Characteristics Comparison of Two Audio Output Devices for Augmented Audio Reality

Kazuhiro Kondo* Naoya Anazawa* and Yosuke Kobayashi*
*Graduate School of Science and Engineering, Yamagata University, Yamagata, Japan
E-mail: {kkondo,ykobahashi}@yz.yamagata-u.ac.jp

***Abstract*—In this paper, we compared two audio output devices for augmented audio reality applications. In these applications, we plan to use speech annotations on top of the actual ambient environment. Thus, it becomes essential that these audio output devices will be able to deliver intelligible speech annotation along with transparent delivery of the environmental auditory scene. Two candidate devices were compared. The first output was the bone-conduction headphones which can deliver speech by vibrating the skull, while normal hearing is left intact for surrounding noise since these headphones leave the ear canal open. The other is the binaural microphone/earphone combo, which is in a form factor similar to a regular earphone, but integrates a small microphone at the ear canal entry. The input from these microphones can be fed back to the earphone along with the annotation speech. In this paper, we compared the speech intelligibility of speech when competing babble noise is simultaneously given from the surrounding environment. It was found that the bone-conduction headphones can deliver speech at higher intelligibility than the binaural combo. However, with the binaural combo, we found that the ear canal transfer characteristics were altered significantly by closing the ear canal with the earphones. If we employed a compensation filter to account for this transfer function deviation, the resultant speech intelligibility was found to be higher than the bone-conduction headphones. In any case, both of these are found to be acceptable as audio output devices for augmented audio reality applications since both are able to deliver speech at high intelligibility even when significant amount of competing noise is present.**

I. Introduction

Recent development in mobile terminal devices has allowed us to bring powerful computing devices on the road. For instance, we can carry powerful smart phones when we walk down the street, typically receiving directions to our destination, or receiving and reading emails. However, current devices give out most of this information in visual form, *i.e.* on displays. This creates a dangerous situation, where the user has his or her eyes on the tiny displays, and may miss cues for possible hazards, *e.g.* obstructions or automobiles coming out from the corner. Accordingly, we are attempting to provide most of this information using speech so that the user does not need to stare at the displays, and keep their eyes on the road. Normally, headphones or earphones are required to provide speech annotations. However, this creates another possibly hazardous situation since we also obtain cues for potential danger using our ears. For example, we may be aware of a motorcycle approaching from behind by hearing its engine, or we may hear a bicycle chime approaching. Thus, we need to keep listening to sound coming from around us at the same time as listening to the speech annotations from the mobile devices. Since we are adding speech in a virtual acoustic space onto an actual audio environment, this forms what we should call an augmented audio reality (AAR)[1], [2]. It is obvious that AAR requires investigation into other forms of audio output devices.

We have identified two possible candidate audio output devices for AAR applications. The first device is the bone-conduction headphones [3], [4] which provide audio output by vibrating the skull with an electromechanical vibrator. Since these headphones can leave the ear canal unobstructed, normal hearing of the environmental noise is left intact.

The other device is the binaural microphone/earphone combo [5]. These are devices that have the same form factor as regular inner-earphones, but have small microphones integrated at the other end facing outwards. The earphones close the ear canals, attenuating much of the environmental sound. However, the environmental sound can be recorded using the integrated microphones, and reproduced along with the added speech annotation. Notice that the microphones are integrated on to earphones on both the left and the right ear, so the environmental sound can be recorded and regenerated separately at both ears.

These devices have their pros and cons. In this paper, we compare the intelligibility of speech annotations when surrounding noise is present at various levels to find out how feasible these devices are in realistic acoustic environment. The noise used in these cases was babble noise, coming from speakers in one of the horizontal directions simulating a busy street. Under these conditions, it was found that both of these devices will show reasonably high speech intelligibility.

This paper is organized as follows. In the next chapter, characteristics of the audio output devices for AAR are described. In chapter III, the conditions for the speech intelligibility experiments are described, followed by the results and its observations in chapter IV. Finally concluding remarks and suggestions for further research is given in V.

