PAPER

Comparison of Output Devices for Augmented Audio Reality

Kazuhiro KONDO[†a)], ***Member***, **Naoya ANAZAWA**[†*], ***Nonmember*****, *and*** **Yosuke KOBAYASHI**[†**b)], ***Member***

SUMMARY 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 are 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 headphone, which can deliver speech signals by vibrating the skull, while normal hearing is left intact for surrounding noise since these headphones leave the ear canals 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 earphones along with the annotation speech. We also compared these devices to normal hearing (*i.e.*, without headphones or earphones) for reference. We compared the speech intelligibility when competing babble noise is simultaneously given from the surrounding environment. It was found that the binaural combo can generally deliver speech signals at comparable or higher intelligibility than the bone-conduction headphones. However, with the binaural combo, we found that the ear canal transfer characteristics were altered significantly by shutting the ear canals closed with the earphones. Accordingly, if we employed a compensation filter to account for this transfer function deviation, the resultant speech intelligibility was found to be significantly higher. However, both of these devices were found to be acceptable as audio output devices for augmented audio reality applications since both are able to deliver speech signals at high intelligibility even when a significant amount of competing noise is present. In fact, both of these speech output methods were able to deliver speech signals at higher intelligibility than natural speech, especially when the SNR was low.

key words: *augmented audio reality, speech intelligibility, mobile audio navigation, bone-conduction headphone, binaural microphone/earphone*

1. 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 or tablets 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 small video displays. This creates a potential hazardous situation, where the user has his or her eyes on extremely small displays on these devices, and may miss cues for possible hazards, *e.g.*, obstructions, automobiles coming out from the corner, bicycles passing by. Accordingly, in order to avoid these hazardous situations, we are attempting to provide most of the information using localized audio, mostly localized speech, so that users do not need to stare at the displays, and keep their eyes on the road. Localization of speech signals is being considered here so that the direction of the speech signal may provide additional cues. For instance, speech signals may be localized towards the point of interest (POI), thereby drawing the attention of the user towards the direction of the POI. Normally, headphones or earphones are required to provide audio and speech annotations. However, this creates another possible 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, the surrounding sound needs to be kept intact, while simultaneously the speech annotations from the mobile devices are played out. Since we are adding speech signals in a virtual acoustic space onto the actual ambient audio environment, this forms what we should call an augmented audio reality (AAR) [1]–[3] environment. It is obvious that AAR systems, especially mobile AAR systems, require investigation into alternate forms of audio output devices.

We have also been investigating the feasibility of AAR systems for mobile audio systems for both pedestrians [4], [5] and cyclists [6], [7]. Both of these applications require ambient noise to be delivered intact so that the users may avoid potential hazards. We have identified two possible candidate audio output devices for these applications. These two devices are the only devices we are aware of to date which are stable, readily available, and practically applicable to AAR systems. However, more advanced devices are constantly being developed, and we plan to compare these devices as well once they become available.

The first device is the bone-conduction headphone [8], [9] which provides 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. It has recently been announced that the much anticipated Google glass will also incorporate bone-conduction headphones as audio output for their applications [10]. Obviously, Google employed this form of audio delivery to implement augmented audio reality applications, although they have not made any announcements of their intentions.

The other device is the binaural microphone/earphone combo [11], [12]. These are devices that have the same form

Manuscript received February 12, 2014.
Manuscript revised April 11, 2014.

[†]The authors are with the Graduate School of Science and Engineering, Yamagata University, Yonezawa-shi, 992–8510 Japan.
[*]Presently, with Hitachi Advanced Digital, Inc.
[**]Presently, with Miyakonojo National College of Technology.
a) E-mail: kkondo@yz.yamagata-u.ac.jp
b) E-mail: ykobayashi@cc.miyakonojo-nct.ac.jp

DOI: 10.1587/transinf.E97.D.2114

factor as regular earphones (in-ear headphones), but have small microphones integrated at the end facing outwards. The earphones seal the ear canals shut, 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 onto the earphones on both the left and the right ear. Therefore, the environmental sound can be recorded and regenerated separately at both ears. Härmä *et al.* [2], [13] as well as some other attempts to implement this type of audio delivery have been reported, with favorable results towards its application to AAR. In a related research, Mori *et al.* have been developing a "smart" hearing aid which uses the binaural microphone/earphone combo to collect mixed speech, and then select and enhance the speech from the target speaker while suppressing other speakers [14]. Their goal is to actively modify the acoustic environment, thereby allowing the speech that the user is attempting to listen to easier to hear. On the other hand, our goal is to maintain the acoustic environment as transparent as possible, while mixing this environment with speech from a virtual environment.

These two types of devices have their pros and cons. The purpose of this paper is to compare the intelligibility of speech annotations when the surrounding noise is present at various levels to find out how feasible these devices are in realistic acoustic environments. As far as we know, no comprehensive comparison tests have been conducted to date with the same conditions, with the purpose of investigating the feasibility of these devices to AAR applications. The noise used in these cases was babble noise, coming from loudspeakers in one of the horizontal directions simulating a busy street. Babble noise is assumed to come from one direction simulating speech signals coming from a group chatting on the street in the direction of the user. This is probably an extreme case where speech signals are masked by a localized directional noise. Most other situations will be less extreme, with a mixture of other types of noise coming from a less localized omnidirectional source due to the reverberations. We previously measured the speech intelligibility with such omnidirectional noise and found that the intelligibility is not significantly different from a directional noise source as long as the signal-to-noise ratio is kept constant, and except in cases where the noise and target speech directions completely match [15]. Thus, directional noise sources represent a much more critical environment. Thus, we will be using these noise sources as the worst case scenario. With these directionally localized noise sources, it was found that both of these devices will show reasonably high speech intelligibility.

