Intelligibility Comparison of Speech Annotation under Wind Noise in AAR Systems for Pedestrians and Cyclists using Two Output Devices

Masanori Miura*, Hideto Watanabe*, Kou Kawai* and Kazuhiro Kondo*
*Graduate School of Science and Engineering, Yamagata University, Yonezawa, Yamagata Japan
E-mail: twm67162@st.yamagata-u.ac.jp Tel: +81-238-26-3312

***Abstract*—Since visual navigation systems using smart phones show information on small screens, user attention is likely to be distracted. Therefore, we are considering a portable navigation system using Augmented Audio Realty. However, if such equipment is used outdoors, very loud wind noise is recorded with the environmental sound. Thus, to reduce this wind noise, we use wind screen (ear muffs) and applied signal processing for wind noise reduction. In this paper, we compared a noise-controlled binaural earphone, and a bone conduction headphone. It was shown that Iterative Wiener filtering improves the speech intelligibility score dramatically with the former device.**

I. Introduction

Navigation systems using portable terminals for comparatively low speed transportation, such as pedestrians or cyclists, have been studied for many years [1]–[3]. Along with the spread of smart phones, these terminals are used daily for portable navigation or AR systems. However, since portability is important, it is difficult to employ large displays in these terminals. Gazing at a small screen is dangerous when moving outdoors and it may become a direct cause of an accident. Accordingly, the authors have been examining an AR system which uses sound information instead of visual information (Augmented Audio Reality, AAR system) [4]. As the presentation method of sound information, an earphone, a headphone, or a bone conduction headphone that vibrates the skull and transmits sound can be considered. However, the most important point is that the environmental sound around a user is not interrupted since they convey essential information about the environment. Moreover, as we shall see, it is essential that the environmental sounds as well as the navigation speech are not interrupted by wind noise.

Although there have been field studies in sound navigation and AAR [1]–[3], there are few examples which focused on the disturbances caused by the wind noise to annotation speech of AAR. Therefore, in this report, speech was reproduced with the bone conduction speaker as well as with the binaural earphone/microphone combo, and speech intelligibility was measured and compared in order to select the optimum device for this application.

II. The Outline Of An AAR System

The AAR system used outdoors needs to be extremely small. In these systems, information is often conveyed using audio due to its size. Audio is delivered using an earphone or a headphone. However, since the ambient noise carries important information about the surrounding environment, such as automobiles approaching, they need to be also delivered. In a previous study [4], we attempted to use a binaural microphone/earphone combo to record ambient noise using the microphone, and add speech annotation mixed with the recorded ambient noise. In this set up, we noticed that the microphone picks up significant amount of wind noise, especially when used by a cyclist. These wind noise needed to be canceled for intelligible speech delivery, and we found that iterative Wiener filter applied to the recorded noise can cancel wind noise while preserving the environmental noise. In this study, we also investigate an alternative device for this purpose. A bone-conduction headphone delivers audio by vibrating the skull. These headphones can deliver annotation speech while leaving the ear canal open, thereby not interfering the ambient noise. However, audio from these headphones are known to be low-pass characteristics due to its conduction path. In this paper, we compare speech intelligibility of speech using these two output devices when wind noise is present.

