SPATIALISATION IN AUDIO AUGMENTED REALITY USING FINGER SNAPS

H. GAMPER and T. LOKKI*

Department of Media Technology, Aalto University,
P.O.Box 15400, FI-00076 Aalto, FINLAND
**E-mail: [Hannes.Gamper,ktlokki]@tml.hut.fi*

In audio augmented reality (AAR) information is embedded into the user's surroundings by enhancing the real audio scene with virtual auditory events. To maximize their embeddedness and naturalness they can be processed with the user's head-related impulse responses (HRIRs). The HRIRs including early (room) reflections can be obtained from transients in the signals of ear-plugged microphones worn by the user, referred to as instant binaural room impulse responses (BRIRs). Those can be applied on-the-fly to virtual sounds played back through the earphones. With the presented method, clapping or finger snapping allows for instant capturing of BRIR, thus for intuitive positioning and reasonable externalisation of virtual sounds in enclosed spaces, at low hardware and computational costs.

Keywords: Audio Augmented Reality; Finger snap detection; Binaural Room Impulse Response; Head Related Transfer Functions.

1. Introduction

Augmented reality (AR) describes the process of overlaying computer generated content onto the real world, to enhance the perception thereof and to guide, assist, or entertain the user.[1,2] In early AR research the focus was primarily on purely visual augmentation of reality,[3] at the expense of other sensory stimuli such as touch and sound. This imbalance seems unfortunate, given the fact that sound 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] and binaural audio over IP.[9,10]

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.[11] 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 the acoustic environment.

The chapter is organised as follows: section 2 describes the real-time acquisition of BRIRs. In section 3 a real-time implementation of the proposed algorithm for spatialisation in audio augmented reality (AAR) is presented. Results from informal listening tests of the real-time implementation are discussed in section 4. Section 5 concludes the chapter.

2. Instant BRIR acquisition

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 hand 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. Applying the BRIRs obtained with the presented method in the actual listening space to virtual auditory sources yields a natural and authentic spatial impression. In a telecommunication scenario with multiple remote talkers, the spatial separation achieved by processing each talker with a separate BRIR can improve the speech intelligibility and speaker segregation.[12,13]

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

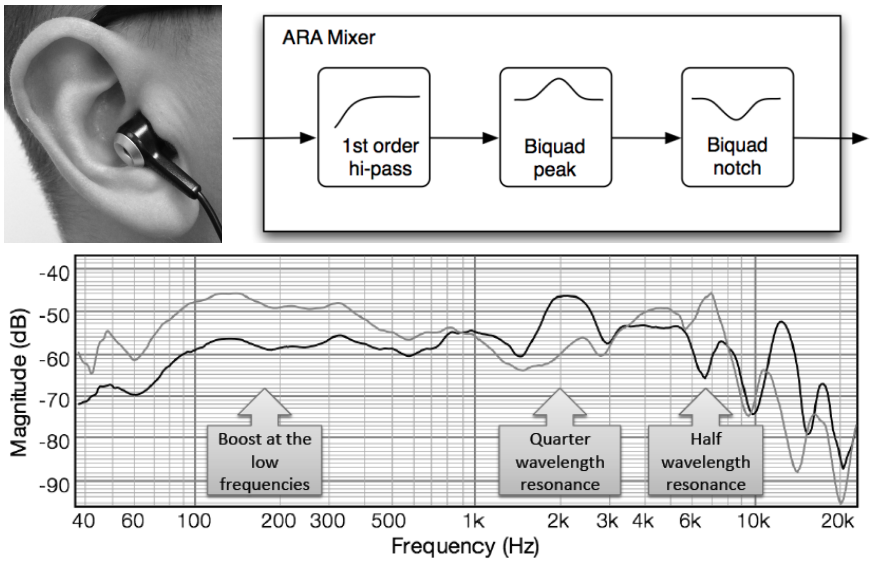

Fig. 1. The MARA headset and the basic principle of the analogue equalisation. Microphones embedded into insert-earphones record the acoustic surroundings at the ears of the MARA user (top figure, left). The bottom graph shows HRTF measurements at the ear drum with earphone (grey line) and without earphone (black line). To compensate for the impact of the earphones on the HRTF, the microphone signals are filtered in the ARA mixer before being played back to the user via the earphones, to ensure acoustic transparency of the transducer setup.[16]

