INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Consonant-emphasis method incorporating robust consonant-section detection to improve intelligibility of bone-conducted speech

Yasufumi Uezu[1], *Sicheng Wang*[1], *Teruki Toya*[2], *Masashi Unoki*[1]

[1]Japan Advaced Institute of Science and Technology, Japan
[2]Kanazawa University, Japan

[1]{y-uezu, s2010025, unoki}@jaist.ac.jp, [2]t-toya@se.kanazawa-u.ac.jp

Abstract

A consonant-emphasis (CE) method was proposed to improve the word intelligibility of presented speech by using bone-conducted (BC) headphones. However, the consonant-section detection (CSD) performance of this method is not robust against certain consonants. Therefore, a CE method with robust CSD is necessary for presented BC speech. We focused on improving the word intelligibility of presented BC speech in noisy environments and propose a CE method with robust CSD that combines the detection processes of voiced and unvoiced consonant sections. The evaluation of CSD procedures showed that more robust CSD procedure outperformed those of the conventional CE method as well as voiced CSD only and unvoiced CSD only. Word-intelligibility tests were also conducted on presented BC speech in noisy environments to compare the proposed and conventional methods, and the proposed method significantly improved word intelligibility over these conventional methods at a noise level of 75 dB.

Index Terms: speech intelligibility, bone-conducted speech, consonant emphasis, consonant section detection

1. Introduction

Bone-conducted (BC) headphones have attracted attention as wearable devices [1, 2, 3, 4, 5]. The advantage of BC headphones is that the user can simultaneously listen to BC sounds while being aware of the surrounding environmental sounds through air conduction (AC) [6]. Therefore, BC headphones are expected to be a useful for safe and secure speech communication in real-time and emergency situations [7]. However, the quality of sound and speech intelligibility deteriorate when listening to BC speech, especially in noisy environments, compared with when listening to AC speech [8, 9, 10]. Thus, using BC headphones for safe and secure speech communication in all types of noisy environments, it is necessary to prevent the deterioration of sound quality and speech intelligibility caused by BC headphones.

Toya et al. [11, 12] proposed two methods of speech emphasis to improve the intelligibility of speech presented through BC headphones (presented BC speech). The first method is the high-frequency emphasis (HFE) method, which was designed to compensate for the limitations of BC headphones, where higher-frequency components of the speech signal are more attenuated [13]. The second method is the consonant-emphasis (CE) method, which was designed to detect consonant sections in the speech signal and apply amplitude emphasis to those sections. Word intelligibility tests for BC presented speech with either the HFE or CE method showed that both methods significantly improved speech intelligibility. However, for low-familiarity words in a high-noise environment, speech intelligibility for both methods remained at around 20%.

The crucial aspect of the CE method is the robustness of the consonant-section detection (CSD) procedure. However, it has been observed that the CSD performance of the CE method [2] is insufficient. If a more robust CSD procedure can be established, it may be possible to further enhance the intelligibility of presented BC speech.

We propose a CE method with a more robust CSD procedure for further enhancing the intelligibility of presented BC speech. We first introduce CSD procedure of the conventional CE method by Toya et al. [11, 12] and its limitations. We then describe the more robust CSD procedure of the proposed CE method. This more robust CSD procedure consists of the following two processing stages: (1) unvoiced or voiced CSD on the basis of the idea of the conventional CSD procedure, and (2) integration of the results of both unvoiced and voiced CSD. We compared the performance of the improved CSD procedure with the conventional one, as well as with unvoiced and voiced CSD. Finally, we conducted a word-intelligibility test of presented BC speech in noisy environments to evaluate the performance of the proposed CE method. We investigated the degree to which the proposed CE method enhances the intelligibility of presented BC speech compared with the conventional CE and HFE methods.

2. Conventional consonant-emphasis method

The flow of the conventional CE method by Toya et al. [11, 12], hereafter referred to as the conventional method, consists of the following four steps for emphasis processing. Note that the sampling frequency is 48 kHz unless stated otherwise.

