24th INTERNATIONAL CONGRESS ON SOUND AND VIBRATION
23–27 July 2017, London

LONG-TERM USER ADAPTATION TO AN AUDIO AUGMENTED REALITY SYSTEM

J. Isaac Engel and Lorenzo Picinali

Imperial College London, Dyson School of Design Engineering, London, UK
email: isaac.engel@imperial.ac.uk

Audio Augmented Reality (AAR) consists in extending a real auditory environment with virtual sound sources. This can be achieved using binaural earphones/microphones. The microphones, placed in the outer part of each earphone, record sounds from the user's environment, which are then mixed with virtual binaural audio, and the resulting signal is finally played back through the earphones. However, previous studies show that, with a system of this type, audio coming from the microphones (or *hear-through* audio) does not sound natural to the user. The goal of this study is to explore the capabilities of long-term user adaptation to an AAR system built with off-the-shelf components (a pair of binaural microphones/earphones and a smartphone), aiming at achieve perceived realism for the hear-through audio. To compensate the acoustical effects of ear canal occlusion, the recorded signal is equalised in the smartphone. In-out latency was minimised to avoid distortion caused by comb filtering effect. To evaluate the adaptation process of the users to the headset, two case studies were performed. The subjects wore an AAR headset for several days while performing daily tests to check the progress of the adaptation. Both quantitative and qualitative evaluations (i.e., localising real and virtual sound sources and analysing the perception of pre-recorded auditory scenes) were carried out, finding slight signs of adaptation, especially in the subjective tests. A demo will be available for the conference visitors, including also the integration of visual Augmented Reality functionalities.

Keywords: augmented reality, adaptation.

1. Introduction

Audio Augmented Reality (AAR) consists in extending a real auditory environment with virtual sound sources [1-3]. One of the biggest challenges in making an AAR system is that it must be acoustically transparent, meaning that it can deliver sound to the ears without modifying the signal or altering the user's natural hearing capabilities (e.g., localisation accuracy) [1, 3, 4]. Different potential approaches exist for AAR, and the present authors plan to undertake long-term research on the most promising ones, being this paper the first step of that path. Future stages of the project will explore technologies such as open-fitting headphones, hearing-aid-based systems, or receiver-in-the-canal devices (as discussed in later sections).

This paper studies an acoustically transparent hearing device based on insert-type earphones with embedded binaural microphones (see Fig. 1). The user's acoustic environment is captured by the microphones and equalised, and the resulting signal (from now on, *hear-through* audio) is mixed with auralised virtual audio and presented through the earphones [1, 2]. As sound is recorded near the entrance of the ear canal, loss of spatial information is minimised [3, 4]. However, the presence of earplugs alters the natural resonance in the ear canals ([5]) and sound may leak through the earphones (particularly at low frequencies), all of which causes that hear-through audio shows 'colouration' with respect to natural hearing [2]. Therefore, to accurately replicate the acoustics of an open ear, the recorded signals must be equalised before mixing them with virtual audio [2, 3, 6-8].

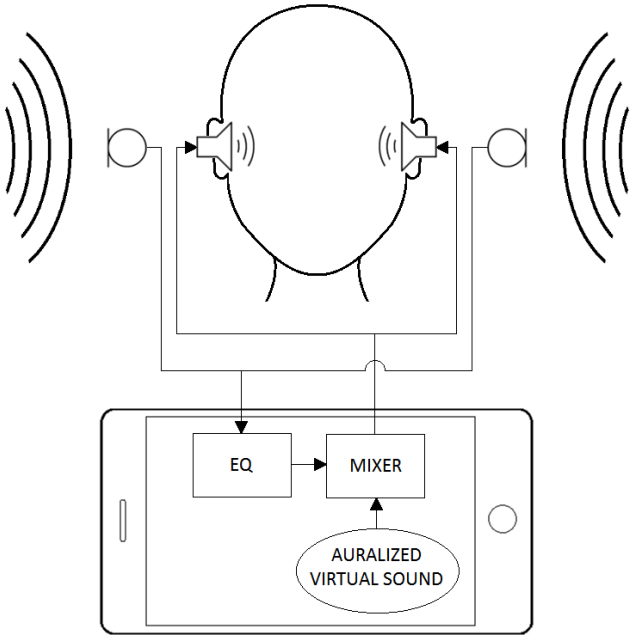

Figure 1: Smartphone-based AAR system, inspired by [1].