must be ensured for 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.[14] It consists of a pair of earphones with integrated microphones and an external mixer (see Fig. 1). The microphones record the real acoustic environment, which is mixed with virtual audio content and played back through the earphones. Analogue equalisation filters in the mixer correct the blocked ear canal response to correspond to the open ear canal response, thus they ensure acoustic transparency of the earphones.[15] This allows for an almost unaltered perception of the real acoustic environment and the augmentation thereof with virtual audio content.

2.2. *Algorithm description*

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 and the reverberation tail contains the information of the surrounding environment. 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 generating a transient in the immediate surroundings of the user, for example by snapping fingers or by clapping, a user can therefore intuitively position a virtual sound source in his or her acoustic environment.

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.[17] By deconvolving the known input signal out of the recorded room response, the impulse response of the room can be derived. If an impulse is used as the excitation signal, the recorded response corresponds to 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 flat, 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.

To facilitate the detection of the snap, the microphone signals are preprocessed: The energy of finger snaps is mainly contained between 1500 and 3500 Hz.[18] 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 (see Fig. 2). To detect transients in the bandpass-filtered microphone signal, a method presented by Duxbury et al.[19] 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 mi-

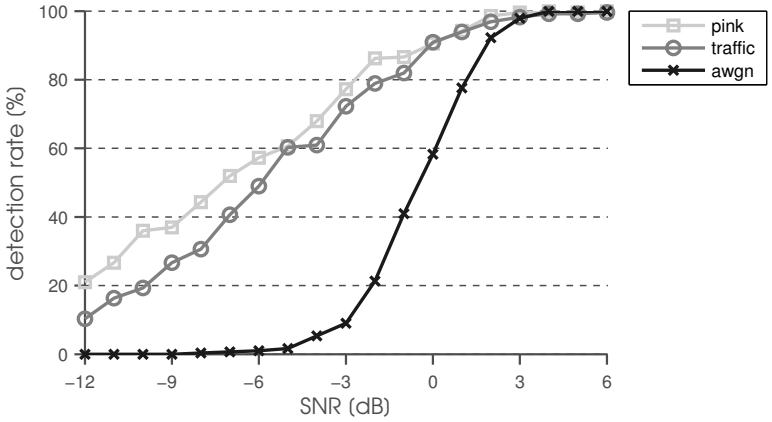

Fig. 2. Finger snap detection. The detection rate of finger snaps in noisy signals is given as a function of the signal-to-noise ratio (SNR), for various noise signals (pink noise, traffic noise, and additive white Gaussian noise). Pink and traffic noise yield higher detection rates, as their power spectral density decreases with frequency, thus less noise energy is present around 2000 Hz, where most of the finger snap energy is concentrated.

crophone 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. The performance of the transient detection in the presence of noise is depicted in Fig. 2.

2.2.2. *Extraction and application of the BRIRs*

The BRIR is extracted by windowing the buffered raw microphone signals around the detected snap. Thus, the BRIR is approximated by the unprocessed finger snap detected in the MARA signals. 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 traveling 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. reflec-

tions from walls and objects inside the room. It is known that inclusion of this room reverberation improves the externalisation of virtual sounds.[4,20] With impulse response lengths of 200 to 400 ms (i.e. window lengths of 8192–16384 samples) 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 by snapping a finger. To reduce the colouration of the BRIR with the finger snap spectrum, inverse filtering could be considered to whiten the BRIR. However, for virtual speech sources the colouration was not found to be disturbing, and postprocessing of the BRIR was thus omitted in the present implementation. A possible application scenario to study the usability of the presented method was implemented in the programming environment Pure Data.[21]