1. CSD

   The power ratio of the high-frequency band (5 to 24 kHz) to the total frequency band (0 to 24 kHz) in the speech signal exceeding -12 dB is used to identify a consonant section.

2. Extension of the section to apply emphasis processing

   Since the formant-transition section is important for the perception of consonants, it is necessary to emphasize both the consonant section and formant-transition section. In this step, the detected consonant section and a 20-ms duration following the end of the consonant section are combined as the emphasis-processing section.

3. Tapering

   To avoid abrupt amplitude changes with and without emphasis, a cosine waveform is applied over a 10-ms duration from the beginning and end of the emphasis-processing section, respectively, resulting in a gentle attenuation of the amplitude.

4. Emphasis processing

10.21437/Interspeech.2023-2568

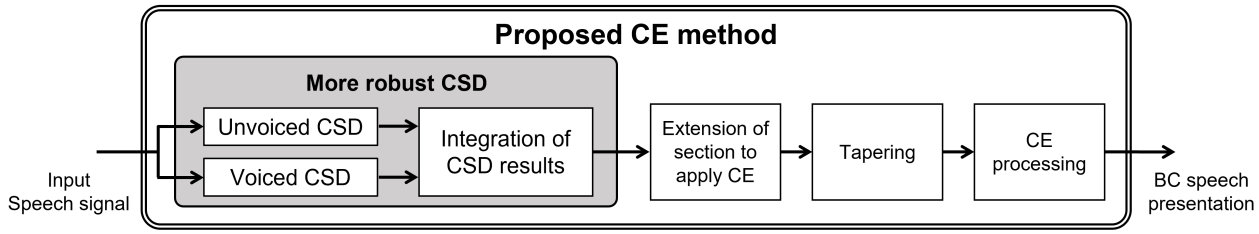

Figure 1: *Schematic of proposed CE method incorporating more robust CSD procedure.*

Emphasis processing is applied to amplify the amplitude of the speech signal by $+12$ dB from the beginning of the emphasis-processing section to the end of the taper-processing section (i.e., the consonant section $+30$ ms).

Toya et al. [11, 12] reported that despite the application of the conventional method, the word intelligibility of low-familiarity presented BC speech in a high noise environment was only about 20%. This result may be attributed to the insufficient CSD procedure of the conventional method. In other words, to further improve the intelligibility of presented BC speech, a CE method with a more robust CSD needs to be developed.

3. Proposed consonant-emphasis method

This section describes the proposed CE method, hereafter referred to as the proposed method, to further enhance the intelligibility of presented BC speech. The proposed method uses emphasis processing through the following two steps.

A. CSD

The proposed method's core is robust CSD in speech signals. This more robust CSD procedure is described in detail in Section 3.1.

B. CE

Following the flow of the conventional method (steps 2.–4. in Section 2), the proposed method emphasizes the detected consonant sections. The main difference is the tapering of the beginning of the detected consonant section. Specifically, emphasis processing is applied to a 10-ms section starting 5 ms before the onset of the consonant while taking into account the minimum duration of the consonant.

3.1. More robust CSD procedure

In this section, we present a more robust CSD procedure based on that by Toya et al. [11, 12]. The more robust CSD procedure includes stages for detecting unvoiced and voiced consonant sections and integrating the results of each. An interval $D_{\mathrm{UC}}(n)$ that satisfies the frequency band power ratio $P_{\mathrm{UC}}(n) > -16$ dB is identified as an unvoiced consonant section, where $P_{\mathrm{UC}}(n)$ is calculated using the following equation,

$$P_{\mathrm{UC}}(n) \quad = \quad 10\log_{10}\frac{e^2_{\mathrm{UC}}(n)}{e^2_{\mathrm{All}}(n)}. \tag{1}$$