II. Audio Output Devices for Augmented Audio Reality

In this section, two audio output devices which may be applied to AAR applications are described. Both devices are capable of delivering annotation speech along with the ambient noise, but its method of delivery is quite different. Both devices

have their strengths and weaknesses which requires careful study in order to make the best choice for AAR.

A. Bone-Conduction Headphones

Humans normally perceive audio through two parallel pathways: air-conduction and bone-conduction. Under normal circumstances, the former is dominant in auditory perception. However, bone-conduction has been utilized for audio communication under hazardous environments for some time. For example, bone-conduction has been used in the military to communicate with personnel who are under extreme amount of noise, and need to wear hearing protection gear. Since the ear canal needs to be completely sealed, bone-conduction was the logical choice of alternate means of auditory communication. Construction workers also have been using these devices for similar purposes.

Bone-conduction devices use transducers which vibrate the skull with the audio signal. The exact path and mechanism which humans perceive sound from these vibrations is still debated. However, it is generally said that much of the perceived sound comes from the vibrations which are converted to sound in the ear canal, while some comes from vibrations reaching the cochlea.

Previous bone-conduction devices suffered very low audio quality, with significant portion of the low frequency region attenuated, and resulting in a "muffled" quality [3]. However, recent improvements in the transducers have significantly improved the audio quality, even to a quality level almost compatible with normal acoustic headphones [4].

Bone-conducting vibration is generated by placing a vibrating transducer on typically the temple or the cheek bone. The quality of the perceived bone-conduction sound seems to differ significantly between individuals depending on how the shape of the transducers fits the listener's contact point, and at what pressure the transducers are applied. The quality also seems to differ for each individual each time the individual wears the bone-conduction device, depending on how well the transducers fit each time. This instability is one of the major drawbacks of this type of audio device.

Currently, it is quite difficult to physically measure the level or the quality of the delivered audio signal in a non-invasive manner. All we can do is to have the listener compare the perceived audio level and quality with normal air-conducted sound subjectively. This also makes the quantitative analysis of the performance of this device difficult if not possible.

B. Binaural Microphone/Earphone Combos

Binaural microphone/earphone combos have recently been manufactured by several vendors for binaural recordings [5], [6]. These devices were mainly targeted to audio hobbyist. Small microphones were placed on earphones, facing outwards at ear canal entry. The recording from these microphones allowed one to experience binaural recordings relatively inexpensively. The earphones were to monitor the recordings in real time, and its original purpose was secondary to the microphones.

On the other hand Härmä *et al.* have been prototyping similar devices. They have been crafting earphones with small microphones they extracted from noise canceling earphones. They devised an analog amplifier and filter for the signal obtained from the microphones in each ear, and fed these back to the earphones mixed with audio from virtual scenes.

We decided that the binaural microphone will serve the same purpose. We chose to use the finished product as is since these small devices were noise-prone, and needed to be housed in a stable chassis so that it will not pick up unwanted sounds, *e.g.* loose wiring rubbing on the chassis, or cross-talk noise, etc. The integrated microphone was found to be surprisingly high quality. All we needed to do was amplify this signal, mix them with speech annotation, and feed back to the earphones. However, we noticed that the fed back ambient noise had an altered quality which seemed somewhat more annoying than natural (*i.e.* heard with open human ears) sound. This alteration seems to be a combination of the microphone frequency characteristics, and the significant acoustic impedance alteration caused by closing the ear canal by the earphone, whereas in the natural state, the ear canals are completely open. Härmä *et al.* also noticed this, and applied a simple analog filter to compensate for this alteration. We will attempt this with a digital filter.