Tikander obtained positive results after testing an equalised hear-through function on a portable AAR system in real life situations, and observed some adaptation after the subjects wore the device for at least 1.5 hours [3]. Previous studies already suggested that humans can adapt to changes in their Head-Related Transfer Function (HRTF) if they are exposed to them for an extended period of time [9-11]. The goal of this study is to further explore the effect of long-term user adaptation to an AAR system built with off-the-shelf components (rather than custom-made ones), aiming at achieve perceived realism in an equalised hear-through function. Two individual case studies were performed where subjects used a AAR headset for several days while performing periodic control tests (localisation accuracy and subjective perception) to check the progress of the adaptation.

The following sections will outline the proposed AAR device, the adaptation test methods, and the results of the experiment. The paper ends with concluding thoughts and ideas for future work.

2. AAR system

2.1 System description

The proposed AAR system uses insert-type earphones and binaural microphones to implement a hear-through function, a concept that was first introduced in [1] and further developed in [2, 3]. Contrary to those approaches, which implemented an AAR mixer/equaliser with custom analog circuits, this device is smartphone-based and only uses commercial off-the-shelf components. Therefore, it is more accessible to the public and can be easily integrated with other applications, such as music listening, hands free calling or games, which is one of the main areas of improvement suggested by test subjects in previous studies [3, 8].

The components (see Fig. 2) are:

- A pair of Roland CS-10EM binaural earphones/microphones.
- A TASCAM iXJ2 audio interface, to power the microphones and provide stereo audio input.
- An Apple iPhone 5 smartphone running iOS 10.

Audio processing and routing is performed by a custom app based on the open source library AudioKit [16]. The app includes a sound equaliser, which consists of four second-order parametric filters that the user can adjust as needed. This approach was chosen over more complex solutions (e.g., a high-order FIR filter) to minimise in-out latency, as discussed in subsection 2.3.

2.2 Frequency response and equalisation

If an unfiltered hear-through function was implemented in the AAR system, the user would perceive a 'coloured' version of the acoustic environment which would sound unnatural. This is due to the ear canal's resonance being modified by the earphones, and to sound leakage (especially at the lower frequency spectrum) [2, 3, 5-8]. To accurately replicate the experience of natural hearing, the hear-through signal must be equalised. To find the right equalisation curve, frequency response was calculated by measuring Head-Related Impulse Responses (HRIR) on a dummy head in two scenarios:

1. Without the AAR system (from now on, *open-ear*).
2. With the AAR system running an unequalised hear-through function.

The result of subtracting the first curve from the second one gives the response of the audio filter that must be implemented in the iOS app. A third scenario was measured later, using the equalised hear-through function.

HRIRs (0° azimuth, 0° elevation) were measured using the sine sweep technique ([12]). An Equator Audio D5 speaker was used to play 4-second-long sweeps from 20 Hz to 20 kHz. It was placed approximately 2 meters away from the dummy head (G.R.A.S. Kemar head and torso simulator; see Fig. 2), at the same height and facing one another. The speaker input and the dummy head output were connected to a computer through a MOTU UltraLite Mk3 audio interface.

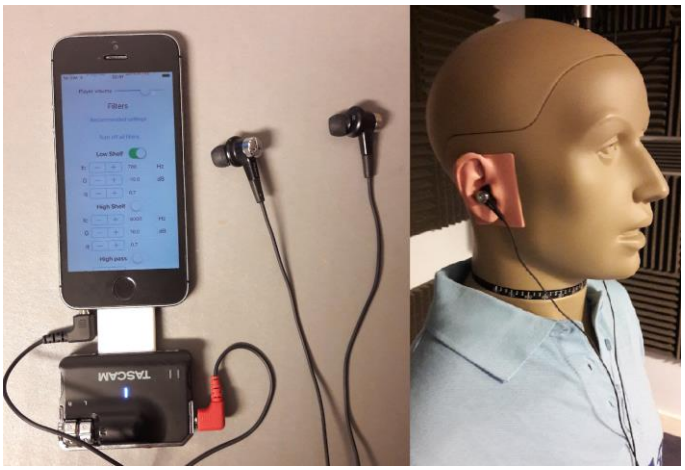

Figure 2: AAR system (left) and dummy head (right).

