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

Manuj Yadav, Densil Cabrera, Ralph Collins and William L. Martens

Faculty of Architecture, Design and Planning, University of Sydney, NSW 2006, Australia

ABSTRACT

When a visual stimulus is not present, the room reflected sound from one's own voice has been shown to be an important cue in determining the characteristics of rooms. The room reflected sound can vary with interaural changes that are sometimes accompanied with head movements. Such head movements, when incorporated in room acoustical simulations, are expected to increase the level of 'presence' within simulations. But whether this headtracking is detectable by talking-listeners hearing a room's response to the projection of their own voice has not been studied. In this pilot study, five participants performed ABX headtracking detection test by projecting their voice in six *real* rooms that were simulated in real-time, with an accurate binaural reproduction of room reflections. The results indicate that headtracking is detectable for the rooms tested, which ranged in volume from 125 m^3 to 7650 m^3, after equalizing for Type 1 and Type 2 errors. Most consistent detection was noticed in a room with the highest early *IACC*.

INTRODUCTION

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, 1999) and other sensory stimuli are not present or augment other sensory inputs when they are present. This 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. In the field of room acoustics, there are many studies that have examined the characteristics of room reflected sound by having a speaking or listening task performed *in situ*. Though it is relatively easy to set up an *in situ* experiment in one or two rooms (McGrath, 1999), to test people in more rooms becomes quite a challenging task (Pop, 2005). One alternative is to measure the impulse responses of a variety of rooms and simulate the acoustical response of these rooms in an anechoic environment using an appropriate reproduction methodology such as the binaural technology (Lindau, 2007, Blauert, 2005). Previously, almost all of the studies belonging to the latter category employed room acoustic simulations to playback anechoic recordings convolved with a room impulse response to listeners (exocentric stimulus); with no provision for the listeners to use their own voice as the stimulus (egocentric stimulus). Such egocentric stimulus, though taken for granted in a *real* room environment is computationally intensive to simulate in a *virtual* or *mixed-reality* (Milgram, 1999) room environment. Cabrera *et al.* (2009) have recently described a technique of measuring the impulse response from the mouth (vocal input) to the two ears (Oral Binaural Room Impulse Response, OBRIR) that can used to simulate the room reflected response of the sound of one's own voice in rooms with minimal latency. Such OBRIR can be measured for a range of head positions over a desired degree of freedom (Figure 1) of head-movements, in a process referred to as binaural room scanning (BRS).

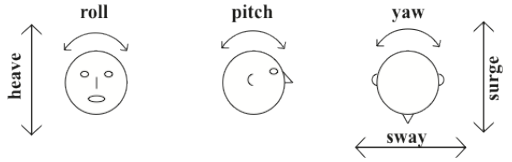

Figure 1. 6 degrees of freedom; 3 angular variations (yaw, pitch, roll) and 3 linear translations (sway, surge, heave)

Hence, in a simulation, 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, in the current study, on a pair of ear-loudspeakers that the talker-listener wears. BRS 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, 1998) within the simulation, implying greater task based performance.

There are however, many aspects of such real-time voice simulations that, due to the nascent nature of research in this field, have not been fully studied. One of the issues is the detectability of headtracking by the talking-listeners using their voice as the stimulus. In other words, it has not previ-

ously 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, 2010) 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, 2008, Welti, 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. As incorporating head-positions within such simulations involves the computationally non-trivial task (Torger & Farina, 2001) of convolution of impulse responses (length varies with reverberation times of rooms and can sometimes go up to 5 s or more) with a talker-listener's voice in real-time, the findings can provide information to the designers of such systems about the practical issues related to implementing headtracking. The findings are also likely to influence future studies that are focusing on real-time simulation of egocentric sounds.

The rest of the paper is organized as follows. In the Method section, we give a brief description of the real-time room acoustic simulation system used in the study and the details of ABX test conducted. The results and discussion of the results follow and we conclude with the scope of future research.

METHOD

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.

Experimental set-up

The real-time room acoustic simulation system used in this experiment was designed by hosting the SIR2 convolution plugin in Max/MSP (buffer size 128 samples) running on a Windows platform. The AD/DA converter used was a RME® ADI-8 QS unit with 48000 Hz sampling rate and 32-bit quantization in a 1-*in*/2-*out* configuration. The electroacoustic latency of this system is about 7.6 ms, which is converted to almost 0 ms with a process described later in this section. The headset microphone used for vocal input was a DPA 4066 and the ear-loudspeakers used were a pair of AKG K1000 (loudspeakers near the ears, without any circumaural cushion or contact with the ears). The presence or absence of the ear-loudspeakers had scarcely any effect on the octave-band gains for microphones (Brüel & Kjær 4101 Binaural Microphone) placed at the entrance of each ear canal for five participants (measured separately) speaking and a Head and Torso Simulator (Brüel & Kjær 4128C) emitting pink noise (Figure 2). The microphone was positioned at a distance of 7 cm from the centre of lips on the right side of the face. This was done to eliminate the detrimental effects associated with plosives and fricatives when the microphone is placed in the direct air-stream from the mouth opening. A similar microphone position has been used in a recent study (Pelegrín-García, 2011) for egocentric sound sources. The headtracker used was a Polhemus Fastrak® unit with a refresh rate of 5 Hz. The headtraker receiver and transmitter were positioned as shown in Figure 3 and the complete experimental set-up is depicted in Figure 4.

