BINAURAL MEASUREMENT AND SIMULATION OF THE ROOM ACOUSTICAL RESPONSE FROM A PERSON’S MOUTH TO THEIR EARS

Densil Cabrera (1), Hayato Sato (2), William L. Martens (1), Doheon Lee (1)

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

2. Environmental Acoustics Laboratory, Faculty of Engineering, Kobe University, Japan

This paper outlines methods to simulate the sound of one's own voice as it is affected by room acoustics, using binaural technology. An oral-binaural room impulse response (OBRIR) measurement can be made of a real room environment from the mouth to the ears of the same head. For simulation, a talker’s voice is convolved in real-time with the OBRIR, so that they can hear the sound of their own voice in the simulated room environment. We show by example how OBRIR measurements can be made using human subjects (by measuring the transfer function of speech) or by a head and torso simulator (HATS), and we illustrate the differences between individualised measurements and HATS measurements. We extend the HATS measurement method through binaural room scanning, which allows the simulation system to produce natural changes in the OBRIR as subjects rotate their heads while listening to their own voice.

INTRODUCTION

There are many situations in which the sound of one’s own voice produces a striking aural effect, for example an unfurnished room, a very small or very large room, an anechoic room, a reverberant room, or rooms with various echo phenomena. In less extreme everyday situations, we may analyse aspects of our environment through such acoustic feedback [1], and the feedback plays a significant role in speaking [2-4] and in playing music [5]. This paper is concerned with techniques that can be used to measure the room acoustical feedback in real rooms from real or artificial speech, for the purpose of simulation. By simulation, we mean a system that allows a speaking person to hear the sound of their voice in real-time in the simulated rooms. The purpose of this could be a tool for the scientific study of self-sound in relevant architectural acoustical contexts (for example, stage acoustics, classroom or lecture theatre acoustics, meeting room acoustics, etc), and could also contribute to virtual reality applications such as teleconferencing and games.

The sound of one’s own voice has three components: corporeal transmission (usually referred to as bone conduction); ‘direct’ airborne transmission from mouth to ear (including body-related acoustic effects, such as shoulder reflections); and reflections from the environment. Nukina and Kawahara [6], Pörschmann [7], and predecessors such as Békésy [8], have studied the first two of these, showing that air-conducted and bone-conducted sound are of similar magnitudes (most similar between 500 Hz and 3 kHz, wherein the acoustic power of the voice is greatest). Outside this range, air-conducted sound is greater than bone-conducted sound. Almost all acoustic radiation is from the mouth. However, in the present paper we are mainly concerned with the third component of the sound of one’s voice: reflection from the environment. Pörschmann [9] has approached this problem using computer modelled virtual environments, but the present paper is concerned with measuring and simulating real environments.

A closely related approach to this problem was taken by Sato et al. [2] in a study of listening difficulty, talking difficulty and conversational speech difficulty. They implemented a system with a microphone 0.1 m from the subject’s mouth, which fed the signal through a two-channel real-time convolver to simulate the room reflections at the subject’s ears (using ear loudspeakers, AKG K1000). A convolver performs an operation equivalent to time-domain convolution of the almost anechoic speech input signal with the room’s binaural impulse response, although the operation is usually implemented (at least partly) by multiplication in the frequency domain. They found that talking and conversing difficulty were much more sensitive to clarity index (C50) than was listening difficulty. Again, the difference between the present study and that study is that we wish to accurately simulate real reverberant environments, whereas Sato et al. did parametric control of the simulation’s reverberation.

This paper restricts its attention to binaural spatial analysis and synthesis. It is also feasible to approach the problem using high order microphone systems for measurement, and high order loudspeaker systems in the sound-field simulation, although that is much more complex to implement (Ueno and Tachibana’s [5] system for stage acoustics simulation for musicians is a simple version of that approach, with six measurement and convolution/reproduction channels, and Favrot and Buchholz [10] have devised a system for real-time auralization of computer-modelled room reflections from a person’s speech using many loudspeakers via high order ambisonics). In room acoustics, the term ‘binaural room impulse response’ (BRIR) is frequently used to denote the impulse response from a source to the two ears of a binaural receiver. In this paper, we use the term ‘oral binaural room impulse response’ (OBRIR) to make clear that we are discussing the room impulse response from a mouth (or mouth simulation) to the ears of the same head.

