International Workshop
on the Principles and Applications of Spatial Hearing
November 11-13, 2009, Zao, Miyagi, Japan

Instant BRIR Acquisition for Auditory Events in Audio Augmented Reality Using Finger Snaps

Hannes Gamper[1] and Tapio Lokki[1*]
[1]*Helsinki University of Technology, Department of Media Technology*

1 Introduction

Augmented reality (AR) aims at enhancing the perception of the real world with virtual content, to guide, assist, or simply entertain the user [1, 2]. To do so, AR systems often rely on purely visual stimuli [3]. Sound, on the other hand, is a key element for conveying information, attracting attention and creating ambience and emotion [4]. Audio augmented reality (AAR) makes use of these properties to enhance the user's environment with virtual acoustic stimuli. Examples of AAR applications range from navigation scenarios [5], social networking [6] and gaming [7] to virtual acoustic diaries [8].

The augmentation is accomplished by mixing binaural virtual sounds into the ear input signals of the AAR user, thus overlaying virtual auditory events onto the surrounding physical space. The position of a real or a virtual sound source is determined by the human hearing based on localisation cues [9]. Encoding them into the binaural signals determines the perceived position of the virtual sounds. In the case of a real sound source, these localisation cues stem from the filtering behaviour of the human head and torso, as well as room reflections. A Binaural Room Impulse Response (BRIR) is the time domain representation of this filtering behaviour of the room and the listener, for given source and listener positions. It contains the localisation cues that an impulse emitted by a source at the given position in the room would carry when reaching the ear drums of the listener. Convolving an appropriate BRIR (for left and right ear) with a monaural virtual sound recreates the listening experience of the same sound as emitted from a real source at the position defined by the BRIR. The BRIR can thus be used to position a virtual source in a virtual room.

2 Instant BRIR acquisition

In this paper we present a simple and cost-effective way to acquire BRIRs on-the-fly and their application to intuitively position virtual sound sources, using finger snaps and/or claps. The BRIRs are obtained in the actual listening space, thus the filtering behaviour of the actual room is contained in them, as well as the filtering behaviour of the actual listener. Considering a telecommunication scenario with remote talkers, spatial separation and thus improved speech intelligibility of the talkers can be achieved by applying instant BRIRs to each talker acquired with differing excitation positions. The BRIRs obtained with the presented method in the actual listening space provide a natural and authentic spatial impression.

2.1 Hardware

Unlike virtual reality (VR) systems, AAR aims at augmenting, rather than replacing reality. This implies that the transducer setup used to reproduce virtual sounds for AAR must allow for the perception of the real acoustic environment. At the same time precise control over the ear input signals must be preserved, to ensure correct playback of the binaural virtual sounds. Using earphones as transducers provides the advantages of excellent channel separation, easily invertible transmission paths and portability.

The transducer setup used in this work is a MARA (mobile augmented reality audio) headset, as proposed by Härmä et al. [10]. It consists of a pair of earphones with integrated microphones and an external mixer. The microphones record the real acoustic environment, which is mixed with virtual audio content and played back through the earphones. Analog equalisation filters in the mixer correct the blocked ear canal response to correspond to the open ear canal response (see Fig. 1), thus they ensure acoustic transparency of the earphones [11]. This allows for an almost unaltered perception of the real acoustic environment and the augmentation thereof with virtual audio content.

2.2 Algorithm description

The proposed algorithm works as follows: If a transient in the microphone signals of the MARA headset is detected, the signals are buffered and the transient is extracted in each channel. These transients are taken as an approximation of the BRIR. A monaural input signal is filtered with this BRIR. The resulting binaural signals carry the same localisation cues as the recorded transient. Thus the monaural input signal is enhanced with the localisation cues of an external sound event at a certain position in the actual listening space. By deliberately generating a transient in the immediate surroundings of the user,