Here, $e_{\mathrm{UC}}(n)$ is the time-amplitude envelope of the signal composed of high-frequency components, and $e_{\mathrm{All}}(n)$ is the time-amplitude envelope of the signal composed of all frequency components. The $e_{\mathrm{UC}}(n)$ and $e_{\mathrm{All}}(n)$ are obtained using the following equations,

$$e_{\mathrm{UC}}(n) \quad = \quad \mathrm{LPF}\left\{\left|\mathrm{Hirbert}\left[\mathrm{HPF}_{\mathrm{UC}}(x(n))\right]\right|\right\}, \tag{2}$$

$$e_{\mathrm{All}}(n) \quad = \quad \mathrm{LPF}\left\{\left|\mathrm{Hirbert}\left[x(n)\right]\right|\right\}. \tag{3}$$

LPF is a low-pass filter with a cutoff frequency of 100 Hz, Hilbert is the Hilbert transform, and $x(n)$ is the input speech signal. The lower limit of the high-frequency band is set to 4 kHz, and $\mathrm{HPF}_{\mathrm{UC}}$ is a high-pass filter with a cutoff frequency of 4 kHz.

An interval $D_{\mathrm{VC}}(n)$ that satisfies the frequency band power ratio $P_{\mathrm{VC}}(n) > -0.12$ dB is identified as a voiced consonant section, where PVC is calculated using the following equation,

$$P_{\mathrm{VC}}(n) \quad = \quad 10\log_{10}\frac{e^2_{\mathrm{VC}}(n)}{e^2_{\mathrm{All}}(n)}. \tag{4}$$

Here, $e_{\mathrm{VC}}(n)$ is the time-amplitude envelope of the signal composed of low-frequency components. The $e_{\mathrm{VC}}(n)$ is obtained using the following equation,

$$e_{\mathrm{VC}}(n) = \mathrm{LPF}\left\{\left|\mathrm{Hirbert}\left[\mathrm{LPF}_{\mathrm{VC}}(x(n))\right]\right|\right\}. \tag{5}$$

The higher limit of the low-frequency band is set to 0.9 kHz, and $\mathrm{LPF}_{\mathrm{VC}}$ is a low-pass filter with a cutoff frequency of 0.9 kHz.

The consonant section is finally determined by taking the logical OR of the results of each judgment of unvoiced or voiced CSD.

$$D_{\mathrm{C}}(n) = D_{\mathrm{UC}}(n) \cup D_{\mathrm{VC}}(n). \tag{6}$$

4. Evaluation

4.1. CSD performance

We compared the performance of the more robust CSD procedure with the conventional CSD procedure, unvoiced CSD only, and voiced CSD only. We used 640 speech data items of four-morae words from the dataset of the familiarity-controlled word lists FW07 [14] for the word-intelligibility test. Two parameters were used to evaluate CSD performance: F-score (0 to 1) and d_{ROC} (0 to 2), where d_{ROC} is the Euclidean distance from the " best solution " point on the Receiver-operating characteristic (ROC) curve. The closer the F-score and d_{ROC} are to 1 and 0, respectively, the better the CSD performance.

4.2. Word-intelligibility test for presented BC speech

To evaluate the word intelligibility of presented BC speech using the proposed method, we conducted a word-intelligibility test using BC headphones in a noisy environment.

Ten native Japanese speakers in their 20s with normal hearing participated in the experiment (five women). We tested four speech-emphasis methods: the proposed method, the conventional method [11, 12], HFE method [11, 12], and with no emphasis. The speech stimuli were categorized on the basis of word-familiarity ranks, 1 (low familiarity) to 4 (high familiarity). Two levels of the background noise were set to 55 and 75