This paper is organized as follows. In the next section, characteristics of the audio output devices for AAR are described. In Sect. 3, the conditions for the speech intelligibility experiments are described, followed by the results and its observations in Sect. 4. Finally concluding remarks and suggestions for further research is given in Sect. 5.

2. 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 investigation in order to make the best choice for AAR systems.

2.1 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 directly.

Previous bone-conduction devices suffered extremely low audio quality, with a significant portion of the low frequency region attenuated, and resulting in a "muffled" quality [8]. However, recent improvements in the transducers have significantly improved the audio quality, even to a quality level almost compatible with normal acoustic headphones [9]. Accordingly, products have become available for joggers and walkers who want to enjoy music while working out.

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

Currently, it is extremely 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 quan-

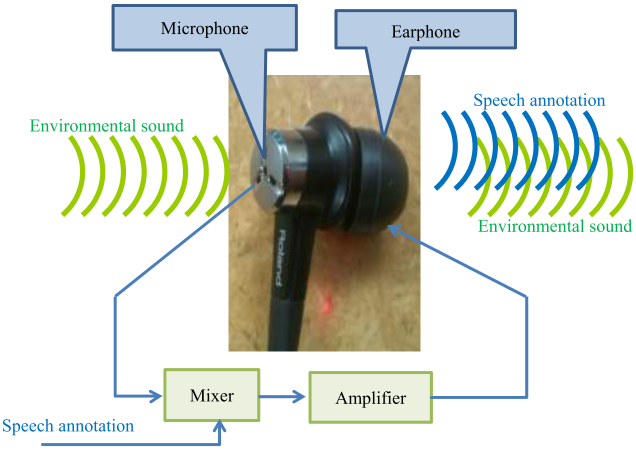


Fig. 1 Configuration of the AAR system using the binaural microphone/earphone combo.

titative analysis of the performance of this device difficult if not possible.

2.2 Binaural Microphone/Earphone Combos

Binaural microphone/earphone combos have recently been manufactured by several vendors for binaural recordings [11], [12]. These devices were mainly targeted to audio hobbyist. Small microphones were placed on the earphones, facing outwards at the ear canal entry. The recording from these microphones allowed one to experience binaural recordings relatively inexpensively. The earphones were primarily used to monitor the recordings in real time, and its original purpose was secondary to the microphones.

On the other hand, Härmä *et al.* [2], [13] have been implementing similar prototypes. 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 (Roland CS-10EM) 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 crosstalk noise. The integrated microphone was found to be surprisingly high quality. All we needed to do was to amplify this signal, mix them with speech annotation, and feedback to the earphones. Figure 1 shows this configuration. 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. Hiipakka *et al.* [16] tried to emulate the outer ear characteristics and measure the transfer characteristics difference caused by the earphones. Härmä *et al.* also noticed this alteration, and applied a simple analog filter to compensate for this alteration. We will attempt this compensation with a digital filter.

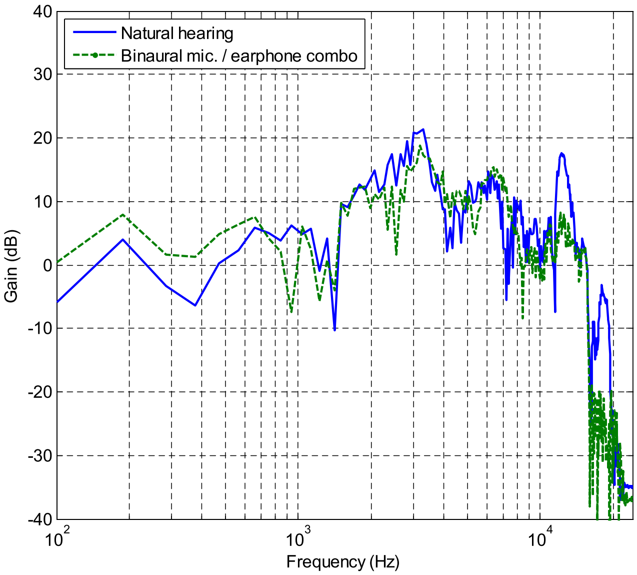


Fig. 2 Spectrum of white noise recorded at the eardrum for both natural hearing and reproduced by the binaural microphone/earphone combo.

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

Thus, we need to compensate for the alteration of the acoustic impedance caused by the binaural microphone/earphone combo according to the discussion in the previous section. Figure 2 shows a comparison of the spectrum for white noise recorded at the eardrum using a probe microphone (Etymotic ER-7C). These microphones measure the sound pressure at the eardrums using a small-diameter probe placed close to the eardrums. The binaural combo used here 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. The overall level difference between these recordings was not compensated for. As can be seen, although the spectrum mostly matches above 750 Hz, there does seem to be some discrepancy at frequencies below.