MEASUREMENT

2.1 Method

In this section, we compare two approaches to measuring the OBRIR: using a head and torso simulator, and using a real person. A head and torso simulator equipped with mouth and ear simulators provides an obvious approach to the measurement of OBRIRs. It is a simple matter to measure the transfer function or impulse response from an input signal (fed to the loudspeaker of the mouth simulator) to the output signals (from the ear microphones). Alternatively, a transfer function can be measured from a microphone near the mouth to the ear microphones. We take the second approach, which has the advantage of removing the response of the mouth simulator from the measurement, and is also well-suited to simulation – as a talking subject can have a microphone positioned similarly near their mouth as part of the simulation system.

We tested this approach to measurement using a Bruel & Kjaer 4128C head and torso simulator (HATS). As shown in Figure 1, the mouth simulator directivity of the HATS is similar to the mean long term directivity of conversational speech from humans, except in the high frequency range [11, *c.f.* 12]. The HATS’ standard mouth microphone position (known as the ‘mouth reference point’) is 25 mm away from the ‘centre of lip’ (which in turn is 6 mm in front of the face surface) [13, 14]. We used a Bruel & Kjaer Type 4939 (1/4″) microphone at the mouth reference point. Rather than using the inbuilt microphones of the HATS (which are at the acoustic equivalent to eardrum position), we used some microphones that are positioned near the entrance of the ear canals (Bruel & Kjaer type 4101). One reason for this is that we could use the same microphones on a real person at equivalent positions. Another reason is that it is desirable to avoid measuring with ear canal resonance, as the strong resonant peaks would need to be inverted in the simulation, which would introduce noise and perhaps latency.

The measurement was made by sending a swept sinusoid test signal to the mouth loudspeaker, the sound of which was recorded at the mouth and ear microphones (Fig 2). The sweep ranged between 50 Hz – 15 kHz, with a constant sweep rate on the logarithmic frequency scale over a period of 15 s. A signal suitable for deconvolving the impulse response from the sweep was sent directly to the recording device, along with the three microphone signals. This yielded the impulse response (IR) from the signal generator to each of the three microphones, and we obtained the transfer function from mouth microphone to ear microphones by dividing the latter by the former in the frequency domain. The procedure for this is, first, to take the Fourier transform of the direct sound from the mouth microphone impulse response, zero-padded to be twice the length of the desired impulse response. The direct sound is identified by the maximum absolute value peak of the mouth microphone IR, and data from -2 to +2 ms around this is used, with a Tukey window function applied (50% of the window is fade-in and fade-out using half periods of a raised cosine, and the central 50% has a constant coefficient of 1).

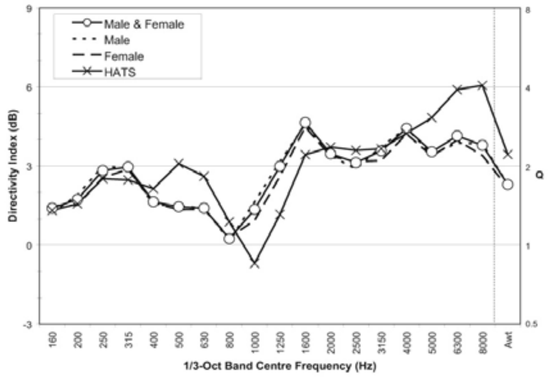

Figure 1. Directivity of a Bruel & Kjaer 4128C head and torso simulator (HATS) compared to the long term directivity of conversational speech (derived from Chu and Warnock’s data [11]).

