaw

When the headtracker switch described in section A.1.1 is turned on, this button can be used to calibrate the head position of the talking-listener (the participant in an experiment) in the simulation to correspond to a yaw angle of 0°. The output of the button is again routed to the subpatcher (not visible to the user in the front-end) that interacts with the headtracking hardware through the computer's serial port (same as section A.1.1), which will be described in the section A.3.

A.1.3 Audio switch

This is a toggle switch that is used to turn the audio input from the talking-listener (which is then convolved with an OBRIR: elaborated in section A.6) on/off. This switch

operates by turning the signal processing in Max/MSP on/off. As seen in the figure, this switch is also labeled with a '(Leave On!)' indicator, because the signal processing in Max/MSP ceases if this switch is turned off. The channel for the audio input can be assigned to any of the available channels, in a process detailed in section A.2.

A.1.4 Control room microphone

This is a toggle switch that can be used to switch on/off the signal from a microphone used by the experimenter in a control room. The output of the control room microphone is routed straight to the ear-loudspeakers that a talking-listener wears and acts as a communication channel between the experimenter and the experiment participant.

A.1.5 Choose the room

This drop-down menu contains a list of rooms that can be simulated through the application. The room selection is routed to the convolver, which will be described in section A.5.

A.1.6 Current Yaw

This display shows the current yaw angle of the talking-listener's head position. It gets this information from the headtracker output, which will be described in section A.3.

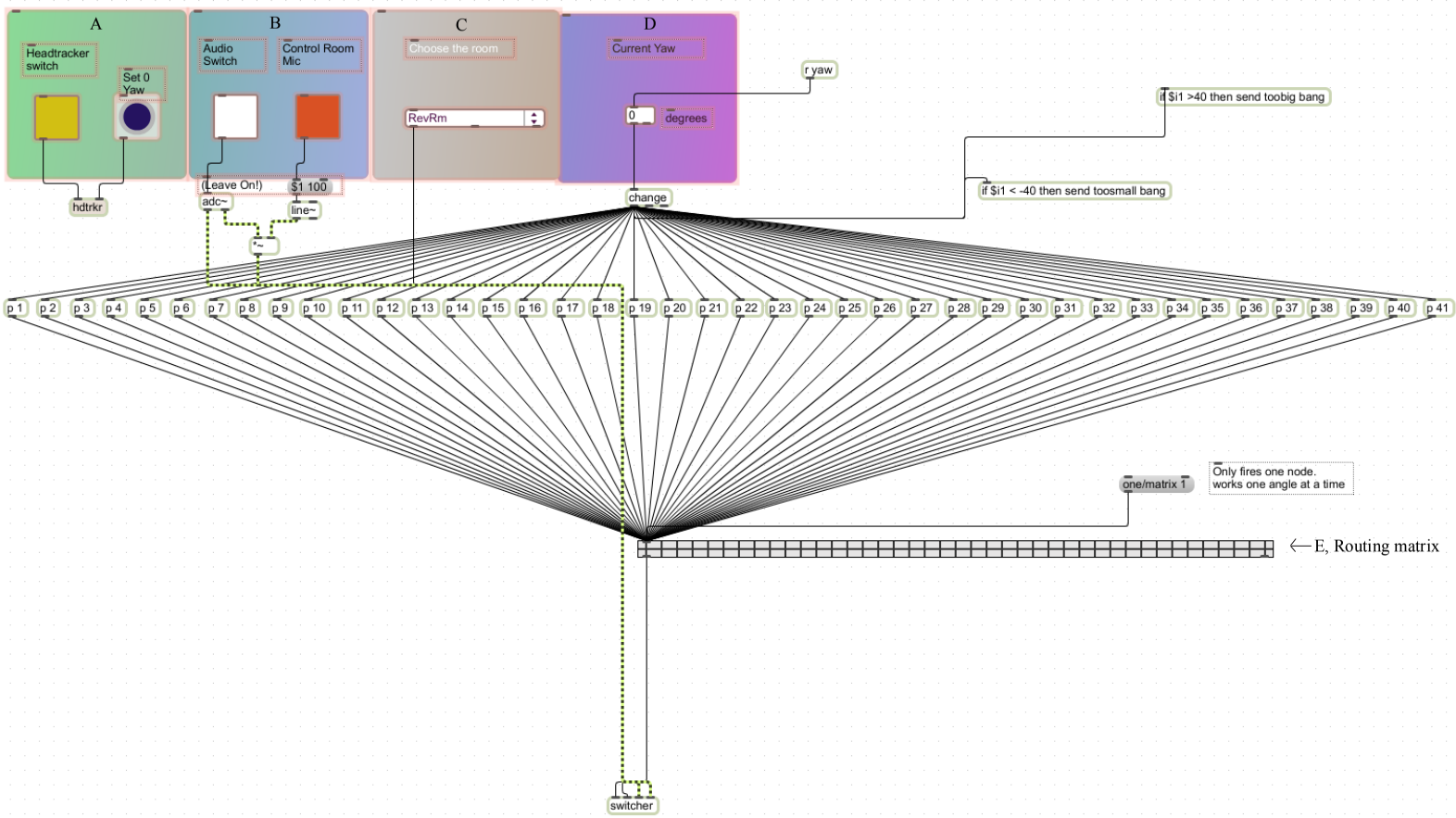


Figure A.2: The patching mode view of the patcher in Figure A.1

A.2 Audio input

Figure A.2 shows the second level of the Max/MSP patcher that is made visible by entering into the patching mode from the presentation mode of Figure A.1. This mode shows the various connections between the Max/MSP objects (Max/MSP objects are described in section 4.1.1 of chapter 4). In the current section, only the audio input is addressed, which consists of the objects in the box labeled B in Figure A.2 and the connections out of these objects.

A.2.1 Input audio routing

The state of the toggle switch A.1.3 (on, corresponding to 1; off corresponding to 0) is sent to an analog to digital converter object (adc~ in Figure A.2). In an on state, the adc~ object converts the analog signal from the microphone that the talker-listener wears, to a digital signal. Each adc~ object has two default outputs (corresponding to the first two output channels in the signal output matrix, unless specified otherwise). The left output from the adc~ object is sent to a subpatcher called 'switcher' in Figure A.2, which is described in section A.5, for further processing. The right output is sent to a signal multiplication object (*~ in Figure A.2), which is also sent to the same subpatcher (switcher) as above, and whose utility is described in the next section (section A.2.2).

In an off state, no signal processing is carried on in Max/MSP.

A.2.2 Control room microphone's audio routing

The state of the control room microphone's toggle switch (A.1.4) determines whether the microphone in the control room that the experimenter uses is in operation. In an on state, the audio signal from the control room microphone is sent to a ramped line~

object. The ramp is attained by the use of a message box (grey box with the text $1 100 in Figure A.2) with the purpose of implementing a fade-in and fade out. Here the $1 text is a placeholder for the signal coming from the A.1.4 toggle switch. An on state of A.1.4 switch sends a 1 to the message box (which sends a message '1 100' to the line~ object) and this corresponds to a fade-in from 0 to 1 in 100 ms for the control room microphone input. An off state of A.1.4 switch sends a 0 to the message box (which sends a message: '0 100' to the line~ object) and this corresponds to a fade-out from 1 to 0 in 100 ms for the control room microphone input. Overall, the output of the ramped line~ is controlled by the control room microphone switch's state and ensures that there are no artifacts when the control room microphone is turned on or off.

The signal from both the right output of the adc~ object (section A.2.1) and the ramped line~ object are sent to the multiplication object (*~) and sent to the 'switcher subpatcher (described in section A.5).

A.3 Headtracker input

As mentioned in section A.1, the signals from the headtracker switch (section A.1.1) and the button to calibrate the talking-listener's head-position (section A.1.2) are routed to the headtracking hardware. This section describes how this is accomplished in Max/MSP.

A.3.1 Headtracking interface

From the box labeled A in Figure A.2, it can be seen that the signals mentioned above are sent to the two inputs of the subpatcher 'hdtrkr', which is where Max/MSP

interacts with the headtracking hardware. The subpatcher ‘hdtrkr’ can be accessed in the application by double-clicking on the hdtrkr box, and is presented in Figure A.3.

The two boxes at the top of Figure A.3 labeled ‘i’ are depicting the inputs into this subpatcher from the patcher seen in Figure A.2. The first input regulates the on/off state of the headtracker through the toggle switch. The second input sets the current head-position of the talking-listener as the 0° yaw angle (it also sets the other degrees of freedom: pitch, roll, sway, heave, and surge to their 0, or reference, position).

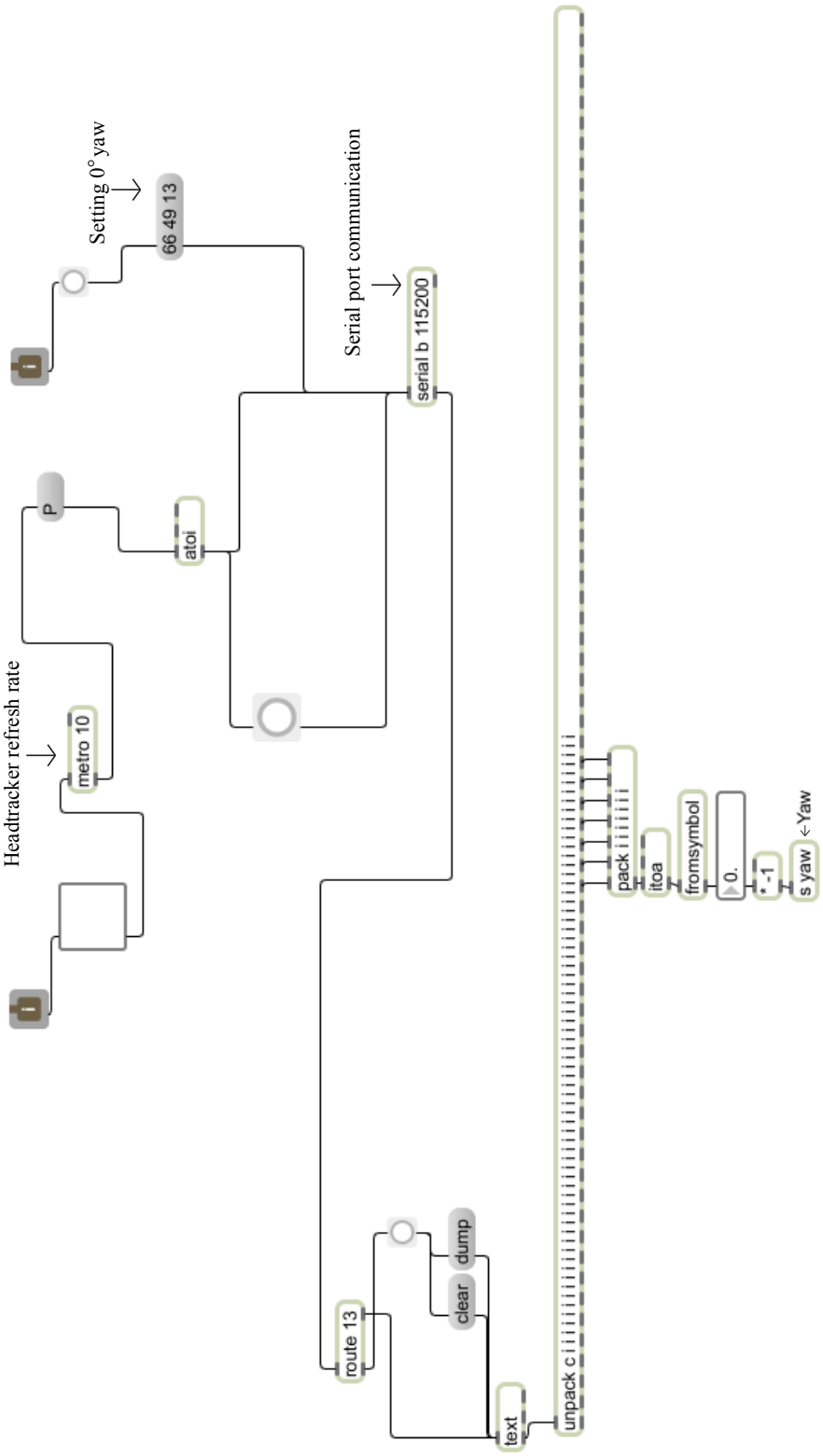


Figure A.3: The ‘hdtkr’ subpatcher

A.3.2 First input routing

Going further in the signal chain, with respect to the first input, the metro (metronome) object sends a signal at regular intervals, determined by value (in ms) of its first argument. In the present case, it is set to 10, which implies that the metro object is sending a bang signal (a trigger signal in Max/MSP) every 10 ms to the message box labeled 'P' (a system message for the Polhemus Fastrak® headtracking unit), where sending the message 'P' to the headtracker hardware through the serial object as seen in Figure A.3 returns the current head-position in all the six degrees of freedom to the max patcher, which can be further routed through the left output of the serial object. So essentially, the argument value of the metro object determines the refresh, or sampling, rate of the headtracker hardware. The first argument of the serial object is the serial port of the computer used (b, or second port in this case) for communicating with the headtracker and the second argument is the baud rate (115200 baud in this case) of the communication.

