



# Method for improving the word intelligibility of presented speech using bone-conduction headphones

Teruki Toya<sup>1</sup>, Wenyu Zhu<sup>1</sup>, Maori Kobayashi<sup>2</sup>, Kenichi Nakamura<sup>3</sup> and Masashi Unoki<sup>1</sup>

<sup>1</sup>Japan Advanced Institute of Science and Technology

<sup>2</sup>Faculty of Human Sciences, Waseda University

<sup>3</sup>Westunitis Co., Ltd.

<sup>1</sup>{t-toya, s2010087, unoki}@jaist.ac.jp, <sup>2</sup>maori-k@aoni.waseda.jp,  
<sup>3</sup>nakamura@westunitis.co.jp

## Abstract

Bone-conduction (BC) headphones enable listeners to hear sounds through BC while leaving the ear canal (EC) open to enable surrounding air-conducted (AC) sound to pass through at the same time. However, the intelligibility of presented speech using BC headphones is degraded by BC transmission, especially in noisy environments. This paper proposes a method for improving the word intelligibility of presented BC speech under noisy conditions. The method consists of two types of emphasis: higher-frequency emphasis and consonant emphasis. In the higher-frequency emphasis, frequency components attenuated due to BC transmission were compensated by the inverse-filtering of the transfer function obtained from the regio-temporalis (RT) vibration or the EC radiated sound. In the consonant emphasis, consonant sections with 20-ms short-formant trajectories of subsequent vowels in speech signals were locally amplified by a constant gain. The results of word intelligibility tests showed that both types of emphasis had significant improvements in comparison with no-emphasis. Moreover, we found that the proposed method had the best improvements under all conditions.

**Index Terms:** Bone-conduction headphones, speech intelligibility, high-frequency emphasis, consonant emphasis

## 1. Introduction

Bone-conduction (BC) hearing is one of the applications for auditory augmentation and hearing aids. While ordinary audio devices generate airborne sound through air-conduction (AC), BC transducers generate vibration in the surrounding bone or skin close to the ear to produce auditory perception. The current growth in precision apparatuses has enabled the development of techniques for BC hearing.

In particular, extrinsic (i.e., transcutaneous) BC transducers conduct vibration to the skull bone via intact skin and soft tissue, and they can be easily used as BC headphones with headbands, glasses, or other mechanisms [1]. BC headphones have a big advantage compared with ordinary AC devices. Listeners can receive sounds through BC while leaving the ear canal (EC) open to receive surrounding AC sound at the same time. They can also continue perceiving sounds through BC even though they receive loud background noise through AC. Based on these advantages, a technique for mixing real and computer-generated audio for augmented reality application using BC devices has been proposed [2].

However, a substantial problem remains regarding poor speech intelligibility when listening to speech through BC headphones. In quiet environments, the intelligibility of speech pre-

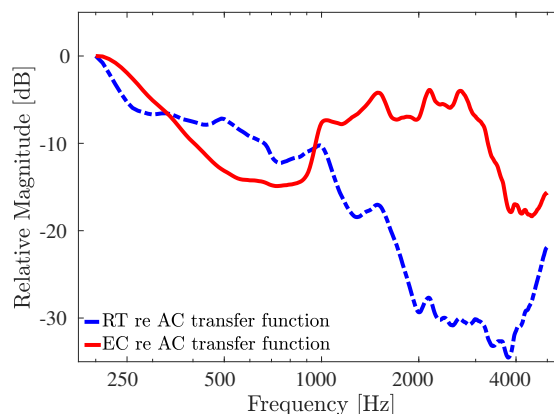


Figure 1: Transfer function of RT vibration relative to AC speech (chained line) and that of EC radiated sound relative to AC speech (solid line) measured by Toya et al. [8]

sented by BC headphones (i.e., BC speech intelligibility) is reported not to be different from that with AC headsets [3]. On the other hand, in noisy environments, BC speech intelligibility is reported to be seriously degraded in comparison with AC [4].

Although speech restoration methods for recorded BC speech with BC microphones have been proposed [5], the improvement in the intelligibility for presented speech using BC headphones with the EC opened is assumed to be much more difficult unless the input speech signal is successfully manipulated beforehand.

BC headphones are subject to signal attenuation especially at high frequencies [1], and this attenuation is assumed to be caused by differences in transmission characteristics between BC and AC pathways. On that basis, we designed a pre-manipulation method for improving BC speech intelligibility under noisy conditions.