*E-mail address: hannes@tml.hut.fi, Tapio.Lokki@tkk.fi

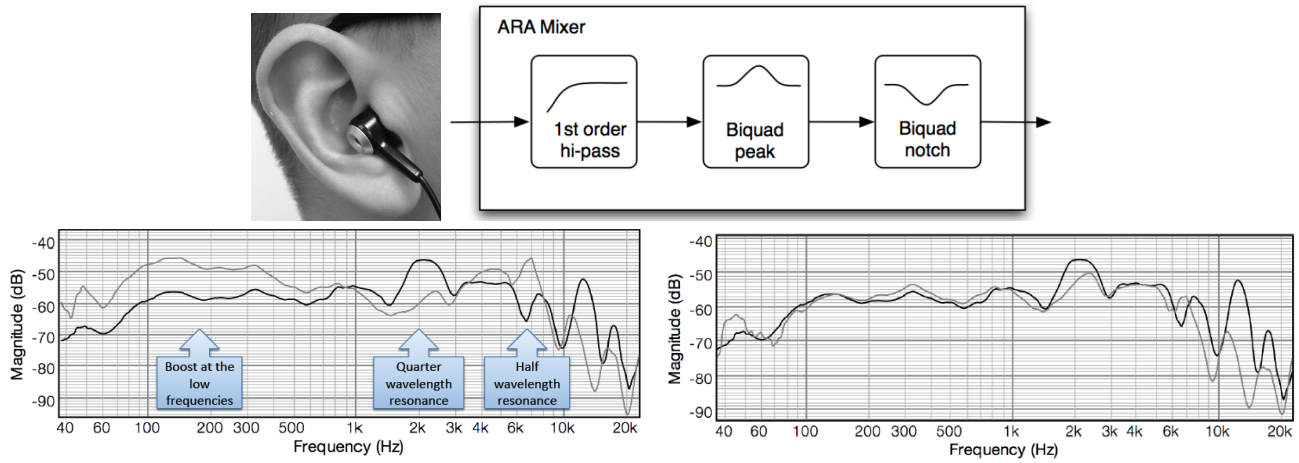

Fig. 1. The MARA headset and the basic principle of the analog equalisation. With three filters the main resonances of the ear canal can be compensated as seen in the right bottom figure (the black line is a measurement at the ear drum without earphone and the grey line is a measurement at the ear drum with earphone and compensation, see details [12]).

for example by snapping fingers or by clapping, a user can therefore intuitively position a virtual sound source in his or her acoustic environment. Details of the algorithm are described in the following sections.

2.2.1 Detection of transients

Room impulse responses are usually measured with a deterministic signal, e.g., with a maximum length sequence (MLS) or a sweep [13]. By deconvolving the known input signal out of the recorded room response, the impulse response of the room can be derived. Using an impulse as the excitation signal, no further processing is needed, as the recorded response is the room impulse response.

In the presented algorithm, a finger snap is taken as the excitation signal to estimate a BRIR on-the-fly. As the spectrum of the finger snap is however not white, the measured frequency response is in fact "coloured" by the snap spectrum. The BRIR derived from a finger snap excitation is thus only the coloured approximation of the real BRIR. The implications of this in the presented usage scenario are discussed in section 3. In theory any impulsive sound could be used as the excitation signal, with the aforementioned effect of colouring the BRIR. For this demonstration however the transient detection is tuned to find finger snaps.

To facilitate the detection of the snap, the microphone signals are pre-processed: The energy of finger snaps is mainly contained between 1500 and 3500 Hz [14]. A bandpass filter with a centre frequency of 2100 Hz and the mentioned bandwidth is applied to the microphone signals to remove frequency components above and below this band. This improves the detection performance in the presence of background noise considerably. To detect transients in the bandpass-filtered microphone signal, a method presented by Duxbury et al. [15] is employed. The energy of the signal is calculated in time frames of 256 samples each with 50 % overlap. Transients in the time domain are characterised by an abrupt rise in the short-time energy estimate. The derivative of the energy estimate is a measure for the abruptness of this rise in energy. If the derivative exceeds the detection threshold, the peak of the derivative is determined and the microphone signals of the MARA headset are buffered. Due to its simplicity the computational cost of the algorithm is very low. The algorithm proved to be quite robust also in the presence of background noise, which is an important criterion especially for mobile AAR applications. An example of the performance of the transient detection in a noisy environment can be seen in Fig. 2.

