2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

Body Motion Noise Reduction of Silent Speech Recognition Using Facial Surface EMG*

Ryosuke Kimoto, *Student Member, IEEE,* Takashi Ohhira, *Member, IEEE,* and Hideki Hashimoto, *Fellow, IEEE*

***Abstract*—This paper proposes a method for reducing body motion noise in silent speech recognition (SSR) systems using facial muscle information. Several SSR methods have been developed that utilize facial muscle information for speech recognition. However, these methods are limited to steady-state conditions. The accuracy of these methods is significantly degraded by body motion noise from daily movements mixed with EMG signals, making word identification impossible under such conditions. To address the body motion problem in SSR systems, this paper proposes an improved SSR system and demonstrates its effectiveness.**

1. INTRODUCTION

The demand for voice recognition systems is increasing annually, and many products now use this technology. Small devices, such as smartphones, can perform web searches or play music using voice commands. Additionally, products that can control home appliances with a single voice command have been developed. Voice recognition systems have been implemented in various applications, becoming a new way of life.

However, speech recognition is associated with several issues. For example, personal or confidential information may be heard by others in public places. Additionally, problems such as inaccurate speech recognition due to loud ambient noise and difficulties faced by individuals with speech impediments can arise. To address these challenges, Silent Speech Recognition (SSR), a method that reads the content of speech without vocalizing it, has gained attention in recent years.

Several methods exist for implementing SSR. One method uses images of the mouth [1]. A camera is positioned in front of the mouth to capture its movement as a video. The lip movements during SSR are then measured, and this information is input into a CNN to recognize the spoken words. However, this method is challenging for everyday use because it requires the camera to be consistently positioned in front of the face.

The second method relies on detecting mouth movements. This approach uses an accelerometer placed under the chin to recognize speech content through a CNN, based on skin acceleration information generated by SSR. A drawback of this method is that the acceleration-based feature values can fluctuate significantly due to daily movements, such as walking, making accurate recognition challenging.

Another method involves tracking tongue movement [2]. In this approach, a dental retainer equipped with a capacitive touch sensor is placed in the mouth to monitor tongue movements during SSR. The recorded tongue movements are then input into a neural network to recognize speech content. The main disadvantage of this method is that it requires the use of custom-fit retainers, which can be problematic due to the need for individual adjustment, as well as concerns related to hygiene and safety.

Another method involves measuring muscle potentials [3]. In this approach, a sensor for detecting EMG potentials is attached to the surface of the face, and the weak EMG potentials generated by SSR are used as inputs to a CNN to recognize speech content. A disadvantage of this method is that the appearance of the face can be affected by the sensor attachment. However, this issue can be mitigated by miniaturizing the sensor using 3D printing or similar technologies to create a more discrete device. An advantage of this method is that muscle potentials are generated approximately 50 milliseconds before actual muscle contraction, making it suitable for real-time recognition.

In this study, we investigated SSR using facial surface EMG due to their relatively fewer disadvantages.

As an example of an SSR application, enabling communication in crowded places or environments where speaking is difficult by converting speech into voice or text. If the device is miniaturized, it can be conveniently used even when wearing a mask. Additionally, there is a growing demand for such devices from a security perspective, as they allow the exchange of confidential and personal information without the risk of leaking conversation or message content to others. This technology can also assist individuals with speech impairments in operating equipment and communicating with others without restricting the number of users. It can be used for rehabilitation of speech impediments that involve complex mouth movements or as a pointing device, similar to a remote control, where simple words such as "1, 2, 3, 4, 5" can correspond to information such as direction and distance.

Current SSR systems need improvement, especially in the recognition of speech content when the user's body is moving. Previous studies often overlook the effects of body motion due to measurements limited to steady-state conditions [4,5]. In this study, we attempt to remove noise caused by body motion using VMD, a technique that decomposes the original

* This work was supported by KEIRIN JKA (2023M-262).

R. Kimoto is with Graduate School of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551 Japan (corresponding author to provide e-mail: a23.5ntb@g.chuo-u.ac.jp).