Thus, we measured the impulse response of both natural hearing, $H_n(\omega)$ and the binaural microphone/earphone combo from the source to the ear drum, $H_b(\omega)$. 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 [17]. A convolution of the recorded waveform with the synchronized time-reversed TSP signal gives the impulse response signal.

We also measured and compared the characteristics for another subject. Unfortunately, some differences in the characteristics were seen by subject, most likely caused by the individuality of the acoustic impedance change due to the earphone, depending on how well the earphones fit each subject. This obviously means the 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 compensation filter configured for one subject (not included in the evaluation). The personalization of the compensation filter and its effect on the intelligibility is an interesting and necessary topic, and will be investigated in the future.

An FIR compensation filter, $H(\omega)$, that transfers the magnitude response of the CS-10EM to approximate the natural sound can be given as follows.

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

$H_n(\omega)$ and $H_b(\omega)$ both include the transfer characteristics from the source to the ear canal entry, as well as the ear canal to the ear drum. Thus, by normalizing $H_n(\omega)$ by $H_b(\omega)$, $H(\omega)$ should only have the inverse characteristics of the CS-10EM, independent of the other transfer characteristics.

Since the mismatch between $H_n(\omega)$ and $H_b(\omega)$ was below 750 Hz, as we have stated, we decided to use compensation on components below this frequency, and use a flat gain above. Figure 3 shows the frequency characteristics of this compensation filter. The phase of this filter was set to a linear phase response. In all experiments, the filter was implemented with 50 taps (at sampling rate 44.1 kHz) for a balance between complexity and filter characteristics.

We decided to implement this filter using the playrec Matlab toolkit [18] 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 in our case), processing delay corresponding to this block is added (approximately 6 ms). However, since the filter is applied to ambient noise, we concluded that this delay will not affect the overall outcome.

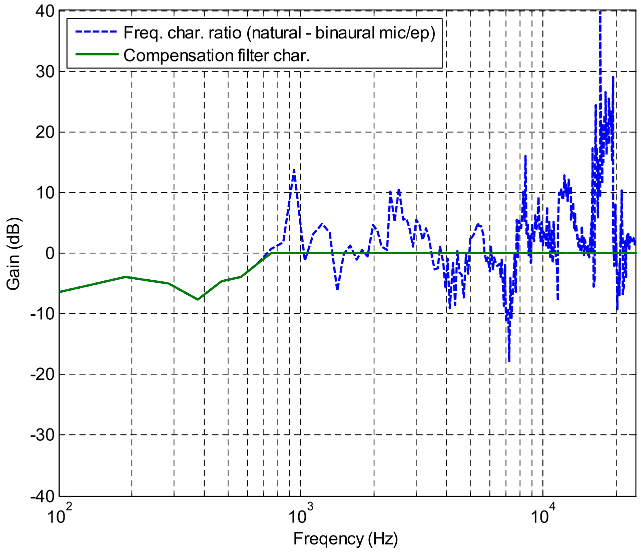


Fig. 3 Frequency characteristics of the compensation filter for the binaural microphone/earphone combo.

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.

3. Speech Intelligibility Measurement

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

3.1 The Diagnostic Rhyme Test

The speech intelligibility was measured using the Japanese Diagnostic Rhyme Test (DRT) [19], [20]. The 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 *et al.* [21], are voicing, nasality, sustention, sibilation, graveness, and compactness. A brief description of this definition along with an example word-pair is shown in Table 1. Ten word-pairs per each of the six features, one pair per each of the five vowel context, were proposed for a total of 120 words [19]. The word-pairs are rhyming 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,

$$S = \frac{N_r - N_w}{N_t} \times 100[\%] \tag{2}$$

Table 1 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

where S is the response rate adjusted for chance (*i.e.*, "true" correct response rate), N_r is the observed number of correct responses, N_w the observed number of incorrect responses, and N_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 in half of the responses to be correct. With the above formula, a completely random response will give an average response rate of 0%.

3.2 Experimental Conditions

We conducted the Japanese DRT test to measure the speech intelligibility when ambient noise is present. Seven 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 application corresponds to the speech annotation. All target speech samples, *i.e.*, 120 DRT words, were read by one female speaker. The target speech was localized at 0° azimuth and on the horizontal plane (0° elevation) by convolving each sample with the corresponding Knowles Electronics Manikin for Acoustics Research (KEMAR) Head Related Transfer Function (HRTF) from MIT [22]. The HRTF for the left pinnacle was used. For the right ear, the angles were mirrored and the same HRTF for the left ear was used, as suggested by the accompanying documentation [23]. This is because the original KEMAR measurements used a different size pinna on each ear. Since the KEMAR mannequin is artificial, its characteristics were assumed to be mostly symmetrical. Thus, the measurement for one of the ears for one of the hemisphere can be used for the opposing ear in the opposing hemisphere. 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°. As stated in the introduction, the localized noise is used here to evaluate a worst-case scenario of the effect of noise on the target speech, as opposed to an omnidirectional noise source. The configuration of this experiment is shown in Fig. 4. 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. Thus, all sound sources, including the target speech was located on the same horizontal plane as the listeners' ears.