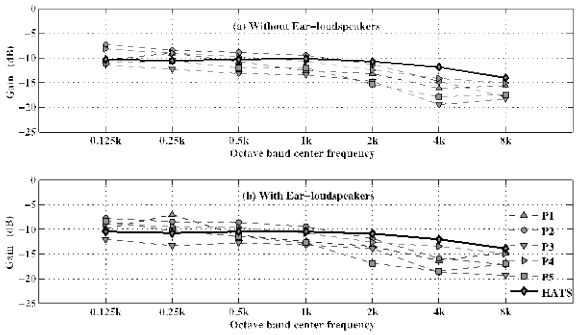

Figure 2: The transfer function from the headset microphone to the right ear microphone (on the same side of the face as the headset microphone was), without (a) and with (b) the 5 participants (and HATS) wearing the ear loudspeakers.

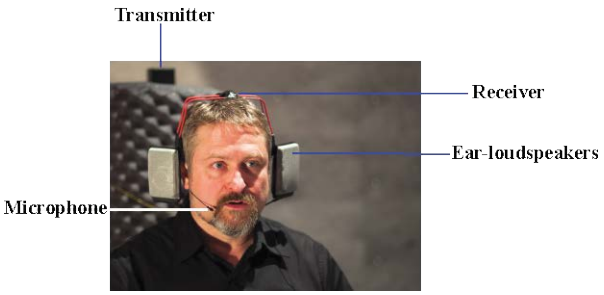

Figure 3: A talking-listener (fourth author; not tested here) wearing the ear-loudspeakers on which the headtracker receiver was mounted, seated on a wooden chair on a partially carpeted floor of an anechoic room. The headtracker transmitter can also be seen in the background.

The method for measuring the rooms with BRS has been described previously by Cabrera *et al*. (2009) and the resulting OBRIRs were band-limited from 100-10000 Hz. 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 (Lee, Cabrera and Mar-

tens, 2009) for individual binaural room impulse responses has been updated to derive multiple OBRIRs to be used for room simulation. Through this process, rooms of any size and reverberation time can be measured and implemented for real-time simulation with headtracking.

For the current experiment, six rooms were implemented within the system above using their OBRIRs over a yaw headtracking range of -40° to +40° with 2° resolution. When implementing the OBRIRs, the samples from the beginning of OBRIR to the floor reflection (about 7.6 ms) including the direct sound are excluded. The floor reflections are provided by adding a carpeted wooden floor to the anechoic room used in the experiment, and as the direct sound is already present with the talking-listener speaking or singing, it is not simulated. The convolved *output* is delayed by the duration of the deleted samples and hence it smoothly follows the direct sound and floor reflection – essentially leading to a latency of 0 ms for the room reflections. There is also no latency in changing from one room to the other, once the experimental set-up is fully operational.

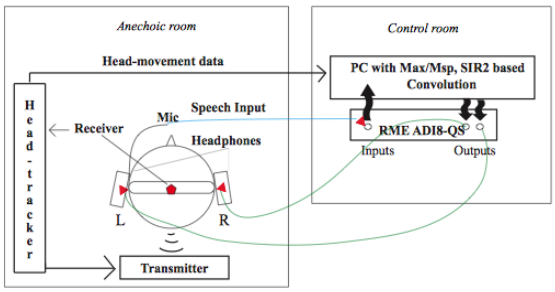

Figure 4: The complete set-up in the Acoustics Laboratory at the Faculty of Architecture, Design and Planning, The University of Sydney.

It must be noted here that the headtracking incorporates changes in head-positions and continually selects the OBRIR to be convolved with the current vocal *input* and the real-time convolution system *outputs* two channel of convolved audio that includes the *output* from the current head position and any other previous head positions (which may still be following a reverberant decay). This provides an auditory scene which is almost the same to the one produced by vocal transduction in *real* rooms for similar head movements.

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 previous study (Cabrera, 2010) and so we are only present their volumes and mid-frequency reverberation times in Table 1. Further details in the previous study include room plans and photographs, and acoustic features such as room gain, interaural level difference and interaural cross-correlation as a function of yaw angles.