## 2. Proposed Method

### 2.1. Concepts for BC speech emphasis

Physiological studies [6, 7] hypothesized that vibration in the skull bone and soft tissue produces auditory perception due to sound pressure induced in the EC, and inertial force in the middle ear ossicles and cochlea.

Toya et al. [8] measured the transmission characteristics of two observable BC pathways: vibration of the regio temporalis (RT) and sound radiated in the EC during vocalization. Figure 1

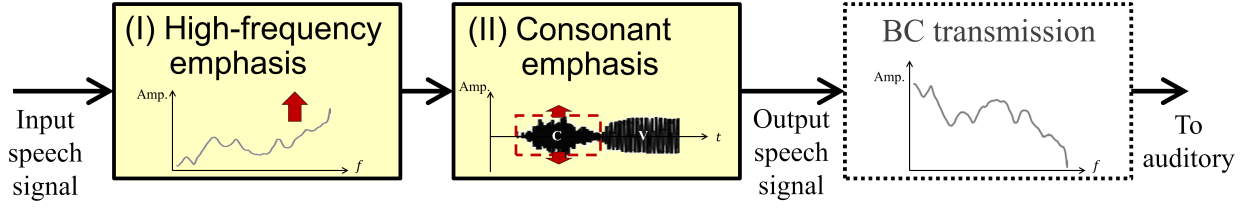


Figure 2: Schematic illustration of proposed method overview for improving intelligibility of bone-conducted speech

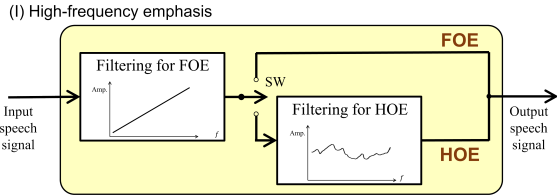


Figure 3: Schematic illustration of procedures of high-frequency emphasis

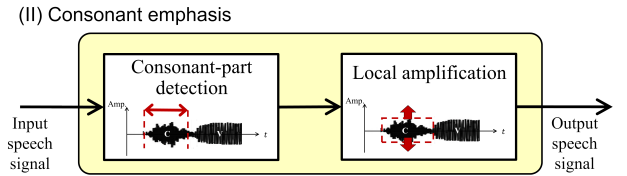


Figure 4: Schematic illustration of procedures of consonant emphasis

shows the measured transfer function of RT vibration relative to AC speech and that of EC radiated sound relative to AC speech. Those measurement results showed the following tendencies: (1) higher-frequency components were linearly attenuated especially for RT vibration, and (2) frequency components below 5 kHz, which are important for speech perception, were modified due to the BC transmission.

Moreover, BC transmission causes poor amplitude of the consonant [9]. Such tendency is assumed to be caused by the attenuation of the higher-frequency components because unvoiced consonants have flat amplitude spectra, while vowels have negative spectral tilts [10]. Considering that the emphasis of the consonants is effective for AC speech intelligibility [11, 12], it may also be effective as pre-processing for improving BC speech intelligibility.

We hypothesized that the compensation of both the spectral attenuation and the consonant-amplitude modification are effective for improving BC speech intelligibility. Figure 2 schematically illustrates the overview of the proposed method, which consists of **(I) Higher-frequency emphasis** and **(II) Consonant emphasis (CE)**.

## 2.2. IIR filter design for higher-frequency emphasis

Figure 3 schematically illustrates the procedure of the high-frequency emphasis. Here, two types of emphasis filters were developed, first-order high-frequency emphasis (FOE) and higher-order high-frequency emphasis (HOE), to separately focus on both a linear decrease in the spectral tilt and local spectral modification. Note that a HOE filter was applied with FOE as a cascade filter during the HOE process.

We used infinite impulse response (IIR) filtering for FOE and HOE. The filter characteristics of FOE  $H_{\text{FOE}}(z)$  and those of HOE  $H_{\text{HOE}}(z)$  are described as follows:

$$H_{\text{FOE}}(z) = 1 + b_1 z^{-1}, \quad (1)$$

$$H_{\text{HOE}}(z) = \frac{\sum_{k=0}^K a_k z^{-k}}{\sum_{k=1}^K b_k z^{-k}}, \quad (2)$$

where  $b_1$ ,  $a_k$  ( $a_0 = 1$ ) and  $b_k$  are the filter coefficients, and  $K$  is the order of the filter. In the case of HOE, the order  $K$  was

set as  $K = 42$ . The filter coefficients in Eq. (1) and (2) were determined using the modified Yule-Walker method (MYW) [13] so that the variance of the error between the desired amplitude and the filter response was minimized.