The filter to be implemented was not direction-dependent, so only frontal HRIRs were measured. For the sake of consistency, two different pairs of Roland CS-10EM were used, performing 10 HRIR measurements with each one in each scenario, for a total of 20 results per scenario, which were then averaged. Considering that the HRIR could vary depending on the earphone fit in the ear, earplugs were removed and reinserted again before each measurement in the hear-through scenario.

The equaliser in the iOS app was adjusted according to the results, and this configuration was used in the adaptation experiments. The same filter was used for left and right channel. Frequency response curves for unfiltered and equalised hear-through functions are shown in Fig. 3.

2.3 Latency and comb filter effect

Total in-out latency of the system is defined as the time lapse between the arrival of sound leaked through the earphone and its hear-through counterpart. This time was measured to be approximately 10 ms. A high latency could cause a phenomenon known as comb filter effect, which would add a 'metallic echo' to the perceived sound, potentially deteriorating the hearing experience [1, 2, 13]. Such issue could be mitigated by reducing latency and/or increasing earphone attenuation [13]. For the Roland CS-10EM, attenuation was measured to be in the range of 5-20dB for frequencies below 1 kHz, and 20-35dB for the rest of the spectrum, meaning that low frequencies were the most affected. Latency was minimised in the iOS app, albeit it being constrained by the phone's ana-

log/digital converters. In practice, it was found that comb filter effect was hard to eliminate completely but it could become barely noticeable when the earphones were well fitted inside the ear, achieving a higher attenuation and therefore less leakage. Ultimately, the subjects did not find it to be a problem during the field test.

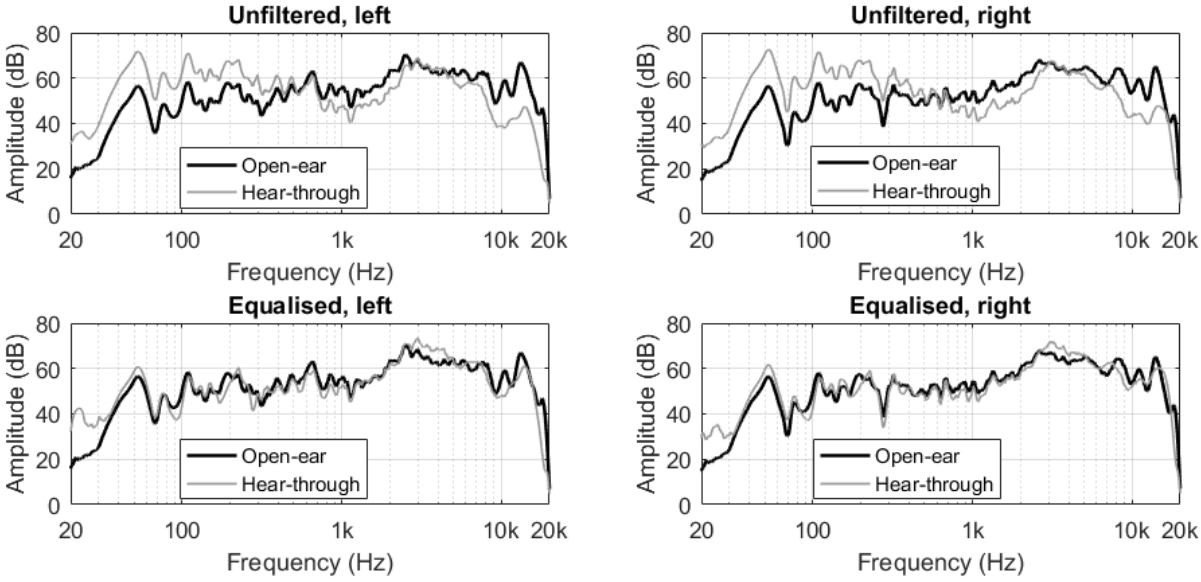

Figure 3: Comparison of open-ear frequency response against unfiltered and equalised hear-through.