III. Output Devices of The AAR System

A. AAR system with earphone/microphone combo

A binaural earphone/microphone combo is an earphone with a small microphone attached to the ear canal entry. Ambient noise collected with these microphones can be fed back to the earphone along with added speech annotation. The advantage of this earphone/microphone combo is that it can deliver high-quality speech. Moreover, they can control the quality and the amount of environmental noise fed back to the user since they block out the ear canal. On the other hand, the disadvantage is that wind hits the microphones directly, and the wind noise is recorded at extremely high levels. Therefore, the wind noise needs to be controlled by a noise reduction filter. The requirement of this filter is small processing delay, while sufficient wind noise reduction is achieved. However, essential environmental noise such as car horns or ambulance sirens must be kept audible. Various techniques were proposed to control wind noise. However, evaluations of speech intelligibility of these methods are required eventually. The microphone/earphone combo is shown in Fig. 1. Since the actual wind noise is very loud, it is difficult to control this noise with a digital filter only. We found that the microphone can be equipped with a wind screen (ear muffs) which is shown in Fig. 2, to significantly reduce wind noise [4]. Moreover, Chebyshev high-pass filter, with stop band frequency at 350 Hz, is applied to the reproduced sound. The stop band frequency was chosen so that the fundamental frequency of a car horn is not filtered out. As for the digital filter, spectrum subtraction (specsub) and Iterative Wiener filtering (Wiener-ite) [6]–[8] were employed. Both the specsub and the Wiener-ite filtering can be expressed as equation (1).

$$S(\omega) = H(\omega)X(\omega) \tag{1}$$

where $S(\omega)$ is the clean environmental sound excluding the wind noise, $X(\omega)$ is the recorded environmental sound including the wind noise. $H(\omega)$ is the transfer function. In the specsub method, this can be obtained using equation (2).

$$H(\omega) = 1 - \frac{W(\omega)}{X(\omega)} \tag{2}$$

$W(\omega)$ is the wind noise predictor.

Wiener-ite applies two iterations of Wiener filtering. The Wiener filter is given by equation (3).

$$H(\omega) = 1 - \frac{P_{XS}(\omega)}{P_{XX}(\omega)} \tag{3}$$

where $P_{XS}(\omega)$, $P_{XX}(\omega)$ are autocorrelation of X and cross correlation of X and S. The clean signal S was estimated as estimated noise W subtracted from the noisy signal X. This filter is applied to the noisy signal, and $H(\omega)$ is re-estimated on the filtered speech iteratively, and the resultant filter is applied to the filtered speech to further reduce the estimated noise.

Filtering of wind noise is performed to both right and left channels independently, with two independent filters. Fig. 3 shows the schematic diagram of digital wind noise filter. The wind noise in the present frame which should be controlled is determined by smoothing past frames, and is given to the filtering algorithm. On the other hand, since car horns or sirens, etc. have frequency higher than wind noise, and since these noise are non-stationary, the present frame is compared with the predicted wind noise spectrum (Fig. 3 (b) – "Snn_base", mixture of recorded short/mid/long term past frames from the present). The current frame is classified as transient or stationary by the amount of change of difference between the past short term average. The mid and long term average is applied to transient frames. Otherwise, weighted short term average is applied in order to obtain the maximum amount of reduction. Wind noise component of stationary frames are selectively filtered out. The frames which have been processed are fed to inverse Fourier transformation, and only the real part is obtained from the result. In this report, frames are processed sequentially by frames at sampling frequency of 16 kHz, 16 bits per sample, and in 512 sample frames. Consecutive frames overlap by 50%. A delay of 100 ms will be added in the actual hardware implementation for causality.

In order to evaluate amount of wind noise rejection, these filters ware applied to pseudo wind noise. This wind noise was generated with the electric fan, and was recorded using a PCM recorder. Fig. 4 shows recording apparatus in a soundproof chamber. A dummy head was equipped with the earphone/microphone, and placed in 0.3 meter front of an electric fan to record samples of relatively stable and controlled wind noise. Fig. 5 shows the reduction of the wind noise by the digital filter, (a) without wind screens, and (b) with wind screens shown in Figure 2. Both spectral subtraction and Wiener-ite was applied to this recording. Note that both filters show large rejection ratio at less than 350 Hz due to the fixed Chebyshev HPF applied prior to noise rejection, and not due to the noise rejection filters itself.