In this study, the inverse characteristics of the transfer functions derived from the RT and the EC (shown in Fig. 1) were utilized as the desired amplitude characteristics. Therefore, a total of four different emphasis filters (RT-FOE, RT-HOE, EC-FOE and EC-HOE) were designed.

## 2.3. Detection of consonant sections and local amplification

Figure 4 schematically illustrates the procedure of CE. CE consists of two steps as follows:

**Detection of consonant sections:** Let the input speech signal be  $x[n]$ . The consonant sections were determined in reference to the relative amplitude level  $P_r[n]$ .  $P_r[n]$  was defined as follows:

$$P_r[n] = 20 \log_{10} \frac{e_h[n]}{e_a[n]}, \quad (3)$$

$$e_h[n] = \text{LPF}|\text{Hilbert}\{\text{HPF}\{x[n]\}\}|,$$

$$e_a[n] = \text{LPF}|\text{Hilbert}\{x[n]\}|$$

where LPF denotes the low-pass filtering with the cut-off frequency 100 Hz, and HPF denotes the high-pass filtering with the cut-off frequency 5000 Hz. In this study, the signal index  $m$  satisfying  $P_r[m] > -12$  dB was regarded as the index for consonant sections.

**Local amplification:** Let the initial and final index of a certain consonant section be  $m_0$  and  $m_{\text{end}}$ , respectively. In addition to the determined consonant sections ( $m_0 \leq m \leq m_{\text{end}}$ ), an additional number of samples  $N_{\text{locus}} = f_s T_{\text{locus}}$  corresponding to short-formant trajectories of subsequent vowels were utilized as the sections for local amplification. Note that  $f_s$  and  $T_{\text{locus}}$  denote the sampling frequency and the time interval for short-formant transitions, respectively. Here,  $T_{\text{locus}}$  was set to 20 ms considering that formant-transitions generally last for approximately 10 ms from the end of the consonant sections [14].

For the determined sections ( $m_0 \leq m \leq m_{\text{end}} + N_{\text{locus}}$ ), the signal  $x[m]$  was locally amplified with a +12-dB gain. Moreover, a cosine-taper manipulation was applied to each subsequent 10-ms time interval from the end of each section to avoid a sudden fluctuation of signals due to the amplification.

### 3. Evaluation

Japanese word intelligibility tests were conducted for the evaluation of the high-frequency emphasis and CE.

#### 3.1. Speech data

The speech data used in the experiment were selected from the dataset of familiarity-controlled word lists FW07 [15]. FW07 contains four-morae Japanese words, divided into four different familiarity ranks (1: low familiarity, 2: lower-middle familiarity, 3: upper-middle familiarity and 4: high familiarity). The speech data were recorded at a 48,000-Hz sampling frequency and 16-bits quantization.

#### 3.2. Apparatus

Figure 5 schematically illustrates the experimental setup for the word intelligibility tests. The experiment was conducted in a soundproof room. The experimental stimuli were presented through a BC transducer (Temco Japan Co., Ltd. KE08-01) with an amplifier (audio-technica AT-HA5000). The transducer was fixed on both sides of the participants' heads (between the temple-regions and pinnas) The input RMS voltage of the BC transducer was 0.37 V, and the maximum absolute input voltage was lower than 1.0 V. Pink noise was presented through a loudspeaker (Eclipse TD508MK3) with a powered amplifier (Yamaha P4050). The loudspeaker was placed 70 cm behind the participants. The presentation of the stimuli was controlled using MATLAB 2014a on a PC (LG Sharkoon, Windows 8.1) and routed through an A/D converter (Fireface UCX). The sampling frequency was 48,000 Hz and the number of quantizing bits were 16. A PC screen and keyboard were placed in front of the participants.