The output of the serial object is then appropriately changed to a degree format by a chain of objects (route, text, unpack, pack, itoa, fromsymbol, float and multiplication objects) and messages (clear and dump) to finally reach the send (s in Figure A.3) object. The send object in Max/MSP, when used in conjunction with the receive object can be used to route values without using any connections, essentially acting as a virtual cord within the same patcher or between patchers of the same Max/MSP application. The argument of the send object can be assigned to have a name (yaw here) and this same name can be used as the argument of the receive object to set up a virtual connection (send ➔ receive).

The other outputs of the unpack object, as seen in Figure A.3, could be used to extract information about the other degrees of freedom in a similar fashion, with the help of the Fastrak® manual to determine the headtracker output format.

A.3.3 Second input routing

The second input sends the message 66 49 13 to the headtracker hardware through the serial object. This message is interpreted in the headtracker hardware as a command to reset all the degrees of freedom to their 0 position which can be used to calibrate the talking-listener's head-position.

A.3.4 Output of hdtrkr

As discussed in section A.3.2, the output of the hdtrkr subpatcher is the yaw angle, in the form of the value in the send object, which can be received anywhere within the current application. In Figure 2, the receive object ('r yaw') is receiving the current yaw angle from the hdtrkr subpatcher, which is displayed in the box labeled 4. This yaw angle value is routed to a change object, which forwards the yaw value only when it is different from the last received value. This is done to eliminate the repetitions of the same yaw angle value being routed, which increases the computational load further down the signal chain.

A.4 Angle zone selection

As was mentioned in section 4.1.2 of chapter 4, the range of headtracking is 81° of yaw angles (-40° to +40°) in the horizontal plane with a 2° resolution. This leads to 41 spatial *zones*, with each *zone* covering 2°. Each such *zone* is implemented in Max/MSP as

a subpatcher, which are labeled from p 1 to p 41 in Figure A.2. When the headtracker output, as described in section A.3.4 falls within this range, it is routed to its corresponding subpatcher for further processing.

As an example, let us assume that the current yaw angle is -22°. This causes the change object mentioned in section A.3.4, to send a bang to the subpatcher labeled ‘p 10’ in Figure A.2. The contents of this subpatcher can be inspected by double clicking its box, and are shown in Figure A.4

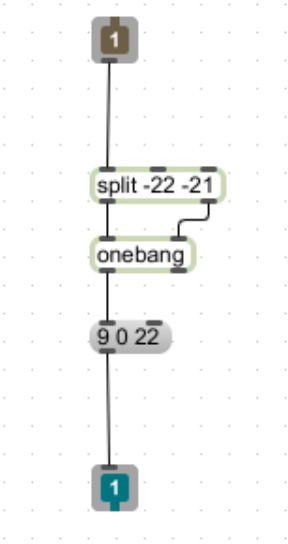


Figure A.4: p 10 subpatcher

In Figure A.4, the two boxes labeled 1 at the top and the bottom represent the input and output, to and from the subpatcher, respectively. This subpatcher sends a message ‘9 0 22’ the routing matrix (labeled as E in Figure A.2), which will be addressed in Section A.5. Similarly, the other subpatchers (p’s) are triggered by a bang when the yaw angle from the headtracker falls within the *zone* they represent, and send a message like the one seen in Figure A.4. There are two important points that must be noted here: a yaw angle of -21 will also trigger the ‘p 10’ subpatcher, as it lies in the *zone* covered by

this subpatcher; and at any instant more than one of the subpatchers out of p 1-p 41 can be active, all sending messages to the routing matrix (E in Figure A.2).

A.5 Routing information to the convolver

A.5.1 Routing matrix

As was mentioned at the end of section A.4, there can be more than one subpatcher sending messages simultaneously to the routing matrix. But the head-position a talking-listener can only be in one place at any instant. The routing matrix (E in Figure A.2) implements this condition (through the 'one/matrix 1' message in Figure A.2) by allowing only the message from the last 'p' subpatcher that was triggered to be routed to the 'switcher' subpatcher, which will be described in section A.5.2.

A.5.2 Switcher subpatcher

The 'switcher' subpatcher has the following four inputs:

(a) From the audio input of the talking-listener's microphone (section A.2.1)

(b) From the multiplication object that route the audio input from both the control room microphone and the talking-listener's microphone (section A.2.1)

(c) From the routing matrix (section A.5.1)

(d) From the 'choose the room' drop down menu (section A.1.5)

The contents of the switcher subpatcher are displayed in two formats: the whole subpatcher is presented in Figure A.5 and the subset of the subpatcher is presented in Figure A.6. Figure A.6 displays all the parts of the subpatcher that are relevant to the

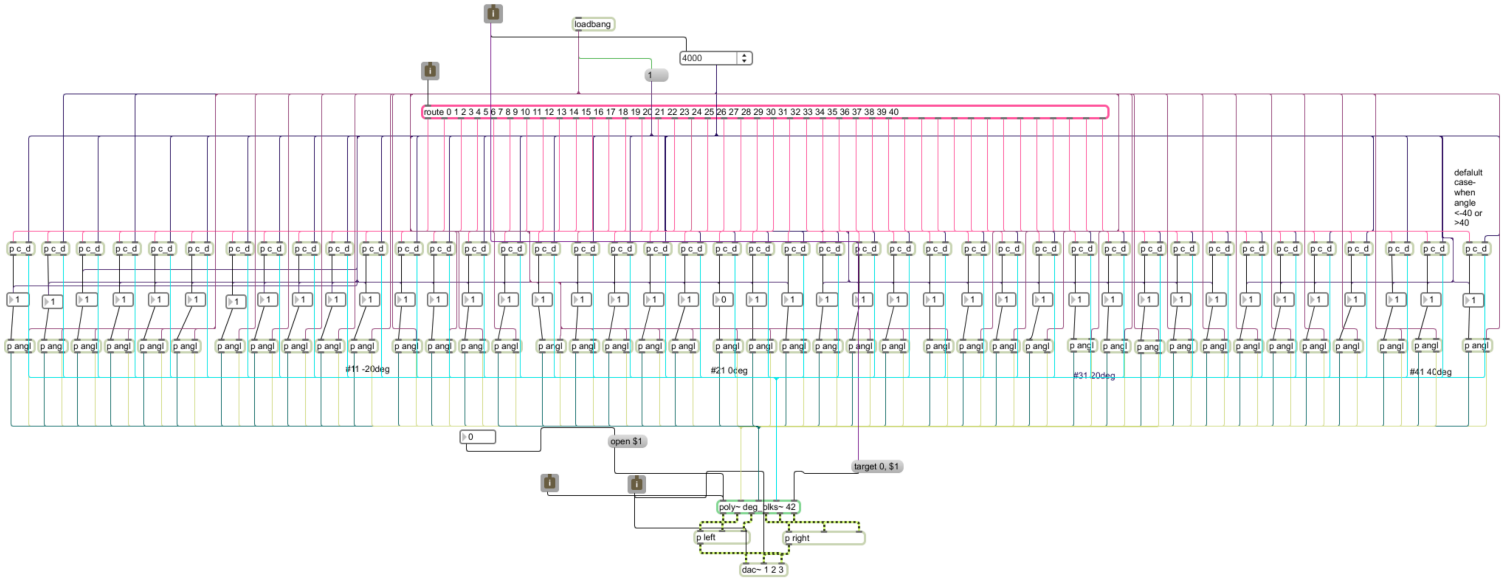


Figure A.5: The ‘switcher’ subpatcher

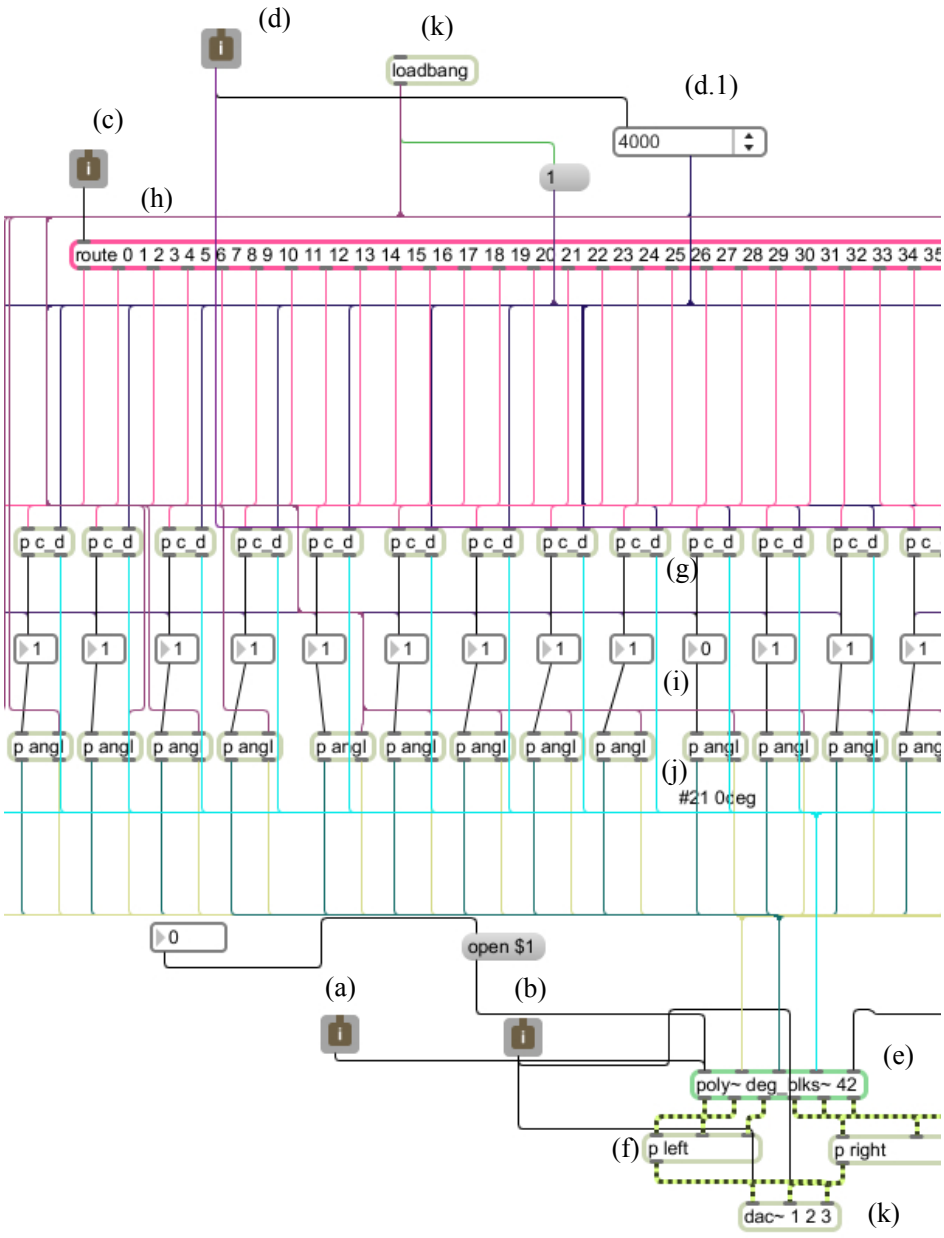


Figure A.6: A subset of the switcher subpatcher displayed in Figure A.5

inputs (a - d) above and the relevant objects in the switcher subpatcher.

The microphone input from the talking-listener (a) is sent directly to the poly~ object, while will be described in section A.6. The control room microphone input (b) is

sent to the two subpatchers labeled as ‘p left’ and ‘p right’, which are described in section A.7. The output from the routing matrix (c) described in section A.5.1 is routed to one out of all the subpatchers labeled ‘p c_d’, depending on the angle *zone* selected, as described in section A.4. The contents of the ‘p c_d’ subpatcher corresponding to the example given in Figure A.4, which in turn corresponds to the yaw angle of -22°, are displayed in Figure A.7.