The same Fourier transform window length is used for each of the ear microphone impulse responses, with the second half of the window zero-padded. The transfer function is obtained by dividing the cross-spectrum (conjugate of mouth IR multiplied by the ear IR) by the auto-spectrum of the mouth microphone’s direct sound. Before returning to the time domain, we bandpass-filter the transfer function to be within 100 Hz – 10 kHz to avoid signal-to-noise ratio problems at the extremes of the spectrum (this is done by multiplying the spectrum components outside this range by coefficients approaching zero). After applying an inverse Fourier transform, we truncate the impulse response (discarding the latter half). The resulting IR for each ear is multiplied by the respective ratio of mouth-to-ear rms values

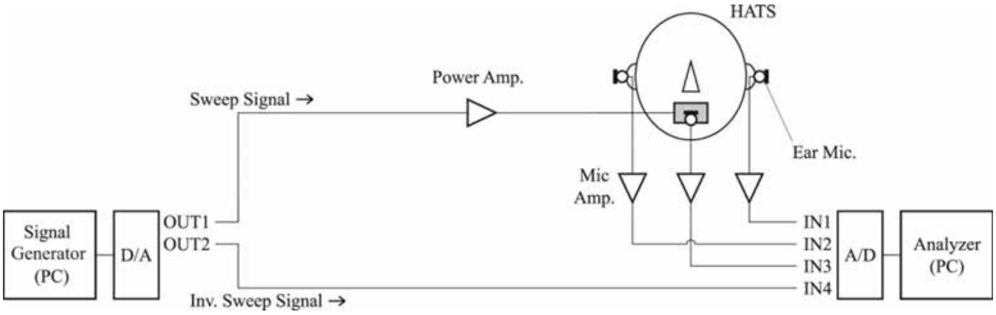

Figure 2. System for measuring OBRIRs using a head and torso simulator (HATS)

of microphone calibration signals (sound pressure level of 94 dB) to compensate for differences in gain between channels of the recording system. To test the process, we made these measurements in an anechoic room and a reverberant room (130 m^3, with a mid-frequency reverberation time of 2.5 s).

Measuring OBRIRs using a real person can be done using a similar microphone arrangement. The sound source could simply be speech, although other possibilities exist. The transfer function is calculated between a microphone near the mouth to each of the ear microphones. This approach was taken by Pörschmann and Nukina and Kawahara in measuring the transfer function from mouth to ear (without room reflections), but it can be used for measuring room reflections too. The advantages of using such a technique (compared to using the HATS) could include matching the individual long term speech directivity of the person; matching the head related transfer functions of the person's ears; and that the measurement system only requires minimal equipment (three microphones). Disadvantages may include effects of time-variance, a poorer signal-to-noise ratio in the measurement, and that some reverberation will be mixed with the direct sound at the mouth microphone (because we cannot isolate the direct sound as we could with an impulse response measurement).

We tested the real person method using the first and second authors. A B&K Type 4939 microphone was positioned near the middle of the mouth (taped to the nose, with a windshield), and the ear microphones were the same as those used in the HATS measurements. The mouth microphone was about 40 mm from the mouth (i.e. further than the HATS microphone). A laser-pointer was attached to the top of the head so that we could maintain an approximately constant head position during prolonged speech utterance (no physical head restraint was used). About ten minutes of continuous speech was recorded, and measurements were made in the anechoic and reverberant rooms as for the HATS. As the authors had different standing head heights, the HATS measurements in the reverberant room had been made to match both heights.

The transfer functions from mouth to ears were derived using the cross-spectrum method [15] with window lengths of 2^{16} samples for the anechoic room, and 2^{18} samples for the reverberant room (sampling rate of 48 kHz), with a Hann window function and a window overlap of 90%. However, we only used the 50% of the windows that had the highest signal level in the mouth microphone, so as to increase the signal-to-noise ratio of the process. The transfer function is estimated from the average cross-spectrum divided by the average auto-spectrum of the mouth microphone signal. The extremes of the spectrum (below the lowest speech fundamental, and above 10 kHz) could not be reliably processed, but indeed are not important for a reflected speech simulation system. There are two limitations to processing in the very high frequency range (above 10 kHz): the signal-to-noise ratio is poor because the voice produces little very high frequency energy at the ears; and the effect of time variance (variable directivity due to varying mouth shape and incidental head movement) is greatest for short wavelengths. We use a 100 Hz – 10 kHz bandpass filter in the same way as for the HATS measurements. The impulse response is obtained by inverse Fourier transform of the transfer function, and the latter half of the impulse response is discarded. An estimate of the reliability of the transfer function estimate is given by the associated coherence function (a value between 0 and 1, which is the squared absolute value average cross-spectrum divided by the product of the two average auto-spectra).