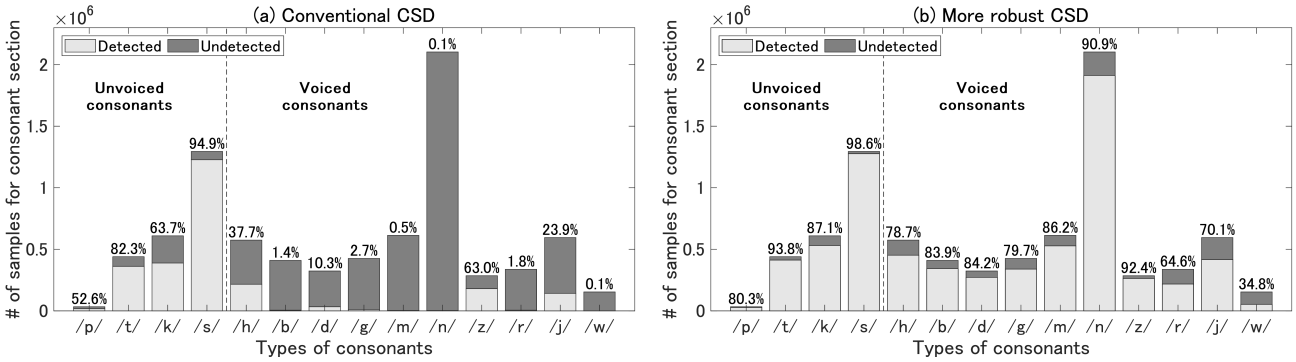

Figure 2: *Detection results of more robust and conventional CSD procedures for each consonant type in FW07's 640 speech data.*

Table 1: *Results of performance evaluation of each CSD procedure.*

CSD types	Precision	Recall	F-score	d_{ROC}
Conventional [11, 12]	0.707	0.315	0.436	0.688
Unvoiced	0.617	0.379	0.469	0.632
Voiced	0.514	0.482	0.498	0.564
Proposed	**0.560**	**0.857**	**0.678**	**0.358**

dB. Thus, a total of 32 combined conditions were tested (four types of emphasis × four familiarity ranks × two noise levels).

Pink noise with a frequency bandwidth of 0–10 kHz was used as the background noise presented through AC For the speech stimulus presented through BC headphones, we used 640 speech data items of four morae words contained in the same FW07 [14] as in Section 2.2. This speech data include 160 words for each of the familiarity ranks. The speech data were randomly sorted into eight blocks for each participant in the following procedure. First, we divided the 160 words from each familiarity rank into 8 groups, each containing 20 words. Next, we created eight blocks by selecting one group from each familiarity rank to avoid duplication then each assigned each block to one of the eight combination conditions (four emphasis types and two noise levels). Finally, we subjected the speech data contained in each block was to the corresponding speech-emphasis processing.

BC transducers (Temco Japan Co., Ltd. KE08-01) and amplifiers (audio-technica AT-HA5000) were used to present the BC speech. The voltage applied to the BC transducer during the sound presentation was adjusted to an average Root-Mean-Square of 0.368 V. The sound-pressure level of the AC sound, which corresponds to the perceived loudness of the BC sound, was approximately 60 dB. A loudspeaker (ECLIPSE TD508MK3) and power amplifier (Yamaha P4050) were used to present the background noise. The loudspeaker was positioned 70 cm behind the participant. The software (Mathworks MATLAB 2014a) on a PC (LG Sharkoon, Windows 8) and an analog-to-digital (A/D) converter (RME Fireface UCX) was used to control the stimulus presentation. A display and keyboard were placed in front of the participant.