2.2.2 Extraction and application of the BRIRs

The position of the highest peak of both captured channels is taken as the position of the finger snap. The BRIR is extracted by windowing the buffered microphone signals around the snap. A flat top hanning window is applied to the buffers, starting 15 to 100 samples (i.e. 0.3–2.3 ms at 44100 Hz sampling rate, depending on the total window length) before the position of the finger snap, to ensure the onset of the transient is preserved. The length of the window is variable. For a short window (128 to 256 samples) only the early part of the impulse response is captured. It contains the direct signal and signal components that arrive 3–6 ms after the direct signal due to travelling an additional path length of up to 1–2 m, e.g. reflections from the shoulders and pinnae. Thus with a short window the room influence is eliminated, and only a coloured HRIR is extracted. Longer windows also include signal components that arrive after 3–6 ms, i.e. reflections from walls and objects inside the room. It is known that inclusion of this room reverberation improves the externalisation of virtual sounds [4, 16]. With impulse response lengths of 200 to 400 ms reasonable externalisation could be achieved.

The BRIR estimated in this way can directly be applied to a monaural input signal, thus enhancing the signal with the localisation cues of the recorded snap. This allows the user to position a virtual source intuitively in his/her environment. A possible application scenario to study the usability of the presented method was implemented in the programming environment Pure Data [17].

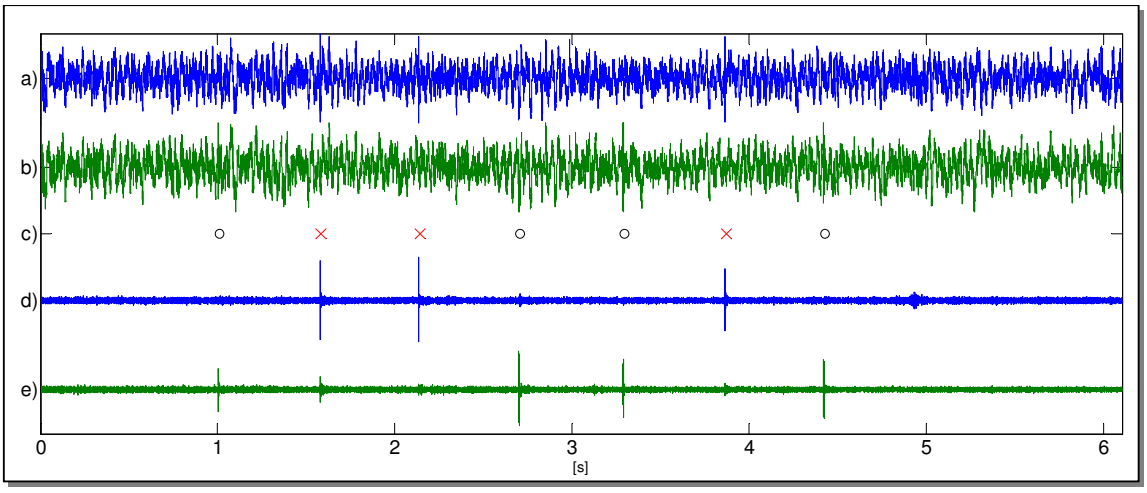

Fig. 2. Finger snap detection. a), b) left and right binaural input signal from the MARA headset; ; d), e) bandpass filtered signal; c) detected finger snaps ($\times$ left, $\circ$ right). For testing purposes, the MARA signals are mixed with binaurally recorded outdoor and traffic noise. The bandpass filtering with centre frequency around 2000 Hz reduces the background noise level considerably, therefore the finger snaps are detected correctly.