C. Compensation of Ear Canal Transfer Function Alteration by the Binaural Microphone/Earphone Combo

We need to compensate for the alteration of the acoustic impedance caused by the binaural microphone/earphone combo. Figure 1 shows a comparison of the spectrum for white noise recorded at the eardrum of one male subject using a probe microphone (Etymotic ER-7C). The binaural combo was the Roland CS-10EM. Spectrum for both natural (open ear) recording and sound reproduced using the binaural microphone/earphone combo with simple loop-back with flat amplification is shown. White noise was played out from a loudspeaker directly in front of the subject at about the height of the subject's ears (1140 mm above the floor) in all cases. Only the spectrum for the left ear is shown for natural recording since the characteristics for the left and right ear were essentially the same. The overall level difference between natural and binaural recordings was not compensated for. As can be seen, although the spectrum mostly matches above 1 kHz, there does seem to be some discrepancy at frequencies below. There is also almost no difference between left and right ear recording with the binaural combo.

We also measured and compared the characteristics for another subject. Unfortunately, some differences in the characteristics was seen by subject, most likely caused by the differences in the frequency characteristics of the pinna, as well as the individuality in the acoustic impedance change due to the earphone, caused by how well the earphones fit each subject. This obviously means personalization of the compensation filter is necessary. However, the measurement and configuration of the compensation filter is a tedious task. Thus, in the following experiments, we will be using the

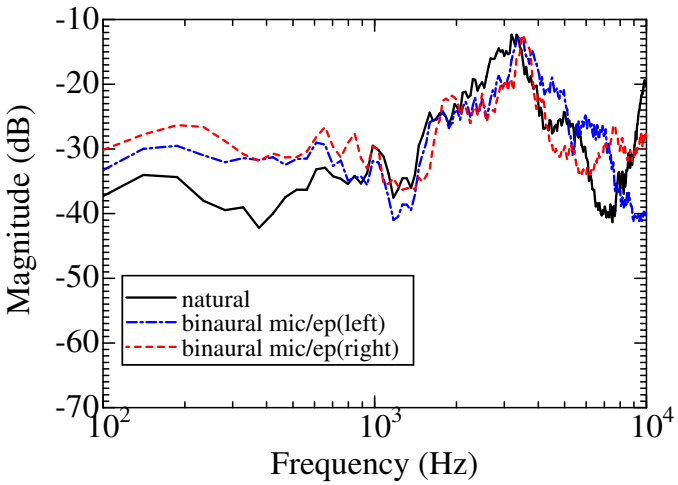

Fig. 1. Spectrum of White Noise Recorded at the Eardrum for Both Natural and Reproduced by Binaural Mic./EP Combo

compensation filter configured for one subject (not included in the evaluation). The personalization of the compensation filter and its affect on the intelligibility is an interesting and necessary topic, and will be investigated in the future.

Following the discussions above, we measured the impulse response of both natural hearing and the binaural microphone/earphone combo from the source to the ear drum. One human subject was used in this measurement. The source was played out from a loudspeaker (Bose Model 101 music monitor) located 1350 mm directly in front of the subject, approximately at the height of the subject's ear in a sitting position, which was about 1140 mm above the floor. The sound was recorded at the eardrum using the probe microphone. The waveform used to calculate the response was the Time-Stretched Pulse (TSP) signal which is basically a chirp signal, but is known to give better SNR than a conventional impulse signal [7]. A convolution of the recorded waveform with the synchronized time-reversed TSP signal gives the impulse response signal.

We measured the impulse response of the natural sound to the ear drum, $H_n(\omega)$ and the sound reproduced through a CS-10EM, $H_b(\omega)$. An FIR compensation filter, $H(\omega)$, with 50 taps (at sampling rate 44.1 kHz) that transfers the magnitude response of the CS-10EM to approximate the natural sound can be given as follows.

$$|H(\omega)| = \frac{|H_n(\omega)|}{|H_b(\omega)|} \tag{1}$$

The phase of this filter was set to a linear phase response.