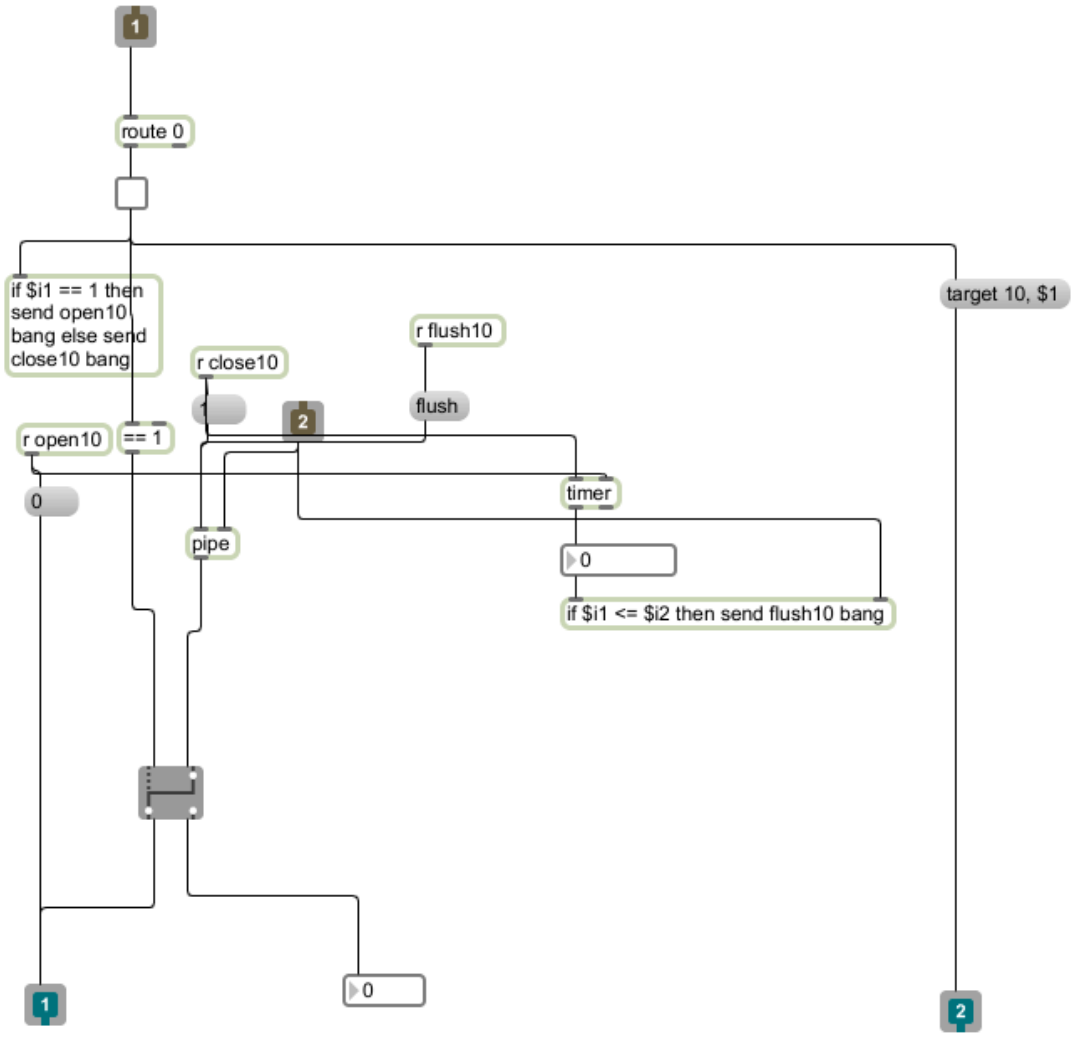


Figure A.7: The c_d subpatcher

This subpatcher receives two inputs: one from the route object labeled (h) in Figure A.6, which routes the angle *zone* opened (10 in this case); and the second from the menu object (d.1), which contains the reverberation time of the room selected, as described in section A.1.5. Without going into the details of the specific objects, the subpatcher can be briefly described as follows. It receives a 1 from the route object (h) in Figure A.6 when the *zone* 10 is selected, and receives a 0 when the *zone* 10 is not selected.

This subpatcher outputs two messages: the left output (green and grey box labeled 1 in Figure A.7) is a '0' message for the duration of the reverberation time of the room selected and a '1' message as soon as the reverberation time of the room is exhausted; and the right output (green and grey box labeled 2 in Figure A.7) is a 'target 10, $1' message, where $1 is a placeholder for a 0 or a 1, depending on the input from the route object (h) in Figure A.6, as described in the last paragraph.

The first output from this subpatcher is displayed in the number box corresponding to it ((i) in Figure A.7), which is routed to the subpatcher labeled 'p angl', described in the next paragraph. A 0 displayed in the number box implies that the signal processing in the corresponding angle *zone* is currently on and *vice-versa*. And depending on the reverberation time of the room being simulated and the velocity of head-movement of the talking-listener, more than one number boxes could be displaying a 0, which implies that the signal processing in more than one angle zones is on. The last two sentences are underlined, as they are one of the main features of this Max/MSP application, which is designed to simulate the room acoustical effects of head-movements

in real rooms as discussed in section 4.1 of chapter 4. The second output is routed to the fourth input of the poly~ object ((e) in Figure A.7, which will be described in section A.6.

Figure A.8 displays the contents of the ‘p angle’ subpatcher for the 0° yaw angle (labeled (j) in Figure A.6). There are two inputs (the brown and grey boxes labeled 1 and 2): the first input is from the corresponding number box ((i) in Figure 6) carrying either a 0 (*zone* open state) or a 1 (*zone* close state) to the message box labeled ‘target 21, $1’, where $1 is placeholder for the input from input 1; the second input is a bang signal from the loadbang object (labeled (k) in Figure A.6), which

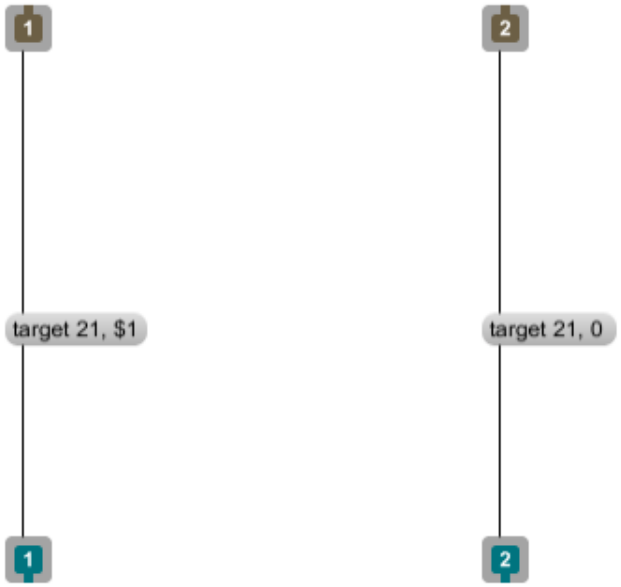


Figure A.8: The ‘angl’ subpatcher corresponding to the 0° yaw angle (j) in Figure A.6

sends a bang the first time the switcher subpatcher is activated with an input. The loadbang object also initializes all the number boxes in Figure A.6 with a value of 1 which implies that all the angle *zones* are closed, except the one that receives a 0 value from the input (c) of Figure A.6.

There are two outputs from the ‘angl’ subpatcher: the first output is the ‘target 21, $1’ message, which is sent to the third input of the poly~ object, which will be described in section A.7; the second output is the message ‘target 21, 0’, which is sent to the second input of the poly~ object, which will be described in section A.7.

As the title of this section suggests, it only describes the routing of signals that act as the inputs to the convolver, which is hosted in the poly~ object (labeled (k) in Figure A.6). The next section describes the operation of the poly~ object and the convolution process.

A.6 Convolver operation

The poly~ object (labeled (e) in figure A.6) is a Max/MSP object that can encapsulate many instances of the same patcher, like the instances of ‘c_d’ and ‘angl’ patchers in Figure A.6. In a sense it is a master patcher object, which encapsulates many subpatchers, and the routing to and from the subpatchers can be organized at the level of this master patcher. The main advantage of the poly~ object from the point of view of the current application is that the inputs can be routed selectively to one subpatcher, while the others are either still undergoing processing (corresponding to the reverberation time of the OBRIR) from when they were initiated, or are muted. This can be likened to playing on a piano, where the player is hitting a note every second, while letting the previously played notes fade away naturally, assuming the notes have a decay time of more than 1 sec here. In fact, the name of the poly~ object is an abbreviation of polyphony, and it is generally employed for implementing polyphony in Max/MSP applications.

The first argument of the poly~ object is the name of the encapsulated patchers, and the poly~ object is also referred by the same name. The second argument refers to the number of instances of the subpatchers. Looking at Figure A.6, the poly~ object (labeled (e)) has 42 instances of the deg_blks~ subpatchers (adapting a different terminology here from Max/MSP, where these are referred to as patchers). Figure A.9 shows the contents of one of the 'deg_blks~' subpatchers. In order to describe the contents of the subpatcher, it is necessary to understand the functions that the poly~ object performs.

A summary of the functionality of the subpatchers associated with the poly~ object in the present Max/MSP application can be found it section 4.1.2 of Chapter 4. This section describes how that functionality is achieved in Max/MSP using the SIR2 plugin for real-time convolutions. Here, instead of explaining the function of each Max/MSP object, various objects are grouped together with respect to their collective functionality.

Based on the summary in section 4.1.2, and the discussion above with respect to the 'switcher' subpatcher, the functioning of the poly~ object will be described with the flowchart in Figure A.10, which is adapted from the flowchart in Figure 4.2 but redefined using the terms of the current section, and Figure A.9.

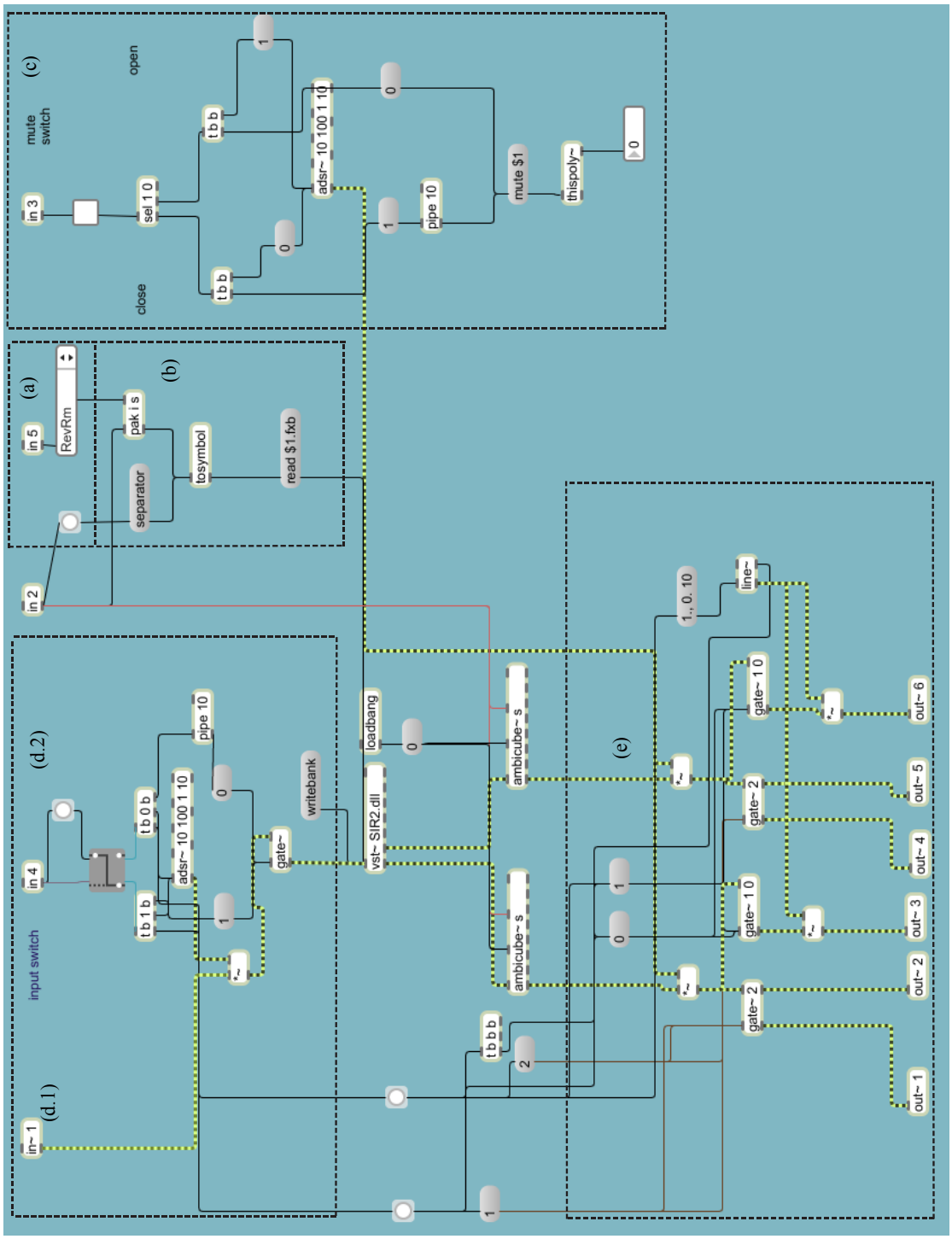


Figure A.9: One of the 42 ‘deg_blks~’ subpatchers

A.6.1 First step

The first step of the flowchart is accomplished in the dashed box labeled (a) in Figure A.9, where the name of the room to be simulated is input from the fifth input to the poly~ object, which in turn is input from the fourth input to the 'switcher' subpatcher that is described in section A.5.2 (labeled (d) in Figure A.6).