2.2 Results

The anechoic measurements (shown in Figure 3) have similar magnitude spectra to those of previous studies. The magnitude of the HATS transfer function is less than that of the human measurements because its microphone is closer to the mouth. The main notches in the spectra are due to the shoulder reflection, and so the tuning of the notches is affected by the mouth microphone position. However, it should be remembered that we are not aiming to simulate direct sound in the present study, so the mouth microphone position should not be critical (so long as it is near the mouth on the median saggital plane).

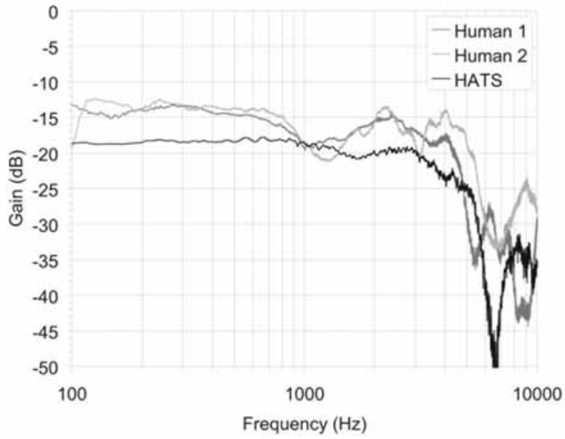

Figure 3. Magnitude of the transfer function from mouth microphone to ear microphone measured from speech (human measurements) or from a swept sinusoid test signal (HATS measurement).

Figure 4 compares the magnitude of the transfer functions (right ear only) for HATS and human measurements in the reverberant room. Differences are greatest in the high frequency range, and it can be observed that the general forms of the curves are similar to the respective anechoic measurements. While the HATS has greater high frequency gain in the reverberant room, this is not seen to the same extent in the human measurements – which may be due to the human mouth's time-varying radiation pattern within the measurement period (indicated by lower associated coherence values).

Comparison of the fine temporal structure of the reverberant room OBRIRs between the HATS and human measurements shows some similarity in peak times and levels up to about 50 ms (Figure 5). Beyond 50 ms, it is difficult to see any relationship between the fine structures. In Figure 5, the normalised values for humans after the direct sound are a little lower than those for the HATS because of the different mouth microphone position.

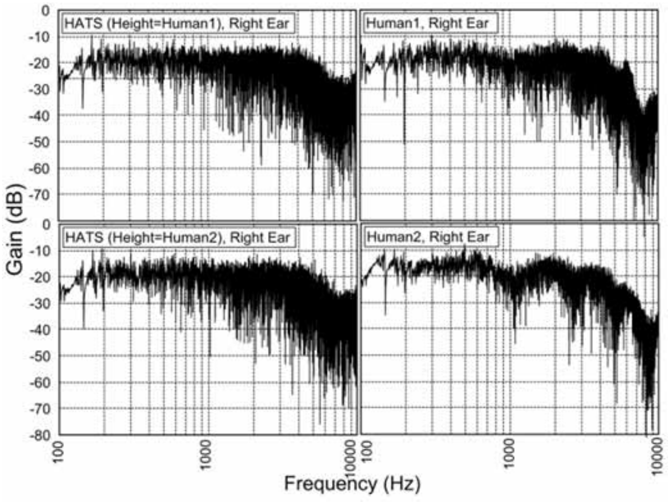

Figure 4. Magnitude of the transfer function of OBRIRs (right ear only) measured in the reverberant room.