#### 3.3. Test (I): Evaluation of various types of filters for higher-frequency emphasis

Test (I) was conducted to evaluate the effectiveness of the various types of filters for the higher-frequency emphasis alone. Here, the speech data filtered with four different filter types (RT-FOE, RT-HOE, EC-FOE and EC-HOE) and without any emphasis (no emphasis) were prepared. Pink noise at three noise levels (sound pressure levels of 55, 65, and 75 dB) for AC presentation was used to mimic various levels of noisy environments. The total number of conditions was 60 (i.e., 5 types of filters  $\times$  4 familiarity ranks  $\times$  3 noise levels).

Ten male students aged 23 to 26 participated in this experiment. All the participants were native Japanese speakers who had normal hearing.

The participants were asked to listen to Japanese four-morae words presented through the BC transducer and to type them using the PC keyboard one by one. The number of trials for each participant in a condition was 20, and a total of 60 conditions (i.e., 1200 trials) were set. The mean correct rate of the words was calculated as the word recognition rate for each type of filter, familiarity rank, and noise level.

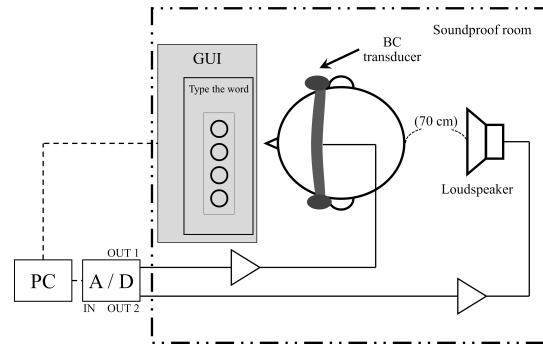


Figure 5: Schematic illustration of experimental setup for word intelligibility tests

#### 3.4. Test (II): Evaluation of CE and both emphasis

Test (II) was conducted to evaluate the effectiveness of CE alone and a combination of two emphases (i.e., the proposed method) under determination of the most effective higher-frequency emphasis filter in Test (I). Here, the speech data processed with three different types of emphasis (higher-frequency emphasis, CE and combined emphasis) and without any emphasis (no emphasis) were prepared. The same pink noise as Test (I) at two noise levels (55 and 75 dB) was used. The total number of conditions was 32 (i.e., 4 types of emphasis  $\times$  4 familiarity ranks  $\times$  2 noise levels).

Ten students (six males and four females) aged 23 to 26 participated in this experiment. All the participants were native Japanese speakers who had normal hearing.

The experiment was conducted using the same procedure as Test (I). The total number of trials was 640. The word recognition rate was calculated for each type of emphasis, familiarity rank, and noise level.

## 4. Results

#### 4.1. Word recognition rate for Test (I)

Figure 6 shows the mean word recognition rate derived from Test (I) for the (a) 55-dB and (b) 75-dB noise levels. The error bar shows the standard error. A three-way repeated measures analysis of variance (ANOVA) showed significant main effects on the word recognition rate for the types of filters [ $F(4, 36) = 11.68, p < 0.01$ ], familiarity ranks [ $F(3, 27) = 190.23, p < 0.01$ ], and noise levels [ $F(2, 18) = 67.5, p < 0.01$ ]. Significant interactions were found between the types of filters and noise levels [ $F(8, 72) = 2.85, p < 0.01$ ]. A post-hoc test showed significant effects of the types of filters on the word recognition rates at the 65-dB and 75-dB noise levels [65 dB:  $F(4, 108) = 5.83, p < 0.05$ ; 75 dB  $F(4, 108) = 12.16, p < 0.01$ ]. The significance cords of the  $p$ -values for a multiple comparison using the Holm-Bonferroni method are shown in Fig. 6 as “\*” ( $p < 0.05$ ) and “\*\*\*” ( $p < 0.01$ ). The word recognition rates of RT-FOE and RT-HOE were significantly higher than those of any other types of filters. No significant differences were evident in the word recognition rate between RT-FOE and RT-HOE. For the upcoming investigation, RT-FOE was utilized as the higher-frequency emphasis.