3. Methodology

Two case studies were performed to test the hypothesis that it is possible for a person to adapt to hearing comfortably through an AAR system. Subjects wore the device for three days, between seven and eight hours per day, performing four daily control sessions: two in the morning, before and after putting the earphones in, and two in the evening, before and after removing them (see Table 1). The following control tests were used (both will be further discussed in later subsections):

1. A localisation test, to measure the subject's accuracy to localise sounds in the surroundings.
2. A subjective evaluation, to estimate changes in the subject's perception of pre-recorded auditory scenes.

Table 1: Test schedule. 'Loc.' stands for 'Localisation test', 'Subj.' stands for 'Subjective evaluation', 'Open' stands for 'Open-ear', and 'HT' stands for 'Hear-Through.

Day	Time		Test ID
1	Morning	Before inserting the earphones (baseline)	Open1 (1 Loc.+1 Subj.)
		After inserting the earphones (null adaptation)	HT1 (1 Loc.+1 Subj.)
	Evening	Before removing the earphones	HT2 (1 Loc.+1 Subj.)
		After removing the earphones	Open2 (1 Loc.+1 Subj.)
2	Morning	Before inserting the earphones	Open3 (1 Loc.+1 Subj.)
		After inserting the earphones	HT3 (1 Loc.+1 Subj.)
	Evening	Before removing the earphones	HT4 (1 Loc.+1 Subj.)
		After removing the earphones	Open4 (1 Loc.+1 Subj.)
3	Morning	Before inserting the earphones	Open5 (1 Loc.+1 Subj.)
		After inserting the earphones	HT5 (1 Loc.+1 Subj.)
	Evening	Before removing the earphones (max. adaptation)	HT6 (1 Loc.+1 Subj.)
		After removing the earphones	Open6 (1 Loc.+1 Subj.)
Total		Open-ear	6 Loc. + 6 Subj.
		Hear-through	6 Loc. + 6 Subj.

Considering the hypothesis, the subjects were expected to adapt to their modified HRTF ([9-11]), so a progressive evolution should be observed along the hear-through sessions, being the first one (HT1) the farthest away from open-ear results, and the last one (HT6) being the closest. On the other hand, all open-ear sessions were expected to obtain similar results, although it would be interesting to observe some degradation due to adaptation to the AAR device.

3.1 Equipment

A surround sound setup of eight Equator Audio D5 speakers was used, disposed as showed in Fig. 1 (positions 1-8) and operated by a computer through a Focusrite RedNet 2 audio interface. The subjects interacted with the system through a user interface made with Cycling ’74 Max 7.

3.2 Localisation test

This test evaluated how accurately could the subject pinpoint the location of surrounding sound sources. A total of 20 sources were defined around the subject (see Fig. 4). Sources 1-8 were the actual speakers, while 9-20 were virtual sources, generated by panning pairs of speakers (e.g., a sound played with the same loudness from speakers 1 and 2 appears to come from source 9).

The subject sat with his head at the same distance of the eight speakers and listened to a sequence of 40 sound samples. The goal was to guess which source was used each time by selecting it from a visual user interface. Sound samples were a combination of white noise and speech, and one second long.

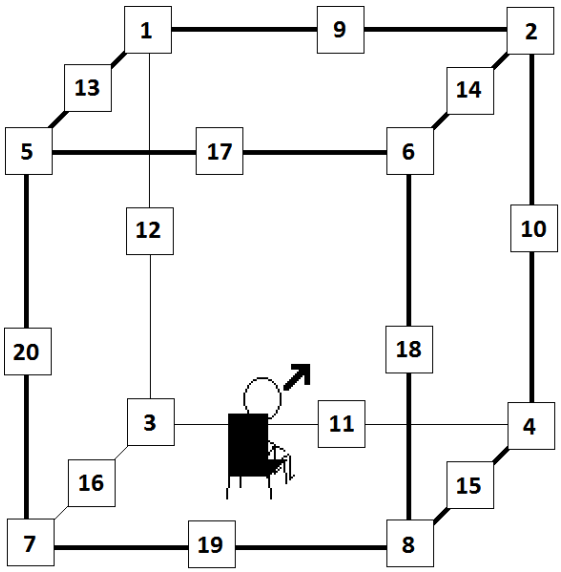

Figure 4: Distribution of sound sources around a subject sitting in the middle of the room. Sources 1-8 are actual speakers, while 9-20 are virtual sources generated by panning pairs of speakers.