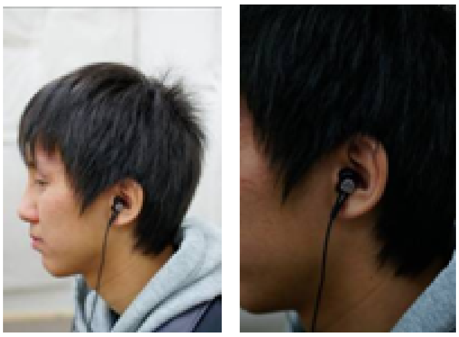


Fig. 1. The binaural microphone/earphone combo.

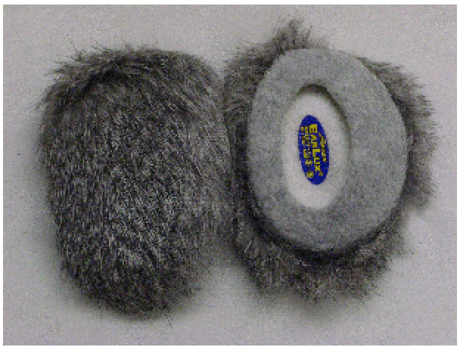
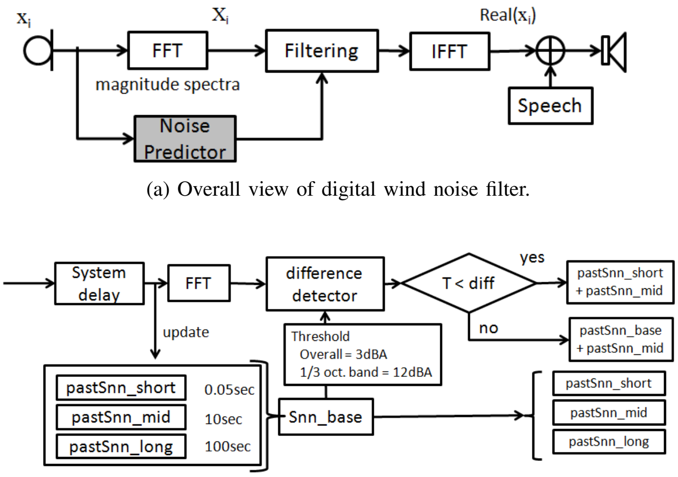


Fig. 2. The wind screen (ear muffs).

With the Wiener-ite, large reduction was obtained in low frequency ranges at 4 kHz or less. On the whole, the amount of reduction obtained was about 30 dB. Only modest reduction was seen with the specsub method. This trend was seen both with and without the wind screens. The wind screens give you additional reduction of 20 to 30 dB across the whole frequency range.

B. AAR system with bone conduction headphones

The bone conduction headphone transmits sound as vibration to the skull, and this stimulates the inner ear directly. The advantage of a bone conduction headphone is that since the ear canal is not plugged, the environmental sound is not interrupted. Moreover, since a microphone is unnecessary, wind noise which occurs when a wind hits a microphone does not occur. Generally, the disadvantage is that it is difficult



(a) Overall view of digital wind noise filter.

(b) Noise predictor.

Fig. 3. Schematic diagram of digital wind noise filter.