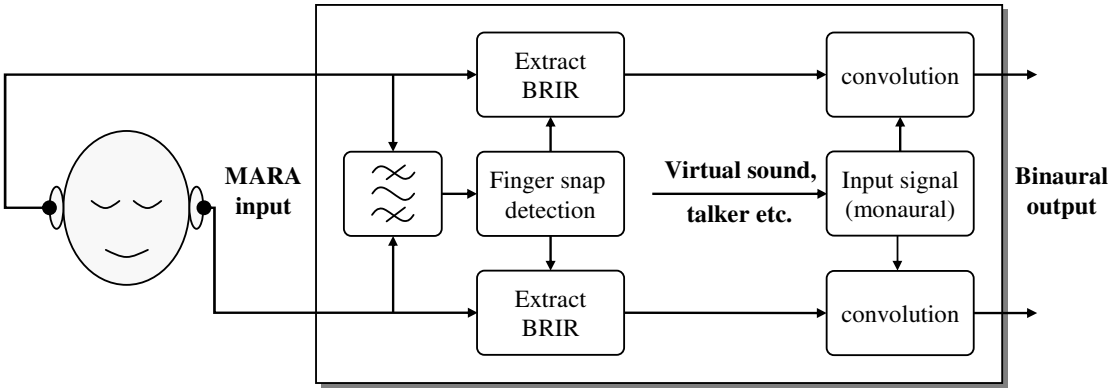

Fig. 3. Structure of the algorithm. If a finger snap is detected, a BRIR is extracted from each microphone channel and convolved with the input signal, i.e. a monaural speech signal of a virtual remote teleconference participant. Convolving each speaker with a separate snap, the participants can be spatially separated.

3 Real-time implementation

The Pure Data implementation of the above algorithm simulates a multiple-talker condition in a teleconference. In the simulated teleconference three participants (two remotes and one local) are discussing. The local participant is wearing the MARA headset. The remote end speakers are simulated by monaural recordings of a male and a female speaker and played back to the local participant over the earphones of the MARA headset. As the simulated remote end speakers are talking simultaneously, a multiple-talker condition arises. The unprocessed monaural signals position both speakers inside the users head, with no spatial separation. When the local participant snaps his or her fingers, the snap is recorded via the microphones of the MARA headset and convolved with the monaural speech signals. Snapping in two different positions, one for each of the remote speakers, allows the local participant to position the speakers in his or her auditory environment. The speakers sound externalised and separated spatially, which improves intelligibility and listening comfort. The structure of the algorithm is depicted in Fig. 3.

As the excitation signal, i.e. the finger snap, is in general non-white, the input signal will be coloured with the snap spectrum after convolution. The colouration can be controlled by the user by varying the spectrum of the transient, e.g. by clapping instead of finger snapping. This was found to be an interesting effect in informal listening tests. Furthermore, as the finger snap energy is mostly contained in a frequency band especially important for speech perception and intelligibility, the effect of the colouration was not found to be disturbing.

4 Discussion

It has been shown that the spatial separation of simultaneously talking speakers improves their intelligibility, a phenomenon known as the "cocktail party effect" [18]. In addition to the implications on speech intelligibility, the externalisation is also considered to "add a pleasing quality" to virtual sounds [19]. In the present work the spatialisation is performed by applying a separate BRIR to each signal. The BRIRs are acquired in the actual environment of the listener and recorded at the ear entrances of the listener.

Informal listening tests suggest that the use of a locally acquired individual BRIR allows for reasonable externalisation of virtual sources, given a sufficient filter length. We believe that there are two main reasons for this result. Firstly, it has been shown that individual HRTFs are in general superior than generic non-individual HRTFs in terms of the localisation performance [20, 21]. As the BRIRs with the presented method are recorded at the user's own ears, the filtering behaviour of the user's own head, torso and pinnae is captured. Applying the BRIR onto virtual sounds simulates the listening experience of that very user when exposed to a real source, leading to localisation cues in the binaural virtual sounds similar to normal listening.

Secondly, the influence of the listening environment on the sound field in the form of reflections and (early) reverberation is preserved in the tail of the BRIR. We assume that spatialising a virtual sound source with a room resembling the actual listening room leads to a more natural and physically coherent binaural reproduction. This is especially beneficial in the context of AAR, where embeddedness and immersion of virtual content in the real surroundings is required. To perceive the virtual and real environment as one, the characteristics of the virtual world have to resemble the ones of the real world.

5 Conclusions and Future Work

Instant individual BRIRs acquired with the described method and applied to monaural speech signals provide reasonable externalisation of virtual talkers. This can be a considerable improvement of intelligibility and listening comfort in multiple-talker conditions in telecommunication. The colouration of speech signals with the non-white input spectrum of a finger snap was not found to be disturbing, and could in fact be seen as an entertaining side effect.