A.6.2 Second step

For the second step, the objects in the box labeled (b), which includes box (a), create messages that are sent to the convolver. The objects are triggered by the second input to the poly~ object, which in turn is the second output of the 'angl' subpatcher that is described in section A.5.2 and Figure A.8. These message are of the format 'read $1.fxb', where $1 contains the location of the VST settings bank on the hard drive of the computer. These setting banks are preconfigured, one per angle *zone*, to contain the OBRIR and the convolver settings for the corresponding angle *zone*. In this step, each subpatcher gets loaded with the corresponding bank when the 'switcher' subpatcher is triggered with a bang from the loadbang object.

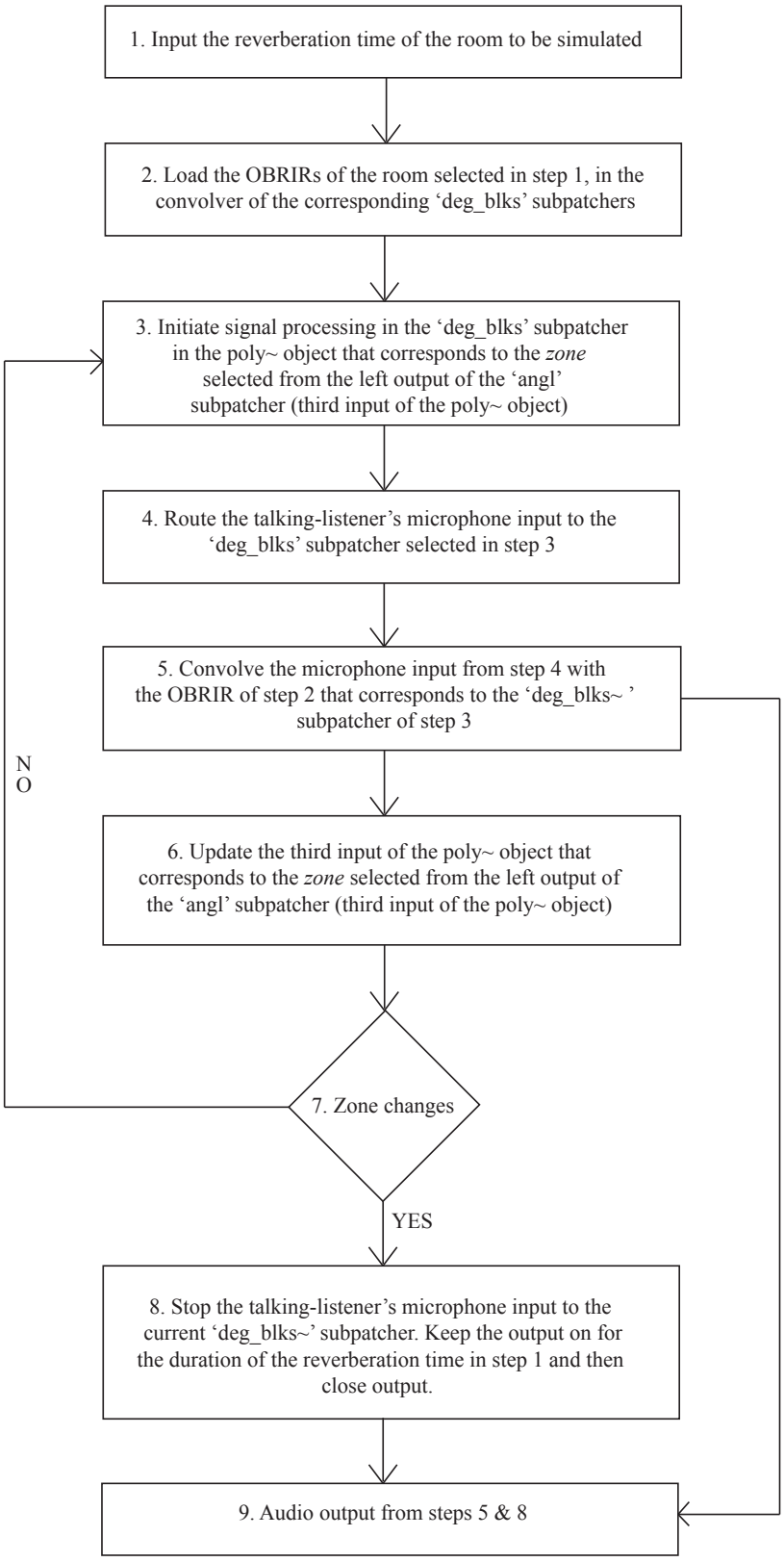


Figure A.10: Flowchart of the operation of the poly~ object

A.6.3 Third step

As soon as an angle *zone* is selected, as described in section A.5.2 and Figure A.7, this information (0, which signifies open) is routed from the right output of the corresponding ‘c_d’ subpatcher to the fourth input of the poly~ object. The left output of the ‘c_d’ subpatcher also sends the same information to the left input of the corresponding ‘angl’ subpatcher, which is routed to the third input of the poly~ subpatcher. These two signals are routed within the poly~ object to the corresponding ‘deg_blks~’ subpatcher. The third input to the corresponding ‘deg_blks~’ subpatcher releases its mute button, as seen in the box labeled (c) in Figure A.9, to start the signal processing in this subpatcher.

A.6.4 Fourth step

The signal processing initiated in the last step causes the input from the talking-listener’s microphone, labeled (d.1) in Figure A.9, being routed to this unmuted ‘deg_blks~’ subpatcher. At the same time, the fourth input to this ‘deg_blks~’ subpatcher, labeled (d.2) in Figure A.9, triggers a ramp from 0 to 1 for 10 ms. This ramp acts as a fade-in for the microphone input.

A.6.5 Fifth step

This step outputs the signal, which results from the convolution of the microphone input from the last step, with the OBRIR (section A.6.2) of the unmuted ‘deg_blks~’ subpatcher of the third step (section A.6.3). Before preceding any further, it

is necessary to briefly describe the functioning of the SIR2 convolver. Figure A.11 shows the SIR2 real-time convolver used, which is loaded with the OBRIR corresponding to the 0° angle *zone* of one of the rooms measured (Recording room at the Faculty of Architecture, Design and Planning, The University of Sydney).

Figure A.11: The SIR2 convolver loaded with an OBRIR in the 'deg_blks~' subpatcher in Max/MSP

In Figure A.11, the OBRIR (only one channel visible here) is visible in the region labeled (a), with the number of samples (11814) per channel (2) and the sampling rate (48 KHz) of the host (Max/MSP). The 'neutral' at the bottom right indicates that the OBRIR is not undergoing ant filtering in the convolver, which can, if required, perform linear phase filtering with the settings in the box labeled 'EQ'. The signal routing represented in

the region labeled (b) indicates that the mono input signal is being convolved with the stereo impulse response (OBRIR in this case). The settings in region (c) indicate that the OBRIR is not being altered in any form within the convolver. The region (d) is displaying the various OBRIR files that would be loaded in their corresponding 'deg_blks~' subpatcher, with the one highlighted loaded in the current subpatcher. Finally region (e) displays the level of the convolved output, with a slider than can be used to change the level, if required.

To determine the accuracy of the convolutions that SIR2 performs, its output was compared with the output in MATLAB, using the same impulse response and input signal. Figure A.12 shows that there is similarity in the fine temporal structure of the first 1000 samples of the two convolved signals. The two signals also have a correlation coefficient of very close to 1 (0.99998), which indicates that the convolution performed in SIR2 is very close to the convolution performed in MATLAB. When the input signal is convolved with the OBRIR in SIR2, the output is routed in real time to the two outputs of the poly~ object, which will be described in section A.7.

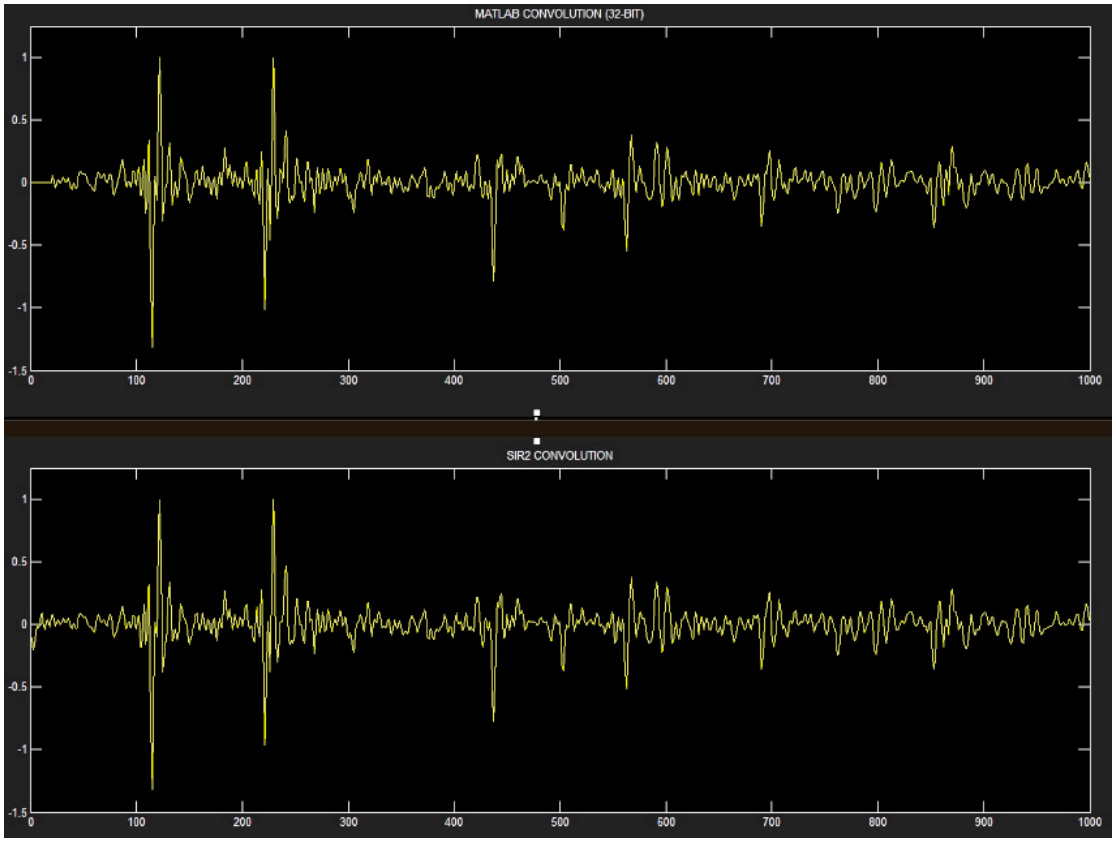


Figure A.12: Comparison of convolution performed in MATLAB and SIR2 using the same signals.

A.6.6 Sixth step

This step marks the beginning of a possible change in the routing of the talking-listener's microphone input. Here, the fourth input to the poly~ object, which is routed to the 'deg_blks~' subpatcher receiving the microphone input, is updated as described in section A.5.2 and Figure A.7 with the message from the right output of the corresponding 'c_d' subpatcher.

A.6.7 Seventh step

This decision box is fed with the update from the last section. There are two cases:

(i) The message from the right output of the ‘c_d’ subpatcher remains the same.

(ii) The message from the right output the ‘c_d’ subpatcher changes.

In the first case, the signal processing status in the corresponding ‘deg_blks~’ subpatcher remains the same and the flow reverts back to the third step in Figure A.12. In the second case, the flow is forwarded to step 8, which is described in the next section.

A.6.8 Eighth step

In this step, there is a change in the first, third and fourth input to the ‘deg_blks~’ subpatcher, which is the opposite of what happens in the third step (section A.6.3). The fourth input causes a ramp (labeled (d.2) in Figure A.9) from 1 to 0 in 10 ms, which acts as a fade-out for the microphone input from the first input. The third input, which controls the mute switch for the subpatcher, now performs two tasks, which are delayed by the duration of the reverberation time of room currently selected. The first task involves routing a fade-out ramp of from 1 to 0 in 1 ms to the convolver output. The second task involves muting the output from the current subpatcher. These two tasks enable the residual audio from the convolver to be output for the duration of the reverberation time of the room, once the input to the convolver is cut, which will be explained in section A.6.4. It must be noted here that, as mentioned in section 4.1.2 of Chapter 4, more than one ‘deg_blks~’ subpatcher could be undergoing the processes of this step. This means that there could be more than one convolver that is outputting a

stereo signal corresponding to its OBRIR. The management of these outputs along with the output from the 'deg_blks~' subpatcher that is receiving microphone input from the talking-listener is the addressed in the next section.