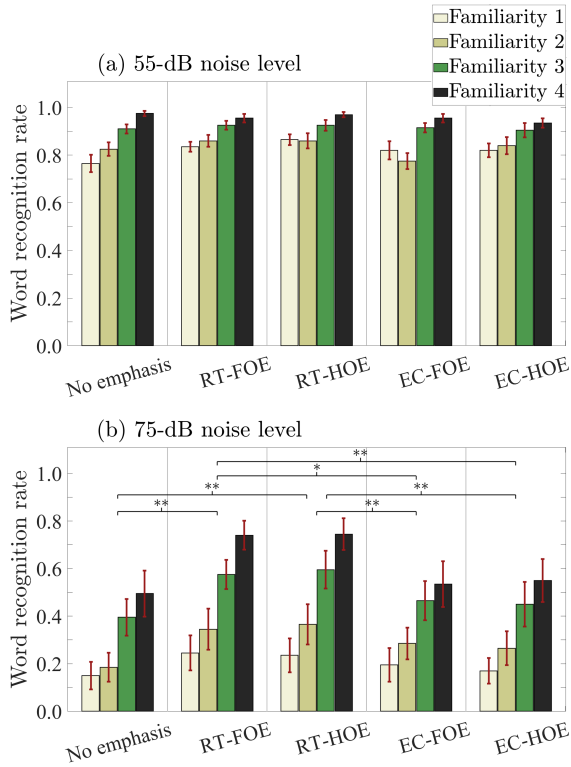


Figure 6: Mean word recognition rate for (a) 55-dB and (b) 75-dB noise levels for each type of filter and familiarity rank derived from word intelligibility tests (I). The error bar shows the standard error. “\*” and “\*\*\*” show significant differences at  $p < 0.05$  and  $p < 0.01$ , respectively.

#### 4.2. Word recognition rate for Test (II)

Figure 7 shows the mean word recognition rate derived from Test (II) for the (a) 55-dB and (c) 75-dB noise levels. A three-way ANOVA showed significant main effects on the word recognition rate for the types of emphasis [ $F(3, 27) = 6.49$ ,  $p < 0.01$ ], the familiarity ranks [ $F(3, 27) = 119.13$ ,  $p < 0.01$ ], and the noise levels [ $F(1, 9) = 348.53$ ,  $p < 0.01$ ]. Significant interactions were found between the types of emphasis and the noise levels [ $F(3, 27) = 4.78$ ,  $p < 0.01$ ]. A post-hoc test showed significant effects of the types of emphasis on the word recognition rates at the 75-dB noise level [ $F(3, 54) = 11.25$ ,  $p < 0.01$ ]. The significance cords of the  $p$ -values for a multiple comparison using the Holm-Bonferroni method are shown in Fig. 7 as “\*” and “\*\*\*.” While the word recognition rate of CE alone was significantly higher than that without emphasis, the word recognition rate of RT-FOE+CE was significantly higher than those of any other types of emphasis.

### 5. Discussion

The results in Test (I) suggest that only RT-FOE and RT-HOE are effective for high-frequency emphasis under noisy conditions. Considering the fact, known as the Lombard effect, that speakers tend to boost the frequency components between 2 to 4 kHz under noisy conditions [16], we hypothesized such components to be strongly related to speech intelligibility for listeners in noisy environments. RT-FOE and RT-HOE were assumed to

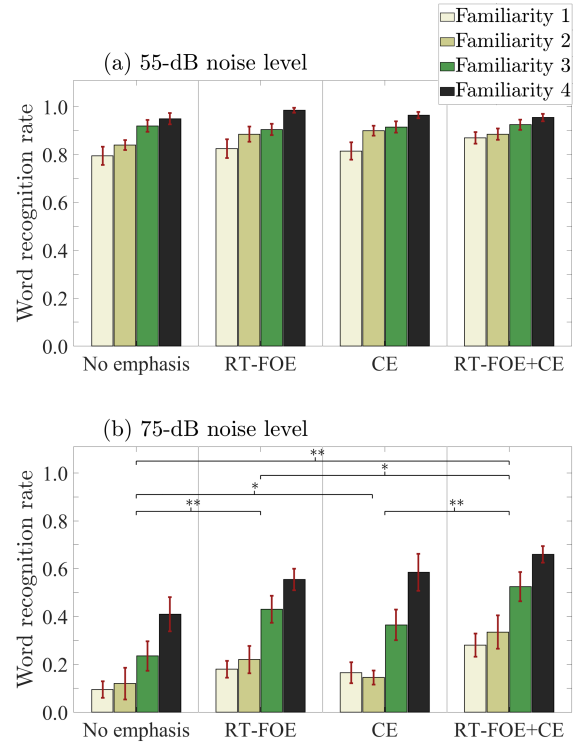


Figure 7: Mean word recognition rate for (a) 55-dB and (b) 75-dB noise levels for each type of emphasis and familiarity rank derived from word intelligibility test (II). The error bar shows the standard error. “\*” and “\*\*\*” show significant differences at  $p < 0.05$  and  $p < 0.01$ , respectively.

compensate successfully for the attenuated components.