The test was divided in two parts:

1. Part one: head movement was not allowed. All 20 sources were used once in a random order.
2. Part two: head movement was permitted to improve accuracy. Again, all 20 sources were used once, but in a different random order than before.

Results were analysed in terms of ratio of wrong answers, front-back confusions and up-down confusions. Both open-ear and hear-through scenarios were compared.

3.3 Subjective evaluation

The aim of this test was to evaluate the evolution of the subject’s spatial hearing perception when presented different auditory scenes during the adaptation process. The scenes were recorded with an Ambisonic microphone (Oktava MK-4012) and decoded to the eight-speaker surround sys-

tem. Two scenarios were used: a quiet office and a busy street in London. A total of 10 audio fragments (see Table 2), all 30 seconds long, were extracted from the recordings. They were selected to be similar to one another to minimise possible subject bias. In each test, all fragments were presented in a random order and the subject was asked to rate using a scale between 1 and 5 for each of the following attributes (inspired by the list proposed in [14], used to evaluate the quality of binaural recordings):

1. Brightness: "How bright the sounds feel, overall" (1 for *very dark* and 5 for *very bright*).
2. Externalisation: "Perception of sounds located outside your head" (1 for *inside the head* and 5 for *outside the head*).
3. Immersion: "Feeling of yourself being located in the middle of the audio scene" (1 for *not immersive at all* and 5 for *very immersive*).
4. Realism: "Feeling of sounds coming from real sources located around yourself" (1 for *very unrealistic* and 5 for *very realistic*).
5. Relief: "Feeling of distance between the closest sound objects and the farthest", (1 for *very compact* and 5 for *very spread out*).

Table 2: Characteristics of the audio fragments.

Audio fragments	Characteristics and location of sounds with respect to the microphone
office1-2	Two people talking (left), plastic bag noise (right), pen writing on paper (right).
office3-6	Two people talking (left), laptop playing classical music (back-right), plastic bag noise (right), mobile phone ringing (right), door slamming (front-right).
street1-4	Three people talking (back) vehicles passing by (all directions).

4. Results and discussion

4.1 Localisation test

Figure 5 shows the evolution of the localisation test results along the 6 sessions, in terms of ratio of wrong answers, ratio of front-back confusions, and ratio of up-down confusions, for both open-ear and hear-through situations.

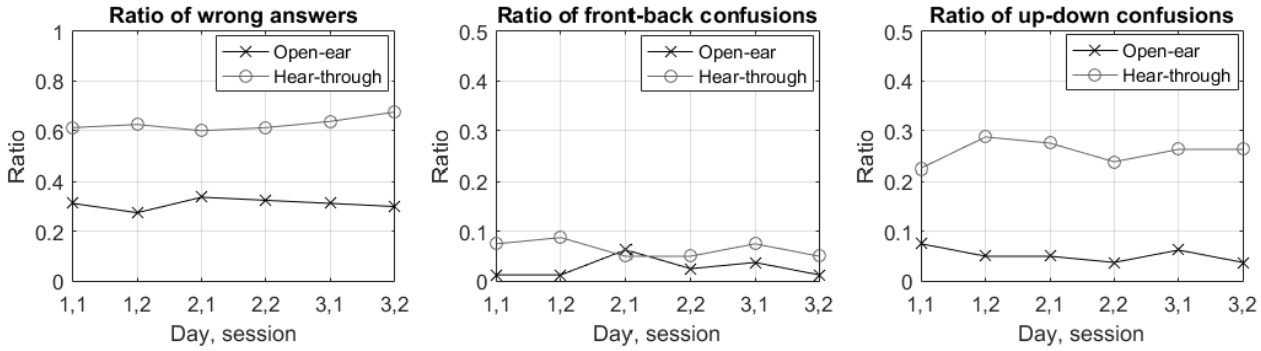

Figure 5: Results of the localisation test.