3.2.1 Experimental Setup with the Bone-Conduction Headphone

Figure 5 shows the configuration of the experiment using the bone-conduction headphones. One controller PC played out both the babble noise and target speech simultaneously at the appropriate timing. This PC also logs all responses (perceived word selection), input from 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 were kept 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 HRTF measured with the KEMAR Manikin, available from MIT [22] (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 per-

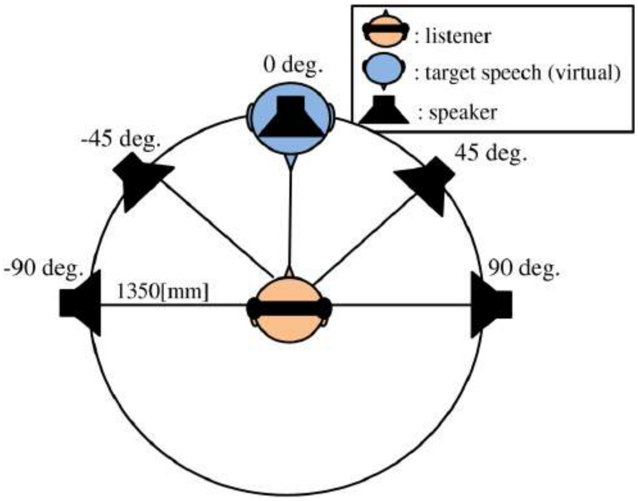


Fig. 4 Location of sound sources.

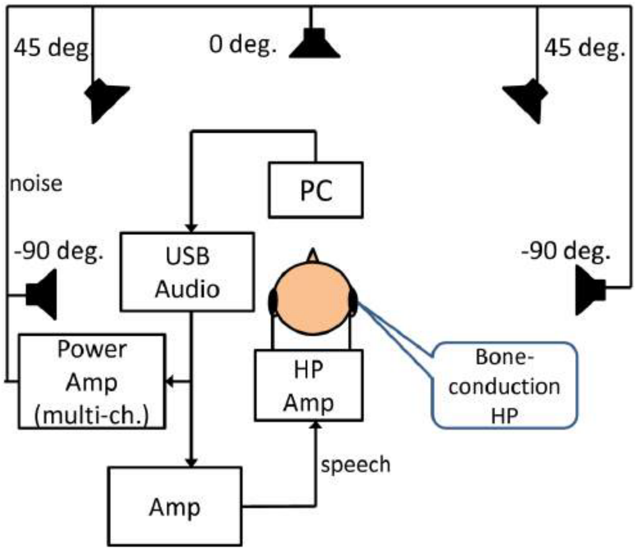


Fig. 5 Configuration of speech intelligibility measurements using the bone-conduction headphones.

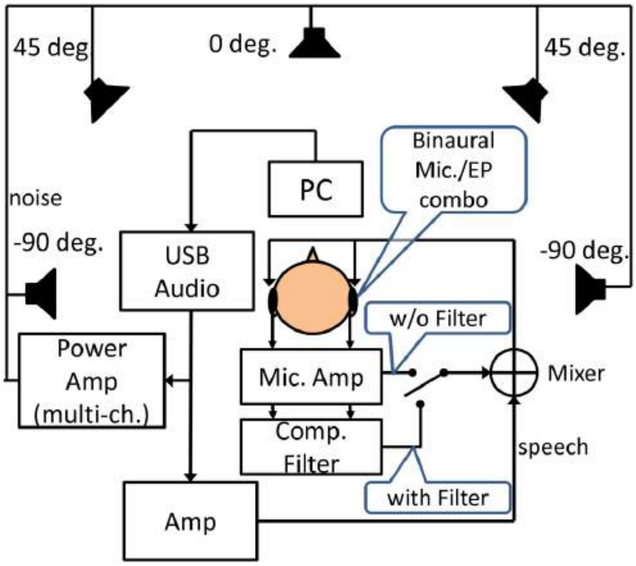


Fig. 6 Configuration of speech intelligibility measurements using the binaural microphone/earphone combo.

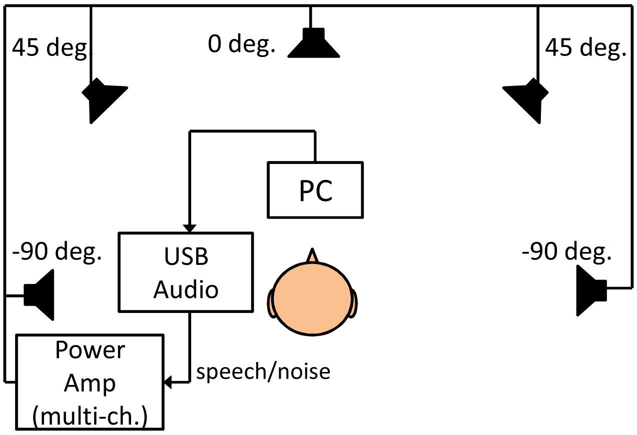


Fig. 7 Configuration of speech intelligibility measurements using loudspeakers for both speech and noise.

ceived 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 at random (0, ±45 and ±90°) at one of the three levels (0, −6 and −12 dB) chosen at 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.