We decided to implement this filter using the playrec Matlab toolkit [8] running on a dedicated computer for its quick prototyping capability. The playrec toolkit, along with the recent powerful computers, allows real-time filtering. Since playrec uses block processing (256 samples), processing delay corresponding to this block is added (approximately 6 ms), but since the filter is applied to ambient noise, we concluded that this delay will not affect the outcome.

TABLE I
JAPANESE PHONETIC TAXONOMY OF THE DRT.

Phonetic Taxonomy	Classification	Example
Voicing	Vocalic and non-vocalic	zai - sai
Nasality	Nasal and oral	man - ban
Sustention	Continuant and interrupted	hashi - kashi
Sibilation	Strident and mellow	jyamu - gamu
Graveness	Grave and acute	waku - raku
Compactness	Compact and diffuse	yaku - waku

Informal listening tests have shown that the compensated sound with the CS-10EM and the compensation filter is much more similar to the naturally heard sound compared to the uncompensated sound using the CS-10EM.

III. Speech Intelligibility Measurement Experiments

We measured and compared the annotation speech intelligibility in noise. Speech was presented using the bone-conduction headphone (TEAC Filltune HP-F200), or the binaural microphone/earphone combo (Roland CS-10EM), the latter with and without the compensation filter described in the previous section.

A. The Diagnostic Rhyme Test

The speech intelligibility was measured using the Japanese Diagnostic Rhyme Test (DRT)[9], [10]. The Diagnostic Rhyme Test (DRT) is a speech intelligibility test that forces the tester to choose one word that they perceived from a list of two rhyming words. The two rhyming words differ by only the initial consonant by a single distinctive feature. The features used in the DRT, following the definition by Jacobson, Fant and Halle [11], are voicing, nasality, sustention, sibilation, graveness, and compactness. A brief description of this definition along with an example word-pair is shown in Table I. Ten word-pairs per each of the 6 features, one pair per each of the five vowel context, were proposed for a total of 120 words [9]. The word-pairs are rhyme words, differing only in the initial phoneme.

The intelligibility is measured by the average correct response rate over each of the six phonetic features, or by the average over all features. The correct response rate should be calculated using the following formula to compensate for the chance level,

$$S = \frac{R - W}{T} \times 100[\%] \tag{2}$$

where S is the response rate adjusted for chance ("true" correct response rate), R is the observed number of correct responses, W the observed number of incorrect responses, and T the total number of responses. Since this test is a two-to-one selection test, a completely random response can be expected to result