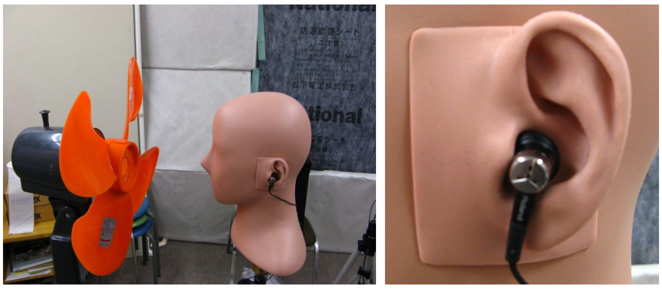
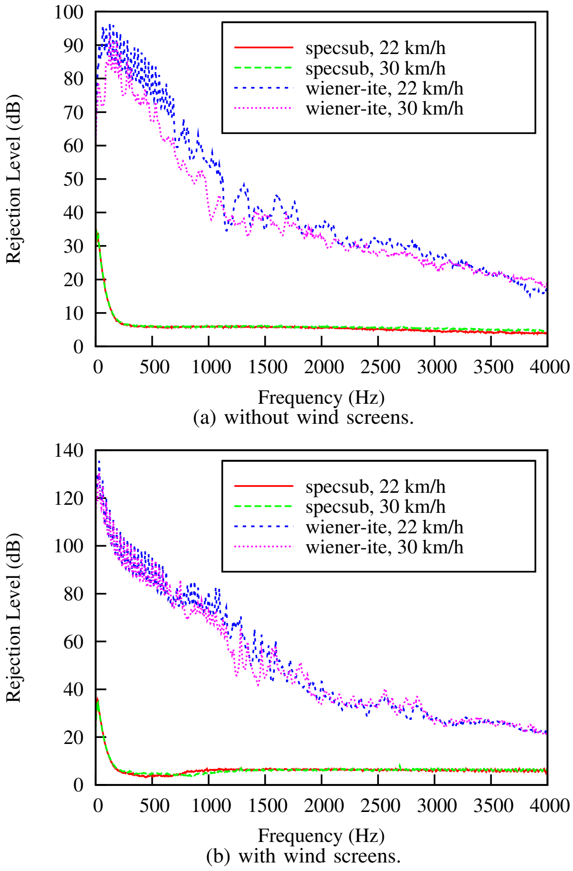


Fig. 4. Recording apparatus in a soundproof chamber.



(a) without wind screens.

(b) with wind screens.

Fig. 5. Wind noise attenuation of AAR system.

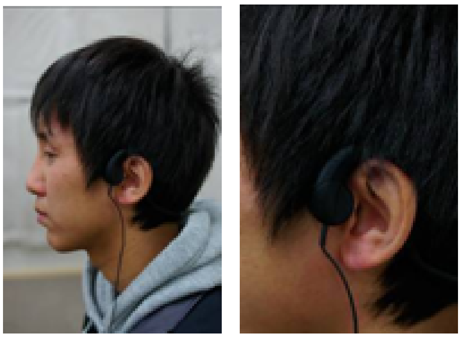


Fig. 6. The bone conduction headphone.

to reproduce high frequency, and also that sounds leaks if a large amplitude vibration (sound pressure) level is reproduced. Moreover, natural wind noise which a user hears because the ear is exposed to wind is not controllable. Also, when localized sound is presented through these headphones, the localized position accuracy tends to decay [9]. In this report, the bone conduction headphone shown in Fig. 6 was used. In order to reduce the wind noise which is made when a wind hits the ear, the same wind screens (ear muffs) with rabbit fur were used as were used with the earphone/microphone combo.

IV. Speech Intelligibility Measurement

We conducted speech intelligibility evaluation in order to measure the influence wind noise has on the speech intelligibility. Both the binaural microphone/earphone and the bone conduction headphone were used in the AAR environment. The Japanese Diagnostic Rhyme Test (JDRT) [10] was carried out in the presence of wind noise. The JDRT uses pairs of 2 mora word which only differ by one phoneme at the beginning of the word. The subject is presented with word speech for one of the word in the pair, and forced to select the correct word. Examples of word pairs are shown in Table I. One male speaker (myu) and one female voice speaker (fao) was used.

The intelligibility is measured using the percentage of correct response adjusted for chance, and is evaluated using equation (4). We will call this the Chance-Adjusted Correct Response (CACR) rate.

$$\begin{aligned}&\text{Chance-Adjusted Correct Response } (CACR)\\&= \frac{\text{Correct Response - Incorrect Response}}{\text{Total Number of Response}}.\end{aligned} \tag{4}$$