3.2.2 Experimental Setup with the Binaural Mic./Earphone Combo

Figure 6 shows a similar configuration for experiments with the binaural microphone/earphone combo (Roland CS-10EM). The configuration of the loudspeakers is exactly the same as previously stated. 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. For experiments with the compensated ambient noise, 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 naturally heard level from the loudspeaker using pink noise. This level adjustment was made for both the ambient noise feedback and the target speech output. After the level configuration, the listener goes through two cycles of evaluation, one with the compensation filter, and one without.

3.2.3 Experimental Setup Naturally-Heard Speech from Loud Speakers

We have also reproduced experimental results for natural speech from [24] for comparison. In this experiment, speech intelligibility of speech signals reproduced from the loudspeakers was measured. Since speech signals were played out from the loudspeaker that is set in the actual direction of the source, no localization (convolution with HRTF) is necessary here. The target speech is played out simultaneously with the competing noise, also played out from one of the loudspeakers. This configuration is shown in Fig. 7. The purpose of including these conditions was to find out how the intelligibility of localized speech played from headphones compare to intelligibility of naturally heard speech, without the use of headphones or earphones. Note that the number of subjects in these experiments was five. The speech level played out from the loudspeakers was adjusted to be comparable to the level played out from the headphones.

In all three configurations (Figs. 5, 6 and 7), measurements were conducted in a sound-proof room with some reverberation control. The reverberation was controlled with rock wool padding on all walls, and rugs on the floors. The ceiling was not acoustically treated.

4. Results and Discussions

Figures 8, 9 and 10 show speech intelligibility for speech played out from various output devices, *i.e.*, bone-conduction headphones, binaural microphone/earphones and loudspeakers (natural hearing), at SNRs 0, −6 and −12 dB, respectively, with competing noise played out from loudspeakers at various azimuths. In most of these figures, there is a dip in the intelligibility for noise at 0° azimuth, which is expected since the target speech is also localized at this angle and so speech is masked the most at this angle compared to all other angles. As the noise moves away from the target speech, the intelligibility generally improves. This is much more apparent at lower SNRs.

At all SNRs, intelligibility with the bone-conduction headphone (HP-F200) is mostly comparable to the binaural microphone/earphone (CS-10EM), without the compen-

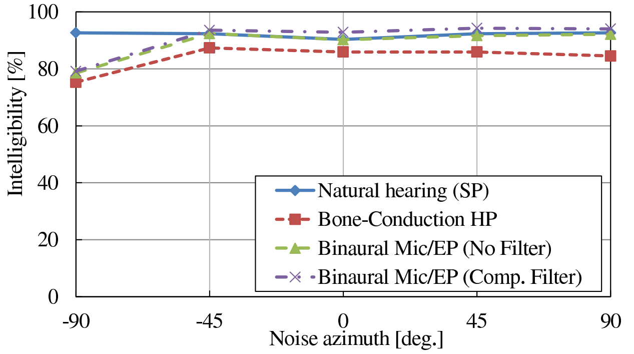


Fig. 8 Noise azimuth vs. intelligibility for various output devices (SNR 0 dB).

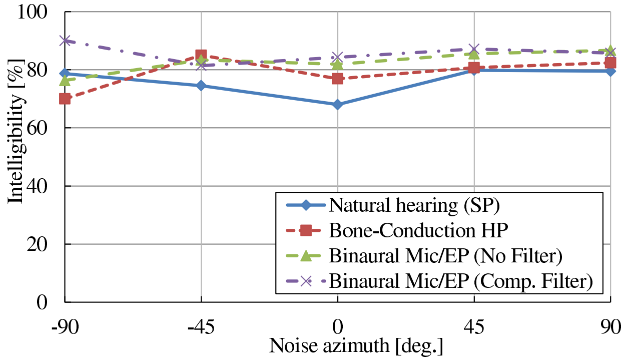


Fig. 9 Noise azimuth vs. intelligibility for various output devices (SNR −6 dB).

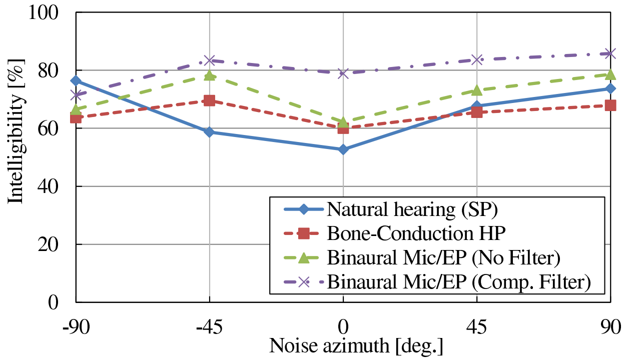


Fig. 10 Noise azimuth vs. intelligibility for various output devices (SNR −12 dB).

sation filter, and are significantly lower than the CS-10EM with the filter. This is again more apparent at lower SNRs. At SNR −12 dB, the CS-10EM with the filter is significantly higher than the other two regardless of the noise azimuth.