The hear-through condition showed a 30% increase in localisation error with respect to open-ear, meaning an overall poorer localisation performance. This result is in line with the findings of Marentakis and Liepens [15]. Up-down confusions were observed to be a major cause of errors when wearing the device, showing a 20% higher ratio than in the open-ear condition. In the case of front-back confusions, they were also found more frequent in hear-through than in open-ear, but the difference was less prominent (less than 8%). In general, the increase of elevation errors and number of front-back confusions in the hear-through condition was arguably due to the position of the microphones being slightly off the entrance of the ear canal, altering the subject's natural localisation cues. On the other hand, azimuth errors were found minimal, arguably because the device did not

cause a modification of the Interaural Time Differences (ITD) and Interaural Level Differences (ILD), which are dominant in lateral localisation.

Neither degradation or improvement were observed for open-ear and hear-through cases over the six sessions. The curves for both hearing conditions do not show a clear tendency to converge, therefore no evident signs of adaptation were found. However, further testing with more subjects and a longer process of adaptation could lead to more conclusive results.

4.2 Subjective evaluation

Figure 6 shows a summary of the results of the subjective evaluation, displayed as the evolution of the average score for each attribute in both the office scene (calculated as the mean of the scores for fragments office1-6), and the street scene (mean of the scores for fragments street1-4).

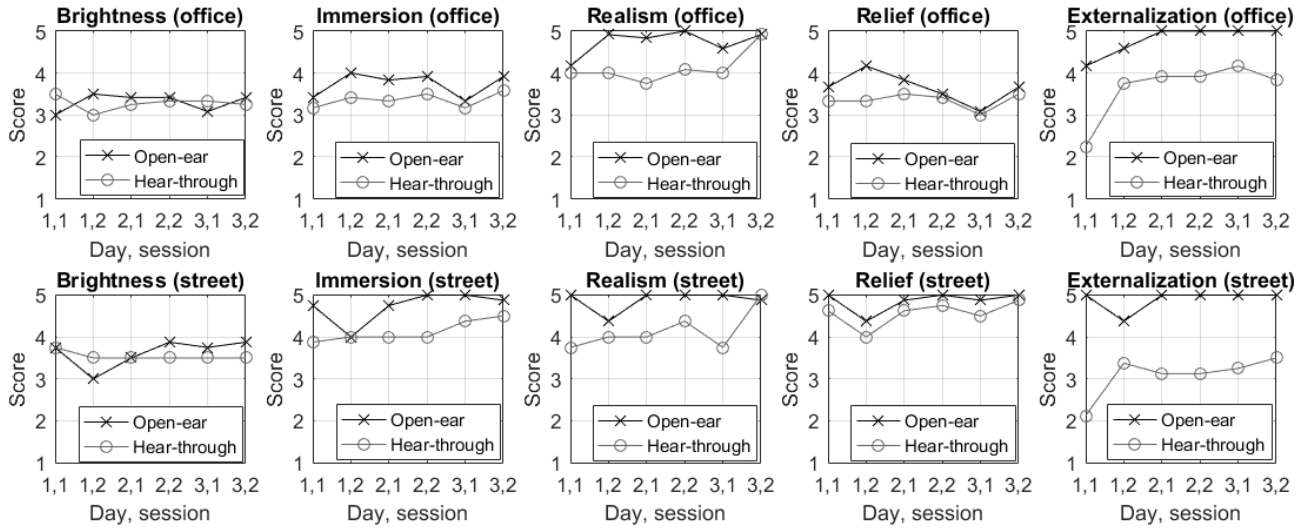

Figure 6: Results of the subjective evaluation. Average results of fragments office1-6 are displayed in the top row, and those of fragments street1-4 are shown in bottom row.