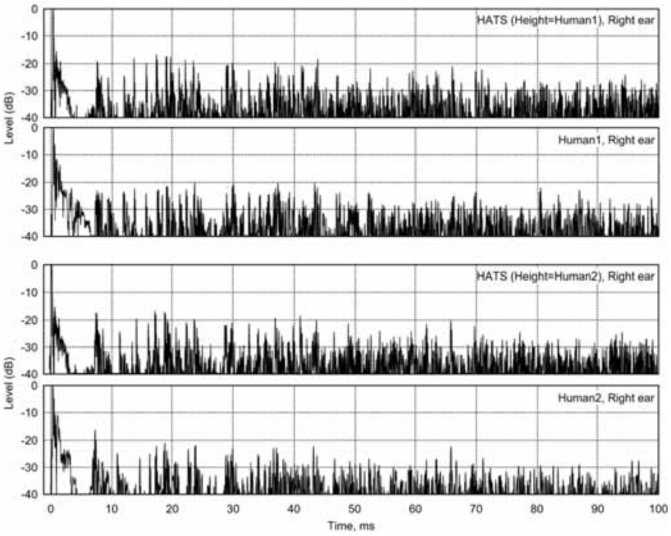

Figure 5. Normalised squared OBRIRs (right ear only) in decibels measured in the reverberant room, showing a comparison between the HATS and human measurements for the first 100 ms of each impulse response.

Comparison of the coarse temporal structure of the reverberant room OBRIRs between the HATS and human measurements can be done using reverberation time. However, since the source and receiver are very close, reverberation time was evaluated between -15 dB and -30 dB in the octave band reverse integration curves, rather than from the standard -5 dB point, because otherwise the reverberation time is artificially reduced due to the large drop in sound level after the direct sound. Results (Table 1) show a similar reverberation time spectrum shape, but with the human measurements reduced by a factor of about 0.82 relative to the HATS. The likely cause of this reduced reverberation time in the human measurements the time-varying directivity of human speech due to changes in the mouth shape and size, as well as minor head movements. Another possible contribution, at least in the low frequency range (where the wavelength is much larger than the distance between mouth and ear), is de-reverberation that could occur from not removing the reverberation from the mouth microphone in the human measurements – although the results do not show greater proportional reduction in reverberation in the low frequency range.

Table 1. Octave band reverberation times measured from the OBRIRs from the HATS and the humans in the reverberant room. Each value is the mean of two head heights and left and right ears. The final row gives the ratio of human to HATS reverberation time values. Values could not be derived reliably for humans in the 8 kHz octave band.

	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	8k
HATS	4.03 s	3.44 s	2.70 s	2.30 s	1.94 s	1.43 s	1.05 s
Humans	3.42 s	2.76 s	2.24 s	1.89 s	1.53 s	1.22 s	
Humans/Hats	0.848	0.803	0.83	0.821	0.789	0.848	

This comparison between HATS and human measurements suggests that, while there might be some advantage in individualising measurements through measurements from real humans, further refinement of the measurement and analysis method would be required to yield results close to measurements from a HATS. Rather than using long duration speech, particular phonemes, including individual vowels over a range of fundamental frequencies, could be used [6]. Restricting the derivation of transfer function to one phoneme would reduce time variance due to the changing mouth shape. It might then be possible to switch impulse responses in a simulation system depending on the phoneme of the talker (although there are considerable practical obstacles to achieving this). However, we tested the concept of single phoneme measurement with the unvoiced phonemes ‘sss’ and ‘shh’ in an effort to obtain increased high frequency coherence, but the results were not substantially better than normal speech.