No significant differences in the word recognition rate between RT-FOE and RT-HOE suggest that the compensation of the global spectral tilt is mainly effective for improving BC speech intelligibility rather than the compensation of the spectral fine modification.

The results in Test (II) suggested that a combination of two emphases is more effective than a single emphasis (RT-FOE or CE alone) in a noisy condition. Here, the compensation of both the spectral attenuation and the consonant attenuation was revealed to be an effective solution for improving BC speech intelligibility.

### 6. Conclusions

This paper proposed a method for improving BC speech intelligibility under noisy conditions. The method consists of two types of emphasis: higher-frequency emphasis and consonant emphasis. The results of word intelligibility tests showed the following: (1) higher-frequency emphasis with the compensation of the RT transfer function relative to the AC is effective, (2) consonant emphasis is effective, and (3) a combination of two types of emphasis enables the best improvements in BC speech intelligibility.

### 7. Acknowledgements

This work was supported by JSPS KAKENHI Grant No. 21H03463 and Westunitis Co., Ltd.

## 8. References

- [1] S. E. Ellsperman, E. M. Nairn and E. Z. Stucken, "Review of bone conduction hearing devices," *Audiology Research*, 11(2), 207-219, 2021.
- [2] R. W. Linderman, H. Noma and P. G. de Barros, "Hear-through and mic-through augmented reality: Using bone conduction to display spatialized audio," *Proc. ISMAR2007*, 2007.
- [3] S. Maeda, K. Kobayashi, H. Nakatani and A. Nakatani, "Comparison of speech intelligibility between normal headsets and bone conduction hearing devices at call center," *Proc. Internoise 2014*, 1-7, 2014.
- [4] L. Pellieux, J. C. Bouy, C. Blancard and A. Guillaume, "Speech intelligibility with a bone vibrator," *Proc. RTO Human Factors and Medicine Panel (HFM) Symposium*, 2005.
- [5] H. S. Shin, H.-G. Kang and T. Fingscheidt, "Survey of speech enhancement supported by a bone conduction microphones," *Speech Communication*, 10, ITG Symposium, 2012.
- [6] S. Stenfelt and R. L. Goode, "Bone-conducted sound: Physiological and clinical aspects," *Otol. Neurotol.*, 26(6), 1245-1261, 2005.
- [7] S. Stenfelt, "Acoustic and physiological aspects of bone conduction hearing," *Adv. Otorhinolaryngol.*, 71, 10-21, 2011.
- [8] T. Toya, P. Birkholz and M. Unoki, "Estimates of transmission characteristics related to perception of bone-conducted speech using real utterances and transcutaneous vibration on larynx," in A. A. Salah, A. Karpov and R. Potapova (Eds.), *Speech and Computer*, 11658, 491-500, Springer Nature Switzerland, AG, Cham, Switzerland, 2019.
- [9] M. S. Rahman and T. Shimamura, "Amplitude variation of bone-conducted speech compared with air-conducted speech," *Acoust. Sci & Tech.*, 40(5), 293-301, 2019.
- [10] R. D. Kent, R. A. Kent and C. Read, *The Acoustic Analysis of Speech*, Singular/Thomson Learning, 2002.
- [11] K. Hoshino, "Articulation scores of spectrum-peak enhanced consonants," *Audiology Japan*, 37(1), 57-63, 1994.
- [12] T. Yasutake and Y. Nakajima, "Quasi-real-time consonant enhancement system," *IEICE Tech. Rep.*, 105(479), 79-84, 2005.
- [13] B. Friedlander and B. Porat, "The modified yule-walker method of ARMA spectral estimation," *Proc. IEEE AES*, AES-20, 2, 158-173, 1984.
- [14] S. Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Am.*, 80(4), 1016-1025, 1986.
- [15] K. Kondo, S. Amano, Y. Suzuki and S. Sakamoto, NTT Tohoku University Familiarity-Controlled Word Lists 2007 (FW07), 2007.
- [16] M. Garnier and M. Henrich, "Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise?," *Comput. Speech Lang.*, 28(2), 580-597, 2014.