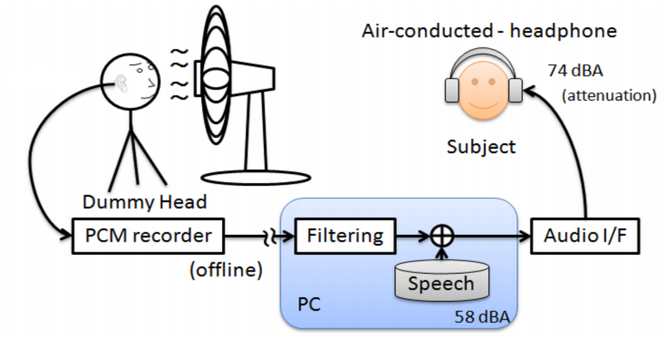


Fig. 7. Schematic diagram of Japanese DRT Test Setup with the binaural earphone/microphone combo.

A. Speech intelligibility measurement using the binaural earphone/microphone combo

In the speech intelligibility measurement using the earphone/microphone combo, wind noise was simulated using an electric fan, and recorded in a soundproof chamber. A dummy head equipped with the microphone was used to record wind noise. The wind velocity was set to 0 km/h, 22 km/h, and 30 km/h. Next, the wind noise reduction filter was applied to this noise, and added to word speech at 58 dBA. The combined filtered wind noise and word speech was presented to the subject through a headphone. However, since the total sound pressure level of the samples became too loud for the subjects to be exposed for the required length of time, we scaled the reproduction volume so that the maximum sound level from the headphone will be about 74 dB.

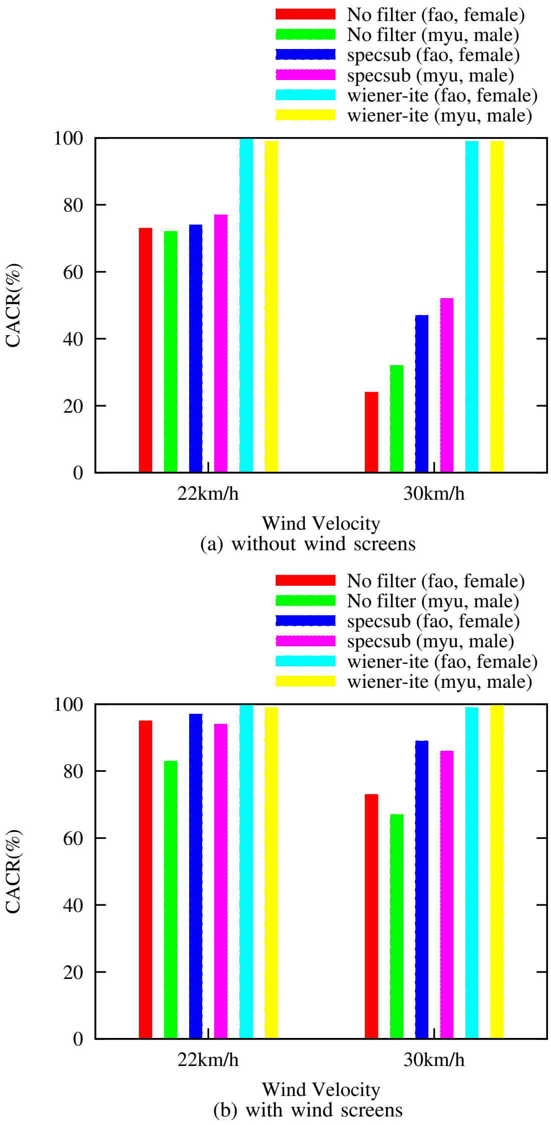
The schematic view of the speech intelligibility measurement is shown in Fig. 7. Seven subjects, all in their twenties participated in the tests. The subjects tested both with and without wind screens (ear muffs).

B. The speech intelligibility measurement using the bone conduction headphone

For intelligibility tests using bone conduction headphones, the subjects sat in front of the same electric fan, and were exposed to wind while listening to word speech played out from the bone-conduction headphones. Since the perceived sound level with the bone conduction headphone differs significantly for each individual, the playout level was compensated using pink noise played out from a loud speaker placed in front of the subjects. The level played out from the speaker was adjusted by the subjects so that the perceived level matches the reference noise from the loud speakers.