SIMULATION

The binaural simulation system is illustrated in Figure 6. A microphone near the mouth is used to obtain the voice signal, which is sent to a real-time convolver. The convolver uses a measured OBRIR, and the resulting convolved speech is presented to the subject via near-ear loudspeakers. The ear loudspeakers that we used are AKG K1000. These are more appropriate than conventional headphones because they provide little occlusion of the ears, thereby allowing the direct airborne sound to arrive from the mouth relatively undistorted. Binaural simulations can usually be improved by implementing a headphone correction filter, which is an inversion of the transfer function from the headphones to the in-ear measurement microphones. We used a 256-sample (sampling rate of 48 kHz) inverse filter (finite impulse response), which was combined with the OBRIR in the real-time convolver. The convolver had a latency of 66 samples (1.375 ms) between input and output, and combining this with the inverse filter yielded a latency of 3.7 ms. The start of the OBRIR was truncated by this system latency so that the simulated room reflections would arrive at the ears correctly delayed. The gain of the simulation system was adjusted to match the relationship between direct and reflected sound that existed in the reverberant room measurement.

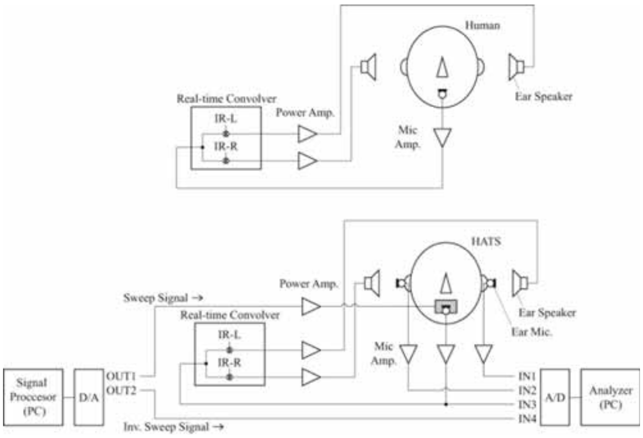

Figure 6. Block diagram of the simulation system (above), and of the test of the simulation system (below).

Measurements of the simulation system were made with the HATS in an anechoic room, as if it were a subject using the simulation system (Fig 7). A swept sinusoid was emitted from the mouth simulator so as to measure an impulse response at the microphone positions. The reverberant room measurement was used for the comparison.

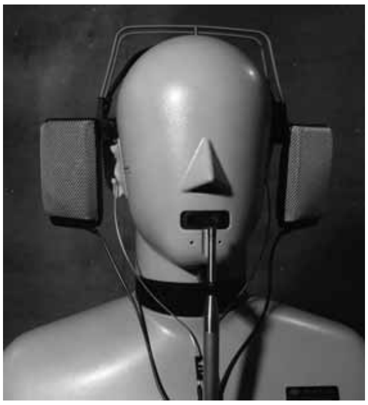

Figure 7. The AKG K1000 ear loudspeakers and the mouth microphone used for the simulation system, along with the head and torso simulator and ear microphones, which were used to test the simulation.

The results of the test show agreement between the simulation and the measured OBRIR, with some minor deviation immediately following the direct sound (which is at least partly due to the acoustic influence of the ear loudspeakers). The deviation at the start is likely to be masked by the direct sound (as it is -25 dB from the peak, and within 20 ms). The impulse response pattern that follows is a close match. Figure 8 shows this comparison for the first 100 ms of the left ear. We have not conducted a formal listening test comparing the original with the simulation, but informal listening has produced very positive responses.

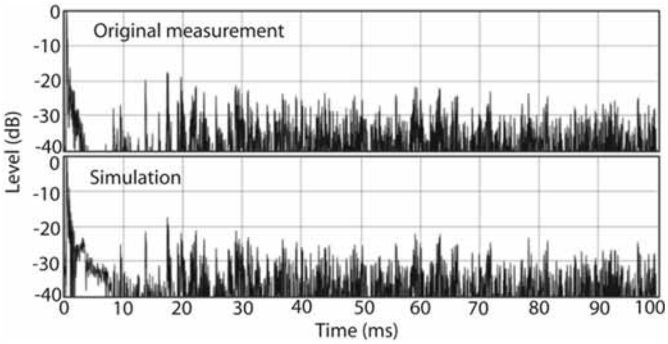

Figure 8. Comparison between the measured and simulated OBRIR (left ear only).