A major improvement of the presented system would be to include head-tracking, to allow for stable externalised sources by dynamically panning them according to the head movements of the user. Another potential enhancement might be to whiten the transient spectrum, thus minimising the colouration, if high fidelity or reproduction of signals other than speech is required. Matched filtering could be applied as an efficient alternative to the proposed transient detection.

Acknowledgements

The research leading to these results has received funding from Nokia Research Center [kamara2009], the Academy of Finland, project no. [119092] and the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. [203636].

References

[1] R. Azuma. A survey of augmented reality. *Presence: Teleoperators and Virtual Environments*, 6:355–385, 1997.
[2] R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, and B. Macintyre. Recent advances in augmented reality. *IEEE Computer Graphics and Applications*, 21(6):34–47, 2001.
[3] M. Cohen and E.M. Wenzel. *Virtual environments and advanced interface design*, chapter The design of multidimensional sound interfaces, pages 291–346. Oxford University Press, Inc., New York, NY, USA, 1995.
[4] R. Shilling and B. Shinn-Cunningham. *Virtual auditory displays*. Handbook of Virtual Environments. Lawrence Erlbaum Associates, Mahwah NJ, 2002.
[5] B.B Bederson. Audio augmented reality: a prototype automated tour guide. In *CHI'95*, pages 210–211, New York, NY, USA, 1995. ACM.
[6] J. Rozier, K. Karahalios, and J. Donath. Hear&there: An augmented reality system of linked audio, 2000.
[7] K. Lyons, M. Gandy, and T. Starner. Guided by voices: An audio augmented reality system. In *Proc ICAD*, Atlanta, Georgia, USA, 2000.
[8] A. Walker, S.A. Brewster, D. McGookin, and A. Ng. A diary in the sky: A spatial audio display for a mobile calendar. In *Proc. BCS IHM-HCI 2001*, pages 531–540, Lille, France, 2001. Springer.
[9] J. Blauert. *Spatial Hearing. The psychophysics of human sound localization*. MIT Press, Cambridge, MA, 2nd edition, 1997.
[10] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, and G. Lorho. Augmented reality audio for mobile and wearable appliances. *Journal of the Audio Engineering Society*, 52(6):618–639, 2004.
[11] M. Tikander. Usability issues in listening to natural sounds with an augmented reality audio headset. *Journal of the Audio Engineering Society*, 57(6):430–441, 2004.
[12] M. Tikander, M. Karjalainen, and V. Riikonen. An augmented reality audio headset. In *in Proc. of the 11th Int. Conference on Digial Audio Effects (DAFx-08)*, pages 181–184, Espoo, Finland, 2008.
[13] S. Müller and P. Massarani. Transfer function measurement with sweeps. *Journal of the Audio Engineering Society*, 49(6):443–471, June 2001.
[14] S. Vesa and T. Lokki. An eyes-free user interface controlled by finger snaps. In *Proc. DAFx-05*, pages 262–265, Madrid, Spain, 2005.
[15] C. Duxbury, M. Davies, and M. Sandler. Improved time-scaling of musical audio using phase locking at transients. In *The AES 112th Convention*, page 5530, Munich, Germany, May 10-13 2002. preprint no. 5530.
[16] U. Zölzer, editor. *DAFX: Digital Audio Effects*. John Wiley & Sons, May 2002.
[17] M. Puckette. Pure data: another integrated computer music environment. In *Proc. ICMC*, pages 37–41, 1996.
[18] A.W. Bronkhorst. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*, 86(1):117–128, January 2000.
[19] B. Kapralos, M. R. Jenkin, and E. Milios. Virtual audio systems. *Presence: Teleoperators and Virtual Environments*, 17(6):527–549, 2008.
[20] H. Møller, M.F. Sørensen, C.B. Jensen, and D. Hammershøi. Binaural technique: Do we need individual recordings? volume 44, pages 451–469, 1996.
[21] H. Møller, C.B. Jensen, D. Hammershøi, and M.F. Sørensen. Evaluation of artificial heads in listening tests. volume 47, pages 83–100, 1999.