It was observed that wearing the AAR device had an immediate effect in the perception of all attributes except for *brightness*, which showed little variation between open-ear and hear-through. The most notable difference between the two hearing conditions was found in *externalisation*, with significantly lower scores in the hear-through case. In fact, subjects seemed to identify this attribute as the defining factor for the overall quality of the 'hearing experience', meaning that better externalisation often translated to more naturalness.

Adaptation was observed to happen to some extent in all attributes (again except *brightness*), showing signs of convergence in the hear-through and open-ear curves, particularly in the case of *realism*, which obtained equal scores for both hearing conditions in the last session. The most notable evolution happened in *externalisation* between the first and second sessions, for both 'street' and 'office' scenes. No evident signs of degradation over time were observed in the open-ear results, meaning that the subjects did not 'reject' their natural hearing condition during the adaptation process.

5. Conclusions

After an analysis of these early outcomes, slight effects of adaptation were observed overall, especially in the subjective evaluation. However, further experimentation is needed to obtain more conclusive results. This is a preliminary study and, in the future, further evaluations will be carried out, increasing the number of subjects and the duration of the field tests, in order to maximise the effects of long-term adaptation.

This work is the first stage of a long-term project that will involve research on other technologies related with Audio Augmented Reality. Approaches based on open-fitting headphones, bone conduction, and receiver-in-the-canal devices, will be considered for future experiments. On later stages, other factors like usability and ergonomics will be investigated for the different prototypes.

6. Acknowledgements

The authors want to thank Dr Chungeun Kim for his participation in this research, and for the always helpful input and discussions.

REFERENCES

1 Härmä, A., Jakka, J., Tikander, M., Karjalainen, M., Lokki, T., Hiipakka, J., and Lorho, G., *Augmented reality audio for mobile and wearable appliances*, J. Audio Eng. Soc., vol. 52, n. 6 (2004).

2 Riikonen, V., Tikander, M., and Karjalainen, M., *An augmented reality audio mixer and equalizer*, AES 124th Convention (2008).

3 Tikander, M., *Usability issues in listening to natural sounds with an augmented reality headset*, J. Audio Eng. Soc., vol. 57, no. 6 (2009).

4 Brungart, D. S., Kordik, A. J., Eades, C. S., and Simpson, B. D., *The effect of microphone placement on localization accuracy with electronic pass-through earplugs*, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (2003).

5 Rona, A., Cockrill, M., Picinali, L., Panday, P., Tripathi, R., and Ayub, M., *Vibro-acoustic response of tympanic-membrane-like models*, 22nd International Congress on Sound and Vibration (2015).

6 Hoffmann, P. F., Christensen, F., and Hammershøi, D., *Insert earphone calibration for hear-through options*, AES 51st International Conference: Loudspeakers and Headphones (2013).

7 Hiipakka, M., Takanen, M., Delikaris-Manias, S, Politis, A., and Pulkki, V., *Localization in binaural reproduction with insert headphones*, AES 132nd Convention (2012).

8 Albrecht, R., Lokki, T., and Savioja, L., *A mobile augmented reality audio system with binaural microphones*, Proceedings of Interacting with Sound Workshop (2011).

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

10 Van Wanrooij, M. M., *Relearning sound localization with a new ear*, J. of Neuroscience, vol. 25 (2005).

11 Mendonça, C., Campos, G., Dias, P., Vieira, J., Ferreira, J. P., and Santos, J. A., *On the improvement of localization accuracy with non-individual HRTF-based sounds*, J. Audio Eng. Soc., vol. 60 (2012).

12 Farina, A., *Advancements in impulse response measurements by sine sweeps*, AES 122nd Convention (2007).

13 Rämö, J., and Välimäki, V., *Digital augmented reality audio headset*, J. Electrical and Computer Engineering, vol. 2012 (2012).

14 Simon, L. S. R., Zacharov, N., and Katz, F. G., *Perceptual attributes for the comparison of head-related transfer functions*, J. Acous. Soc. of America, vol. 140 (2016).

15 Marentakis, G., and Liepins, R., *Evaluation of hear-through sound localization*, SIGCHI Conference on Human Factors in Computing Systems (2014).

16 AudioKit library. [Online.] available: http://audiokit.io