BINAURAL ROOM SCANNING

Binaural room scanning refers to a process of collecting and reproducing room impulse responses for a range of head orientations in a room [16]. Assuming that a subject will only make relatively constrained movements, measurements are made for horizontal rotations of a binaural recording device, at 2º intervals between -60º and +60º from the direction that is nominally straight ahead. The resulting sixty-one OBRIRs are switched in the real time convolver, with the selected OBRIR determined by the horizontal rotation of the subject's head. A head-tracking device is used to provide real time data to the computer so that the OBRIR is continually updated for the convolver.

Binaural room scanning has been used previously to simulate sound sources (such as loudspeakers) in rooms, with the listener at some distance from the source. In that application, maintaining the exocentric position of the sound source (independent of the head position) provides a great advantage in realism compared to simple head-locked binaural reproduction (where not using head tracking means that the auditory space moves with the head). The purpose of binaural room scanning is not to encourage large head movements, but rather to account for incidental head movements – and in this way it subtly provides a dramatic improvement in externalisation and realism. Only accounting for horizontal rotations is an approximation which is nevertheless effective because the predominant incidental head movements that strongly affect binaural hearing are horizontal rotations, which are also larger than head rotations typically observed around other axes. Measurements of the typical extent of incidental head rotations were collected for five human subjects engaged in talking, and the values sampled over a 3-second sampling period using a Polhemus FASTRAK system showed a standard deviation of 6.7º for the concatenated horizontal rotation data. Of course, during talking the human head is continuously shifting in orientation along its other two degrees of freedom, most often termed roll and pitch. Compared to the standard deviation of the measured horizontal rotation values, it was found that there was less than half that variation in head roll over the same 3-second sampling periods (the roll standard deviation was 2.3º). Although others have included a coupling of convolution with both head rotation and head pitch [17], coupling only horizontal rotations was included in our current implementation of head-tracked OBRIR reproduction, so that the collected room responses could be limited to a single sequence of only sixty-one OBRIRs.

Using binaural room scanning for OBRIRs is a little different to its conventional implementation, because the direct sound is not simulated (i.e., all that is being simulated is room reflections from the mouth source to the ears). Another difference is that the mouth and ears are all being rotated in the room, so that effects of moving the directional voice can be simulated. While changes in voice direction are probably only clearly discerned over large rotations (in environments with an uneven reflected soundfield, such as an auditorium stage) binaural room scanning provides a compelling reinforcement of externalization in the perception of the soundfield, and so is a cost-effective solution to rendering OBRIRs. At the time of writing, we have made binaural room scanned measurements of ten rooms from small to large for the purpose of experimental study of room acoustical features. Figure 9 gives an example of OBRIRs measured using binaural room scanning in a reverberant room, illustrating how the timing and strength of early reflections at each ear vary.

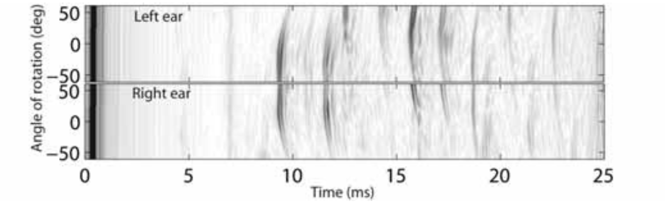

Figure 9. The first 25 ms of OBRIRs measured in a reverberant room using binaural room scanning from -60 to +60 degrees. The values shown are the absolute value of Hilbert-transformed waveforms, where black indicates high magnitude. The direct sound is seen just after 0 ms, and the floor reflection is a faint trace in the vicinity of 7 ms (ear height of 1.2 m).

CONCLUSIONS

This paper has outlined methods for measuring and simulating room acoustics using oral-binaural room impulse responses. Measurements can be made with minimal equipment from a speaking person, but using a head and torso simulator provides greater repeatability, a greatly shortened measurement time, removal of reverberation from the mouth microphone, and the possibility of implementing binaural room scanning. A simulation system using a mouth microphone, real-time convolver and ear loudspeakers produces signals at the ears that are close to those recorded in a real room. For such a system, the direct sound in the OBRIR is removed, which allows for a few milliseconds of latency in the simulation system.