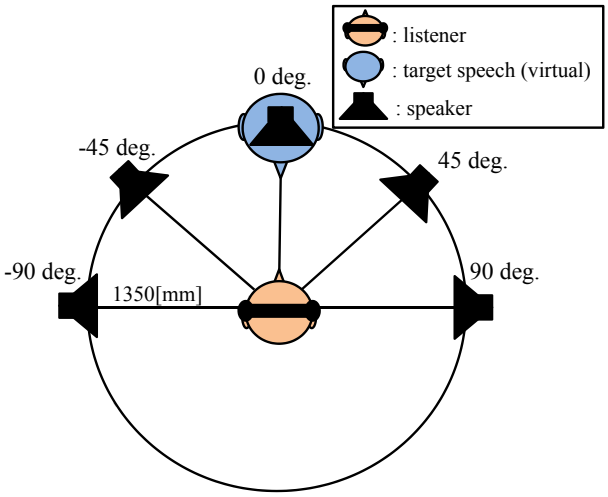

Fig. 2. Location of Sound Sources

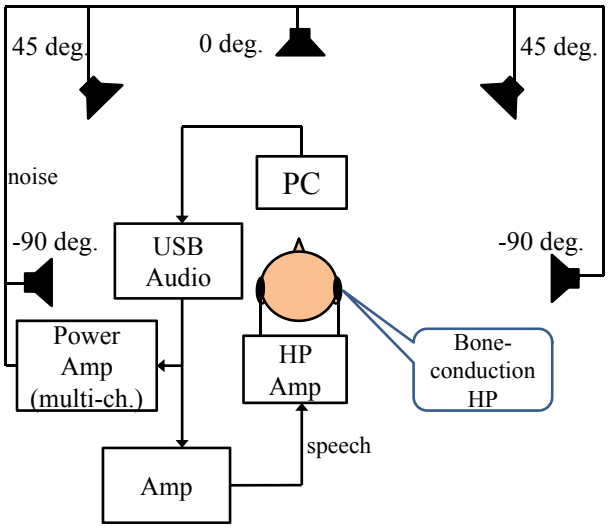

Fig. 3. Configuration of Speech Intelligibility Measurements using the Bone-Conduction Headphones

in half of the responses to be correct. With the above formula, a completely random response will give average response rate of 0%.

B. Experimental Conditions

We conducted the Japanese DRT test to measure the speech intelligibility when ambient noise is present. Ten subjects, all in their early twenties with normal hearing, participated and rated all samples. We used either the bone-conduction headphone (TEAC Filltune HP-F200) or the binaural microphone/earphone combo (Roland CS-10EM) to play the target DRT word speech, which in the actual applications corresponds to the speech annotation. All target speech samples, *i.e.* 120 DRT words, were read by one female speaker. The ambient noise was simulated using babble noise, and will be played out from one of the five loudspeakers (Bose model 101 music monitors) placed in front of the listener, at azimuths ± 90, ± 45, and 0°. The configuration of this experiment is shown in Fig. 2. The loudspeakers were all located in a circle with radius 1350 mm, and were at a height of 1140 mm from the floor, which is roughly the height of the listeners' ears in a sitting position.

Figure 3 shows the configuration of the experiment using the bone-conduction headphones. One controller PC will play out both the babble noise and target speech simultaneously at the appropriate timing. This PC will also log all responses (perceived word selection), input by the listener. The noise is output to a multi-channel audio interface (Edirol UA101), where only one randomly-chosen channel is actually fed the babble noise, and the rest of the channels are silent. Each of the output channels is connected to one of the five loudspeakers, and so the orientation of the noise output is switched randomly. The outputs of all loudspeakers were adjusted so that their levels become 54 dBA at the head location. This noise level is designated as 0 dB. Noise was also played out at half (-6 dB) or quarter (-12 dB) of this level at random. The target speech was convolved with the Head-Related Transfer Function (HRTF) measured with the KEMAR Mannequin, available from MIT [12] (large pinna). In all experiments described here, the target speech was localized at 0° azimuth and elevation, *i.e.*, directly in front. The localized target speech was fed to another amplifier, and fed to the HP-F200 at the same perceived level as the 0 dB noise. In other words, the level of the HP-F200 output was adjusted so that the listener perceived the same level as the output from the loudspeaker in front (0°). Pink noise was used in this level adjustment phase. Once the output levels are configured, the listener hears one of the 120 target words from the HP-F200, while simultaneously hearing babble noise coming from one of the loudspeakers in random (0, ± 45 and $\pm 90^\circ$) at one of the three levels (0, -6 and -12 dB) chosen in random. The listener selects one of the two words shown on the PC display in response. This cycle is continued until all samples are exhausted.

Figure 4 shows a similar configuration for the binaural microphone/earphone combo (Roland CS-10EM). The configuration of the loudspeakers is exactly the same. With the CS-10EM, however, the binaural microphone output (which records the noise) is fed to a stereo amplifier, and then mixed with the target speech. The amplified microphone output is also fed to the compensation filter (a dedicated PC) and mixed with the target speech. The relative level of the CS-10EM is also adjusted beforehand to match the loudspeaker output using pink noise. After the level configuration, the listener goes through two cycles of evaluation, one with the compensation filter, and one without.