3. Real-time implementation

A real-time implementation of the proposed algorithm for spatialisation in audio augmented reality was presented at the IWPASH 2009 (International Workshop on the Principles and Applications of Spatial Hearing) conference in Japan.[22] The Pure Data implementation of the described 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 speech signals are perceived inside the 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 remote speakers are externalised and spatially separated, 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, does not have a flat spectrum, 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

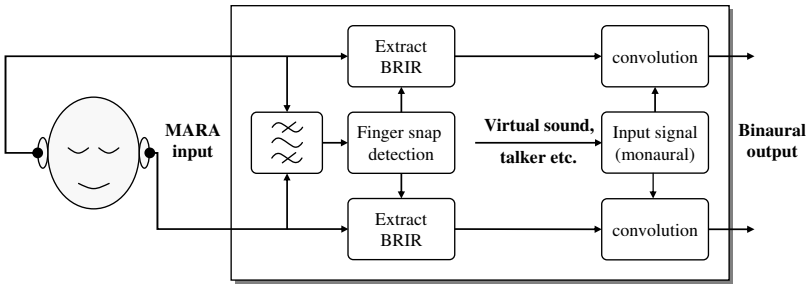

Fig. 3. Structure of the algorithm. If a finger snap is detected, a BRIR is extracted from each microphone channel and convolved with the input signal, i.e. a monaural speech signal of a virtual remote teleconference participant. Convolving each speaker with a separate snap, the participants can be spatially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

found to be an interesting effect in informal listening tests. Furthermore, as the finger snap energy is mostly contained in a frequency band in particular important for speech perception and intelligibility, the colouration was not found to deteriorate the communication performance.

4. Discussion

It has been shown that the spatial separation of simultaneously talking speakers improves their intelligibility. This is a phenomenon known as the "cocktail party effect".[23] In addition to the implications on speech intelligibility, the externalisation is also considered to "add a pleasing quality" to virtual sounds.[24] 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 canal entrances of the listener.

Informal listening tests suggest that the use of a locally acquired individual BRIRs 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.[25,26] As the BRIRs are recorded at the user's own ears with the presented method, the filtering behaviour of the user's own head, torso and pinnae is captured. Applying the BRIRs to 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 real-time implementation of the system was presented at the IWPASH 2009 (International Workshop on the Principles and Applications of Spatial Hearing) conference in Japan.[22] During the demo session, it was found that the described method of BRIR acquisition using finger snaps or clapping provides a very intuitive and straightforward way of defining the positions of virtual auditory events.

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 Sev-

enth Framework Programme (FP7/2007-2013) / ERC grant agreement no. [203636].