ANOVA has been attempted on these results with a significance level of 5%. However, at SNR −6 dB, Levene's Test of Homogeneity of Variances were $F(19, 110) = 3.294, p = 0.036$, and the null hypothesis (no variance difference) was rejected. Thus, we cannot apply ANOVA at this SNR. At SNR=0 dB, however, the null hypothesis was not rejected ($F(19, 110) = 1.143, p = 0.320$), and ANOVA showed that the effect of audio presentation mode (Natural hearing, HP-F200, CS-10EM without and with the compensation filter) on intelligibility is significant ($F(3, 110) = 10.577, p < 0.001$). Post hoc comparisons with the Tukey HSD test shows that the mean for the HP-F200 is significantly different from the others (significantly lower), while no significant difference exist between the remaining three (Natural, CS-10EM with and without filter). At SNR=−12 dB, the null hypothesis was also not rejected ($F(19, 110) = 1.482, p = 0.106$), and the effect of audio presentation mode on the intelligibility is again significant ($F(3, 110) = 11.724, p < 0.001$). Post hoc comparisons using the Tukey HSD test showed that mean intelligibility is significantly different between CS-10EM with the filter and the other three modes (Natural, HP-F100, and CS-10EM without filter), but there was no difference among these three. In other words, CS-10EM with the filter was significantly higher than the other modes.

From Figs. 9 and 10, it seems that the compensation filtering helps the intelligibility of the target speech slightly at SNR=−6 dB, and significantly at −12 dB. It also 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 signal 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 lower intelligibility than the CS-10EM, especially with the filter. This can be attributed to the frequency characteristics of the bone-conduction path of the HP-F200, which is known to have poor low frequency range gain [8], and result in somewhat "muffled" quality speech, which may lower the intelligibility, with or without competing noise. However, it should be emphasized again that the sound quality of bone-conduction headphones have 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 the CS-10EM show high intelligibility, above 70% in most cases (above 80% for the CS-10EM with the filter in most cases). This is even true at SNR −12 dB, which is quite noisy. However, it seems that both the HP-F200 and the CS-10EM (with or without the filters) are well over acceptable quality for AAR applications in realistic acoustic environments, achieving acceptable levels of the annotation speech intelligibility.

The CS-10EM, which can potentially deliver higher quality speech signals, needs additional hardware for the compensation filter (50 taps at a sampling frequency of 44.1 kHz in this experiment) 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 speech quality. However, novel transducers with higher quality are constantly being manufactured, and this soon

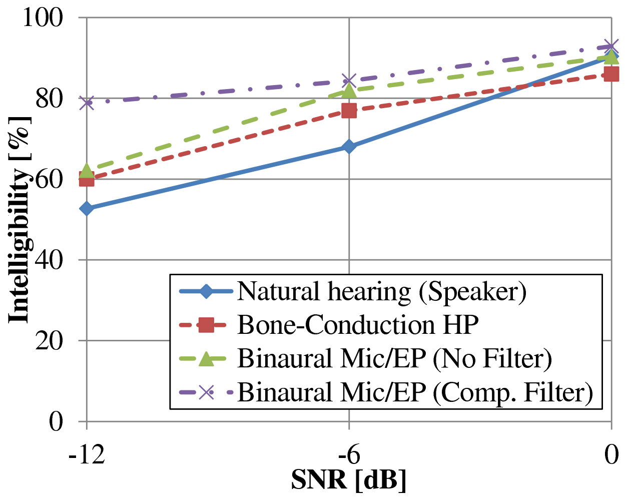


Fig. 11 SNR vs. intelligibility for various output devices (noise azimuth 0°).

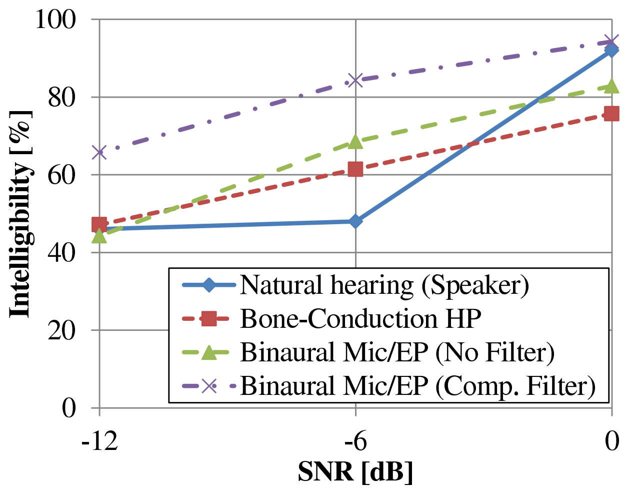


Fig. 12 SNR vs. intelligibility for various output devices (noise azimuth 0°, nasality).

may not be a problem. We have seen that the quality of the delivered speech signal 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 each user and contact the skin firmly at the temple. We may be able to equalize the transfer characteristics using an individualized equalizer. However, the conduction path of this headphone is still being debated. It is generally said that some of the vibrations travel through the skull into the inner ear duct, where it is converted to audible sound waves, while others travel directly to the ear drum, and still others directly to the cochlea. Thus, a single equalizer cannot compensate for all conduction paths. Moreover, it is not clear how to measure the reference signal of the conducted vibration on which the equalizer characteristics design will be based on. In any case, the equalization of the bone-conducted sound is a difficult issue and is out of the scope of this paper. Also, in order to achieve firm 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.