The word-intelligibility test was conducted using the following protocol. Participants were seated in a chair and wore a BC transducer on their head during the test. Participants were presented with speech via the BC transducer and background noise through a loudspeaker simultaneously. Following the presentation of the stimuli, participants typed the words they heard in katakana using a keyboard. All participants completed eight blocks, beginning with four blocks at a noise level of 55 dB,

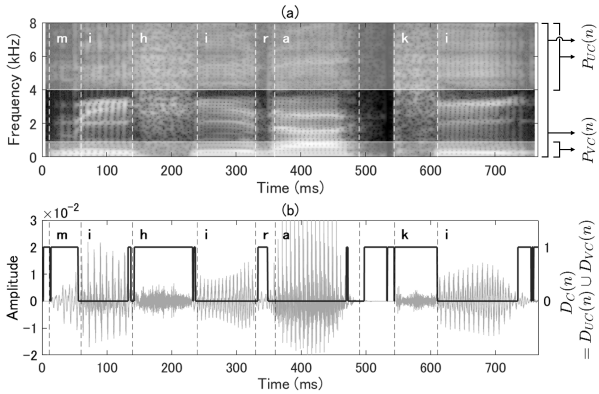

Figure 3: *(a) Power spectrogram of speech signal /mihiraki/. White shaded areas above 4 kHz and below 0.9 kHz indicate frequency bands used for calculating unvoiced and voiced CSD, respectively. (b) Speech waveform (gray line) and detection results of using more roust CSD procedure (black line, where 1 and 0 represent consonant and non-consonant, respectively).*

followed by four blocks at a noise level of 75 dB. A sufficient break was provided between blocks.

We calculated the correct-word rate on the basis of the participants' responses to evaluate word intelligibility. A trial was deemed correct when the participant accurately responded to all four morae included in the speech sounds presented within the trial. For each participant in the experiment, we calculated the correct-word rate for all 32 conditions. We then conducted a $(4 \times 4 \times 2)$ analysis of variance (ANOVA) and a multiple comparison test using the Holm-Bonferroni method.

5. Results

Table 1 lists the results of the CSD performance evaluation. The F-score and d_{ROC} indicate that the more robust CSD procedure performed the best compared with the conventional CSD procedure. Figure 2 shows the detection results with the more robust and conventional CSD procedures for each consonant type in FW07's 640 speech data. The vertical axis indicates the number of samples in the consonant section of the speech. The light-gray and dark-gray bars indicate the number of samples that were identified as consonant or non-consonant sections with both CSD procedures, respectively. The sum of the light-gray and dark-gray bars indicates the total number of samples in the consonant section for each consonant. The number at the top of each bar represents the detection rate for each consonant.

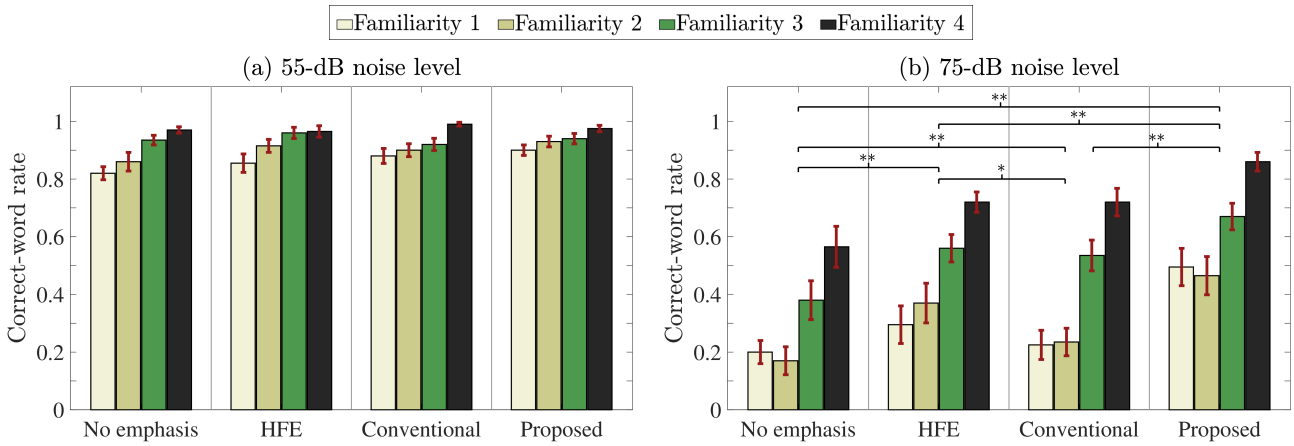

Figure 4: *Mean correct-word rate for (a) 55-dB and (b) 75-dB noise levels for each type of emphasis and familiarity rank derived from Word-intelligibility test. Error bar shows standard error. "*" and "**" show significant differences at $p < 0.05$ and $p < 0.01$, respectively.*