A.6.9 Ninth step

The handling of the output in the 'deg_blks~' subpatchers is displayed in Figure A.9 in the region labeled (e). There are three outputs per channel (2 channels in total for the left and right ear), in the form of Max/MSP out~ objects, leading to a total of six outputs from the poly~ object. The following describes the output routing for only one channel, which is the same for the other channel.

The output from the object labeled 'out~ 1' is active when the corresponding 'deg_blks~' subpatcher is receiving the input from the talking-listener's microphone. As soon as the microphone input is routed to another 'deg_blks~' subpatcher, as described in section A.6.6-A.6.8, the output switches to the object labeled 'out~2' becomes active, which is faded out with a ramp of 1 to 0 in 10 ms. And 10 ms before the subpatcher is to be muted, the output switches to 'out~ 3', which is again faded out with a ramp from 1 to 0 in these 10 ms. The reason for switching the outputs in this way is to eliminate any artifacts that might accompany if the outputs were abruptly cut off.

A.7 Output routing

This is the final step in the description of the software module of the room acoustical simulation system. The three outputs per channel for all the active 'deg_blks~' subpatcher, as described in section A.6, are routed from the poly~ object to the respective subpatcher: 'p left' and ' p right'. The 'p left' subpatcher ('p right' subpatcher has the

same routing structure as ‘p left’) is displayed in Figure A.13, where the three inputs are combined into a single signal, that is output to the first input of the dac~ object, labeled (k) in the ‘switcher’ subpatcher displayed in Figure A.6.

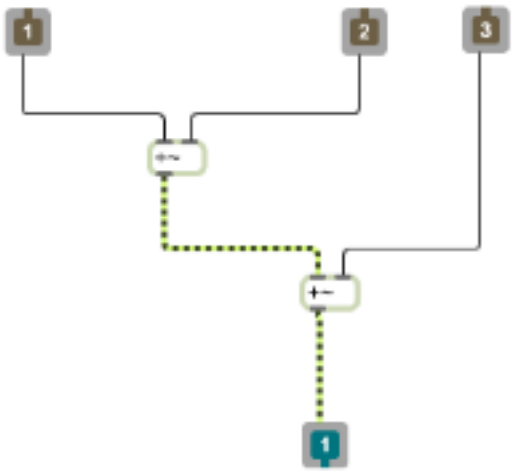


Figure A.13: The ‘p left’ subpatcher

The first two inputs of the dac~ object (converts signals from the digital to analog domain) represents the left and right channels that are output to the ear-loudspeakers that the talking-listener wears, as described in section 4.2 of chapter 4. The third input to the dac~ object is the output from both the left and right channels, which is routed to a monitoring channel for the experimenter through the routing matrix of the A/D converter.

Appendix B

Description of the simulated room conditions

This appendix contains information about the room conditions that were simulated with the room acoustical simulation system described in this thesis. The simulated room conditions were used in the case studies that are described in Chapters 5 and 6 of this thesis. The contents of this appendix complement the information presented in those chapters, and consulting both (the appendix and the chapters).

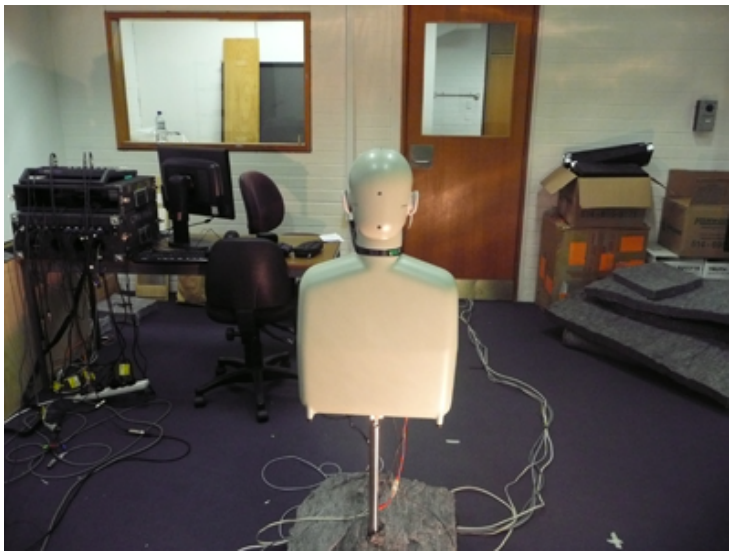
All the room conditions, except one (room condition 6), are in the building of the Faculty of Architecture, Design and Planning at the University of Sydney. Here, each room condition has the same number with which it are referred to in Chapters 5 and 6, and it is accompanied by its volume and mid-frequency reverberation time in brackets. The description of each room condition is supplemented with a figure with three subfigures which are labeled as (a), (b) and (c). The first subfigure (a) is a photograph of the measurement apparatus in the room and the second (b) the floor plan of the room, which also shows the location and orientation of the apparatus. The third (c) is the envelope of the first 25 ms of the OBRIR (L and R here refer to the left and right ears) as a function of time and angle of rotation of the apparatus, where the color black in the subfigure represents an instantaneous sound level 100 dB less than the white.

B.1 Room condition 1 (125 m^3; 0.6 s)

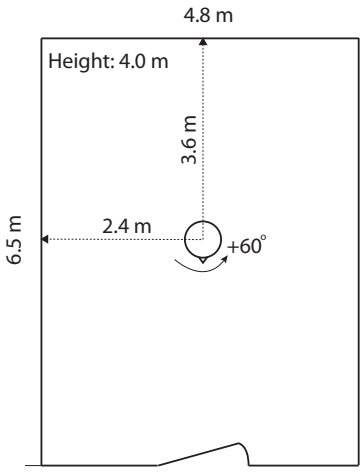
This room, called the Listening room, is rectangular in shape and had a disproportionately high ceiling (4 m) at the time of the OBRIR measurement. The main sources of reflection are the wall with the door, as seen in Figure B.1, when the HATS is facing it in and around its 0° orientation; the side walls; and some furniture around the measurement apparatus.

The first wall reflection arrives 12.8 ms after the direct sound. The ceiling reflection is seen almost 16 ms after the direct sound, while the reflections from the furniture are weak.

(a)



(b)



(c)

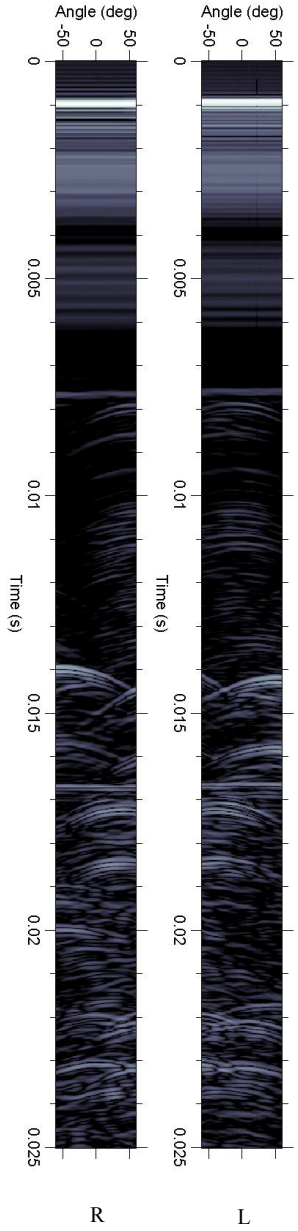


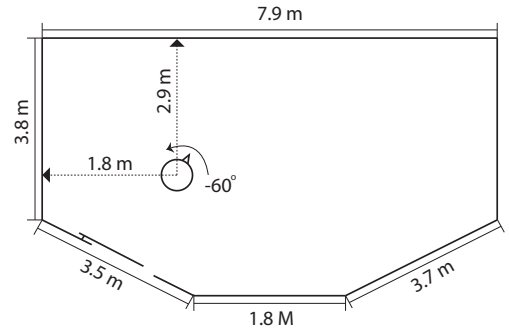
Figure B.1: Listening room

B.2 Room condition 2 (152 m^3; 0.35 s)

This room, called the Control room, is part of a Recording studio and has a relatively high broadband sound absorption for its volume. This is due to the acoustic treatment on its rear and side walls, which also yields diffuse reflections as seen in the relatively high early reflection density in the subfigure (c) of Figure B.2. The doors and the remainder of the room's walls are not treated for absorption and diffusion. The first major non-floor reflections (around 7.8 ms after the direct sound) come from the glass doors behind the HATS. There are also some minor reflections from the console furniture to the right of the HATS.

(a)

(b)



(c)

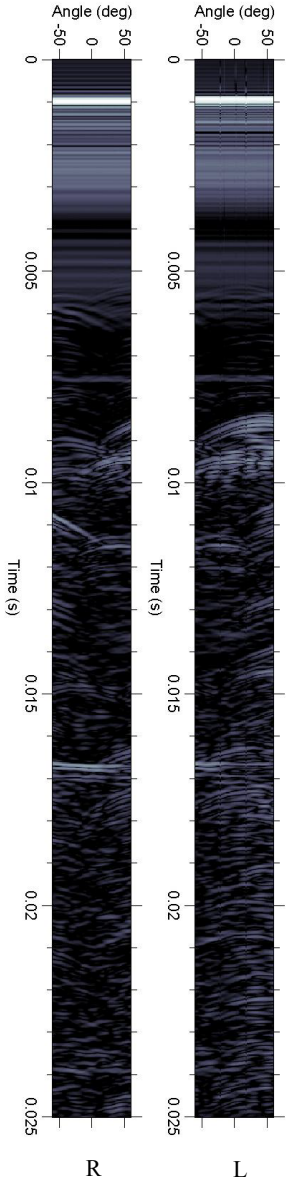


Figure B.2: Control room

B.3 Room condition 3 (170 m^3; 0.4 s)

This room, called the Recording room, is also part of the same recording studio that room condition 2 is a part of. It is characterized by a sloping ceiling, with absorptive and diffusive treatment on many of the walls. The main sources of reflection here is the nearby wall, which has no acoustic treatment.

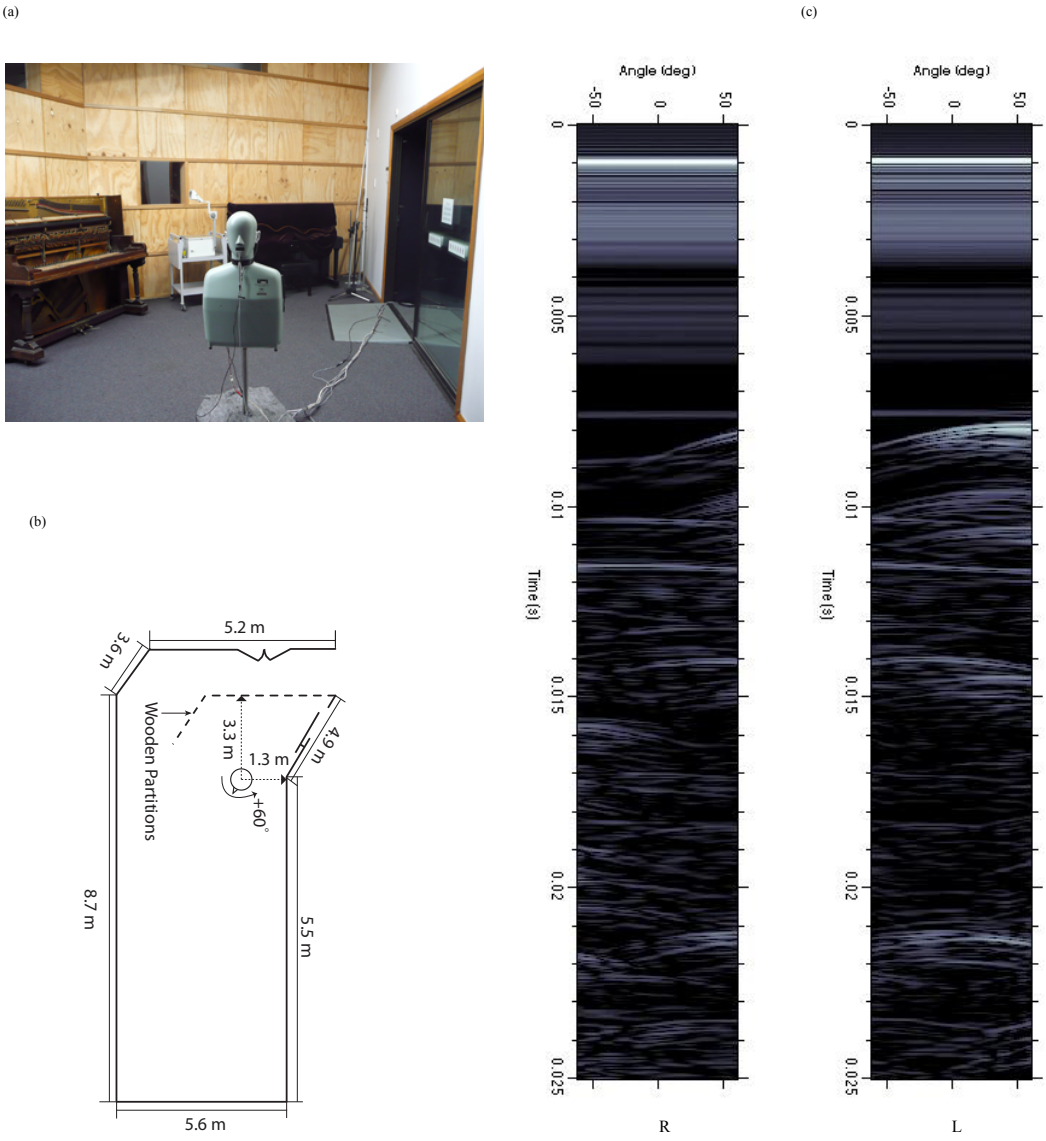


Figure B.3: Recording room

B.4 Room condition 4 (188 m^3; 0.9 s)

This room, called the Photometric laboratory, contains various photometric measurement devices, as seen in Figure B.4 (a). All the surfaces of the rooms are hard, and there is a lot of equipment near the walls and in the shelves near the HATS. These features are represented in the subfigure (c) of Figure B.4, with a relatively high reflection density.