Interestingly, natural hearing shows the high intelligibility when the SNR is high, but degrades significantly at lower SNR. One of the factors contributing to this degradation seems to be that speech signal played out from the loudspeakers will include its reverberation along with the noise, and so the speech signal may be masked by this mixed reverberation at low SNR. However, other factors may be also contributing, and so further investigation is needed to determine the factors contributing to this degradation.

Figures 8, 9 and 10 are replotted in Fig. 11 as SNR vs. intelligibility at noise azimuth 0°, in which the noise is heard from the same direction as the speech signal. At this critical azimuth, the intelligibility is affected most by the noise. From this figure, it seems that the CS-10EM with the compensation filter is significantly better compared to the HP-F100 or the CS10EM without the filter. This difference becomes more significant as the SNR becomes lower. Natural speech clearly shows even lower intelligibility than these two, especially at lower SNR. The CS-10EM without the filter especially seems to show intelligibility comparable to the bone-conduction headphones. We were not able to perform the ANOVA test on this condition since the assumption of variance homogeneity was not met.

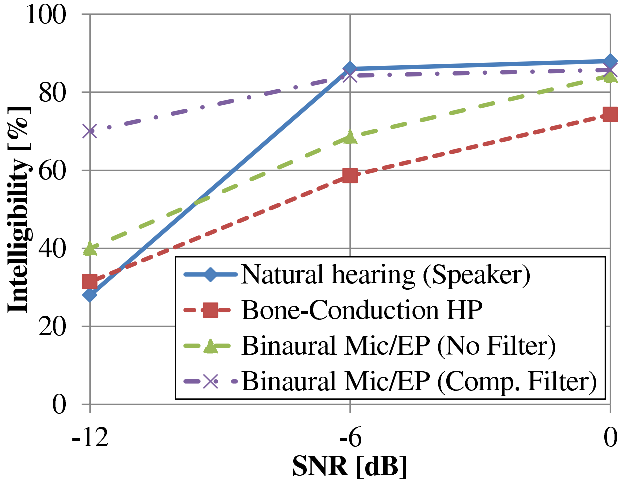


Fig. 13 SNR vs. intelligibility for various output devices (noise azimuth 0°, graveness).

Figures 12 and 13 shows the same SNR vs. intelligibility for the nasality and the graveness feature. These two features showed the largest difference in the intelligibility by audio presentation mode compared to other features. The intelligibility for the nasality feature seems to show even more advantage of the binaural combo with the filtering compared to all other output. On the other hand, the graveness feature shows natural speech with the highest intelligibility at SNR 0 and −6 dB, comparable with the binaural combo with the filters, but sharp degradation at SNR

−12 dB. Thus, the intelligibility difference does show some phonetic feature dependencies. The binaural combo with filtering seems to achieve the highest intelligibility for all features shown here. ANOVA tests with significance level of 5% was performed on nasality, and significant effect of the audio presentation mode was shown on intelligibility $(F3, 66) = 3.697, p = 0.016)$. Post hoc comparisons with the Tukey HSD showed that the difference in the intelligibility mean for HP-F100 and the CS-10EM with the filter was significant, but all other combinations were not. It was not possible to apply ANOVA tests to results for graveness since again the variance homogeneity assumption was not met according to Levene's test.

To summarize, the CS-10EM with the compensation filter shows the highest intelligibility of all the audio output devices tested, especially when the SNR is low, and thus seems to be the choice for speech annotation in mobile augmented audio reality systems. However, the HP-F100 can provide just as high intelligibility speech, and also can be used for AAR systems.

5. 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 cues 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 by 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. We also compared these two devices with natural hearing (no headphones) for reference.

We played word speech localized from the front and 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 comparable speech intelligibility with the binaural microphone/earphone combos without compensation filters, but 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 a significant amount of noise. In fact, this intelligibility was even higher than natural speech, especially at low SNR levels. Thus, we conclude that both of these outputs are 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 their 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.

Acknowledgments

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,” Proc. 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,” J. Audio Engineering Society, vol.52, no.6, pp.618–639, June 2004.

[3] A. Martin, C. Jin, and A.V. Schaik, “Psychoacoustic evaluation of systems for delivering spatialized augmented-reality audio,” J. Audio Engineering Society, vol.57, no.12, pp.1016–1027, 2009.

[4] N. Anazawa, Y. Kobayashi, Y. Yagyu, H. Kanda, and K. Kondo, “Evaluation of localized speech intelligibility from bone-conduction headphones with competing noise for augmented audio reality,” Proc. 40th International Congress and Exhibition on Noise Control Engineering (Inter-noise 2011), Osaka, Japan, 2011.

[5] T. Kanda, H. Yagyu, Y. Kobayashi, K. Kondo, and K. Nakagawa, “Comparison of localized speech intelligibility with competing noise using regular and bone-conduction stereo headphones,” Proc. International Workshop on Principles and Applications of Spatial Hearing (IWPASH), Miyagi, Japan, Nov. 2009.

[6] M. Miura, H. Isaka, and K. Kondo, “Sound presentation of audio reality systems in environment with wind noise,” Proc. 40th International Congress and Exhibition on Noise Control Engineering (Inter-noise 2011), Osaka, Japan, Sept. 2011.