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).

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 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.* (2010), followed by their volume and mid-frequency reverberation times.

RESULTS AND DISCUSSION

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 use a vocal transduction 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 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.*, 2010).

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 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

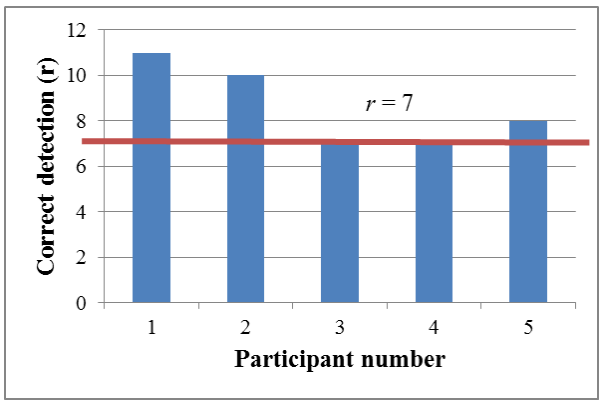

Figure 5: Number of correct detections (r) in the ABX test plotted for each participant. r=7, the detection threshold is represented as the thick red line on the plot

It must be emphasized here that even though the significance level of 0.39 and 0.33 is much higher than 0.05 or 0.01 that 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.

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, we prefer the previous line of analysis, as it provides a clearer picture of detection performance across the participants.

Detection for each room

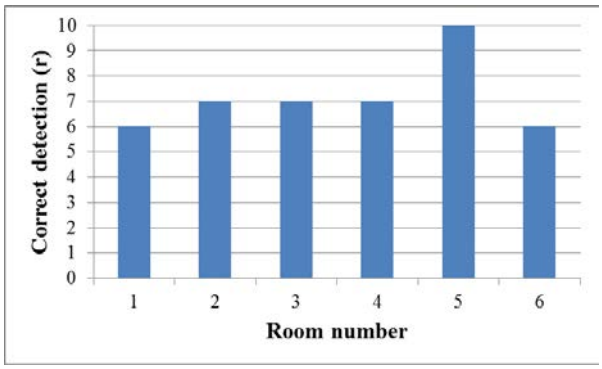

Figure 6: Number of correct detections (r) out 10 in the ABX test for the concatenated participant results for the six rooms that were simulated arranged in ascending order of physical volume in m^3

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.* (2010).

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.

CONCLUSIONS

Headtracking was shown to be detectable by five participants performing an ABX detection test for an egocentric sound source (one's own voice), for six measured real rooms simulated using a real-time room acoustic system that incorporates headtracking. Type 1 and Type 2 errors were equalized to be the maximum for a given number of trials (12) and probability (0.6) of correct detection. 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 for a range of 81º of yaw angles, 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.

ACKNOWLEDGMENTS

The authors wish to thank all the participants for their patience and Ken Stewart for technical assistance.

REFERENCES

Blauert, J 2005, 'Communication acoustics', Berlin: Springer-Verlag.

Cabrera, D, Lee, D, Collins, R, Hartmann, B, Martens, WL, & Sato, H, 2010, 'Characterising the variation in oral-binaural room impulse responses for horizontal rotations of a head and torso simulator', *In: Proceedings of the International Symposium on Room Acoustics.* Melbourne, Australia.

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

Lee, D, Cabrera, D, Martens, WL, 2009, 'Equal reverberance matching of music', *In: ACOUSTICS 2009.* Adelaide, Australia.

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, SY, 2007, 'Binaural resynthesis for comparative studies of acoustical environments', *In: 122nd Convention of Audio Engineering Society*, 2007 May 5–8 2007 Vienna, Austria.

McGrath, R, Thomas, W, & Fernstrom, M, 1999, 'Listening to rooms and objects', *In: 16th International Conference of the Audio Engineering Society,* Rovaniemi, Finland.

Milgram, P, & Colquhoun, H, 1999, 'A Taxonomy of Real and Virtual World Display Integration', *In:* Ohta, Y., & Hideyuki, T. (ed.) *Mixed Reality - Merging Real and Virtual Worlds,* Berlin: Springer-Verlag.

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

Pop, CB, & Cabrera, D, 2005, 'Auditory room size perception for real rooms' *In: ACOUSTICS 2005*. Busselton, Western Australia.

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

Welti, T, & Zhang, X. 2010, 'Angular resolution requirements for binaural room scanning'. *129th Convention of Audio Engineering Society*, San Francisco, CA, USA.

Witmer, BG, & Singer, MJ 1998, 'Measuring Presence in Virtual Environments: A Presence Questionnaire', *Presence: Teleoperators, Virtual Environments,* 7**,** 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.