This room also has a black curtain on one off the walls, as seen in Figure B.4 (a), which was drawn for another set of OBRIR measurement in this room, which was used as a control in the experiments mentioned in Chapters 5 and 6. Since the two conditions are very similar, the control condition is not described here.

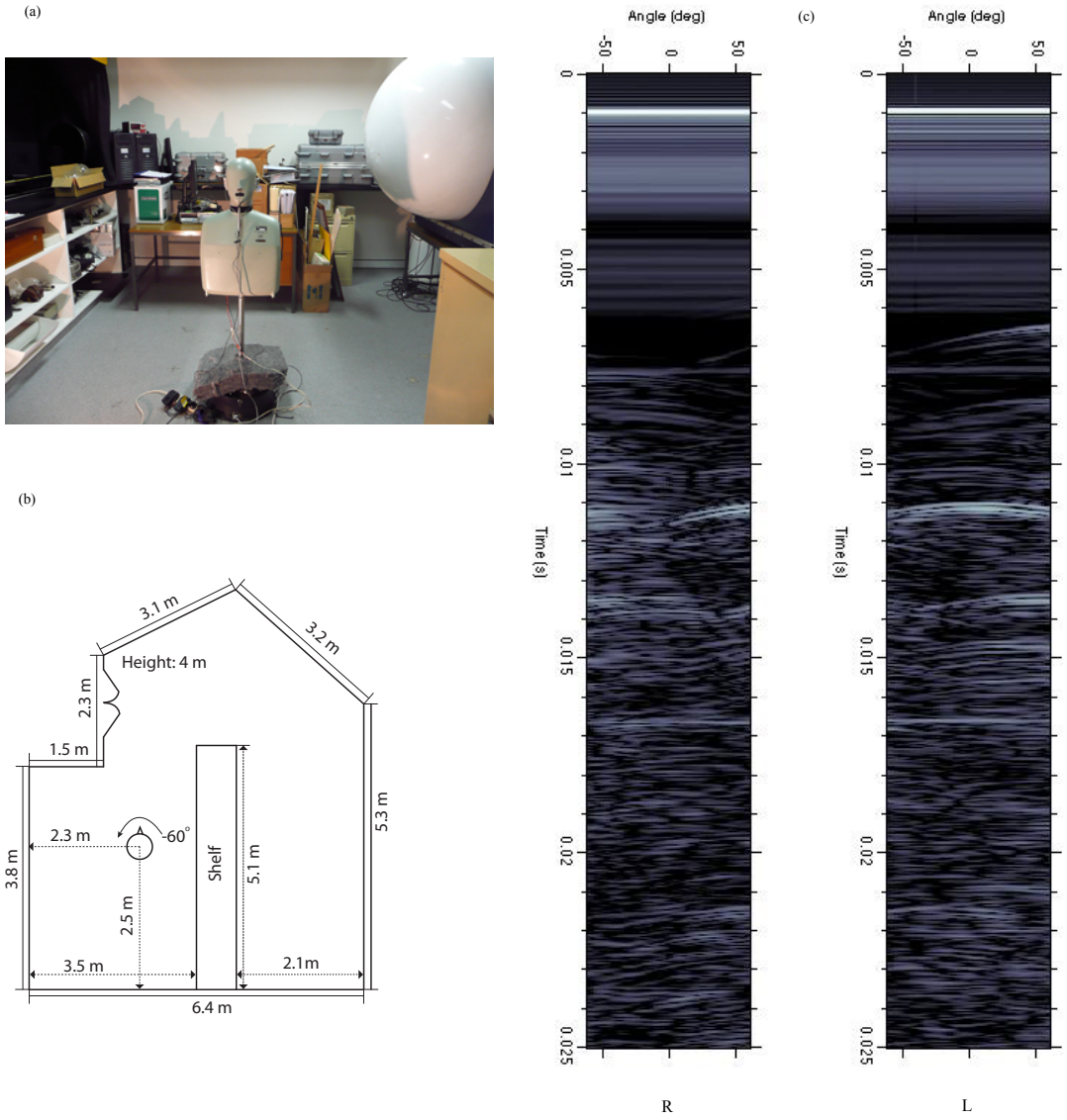
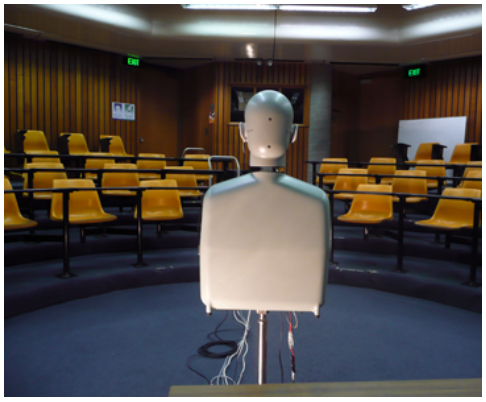


Figure B.3: Recording room

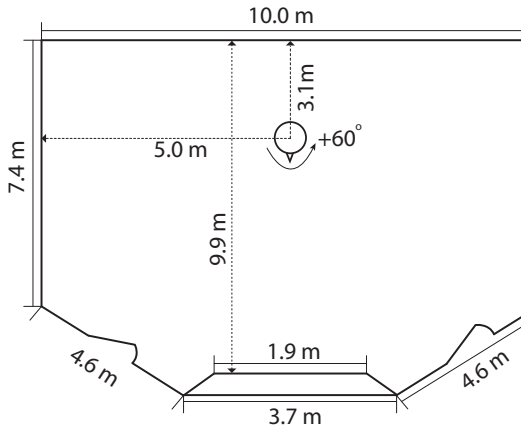
B.5 Room condition 5 (310 m^3; 0.5 s)

This room, called Lecture theatre 2, can seat 70 people. The seats in the lecture theatre are arranged in raked concentric circles, as seen in Figure B.5 (a). The HATS was positioned in the middle of the front of this room, where a lecturer is likely to stand while addressing the class. Here the first main reflection after the floor reflection is from the wall behind.

(a)



(b)



(c)

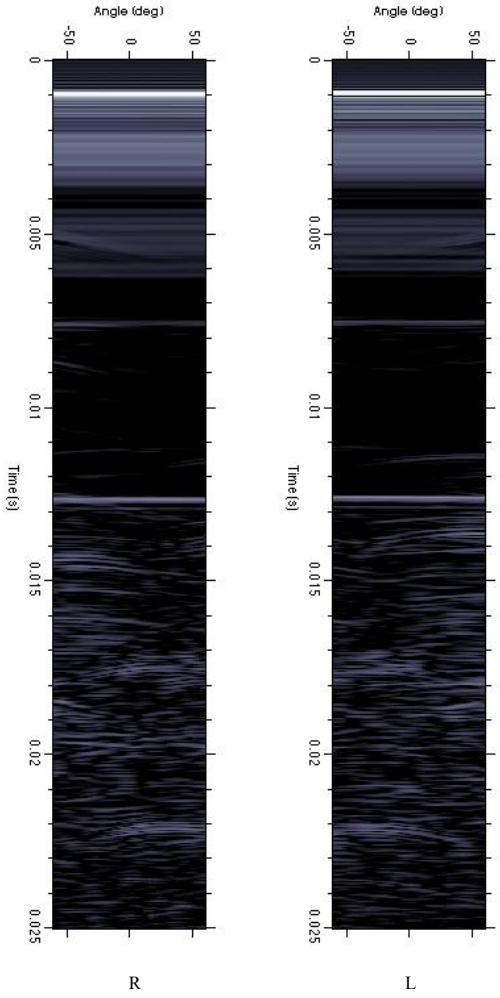


Figure B.5 Lecture Theater 2

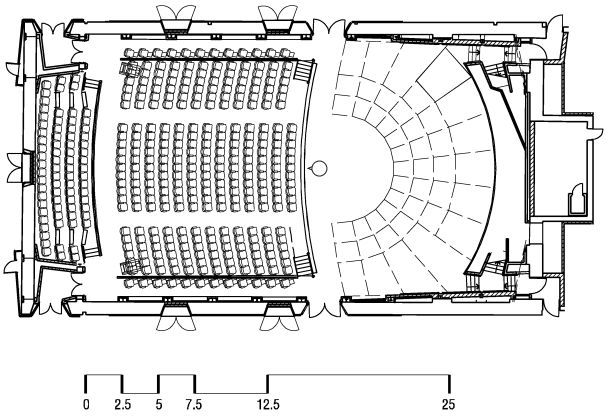
B.6 Room condition 6 (7650 m^3; 1.7 s)

The room, called Verbrugghen Hall, is a music auditorium within the Conservatorium of Music at the University of Sydney. It has a large stage area and the HATS was positioned downstage, halfway across the stage width. The HATS's transducer height here was 1.5 m from the floor, unlike 1.2 m (sitting height) in the other room conditions. This was due to the consideration that a singer or someone delivering a speech is unlikely to do so while sitting. Apart from being the largest of all the room mentioned here, this room also had the most absorption.

The OBRIRs that were collected from room had sufficient signal-to-noise ratio, which shows the efficacy of the OBRIR measurement method for large rooms.

(a)

(b)



(c)

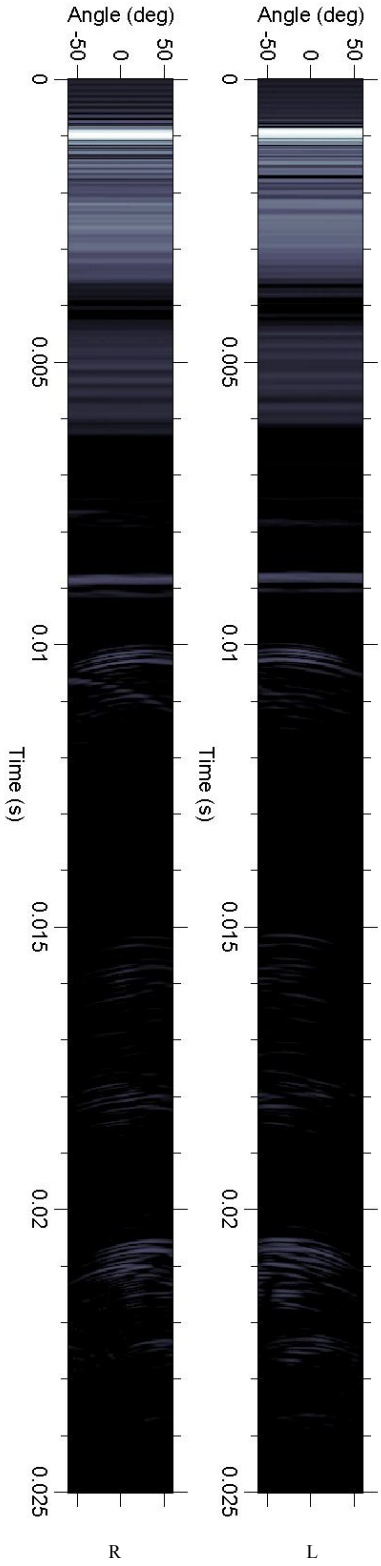


Figure B.6: Verbrugghen Hall

Appendix C

Review of existing room acoustic simulation systems

This appendix discusses a selection of existing room acoustic simulation systems that are classified as auditory virtual environment (AVE) generators by their authors and experiments conducted in-situ, with or without a virtual (or computer generated) component. In Chapter 4, it was suggested that the system presented in this thesis could be more appropriately classified as a mixed-reality system. This is due to the observation that the current system has a real component: the talking-listener's own voice; and a virtual component: the presentation of simulated room environments, as illustrated in Figure C.1, where the real and virtual components act are seamlessly integrated to create a sense of immersion in the mixed-reality environment.