[7] M. Miura, H. Watanabe, K. Kawai, and K. Kondo, “A comparison of the noise control earphone and the bone conduction headphone in outdoor audio augmented reality for pedestrian and cyclists,” Trans. J. Soc. Mech. Eng., vol.79, no.805, pp.2992–3001, Sept. 2013. (In Japanese)

[8] J. MacDonald, P. Henry, and T. Letowski, “Spatial audio through a bone conduction interface,” International J. Audiology, vol.45, pp.595–599, 2006.

[9] TEAC Corporation, “Filltune bone-conduction headphones HP-F200,” http://www.teac.jp/product/hp-f200/

[10] Google Inc., “Google glass tech specs,” https://support.google.com/glass/answer/3064128?hl=en, 2013.

[11] Adphox Co., "Binaural microphone & earphone bme-200," http://www.adphox.co.jp/microphone/sound-eng.html, 2010.

[12] Roland Co., "Binaural microphone/earphone cs-10em," http://www.roland.com/products/en/CS-10EM/

[13] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, H. Nironen, and S. Vesa, "Techniques and applications of wearable augmented reality audio," Proc. 114th Convention of the Audio Engineering Society, Amsterdam, Netherlands, 2003.

[14] Y. Mori, T. Takatani, H. Saruwatari, K. Shikano, T. Hiekata, and T. Morita, "High-presence hearing-aid system using DSP-based real-time blind souce separation module," Proc. IEEE Intern. Conf. on Acoustics, Speech and Sig. Process., April 2007.

[15] N. Anazawa and K. Kondo, "Intelligibility comparison of bone-conducted speech by noise presentation mode," Proc. Tohoku Section Joint Conv. of Institutes of Electrical and Information Engineers, p.2E13, Aug. 2011. (In Japanese)

[16] M. Hiipakka, M. Tikander, and M. Karjalainen, "Modeling of external ear acoustics for insert headphone usage," J. Audio Engineering Society, vol.58, no.4, pp.269–281, 2010.

[17] Y. Suzuki, D. Asano, H.Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," J. Acoust. Soc. Am., vol.97, no.2, pp.1119–1123, 1995.

[18] R. Humphrey, "Playrec: Multi-channel matlab audio," http://www.playrec.co.uk

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

[20] K. Kondo, Subjective Quality Measurement of Speech - Its Evaluation, Estimation and Applications, Signals and Communication Technology, Springer, Heidelberg, Germany, 2012.

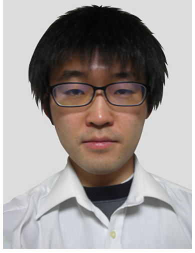
[21] R. Jakobson, C.G.M. Fant, and M. Halle, "Preliminaries to speech analysis: The distinctive features and their correlates," Tech. Rep. 13, Acoustics Laboratory, MIT, 1952.

[22] B. Gardner and K. Martin, "HRTF measurements of a KEMAR dummy-head microphone," http://sound.media.mit.edu/resources/KEMAR.html, May 1994.

[23] B. Gardner and K. Martin, "HRTF measurements of a KEMAR dummy-head microphone," http://sound.media.mit.edu/resources/KEMAR/hrtfdoc.txt, May 1994. MIT Media Lab Perceptual Computing - Technical Report #280.

[24] K. Kondo, T. Chiba, Y. Kitashima, and N. Yano, "Intelligibility comparison of Japanese speech with competing noise spatialized in real and virtual acoustic environments," Acoustical Science & Technology, vol.31, no.3, pp.231–238, 2010.

Kazuhiro Kondo received the B.E., the M.E., and the Ph.D. degrees from Waseda University in 1982, 1984, and 1998, respectively. From 1984 to 1992, he was with the Central Research Laboratory, Hitachi Limited, Kukubunji, Tokyo, Japan. During this time, he was engaged in research on speech and video coding systems. From 1992 to 1995, he was with the Texas Instruments Tsukuba R & D Center Limited, Tsukuba, Ibaraki Japan. From 1995 to 1998, he was with the DSP R & D Center, Texas Instruments Inc., Dallas, Texas, USA. During this time, he worked on speech recognition systems and multimedia signal processing. In 1999, he joined the Faculty of Engineering at Yamagata University, Yonezawa, Yamagata, Japan. His current interests include broad aspects of speech and audio signal processing, multimedia signal processing, and speech and audio quality evaluation methods. He has authored one book, edited another, and contributed chapters to 7 books. Dr. Kondo is a member of the Acoustical Society of Japan, IEEE, and the Audio Engineering Society.



Naoya Anazawa received the B.E., and the M.E. from Yamagata University in 2011 and 2013, respectively. Since 2013, he is with Hitachi Advanced Digital Inc. While at Yamagata University, he conducted research on audio output devices for augmented audio reality applications.

Yosuke Kobayashi received the B.E, the M.E., and the Ph.D. degrees from Yamagata University in 2008, 2010, and 2013, respectively. In 2013, he was with the Faculty of Engineering, Yamagata University, Yonezawa, Yamagata, Japan, and since 2014, he has been with Miyakonojo National College of Technology, Miyakonojo, Miyazaki, Japan. His current research interests include estimation of speech intelligibility using machine learning, spatial audio systems, and speech privacy systems. He has contributed one chapter to a book. Dr. Kobayashi is a member of the ASJ, IEEE, IEEJ, and the AES.