IV. Results and Discussions

Figures 5, 6 and 7 show speech intelligibility for each of the output device at SNR 0, -6 and -12 dB, respectively. The error bars in these figures show the 95% confidence intervals. In most of these figures, there is a dip in the intelligibility for

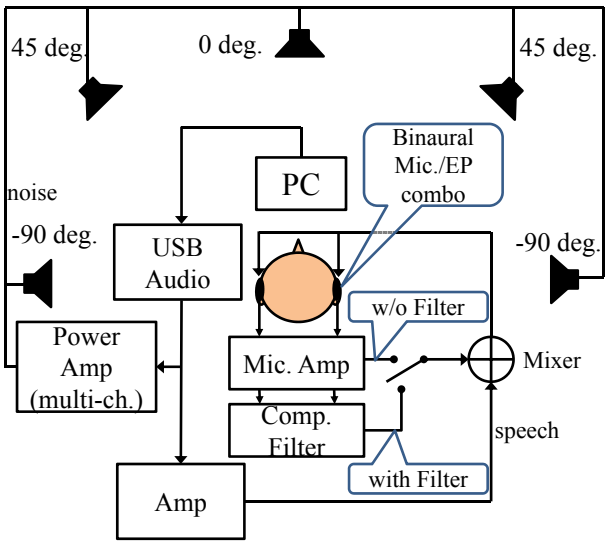

Fig. 4. Configuration of Speech Intelligibility Measurements using the Binaural Microphone/Earphone Combo

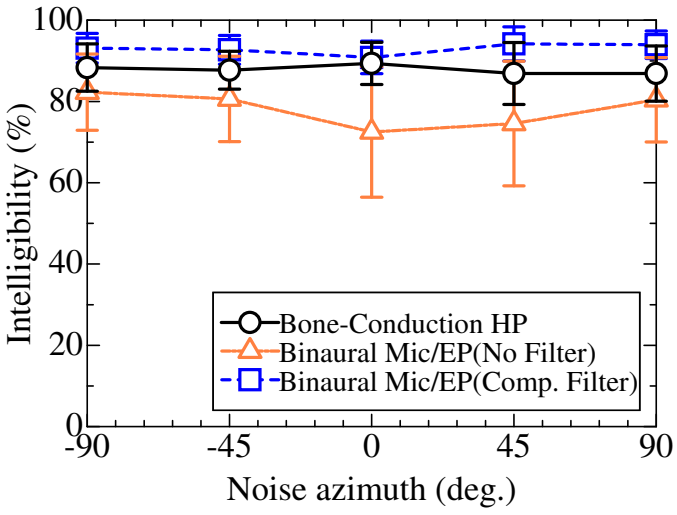

Fig. 5. Noise Azimuth vs. Intelligibility for Various Output (SNR 0 dB)

noise at 0° azimuth, which is expected since the target speech is also localized at this angle, and so is masked the most at this angle. As the noise moves away from the target speech, the intelligibility improves. This is much more visible at lower SNRs.

At all SNRs, intelligibility with the bone-conduction headphone (HP-F200) is higher compared to the binaural microphone/earphone (CS-10M), without the compensation filter, but is slightly lower than the CS-10M with the filter. This is more apparent at lower SNRs. Thus, it seems that the compensation significantly helps the intelligibility of the target speech. It seems that without this filter, the essential frequency range (1 to 2 kHz) is emphasized by the ear canal characteristics alteration, and tend to mask the speech at a higher level. The compensation filter seems to de-emphasize this region and help lower the masking efficiency of the noise.

The HP-F200 shows slightly lower intelligibility than CS-10EM with the filter. This can be attained to the frequency characteristics of the bone-conduction path of the HP-F200, which is known to have poor low frequency range gain [3], and result in somewhat "muffled" quality speech, which may lower the intelligibility, with or without competing noise. However, it should be noted that the sound quality of HP-F200 has improved compared to older bone-conduction headphones, to a quality level almost equal to regular air-conduction headphones.