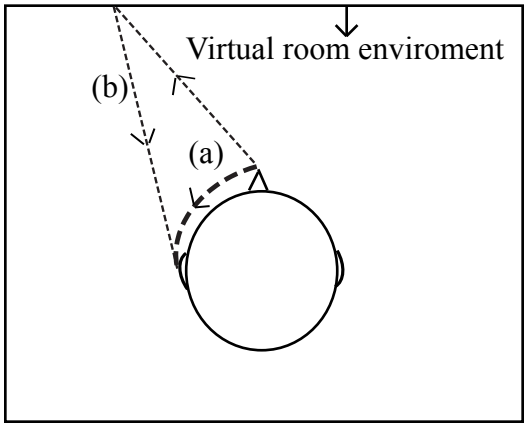


Figure C.1: A talking-listener in a mixed-reality environment. Here (a) is the real component comprising of the talking-listener's own voice; and (b) is the virtual component, simulated with the system described in the thesis, for the representation of the room reflections from one's own voice (component (a)).

Some of the systems that are described in what follows can also be classified as mixed-reality systems but instead of discussing the differences in classification, this appendix focuses on the various approaches that have been adopted in the previous system to implement the room acoustic simulation.

A review of room acoustic simulation systems can focus on the different methods to measure the room impulse responses (RIRs) for a certain source-receiver configuration (for such a review, refer to Novo (2007)) however, it is more informative to discuss the various reproduction formats being employed by the existing systems and how the real-time convolution is implemented: to address the salient features of the current system in the respective contexts. Table 1 provides the summary of the systems to be described in this appendix.

VAE System	**Reproduction format**	**Room measurement**
Pelegrín-García *et al.* (2011a)	Customized earphones	Synthesized RIR
Pörschmann and Pellegrini (2010)	Customized headphones	Software based RIR
Ueno and Tachibana (2010)	Loudspeakers	Measured real rooms
LoRA (2010)	Loudspeakers	Software based RIR
Wefers and Vorländer (2010)	Headphones/Loudspeakers	Software based RIR
Current system	Headphones (ear-loudspeakers)	Measured real rooms

Table1: The various VAE systems discussed above on the basis of the reproduction format used.

C.1 First system

The first system in this review, besides being the most recent, is also the closest to the current system in terms of how the convolution is implemented and the microphone

positioning used. The system described by Pelegrín-García *et al.* (2011a) is implemented by hosting a convolver (jconvolver) for Linux platforms. In the complete set-up, a talking-listener's voice is picked up by a DPA 4066 microphone (same as the one used in the system described in this thesis; see Chapter 4, section 4.2 for more detail) and the room reflections (they used synthetic impulse responses in their study) are reproduced on a pair of custom designed earphones, with the required AD/DA conversion.

The microphone is positioned at a distance of 5 cm from the edge of lips of a talking-listener on the right side, which is similar to the microphone positioning of 7 cm from the center of the lips in the system described in this thesis. The earphones are designed to not affect the direct and bone conducted pathways significantly, and offer a relatively cheap solution compared to the ear-loudspeakers used in the system described in thesis: which are expensive. It is not clear whether there is any difference in the two reproduction formats at this point.

The convolution operation in their set-up has been reported to have a latency of 11.5 ms at a sampling rate of 44.1 kHz, though it is not clearly specified what the electroacoustic latency of the system is, when the AD/DA conversion is included.

Currently, there is also no provision for headtracking using BRS, which limits this system's capabilities in terms of designing experiments that requires the auditory image to change in accordance with the talking-listener's head movements.

C.2 Second system

One of the determining factors in choosing a reproduction format, apart from the ones native to the design methodology employed, can be the intended use of the system, e.g. a system catering to mobile communication devices could benefit more from a

headphone based reproduction format. An example of such a system is the one described by Pörschmann (2007), which was further elaborated by Pörschmann and Pellegrini (2010). In the latter system, the closest to the current system in terms of its focus on ones' own voice and BRS capabilities, the authors have divided the system design into two components. The division enables them to auralize a simulated room environment based on a head-tracked 'Binaural Room Scanning' (BRS) module and a 'Perception of the Own Voice' (POV) module.

The BRS module is an extension of the system described in Mackensen *et al.* (2000), which enables the reproduction of the simulated room reflections from 5 virtual sound sources, where each sound source can perform convolution with impulse response of upto 80 ms. The simulated room reflections change in accordance with the talking-listener's head movements, and its incorporation is shown to provide a higher degree of 'presence' experienced by the users (Pörschmann and Pellegrini, 2010). Here, the BRS module does not address the direct sound (done in the POV module) and the sound from this module is filtered to reproduce the room reflected component of the talking-listener's voice from the sound sources which are located in his/her far-field, as described in detail by Pörschmann (2001).

The POV module is a reproduction the direct sound component of the talking-listener's own voice in his/her near-field, which is picked up by a microphone at a distance of 40 mm from the lips. As the talking-listener wears closed headphones, the sound reproduced through this module is filtering to take into account the headphone's insertion loss, and to compensate for the already present bone conducted sound.

The complete system is a summation of the POV and BRS modules played back to the talking-listener in real-time as they speak. Besides the differences in the microphone positioning and the reproduction format which involves real-time filtering, in comparison to their system, the system described in this thesis does not simulate the direct sound (no POV module), as it is already present as the talker's own voice, which leads to a simplification in the overall model. Also, the current system can perform real-time convolution of impulse responses of duration much longer than 85 ms, which is the limit of their system, and hence can be used to simulate a much broader range of rooms. As the rooms used in their system are simulated using the tool described in Lehnert and Blauert (1992), they are easier to render than measuring real rooms as in the current study. Also their system can be used to auralize one's own voice from sound sources that are located in the far-field of the talking-listeners which, though possible, is currently not in scope of the system described in this thesis.

C.3 Third system

Compared to the last system, the system described by Ueno and Tachibana (2010) illustrates a case where a loudspeaker based reproduction system is used. Their system simulates the acoustic conditions of a concert hall, for an ensemble performance of two musicians. It could be argued that in this case the loudspeaker based reproduction format for both the musicians would be a better choice, compared to the headphone based reproduction format, which might hamper the more natural state of a musician playing his/her instrument without any cumbersome appendages.

The design of their system was inspired by the system described by Gade (1989), which was used in studies of stage support for musicians. In Gade's system, a

microphone in an anechoic chamber picked the sound of a musician's instrument, and the signal was routed to loudspeakers in a real reverberation chamber. The next step in the signal chain was to route the sound recorded in the reverberation chamber from the last step, back to the musician's room in real-time, which was reproduced through loudspeakers in the same arrangement as the reverberation room.

In the system described by Ueno and Tachibana (2010), the task performed by the routing of the signal to a reverberation room in Gade's system is performed by the hardware-based Huron convolution platform (D. McGrath, 1995), one per the two musicians in an ensemble. Their system is currently performing real-time convolution of dry anechoic signals from the musician's instrument, with the measured 24-channel impulse response of a particular concert hall. The convolved signal is being played back to the musicians in real-time through 12 loudspeakers, 6 per musician.

In comparison, in the system described in this thesis, the virtual sound field reproduction is attained without employing a large number of loudspeakers. Also, as the convolution is software-based, the current system is cost effective, easier to implement and flexible in terms of incorporating more channels with head-tracking. Finally, the current system is a more accurate system for simulating self-speech, whereas the system described by Ueno and Tachibana (2010) is suited for a musician playing an instrument (like flute, violin, etc.).

C.4 Fourth system

As an example of a system, where the design methodology includes a loudspeaker-based reproduction format, Favrot and Buchholz (2010) describe an Ambisonics based sound field reproduction. Their Loudspeaker-based Room

Auralization system (LoRA) can be used to auralize room acoustic models generated by commercially available software, using higher-order Ambisonics. The sound field generated by their system is limited by the Ambisonic order applied, which in turn is limited by the number of loudspeakers, according to the following relation,

$$N \geq (M + 1)^2$$

for a three-dimensional reproduction (Ward and Abhayapala, 2001), where N is number of loudspeakers and M is the ambisonic order to be applied. Therefore, to ensure accurate localization cues for the direct sound, which would require ambisonic order of greater than or equal to 4, the loudspeaker number would be more than or equal to 25 (Favrot and Buchholz, 2010). Besides the logistic issues involved in setting up large number of loudspeakers, a toolbox such as LoRA has not been designed to simulate self-sounds in real-time yet. Here the convolution is software-based and performed off-line.

C.5 Fifth system

Featuring a hybrid method of combining image source and stochastic ray-tracing, the system described by Wefers & Vorländer (2010) has been implemented within the CAVE system (Schröeder et al., 2010) to incorporate direct acoustical feedback from the user into the virtual acoustic domain. It is more flexible than most of the other systems described in this section as it can provide real-time feedback to the user either through headphones or loudspeakers with acceptable latencies, with headtracking (Wefers & Vorländer, 2010). The convolution process here is software-based and performed in real-time.

C.6 In-situ studies

To conduct a room acoustic experiment in real room, though not entirely a simulated environment, is also an option. Here, the talking-listeners could either be listening to the sound of their own voices in rooms without any electronic reproduction; or listening to exocentric sounds that are reproduced mostly through loudspeakers. The latter case is similar to a real-world counterpart of the system described in section C.2 but without the POV module; while the former is the most natural mode of room acoustic experiments that generally, but not always, requires some means to limit visual information about the rooms to the talking-listener, such as blindfolding, selective lighting, etc. It must be noted here that the latter case can be classified as a mixed-reality environment, where the contribution of the reality or virtuality to the whole environment can be easily modified.

The in-situ studies are mentioned here in the appendix for the following reasons.

(a) They are the most natural mode of conducting room acoustic experiments.

(b) To compare it to the studies conducted using the system described in this thesis.

(c) They could be used for perceptual validation of the system described in this thesis

The first reason (a) has been discussed at the beginning of this subsection. Two examples of in-situ studies include the experiments conducted by McGrath *et al.* with blind and sighted but blindfolded participants, and the experiment conducted by Pop & Cabrera (2005), with sighted but blindfolded participants. As both of these experiments involved auditory room size perception, it is instructive to compare their set-up with the study described in Chapter 6.

The first study by McGrath *et al.* (1999) involved participants using egocentric (sound of their own voices, with natural or exploratory head movements) and exocentric (the experimenter producing natural sounds with various devices) stimuli in two rooms: one small (a soundproof recording studio) and one large (a concert hall). Without going into detail about the experiments, a few points need to be mentioned here that address the characteristics of this type of in-situ studies. Firstly, the scope of these studies is always confounded by the process of moving the participants from room to room, where the acoustic characteristics of the path from room to room, or indeed between positions in the same room could influence the auditory perception of participants. As an example, most of the blind participants in their study could determine the characteristics of the room (such as size) by listening to the sound of their footsteps, before even speaking in the room. And even though it is an interesting phenomenon, similar tasks could influence the participants' judgments. Also, it is a cumbersome process to move the participant' from room to room and it is generally hard to change the acoustic characteristics of the rooms being tested in real-time. So even though egocentric studies in one real room could be easy to set-up, testing in a large number of rooms, which is required for a broader understanding of a phenomenon, is easier to organize with the system described in this thesis, where the acoustic characteristics of the rooms could be changed in real-time within the software.

Similarly exocentric tasks performed in the study by McGrath *et al.* (1999) were confounded to a degree by the presence of the experimenter, which was necessary to produced sounds from objects, such as bouncing a ball. An example of a way around this problem is the approach adopted by Pop & Cabrera (2005), in which exocentric sounds

were produced by loudspeakers in the far-field of the participants who judged the size of the room from these sounds, in a number of rooms. But the experimenters commented that, apart from setting up the loudspeakers in a number of rooms, it is hard to work out the logistics of moving the participants (especially if more than one) from room to room without the path between rooms influencing the experiment. And it is harder, if not impossible, to move from rooms in one building to another, without seriously affecting the judgments of participants: an issue easily solved with the system described in this thesis, where rooms from different buildings or even computer simulated rooms (which can be designed to be 'unrealistic' in some respect) can be measured with BRS and simulated with the option to switch in real-time between these rooms.