TABLE I
EXAMPLE WORD PAIR OF THE JDRT

Phoneme feature	With	Without
Voicing	zai	sai
Nasality	nan	ban
Sustention	hashi	kashi
Sibilation	jamu	gamu
Graveness	waku	raku
Compactness	yaku	waku



(a) without wind screens

(b) with wind screens

Fig. 8. Speech intelligibility using the binaural combo with/without wind screens.

Ten subjects, all in their twenties participated in this portion of the tests. The subjects tested both with and without wind screens (ear muffs).

V. EXPERIMENTAL RESULTS AND DISCUSSION

A. Speech intelligibility of binaural combo with noise reduction filters

CACR with and without noise reduction filters, i.e. spectrum subtraction and Wiener-ite, is shown in Fig. 8, (a) without wind screens, and (b) with wind screens. As can be seen, the basic trend in the intelligibility is the same for both male and female speech.

As can be seen, wind screens can enhance speech intelligibility significantly in all cases. However, the Wiener-ite filter is so effective that speech intelligibility is almost 100% regardless of the use of wind screens. In fact, with the Wiener-ite filter, the wind noise becomes almost silent. The spectrum subtraction filter can give modest improvement over speech with no filters, but the wind noise is still quite audible, and musical noise also becomes audible, which may be degrading the intelligibility.

B. Comparison of bone conduction headphone and an earphone/microphone combo

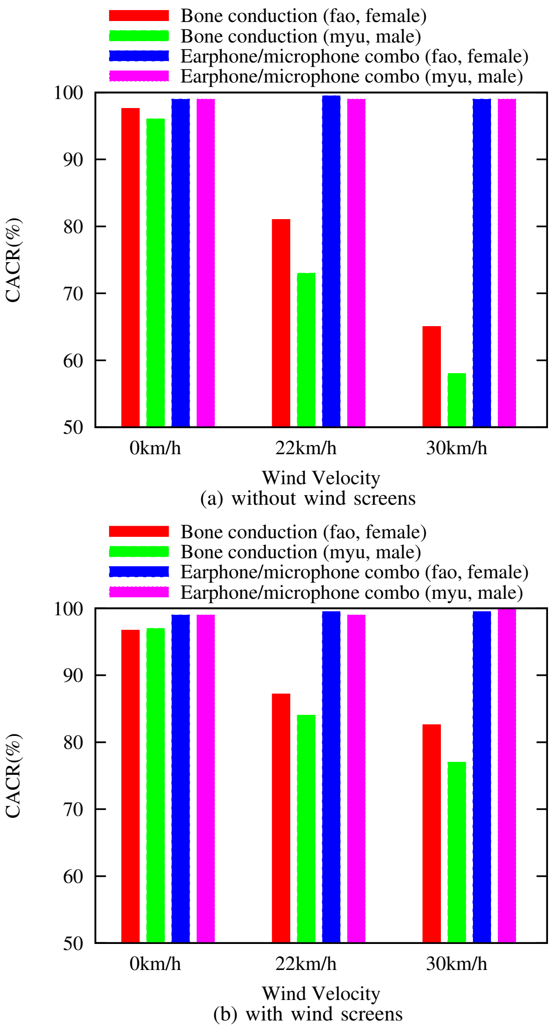
Fig. 9 compares the intelligibility (CACR) using bone conduction headphones and the binaural combo with Wiener-ite filter, (a) without wind screens, and (b) with wind screens. As can be seen, bone conduction headphones can deliver speech at relatively high intelligibility, but wind screens can help. In fact, by using the wind screens, the intelligibility can be kept above 80% even when the wind velocity is 30 km/h. However, the binaural combo with the Wiener-ite filters can achieve almost 100% intelligibility at all wind velocities, with and without wind screens.