Measurement and realistic simulation of OBRIRs can be used for the scientific study of the perception of room acoustics (for example, of loudness, clarity, stage support, speaking difficulty or room size). While binaural room scanning of horizontal head rotation provides some support for dynamic binaural perception, this is probably inadequate for some room acoustical studies, such as of human echo-location. As suggested by Pörschmann, there are also applications of such simulation systems beyond scientific studies, for example, in teleconferencing.

REFERENCES

[1] R. McGrath, T. Waldmann and M. Fernström, "Listening to rooms and objects" *Proc.16th Audio Engineering Society Int. Conf.*, Rovaniemi, Finland (1996)

[2] H. Sato, M. Morimoto and K. Fukunaga, "Effects of reverberation sounds on conversing difficulty in living rooms" *Tech. Rep. Architectural acoustics Acoust. Soc. Jpn.* No. AA2008-52 (2008) pp. 1-8 (in Japanese)

[3] M. Kob, G. Behler, A. Kamprolf, O. Goldschmidt and C.N. Rube, "Experimental investigations of the influence of room acoustics on the teacher's voice" *Acoustical Science and Technology* **29** pp. 86–94 (2008)

[4] J. Brunskog, A.C. Gade, G.P.Bellester and L. Calbo, "Increase in voice level and speaker comfort in lecture rooms" *J. Acoust. Soc. Am.* **125** pp. 2072-2082 (2009)

[5] K. Ueno and H. Tachibana, "Experimental study on the evaluation of stage acoustics by musicians using a 6-channel sound simulation system" *Acoustical Science and Technology* **24** pp. 130–138 (2003)

[6] M. Nukina and H. Kawahara, "Transfer characteristics of speech sounds around speaker's head" *J. Acoust. Soc. Jpn.* **59** pp. 256-260 (2003) (in Japanese)

[7] C. Pörschmann, "Influences on bone conduction and air conduction on the sound of one's own voice," *Acta Acustica united with Acustica* **86** pp. 1038-1045 (2000)

[8] G. v. Békésy, "The structure of the middle ear and the hearing of one's own voice by bone conduction" *J. Acoust. Soc. Am.* **21** pp. 217-232 (1949)

[9] C. Pörschmann, "One's own voice in auditory virtual environments" *Acta Acustica united with Acustica* **87** pp. 378-388 (2001)

[10] S. Favrot and J.M. Buchholz, "LoRA – A loudspeaker-based room auralisation system," *Acta Acustica united with Acustica*, forthcoming

[11] Chu, W.T. and A.C.C. Warnock, *Detailed Directivity of Sound Fields around Human Talkers,* National Research Council of Canada, Report IRC-RR-104 (2002)

[12] T. Halkosaari, M. Vaalgamaa and M. Karjalainen, "Directivity of artificial and real human speech" *J. Audio Eng. Soc.* **53** pp. 620-631 (2005)

[13] Bruel & Kjaer Product Data: Head and Torso Simulators – Types 4128-C and 4128-D

[14] International Telecommunication Union, ITU-T P.58 (08/96) *Head and Torso Simulator for Telephonometry* (2008)

[15] C.H. Knapp and G.C. Carter, "The generalized correlation method for estimation of time delay" *IEEE Transactions on Acoustics, Speech and Signal Processing* **24** pp. 320-327 (1976)

[16] S.E. Olive, T. Welti and W.L. Martens, "Listener loudspeaker preference ratings obtained in situ match those obtained via a binaural room scanning measurement and playback system" *Proceedings of the 122nd Audio Engineering Society Convention*, Vienna, Austria (2007)

[17] M. Vorländer, "Virtual Acoustics - Opportunities and limits of spatial sound reproduction" *Archives of Acoustics* **33**, pp. 413-422 (2008)