In any case, both the HP-F200 and CS-10EM show high intelligibility, above 70% in most cases (above 80% for the CS-10EM with filter in most cases). This is even true at SNR -12 dB, which is quite noisy. Thus, it seems both of these are well over acceptable quality for AAR applications in realistic acoustic environments in terms of the annotation speech intelligibility.

The CS-10EM, which can potentially deliver higher quality speech, needs additional hardware for the compensation filter for ambient noise, which can be expensive. The use of lower quality compensation filter with simplified hardware may compromise the intelligibility. A good balance between intelligibility and hardware complexity may need to be investigated.

On the other hand, the HP-F200 does not require this additional hardware, but still suffers from somewhat inferior quality speech. However, novel transducers with higher quality are constantly being manufactured, and this soon may not be a problem. We have seen that the quality of the delivered speech does have some individuality, *i.e.*, some users enjoy high quality while some users suffer lower "muffled" quality. This seems to be dependent on how well the transducers fit and make good contact with the skin at the temple. Also, in order to make good contact, the transducers need to be applied using some pressure, which some users reported as uncomfortable, especially when worn for a long period. Some ergonomic design may be in order here.

V. Conclusion

We compared two audio devices for augmented audio reality (AAR) applications, for example mobile audio navigation systems. In these applications, speech annotation needs to be delivered at high speech intelligibility, while the ambient noise also needs to be delivered since the noise can give cue to potential hazards such as an automobile approaching. We compared the bone-conduction headphones, which deliver audio by vibrating the skull with a transducer placed at the temple or the cheek bone, and the binaural microphone/earphone combo, which is an earphone with a tiny microphone at the ear canal entry. The ambient noise picked up with the microphone can be fed back to the earphone to reproduce the ambient environment. It was observed that the acoustic impedance change with the earphones change the quality of the ambient noise, and a compensation filter to equalize the impedance change is required.

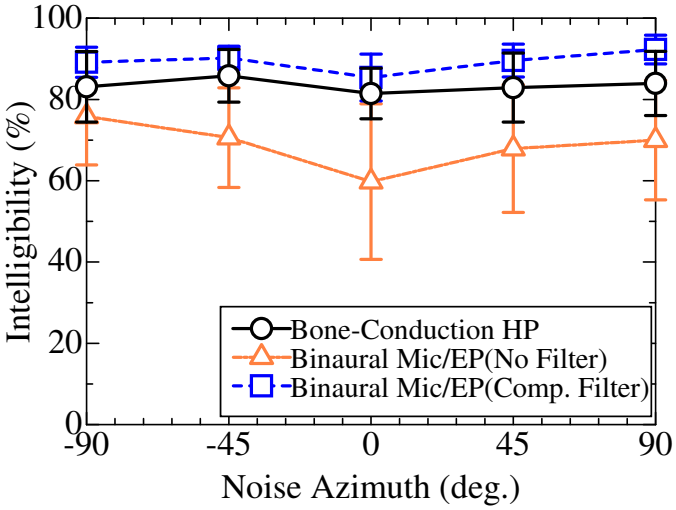

Fig. 6. Noise Azimuth vs. Intelligibility for Various Output (SNR −6 dB)

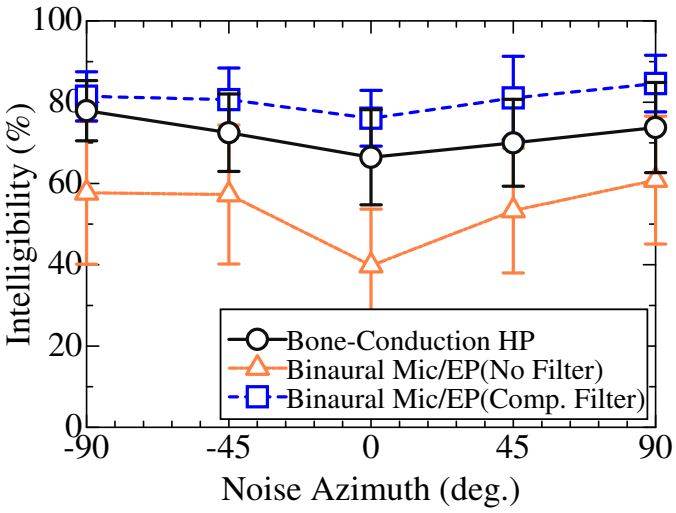

Fig. 7. Noise Azimuth vs. Intelligibility for Various Output (SNR −12 dB)