Overall, the more robust CSD outperformed the conventional one. Notably, the detection rate of the more robust CSD procedure for voiced consonants was markedly higher than that of the conventional CSD procedure. These findings indicate that the more robust CSD procedure exhibits better performance, irrespective of the consonant type.

Figure 3(a) shows the power spectrogram of the speech signal /mihiraki/. The vertical dashed lines indicate the boundaries of each phoneme segment in the speech signal. The white shaded areas above 4 kHz and below 0.9 kHz indicate the frequency bands used for the calculation of unvoiced only and voiced only CSD, respectively. Figure 3(b) shows the speech waveform (indicated with a gray line) and the detection results of using the more robust CSD (indicated with a black line, where 1 and 0 represent consonant and non-consonant, respectively).

Figure 4 shows the correct-word rate for each emphasis type and familiarity rank obtained from the word-intelligibility test. Figures 4(a) and 4(b) show the correct-word rate at noise levels of 55 and 75 dB, respectively, while the error bars represent the standard errors. The results of the three-way ANOVA indicated significant main effects of emphasis type, familiarity rank, and noise level on correct-word rate. Moreover, an interaction effect between the emphasis type and noise level was observed. Specifically, at a noise level of 75 dB, a significant simple main effect of emphasis type was observed for the interaction between emphasis type and noise level. The results of the multiple comparisons between emphasis types at a noise level of 75 dB are shown in Fig. 4(b). At a noise level of 75 dB, the proposed method resulted in a significantly higher correct-word rate compared with the other methods. Moreover, the proposed method improved the correct-word rate with high familiarity ranks (3 and 4) by approximately 10%, and for words with low familiarity ranks (1 and 2) by approximately 30%, compared with the conventional method.

6. Discussion

Significant differences were observed in the correct-word rate between the proposed method and conventional method in a high-noise environment. This difference in performance may be attributed to the enhanced robustness of the CSD procedure incorporated into the proposed method. The proposed method can more effectively detect and emphasize consonant sections in speech signals, leading to improved intelligibility of presented BC speech.

Significant differences were observed in the correct-word rate between the proposed method and HFE method in a high-noise environment. These findings suggest that emphasizing consonant sections in the speech signal is more effective in enhancing the intelligibility of presented BC speech in high-noise environments than uniformly emphasizing the entire speech signal at high frequencies.

An interaction between emphasis type and noise level was observed for the correct-word rate, indicating that the proposed method is particularly effective in improving the intelligibility of presented BC speech in high-noise environments. Additionally, no significant interaction was observed between emphasis type and familiarity rank for the correct-word rate, suggesting that the effectiveness of the proposed method is independent of word familiarity rank.

This method focuses on the power in specific frequency bands of voiced and unvoiced consonants, which may be applicable to languages that have similar acoustic characteristics to Japanese. The results of the study are expected to hold for tonal languages as well, since pitch changes are unlikely to significantly affect the power ratio.

7. Conclusion

We proposed a CE method that incorporates a robust CSD procedure to improve word intelligibility in presented BC speech in noisy environments. The efficacy of the proposed method was evaluated through a word-intelligibility test of BC speech. The results indicate that the proposed method significantly enhances the intelligibility of BC speech in high-noise environments compared with other speech-emphasis methods. For future work, we plan to investigate whether combining the proposed method with the HFE method can further improve the intelligibility of presented BC speech.