T. Ohhira and H. Hashimoto are with Department of Electrical, Electronic, and Communication Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551 Japan (e-mail: { ohhira, hashimoto }@elect.chuo-u.ac.jp).

signal into several modes based on the relationship between time and frequency and reconstructs the signal by selecting the mode with the least noise among them. However, the number of modes to decompose and the mode to select must be set by oneself. In a previous study by [6], VMD is used to remove body movement noise, but the method of measuring EMG signals is not clear and does not guarantee sufficient performance. Body movement noise is mainly caused by friction between the electrodes and the skin due to the shaking of the cord of the EMG device caused by daily movements. Therefore, the method of measuring myopotential signals must be clarified.

Therefore, this paper describes a method to fix the measurement position and proposes an effective selection method for mode of VMD to reduce the effect of body motion noise. The proposed method aims to minimize body motion noise and increase the potential of SSR technology.

2. EMG

An EMG (electromyogram) is an action potential generated when muscle fibers contract, beginning with the excitation of alpha motor neurons in the spinal cord. This excitation is transmitted to the muscle, where the release of acetylcholine at the neuromuscular junction depolarizes the muscle fibers, generating an action potential. An electromyogram is a time-sequence record of the combined action potentials from multiple muscle fibers. The Mechanism of EMG Signal Generation is shown in Fig. 1.

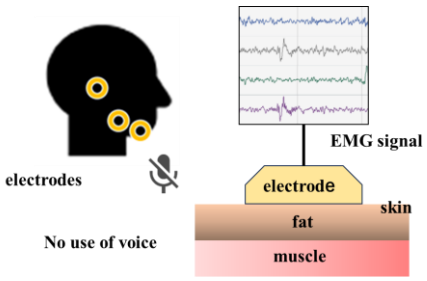

Fig. 1: Mechanism of EMG Signal Generation

3. MEASUREMENT METHOD

In this study, an OpenBCI [7] sensor capable of measuring EEG, EMG, and cardiac electrical activity was used, as shown in Figure 2. In the measurement setup, three signal electrodes were attached to the face, one under the chin, and one reference and one bias electrode behind each ear. The potential difference between the electrodes was measured to capture mouth movements. The positions of the electrodes were determined by trial and error based on the distribution of facial muscles [8] and findings from previous studies.

The measurement setup is shown in Figure 3. As shown in Figure 4, the length of the cord from the electrode attachment was minimized and the device was fixed to the cap to prevent sagging during measurement. This configuration maximizes the prevention of body noise generation due to friction between the electrode and skin caused by the swinging of the cord.

Although costly to produce, creating the sensor device with a 3D printer would reduce the size of the device and eliminate the need for sticker electrodes on the facial skin, thereby reducing discomfort for the user. However, the measurements in this study are performed using the fixation method shown in Figure 3 to measure EMG potentials in steady state and body motion conditions.

The myopotential information obtained from each sensor was used for subsequent processing.

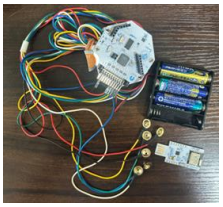

Fig. 2: Measurement Device (OpenBCI)

Fig. 3: Measurement Method

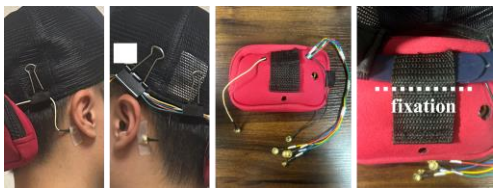

Fig. 4: Fixing method with a cap

4. SSR IN STEADY-STATE

4.1. Pretreatment

The signals measured by the OpenBCI sensor were raw data, and the value scales of each channel varied. The different scales made it difficult to compare features between channels. To address this issue, standardization was applied to equalize the signal scales.