References

1. R. Azuma. A survey of augmented reality. *Presence: Teleoperators and Virtual Environments*, pp. 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**, pp. 34–47 (2001).
3. M. Cohen and E.M. Wenzel. The design of multidimensional sound interfaces. In W. Barfield and T.A. Furness, editors, *Virtual environments and advanced interface design*, pp. 291–346 (Oxford University Press, Inc., New York, NY, USA, 1995).
4. R. Shilling and B. Shinn-Cunningham. Virtual auditory displays. In K. Stanney, editor, *Handbook of Virtual Environments*, pp. 65–92 (Lawrence Erlbaum Associates, Mahwah NJ, 2002).
5. B.B Bederson. Audio augmented reality: a prototype automated tour guide. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 210–211 (New York, NY, USA, 1995).
6. J. Rozier, K. Karahalios, and J. Donath. Hear&there: An augmented reality system of linked audio. In *Proceedings of the International Conference on Auditory Display (ICAD)*, pp. 63–67 (Atlanta, Georgia, USA, 2000).
7. K. Lyons, M. Gandy, and T. Starner. Guided by voices: An audio augmented reality system. In *Proceedings of the International Conference on Auditory Display (ICAD)*, pp. 57–62 (Atlanta, Georgia, USA, 2000).
8. A. Walker, S.A. Brewster, D. McGookin, and A. Ng. Diary in the sky: A spatial audio display for a mobile calendar. In *Proceedings of the 15th Annual Conference of the British HCI Group*, pp. 531–540 (Lille, France, 2001. Springer).
9. T. Lokki, H. Nironen, S. Vesa, L. Savioja, and A. Härmä. Problem of far-end user's voice in binaural telephony. In *the 18th International Congress on Acoustics (ICA'2004)*, volume II, pp. 1001–1004 (Kyoto, Japan, April 4-9 2004).
10. T. Lokki, H. Nironen, S. Vesa, L. Savioja, A. Härmä, and M. Karjalainen. Application scenarios of wearable and mobile augmented reality audio. In *the 116th Audio Engineering Society (AES) Convention* (Berlin, Germany, May 8-11 2004). paper no. 6026.
11. J. Blauert. *Spatial Hearing. The psychophysics of human sound localization*, pp. 36–200. (MIT Press, Cambridge, MA, 2nd edition, 1997).
12. R. Drullman and A. W. Bronkhorst. Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. *Journal of the Acoustical Society of America* **107**, pp. 2224–2235 (2000).
13. H. Gamper and T. Lokki. Audio augmented reality in telecommunication through virtual auditory display. In *Proceedings of the 16th International*

Conference on Auditory Display (ICAD), pp. 63–70 (Washington, DC, USA, 2010).
14. 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**, pp. 618–639 (June 2004).
15. M. Tikander. Usability issues in listening to natural sounds with an augmented reality audio headset. *Journal of the Audio Engineering Society* **57**, pp. 430–441 (June 2009).
16. M. Tikander, M. Karjalainen, and V. Riikonen. An augmented reality audio headset. In *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08)*, pp. 181–184 (Espoo, Finland, 2008).
17. S. Müller and P. Massarani. Transfer function measurement with sweeps. *Journal of the Audio Engineering Society* **49**, pp. 443–471 (June 2001).
18. S. Vesa and T. Lokki. An eyes-free user interface controlled by finger snaps. In *Proceedings of the 8th International Conference on Digital Audio Effects (DAFx-05)*, pp. 262–265 (Madrid, Spain, 2005).
19. C. Duxbury, M. Davies, and M. Sandler. Improved time-scaling of musical audio using phase locking at transients. In *the 112th Audio Engineering Society (AES) Convention* (Munich, Germany, May 10-13 2002). preprint no. 5530.
20. U. Zölzer, editor. *DAFX: Digital Audio Effects*, pp. 151–153. (John Wiley & Sons, May 2002).
21. M. Puckette. Pure data: another integrated computer music environment. In *Proceedings of the International Computer Music Conference (ICMC)*, pp. 37–41 (Hong Kong, 1996).
22. IWPASH Organizing Committee. IWPASH 2009 International Workshop on the Principles and Applications of Spatial Hearing. `http://www.riec.tohoku.ac.jp/IWPASH/` (November 2009).
23. A.W. Bronkhorst. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica* **86**, pp. 117–128 (January 2000).
24. B. Kapralos, M. R. Jenkin, and E. Milios. Virtual audio systems. *Presence: Teleoperators and Virtual Environments* **17**, pp. 527–549 (2008).
25. H. Møller, M.F. Sørensen, C.B. Jensen, and D. Hammershøi. Binaural technique: Do we need individual recordings? *Journal of the Audio Engineering Society* **44**, pp. 451–469 (1996).
26. H. Møller, C.B. Jensen, D. Hammershøi, and M.F. Sørensen. Evaluation of artificial heads in listening tests. *Journal of the Audio Engineering Society* **47**, pp. 83–100 (1999).