8. Acknowledgements

This work was supported by the Fund for the Promotion of Joint International Research (Fostering Joint International Research (B)(20KK0233) and Research found of Shibuya Science Culture and Sports Foundation.

9. References

[1] B. N. Walker and J. Lindsay, "Navigation performance in a virtual environment with bonephones," in *Proc. ICAD 05 – 11th Meeting of the International Conference on Auditory Display*, Limerick, Ireland, Jul. 2005, pp. 260–263.

[2] H.-W. Park, M.-S. Kim, and M.-J. Bae, "Improvement of voice quality and prevention of deafness by a bone-conduction device," *Biotechnology & Biotechnological Equipment*, vol. 28, no. S1, pp. S14–S20, 2014.

[3] C. Manning, T. Mermagen, and A. Scharine, "The effect of sensorineural hearing loss and tinnitus on speech recognition over air and bone conduction military communications headsets," *Hearing Research*, vol. 349, pp. 67–75, 2017.

[4] A. Barde, M. Ward, R. W. Lindeman, and M. Billinghurst, "Less is more: Using spatialized auditory and visual cues for target acquisition in a real-world search task," in *Proc. ISMAR-adjunct – 2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct*, Beijing, China, Oct. 2019, pp. 340–341.

[5] S. E. Ellsperman, E. M. Nairn, and E. Z. Stucken, "Review of bone conduction hearing devices," *Audiology research*, vol. 11, no. 2, pp. 207–219, 2021.

[6] B. N. Walker and R. M. Stanley, "Thresholds of audibility for bone-conduction headsets," in *Proc. ICAD 05 – 11th Meeting of the International Conference on Auditory Display*, Limerick, Ireland, Jul. 2005, pp. 218–222.

[7] Z. J. Lim and J. Claydon, "Use of bone conduction headsets to improve communication during the COVID-19 pandemic," *Emergency Medicine Australasia*, vol. 32, no. 5, pp. 903–904, 2020.

[8] M. Gripper, M. McBride, B. Osafo-Yeboah, and X. Jiang, "Using the Callsign Acquisition Test (CAT) to compare the speech intelligibility of air versus bone conduction," *International Journal of Industrial Ergonomics*, vol. 37, no. 7, pp. 631–641, Jul. 2007.

[9] B. Osafo-Yeboah, X. Jiang, M. McBride, D. Mountjoy, and E. Park, "Using the Callsign Acquisition Test (CAT) to investigate the impact of background noise, gender, and bone vibrator location on the intelligibility of bone-conducted speech," *International Journal of Industrial Ergonomics*, vol. 39, no. 1, pp. 246–254, 2009.

[10] R. M. Stanley and B. N. Walker, "Intelligibility of bone-conducted speech at different locations compared to air-conducted speech," in *Proc. HFES – Human Factors and Ergonomics Society Annual Meeting*, vol. 53, 2009, pp. 1086–1090.

[11] T. Toya, W. Zhu, M. Kobayashi, K. Nakamura, and M. Unoki, "Method for improving the word intelligibility of presented speech using bone-conduction headphones," in *Proc. INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, Incheon, Korea, Sep. 2022, pp. 759–763.

[12] T. Toya, M. Kobayashi, K. Nakamura, and M. Unoki, "Methods for improving word intelligibility of bone-conducted speech by using bone-conduction headphones," *Applied Acoustics*, vol. 207, p. 109337, 2023.

[13] 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 *Proc. SPECOM 2019 – 21st International Conference of Speech and Computer*, Istanbul, Turkey, Aug. 2019, pp. 491–500.

[14] S. Sakamoto, T. Yoshikawa, S. Amano, Y. Suzuki, and T. Kondo, "New 20-word lists for word intelligibility test in japanese," in *Proc. INTERSPEECH 2006 - ICSLP – 9th International Conference on Spoken Language Processing*, Pittsburgh, PA, USA, Sep. 2006, pp. 2158–2161.