It must be emphasized here that in-situ experiments still hold an advantage over virtual or mixed-reality experiments, in being the mode where the participants can behave most naturally and potentially perform a wider range of activities, some of which are presently very hard to simulate with accuracy within room acoustic models. Also, in-situ experiments can be crucially important for any virtual or mixed-reality environment, as the validity of the results of experiments conducted in these environments, or the degree of 'presence' (Witmer & Singer, 1998) within them, can be ascertained by conducting the same experiment (with appropriate modifications) in the real environment or comparing the 'presence' in the more virtual environments with the real environment.

Appendix D

List of publications arising from this thesis

1. Yadav, M., Cabrera, D., & Martens, W.L. (2011) A system for simulating room acoustical environments for one's own voice. *Applied Acoustics,* doi: 10.1016/j.apacoust.2011.10.001
2. Yadav, M., Cabrera, D., & Martens, W.L. (accepted, 2011) Auditory room size perceived from a room acoustic simulation with autophonic stimuli. *Acoustics Australia.*
3. Yadav, M., Cabrera, D., Collins, R., Martens, W.L. (2011) Detection of headtracking in room acoustic simulations for one's own voice. *ACOUSTICS 2011*. Gold Coast, Australia.

References

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

Brunskog, J., Gade, A. C., Bellester, G. P., & Calbo, L. R. (2009). Increase in voice level and speaker comfort in lecture rooms. *The Journal of the Acoustical Society of America*, *125*, 2072-2082.

Cabrera, D. (2007). Control of perceived room size using simple binaural technology. *13th International Conference on Auditory Display*. Montréal, Canada.

Cabrera, D., Davis, P., Barnes, J., Jacobs, M., & Bell, D. (2002). Recording the operatic voice for acoustic analysis. *Acoustics Australia*, *30*(3), 103-108.

Cabrera, D., Lee, D., Collins, R., Hartmann, B., Martens, W.L, & Sato, H. (2011). Variation in Oral-Binaural Room Impulse Responses for Horizontal Rotations of a Head and Torso Simulator. *Building Acoustics*, *18*(1), 277.

Cabrera, D., Sato, H., Martens, W. L., & Lee, D. (2009). Binaural measurement and simulation of the room acoustical response from a person's mouth to their ears. *Acoustics Australia*, *37*(3), 98-103.

Cabrera, D., Jeong, D., Kwak, H. J., & Kim, J.-Y. (2005). Auditory Room Size Perception for Modeled and Measured Rooms. Presented at the Inter-noise, The 2005 Congress and Exposition on Noise Control Engineering, Rio de Janeiro.

Cycling'74. (2011). Max/MSP. http://cycling74.com/

Dunn, H. K., & Farnsworth, D. W. (1939). Exploration of Pressure Field Around the Human Head During Speech. *The Journal of the Acoustical Society of America*, *10*, 184-199.

Favrot, S., & Buchholz, J. (2010). LoRA: A Loudspeaker-Based Room Auralization System. *Acta Acustica united with Acustica*, *96*, 364-375.

Gade, A.C. (1989). Investigations of Musicians' Room Acoustic Conditions in Concert Halls. Part I: Methods and Laboratory Experiments. *Acustica*, *69*, 193-203.

Hameed, S., Pakarinen, K., Valde, K., & Pulkki, V. (2004). Psychoacoustic Cues in Room Size Perception. *116th Audio Engineering Society Convention*. Berlin, Germany.

Hutchings, N., Irving, E. L., Jung, N., Dowling, L. M., & Wells, K. A. (2007). Eye and head movement alterations in naïve progressive addition lens wearers. *Ophthalmic and Physiological Optics*, *27*(2), 142-153.

Knapp, C., & Carter, G. (1976). The generalized correlation method for estimation of time delay. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, *24*(4), 320-327.

Knufinke, C. (2010). SIR2. http://www.knufinke.de/sir/sir2.html

Kutruff, H. (2009). *Room Acoustics*. London & New York: Spon Press/Taylor and Francis.

Lane, H. L., Catania, A.C., & Stevens, S.S. (1961). Voice Level: Autophonic Scale, Perceived Loudness, and Effects of Sidetone. *The Journal of the Acoustical Society of America*, *33*, 160-167

Lee, C. (1999). Eye and head coordination in reading: roles of head movement and cognitive control. *Vision Research*, *39*(22), 3761-3768.

Lee, D., Cabrera, D., & Martens, W. L. (2009). Equal reverberance matching of music. *Proceedings of ACOUSTICS*. Adelaide, Australia.

Lehnert, H., & Giron, F. (1995). Vocal comminication in virtual enviroments. *Virtual Reality World*, 279-293.

Lehnert, H., & Blauert, J. (1992). Principles of binaural room simulation. *Applied Acoustics*, *36*(3-4), 259-291.

Leventhal, L. (1986). Type I and Type 2 Errors in the Statistical Analysis of Listening Tests. *Journal of the Audio Engineering Society*, *34*(6), 437-664.

Lindau, A., Hohn, T., & Weinzierl, S. (2007). Binaural resynthesis for comparative studies of acoustical environments. *AES 122nd Convention*. Vienna, Austria.

Lokki, T., Pätynen, J., Kuusinen, A., Vertanen, H., & Tervo, S. (2011). Concert hall acoustics assessment with individually elicited attributes. *The Journal of the Acoustical Society of America*, *130*(2), 835.

Mackensen, P., Fruhmann, M., Thanner, M., & Theile, G. (2000). Head-Tracker Based Auralization Systems: Additional Consideration of Vertical Head Movements. *AES 108th CONVENTION*. Paris, France.

McConkie, G. W., Reddix, M. D., & Zola, D. (1992). Perception and cognition in reading: where is the meeting point? *Eye movements and visual cognition* (pp. 293-303). New York: Springer.

McGrath, D. (1995). Huron - A Digital Audio Convolution Workstation. *5th Australian Regional Convention of the Audio Engineering Society*. Sydney, Australia.

McGrath, R., Waldmann, T., & Fernstrom, M. (1999). Listening to rooms and objects. *16th Audio Eng. Soc. Int. Conf.* Rovaniemi, Finland.

Mershon, D. H., & Bowers, J. N. (1979). Absolute and relative cues for the auditory perception of egocentric distance. *Perception*, *8*(3), 311-322.

Mershon, D. H., & King, L. E. (1975). Intensity and reverberation as factors in the auditory perception of egocentric distance. *Perception & Psychophysics*, *18*(6), 409-415.

Milgram, P., & Colquhoun, H. (1999). A Taxonomy of Real and Virtual World Display Integration. *Mixed Reality - Merging Real and Virtual Worlds* (pp. 1-16). Berlin: Springer-Verlag.

Moldrzyk, C., Ahnert, W., Feistel, W., Lentz, T., & Weinzierl, S. (2004). Head-Tracked Auralization of Acoustical Simulation. *AES 117th Convention*. San Francisco, CA, USA.

Novo, P. (2005). Auditory virtual environments. In Blauert, J. (Ed.), *Communication acoustics* (pp. 277-297). Berlin: Springer-Verlag.

Pelegrín-García, D. (2011). Comment on "Increase in voice level and speaker comfort in lecture rooms" [J. Acoust. Soc. Am. 125, 2072–2082 (2009)] (L). *The Journal of the Acoustical Society of America*, *129*(3), 1161-1164.

Pelegrín-García, D., Fuentes-Mendizábal, O., Brunskog, J., & Jeong, C-H. (2011a). Equal autophonic level curves under different room acoustics conditions. *Journal of the Acoustical Society of America*, *130*(1), 228–238.

Pelegrín-García, D., Smits, B., Brunskog, J., & Jeong, C-H. (2011b). Vocal effort with changing talker-to-listener distance in different acoustic environments *Journal of the Acoustical Society of America*. *129*(4), 1981-1990.

Pop, C. B., & Cabrera, D. (2005). Auditory room size perception for real rooms. *ACOUSTICS 2005*. Busselton, Western Australia.

Pörschmann, C. (2000). Influences of Bone Conduction and Air Conduction on the Sound of One's Own Voice. *Acta Acustica united with Acustica*, *86*(6), 1038-1045.

Pörschmann, C. (2001). One's Own Voice in Auditory Virtual Environments. *Acta Acustica united with Acustica*, *87*, 378-388.

Pörschmann, C. (2007). 3-D Audio in Mobile Communication Devices: Methods for Mobile Head-Tracking. *Journal of Virtual Reality and Broadcasting*, *4*(13).

Pörschmann, C., & Pellegrini, R. S. (2010). 3-D Audio in Mobile Communication Devices: Effects of Self-Created and External Sounds on Presence in Auditory Virtual Environments. *Journal of Virtual Reality and Broadcasting*, *7*(11).

Proudlock, F. A., Shekhar, H., & Gottlob, I. (2003). Coordination of Eye and Head Movements during Reading. *Investigative opthamology and visual science*, *44*(7), 2991-2998.

Sandvad, J. (1999). Auditory perception of reverberant surroundings. *The Journal of the Acoustical Society of America*, *105*, 1193.

Sato, Hiroshi, Bradley, J. S., & Morimoto, M. (2005). Using listening difficulty ratings of conditions for speech communication in rooms. *The Journal of the Acoustical Society of America*, *117*(3), 1157-1167.

Schröder, D., Wefers, F., Pelzer, S., Rausch, D., Vorländer, M., & Kuhlen, T. (2010). Virtual Reality System at RWTH Aachen University. *International Symposium on Room Acoustics*. Melbourne, Australia.

Seo, H., & Lee, C. (2002). Head-free reading of horizontally and vertically arranged texts. *Vision Research*, *42*(10), 1325-1337.

Shabtai, N. R., Zigel, Y., & Rafaely, B. (2010). Room volume classification from room impulse response using statistical pattern recognition and feature selection. *The Journal of the Acoustical Society of America*, *128*(3), 1155.

Shanefield, D., Clark, D., Nousaine, T., Leventhal, L. (1987). Comments on "Type 1 and Type 2 Errors in the Statistical Analysis of Listening Tests" and Author's Replies. *Journal of the Audio Engineering Society*, *35*(7/8), 567-572.

Silzle, A., Novo, P., & Strauss, H. (2004). IKA-SIM: A System to Generate Auditory Virtual Environments. *116th Convention of Audio Engineering Society*. Berlin, Germany.

Tajadura-Jiménez, A., Larsson, P., Väljamäe, A., Västfjäll, D., & Kleiner, M. (2010). When room size matters: Acoustic influences on emotional responses to sounds. *Emotion*, *10*(3), 416-422.

Tonndorf, J. (1962). Compressional Bone Conduction in Cochlear Models. *The Journal of the Acoustical Society of America*, *34*, 1127.

Torger, A., & Farina, A. (2001). Real-time partitioned convolution for Ambiophonics Surround Sound. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. New Paltz, New York.

Ueno, K., & Tachibana, H. (2010). A consideration on acoustic properties on concert-hall stages. *Proceedings of the International Symposium on Room Acoustics*. Melbourne, Australia.

Ueno, K., Kato, K., & Kawai, K. (2010). Effect of Room Acoustics on Musicians' Performance. Part I: Experimental Investigation with a Conceptual Model. *Acta Acustica united with Acustica*, *96*, 505-515.

v. Békésy, G. (1949). The Structure of the Middle Ear and the Hearing of One's Own Voice by Bone Conduction. *The Journal of the Acoustical Society of America*, *21*, 217.

Ward, D. B., & Abhayapala, T. D. (2001). Reproduction of a plane-wave sound field using an array of loudspeakers. *IEEE Transactions on Speech and Audio Processing*, *9*, 697-707.

Wefers, F., & Vorländer, M. (2010). Interactive acoustic feedback into virtual acoustic scenes. *Congress on sound and vibration*. Ljubljana, Slovenia.

Welti, T., & Zhang, X. (2010). Angular Resolution Requirements for Binaural Room Scanning. *129th Convention of Audio Engineering Society*. San Francisco, CA,USA.

Wenzel, E. M. (2001). Effect of Increasing system latency on localization of virtual sounds with short and long duration. *International Conference on Auditory Display*. Espoo, Finland.

Wenzel, E. M., & Foster, S. H. (1993). Perceptual consequences of interpolating head-related transfer functions during spatial synthesis. *Applications of Signal*

Processing to Audio and Acoustics, 1993. Final Program and Paper Summaries., 1993 IEEE Workshop on (pp. 102-105).

Wenzel, Elizabeth M., Arruda, M., Kistler, D. J., & Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America*, *94*(1), 111-123.

Witmer, B., & Singer, M. J. (1998). Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoperators and Virtual Environments*, *7*(3), 225-240.

Yairi, S., Iwaya, Y., & Suzuki, Y. (2008). Influence of Large System Latency of Virtual Auditory Display on Behavior of Head Movement in Sound Localization Task. *Acta Acustica united with Acustica*, *94*, 1016-1023.

Zahorik, P. (2009). Perceptually relevant parameters for virtual listening simulation of small room acoustics. *The Journal of the Acoustical Society of America*, *126*(2), 776-791.