We played word speech localized from the front, and played babble noise from one of the five locations towards the front to simulate ambient noise commonly seen in the real environment. Speech intelligibility was measured in this configuration. It was found that the bone-conduction headphones show significantly higher intelligibility compared to binaural microphone/earphone combos without compensation filters, but slightly lower intelligibility than the binaural combos with the filters. However, both the bone-conduction headphone and the binaural combo showed relatively high intelligibility, above 70% in most cases even with significant amount of noise. Thus, we conclude that both of these output is applicable for AAR applications.

We still may need to confirm how well the localization of the ambient environment is preserved with the binaural combos with the compensation filters. Accurate localization is crucial since the whole purpose of feeding back the ambient noise is to give cues to the location of the hazards, as well as its severity.

We would also like to implement an actual AAR system with one of the acoustic output device, and do a field trial or test. With the binaural combo, the compensation filter, which we have shown is required, needs to be made into a much more compact form. Perhaps a battery-operated implementation using FPGAs or other small-factor programmable devices is needed. On the other hand, the bone-conduction headphones need to be improved for comfort since these devices need to be worn for a long time for realistic field trials.

ACKNOWLEDGMENT

H. Yagyu (now with Tohoku University) and T. Kanda (formerly with Tohoku University, now with Rion Co. Ltd.) contributed to the earlier portion of this work with bone-conduction headphones. The authors also thank many anonymous subjects who tolerated the long intelligibility evaluation sessions.

REFERENCES

[1] J. Rozier, K. Karahalios and J. Donath"Hear&There: An Augmented Reality System of Linked Audio," *Proceedings of the International Conference on Auditory Display,* Atlanta, GA, April, 2000.

[2] A. Härmä, J. Jakka, M. Tikander, and M. Karjalainen, "Augmented Reality Audio for Mobile and Wearable Appliances," *Journal of the Audio Engineering Society,* vol. 52. no. 6, June, 2004, pp.618-639.

[3] J. MacDonald, P. Henry, and T. Letowski, "Spatial audio through a bone conduction interface," *International Journal of Audiology,* vol. 45, 2006, pp.595-599.

[4] TEAC Filltune Bone-Conduction Headphones HP-F200, http://www.teac.jp/product/hp-f200/. In Japanese only.

[5] Roland Binaural Microphone/Earphone CS-10EM, http://www.roland.com/products/en/CS-10EM/.

[6] Adphox binaural microphone & earphone BME-200, http://www.adphox.co.jp/microphone/sound-eng.html.

[7] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," Journal of the Acoustical Society of America, vol.97, no. 2, 1995, pp.-1119-1123.

[8] R. Humphrey, "Playrec: Multi-Channel Matlab Audio," http://www.playrec.co.uk/.

[9] M. Fujimori, K. Kondo, K. Takano, and K.Nakagawa, "On a revised word-pair list for the Japanese intelligibility test," Proceedings of the International Symposium on Frontiers in Speech and Hearing Research, 2006.

[10] K. Kondo, "Subjective Quality Measurement of Speech - Its Evaluation, Estimation and Applications," Heidelberg: Springer, 2012.

[11] R. Jakobson, C. G. M. Fant, and M. Halle, "Preliminaries to speech analysis: The distinctive features and their correlates," Acoustics Laboratory, MIT, Technical Report, 13, 1952.

[12] B. Gardner and K Martin, "HRTF Measurements of a KEMAR Dummy-Head Microphone," http://sound.media.mit.edu/resources/KEMAR.html, 2000.