Thus, the binaural earphone/microphone combo combined with the Wiener-ite wind noise reduction filter seems to be the choice for AAR applications, especially when high speech intelligibility is the primary goal. However, we observed that the Wiener-ite filters are so effective that it not only reduces the wind noise, but can also reduce necessary environmental noise, such as car horns and sirens. This was not always the case, but does seem to occur on some occasions. Thus, some measure to mitigate this effect is necessary. Perhaps a dedicated horn or siren event detection is necessary. Its detection can be used to control the filter gain of the Wiener-ite filter, so that the detected events are not completely reduced.

VI. Conclusion

We investigated on audio devices to be used for mobile Augmented Audio Reality (AAR) systems. In these systems, environmental noise needs to be fed back to the user, along with the virtual sounds, typically annotation speech. However, we found that significant wind noise is mixed into the environmental noise in these systems, and need to be dealt with. We compared a binaural earphone/microphone combo with a bone conduction headphone. The binaural combo is an earphone with a small microphone embedded at ear canal entry. The recorded environmental sound can be fed back to the earphone along with the annotation speech. The bone conduction headphone can reproduce annotation speech by vibrating the skull. The ear canal is not covered, and so environmental noise can be heard intact.

We found that the binaural combo picks up significant amount of wind noise, and so noise reduction filters were applied. Iterative Wiener filters were found to be able to enhance speech intelligibility to almost 100% at all wind velocity tested. Bone conduction headphones were also able to deliver speech at high intelligibility levels when wind screens are employed, although at still somewhat lower level than the binaural combo. However, the Wiener filters also reduce important environmental sound in some cases, and needs some modifications to preserve these sounds.



(a) without wind screens

(b) with wind screens

Fig. 9. Speech intelligibility, bone-conduction/headphone.

References

[1] S. Holland and D. R. Morse, “Audio GPS: Spatial Audio in a Minimal Attention Interface,” Mobile HCI 01 Proc., 2001.

[2] A. Härmä and J. Jakka, “Augmented Reality Audio for Mobile and Wearable Appliances,” JAES Volume 52 Issue 6, 2004, pp.618-639.

[3] J. Mantell, J. Rod, Y. Kage, F. Delmotte and J. Leu, “NAVINKO: Audio Augmented Reality-Enabled Social Navigation for City Cyclists,” Programme, Workshop Pervasive 2010 Proc., 2010.

[4] M. Miura, K. Kondo and H. Isaka, “Sound presentation of audio reality system in environment with wind noise,” Proc. Inter-Noise 2011, 2011.

[5] R. W. Lindeman, H. Noma, and P. G. Barros, “An Empirical Study of Hear-Through Augmented Reality: Using Bone Conduction to Deliver Spatialized Audio,” IEEE Virtual Reality 2008, 2008, pp.35-42.

[6] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction, Second Edition*, JohnWiley & Sons, 2000.

[7] M. A. Abd El-Fattah, M. I. Dessouky, S. M. Diab and F. E. -S. Abd El-Samie, “Speech enhancement using an adaptive wiener filtering approach.” Progress In Electromagnetics Research M, Vol. 4, 2008, pp.167-184.

[8] K. Funaki, “Speech Enhancement based on Wiener-ite using Complex Speech Analysis,” EUSIPCO-2008, 2008.

[9] D. Schonsteina, L. Ferréb, and B. F. G. Katzb, "Comparison of headphones and equalization for virtual auditory source localization," Acoustic 2008 Proc., 2008, pp.4617-4622.

[10] K. Kondo, R. Izumi, M. Fujimori, R. Kaga and K. Nakagawa, "Two-to-one selection-based Japanese speech intelligibility test," Journal of the Acoustical Society of Japan, 63(4), 2007, pp.196-205.

[11] M. Miura, H. Watanabe and K. Kondo, "Evaluation of the portable AAR environmental sound source accuracy," Tohoku-section Joint Convention of Institutes of Electrical and Information Engineers, Japan , 2012, 1A06.