Standardization was employed to facilitate comparison and analysis of the data across different scales. If each data element is x_i, each standardized element is x'_i, the mean value of all data is $\bar{x}$, and the standard deviation of all data is σ, the standardization formula can be shown in equation (1).

$$x'_i = \frac{x_i - \bar{x}}{\sigma} \quad (i = 1 \cdots n) \tag{1}$$

Applying this formula provides signals on a uniform scale. Fig. 5 displays the real-time signals before and after standardization.

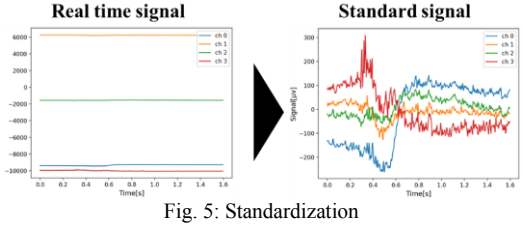

Fig. 5: Standardization

However, real-time signals contain AC power-supply noise and weak noise emitted from the body, which must be removed to accurately capture the characteristics of the spoken content.

The AC power-supply noise operates at 50 Hz and its harmonics. A study [9] demonstrated that a notch filter is effective for removing AC power-supply noise. Another type of noise, heart rate variability (HRV), is a biological signal. Spontaneous heart rate variability, observed in the resting state, comprises waves of various speeds and affects the frequency band of 0.04–0.5 Hz.

Figure 6 shows the amplitude spectrum of the real-time signal before and after signal processing, indicating that the facial surface EMG at the sensor location used in this study is concentrated in the 1-5 Hz frequency band. Therefore, if we can focus on this frequency band and eliminate all but the frequency band, we can ignore the effects of AC power supply noise and natural variations. In this study, we used a bandpass filter that satisfies this condition. By processing these signals, distinct features can be extracted from the myopotential signal of each word, as shown in Figure 7. The measurement time is 1.6 seconds.

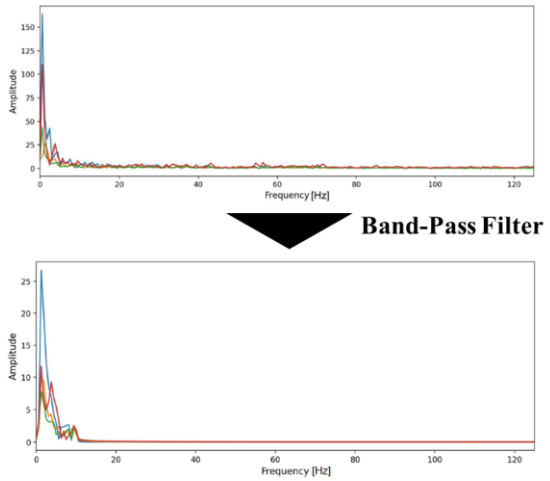

Fig.6: Band-Pass filter

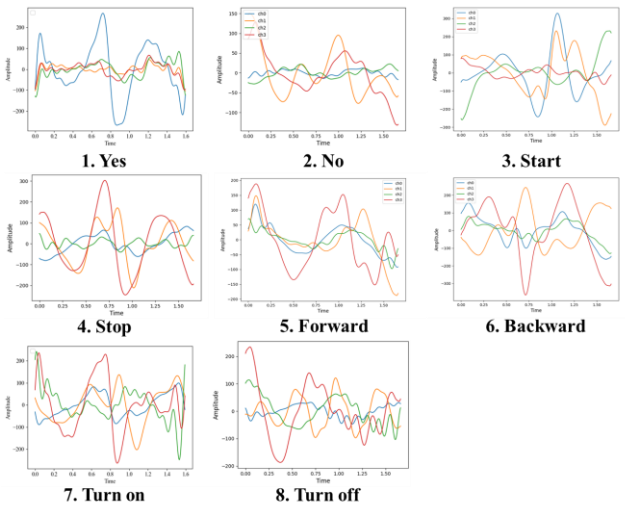

Fig. 7: Features of EMG

4.2. SSR-system (Steady-state)

The SSR recognition model is trained using a one-dimensional convolutional neural network (1D CNN) [10]. Figure. 8 presents a flowchart illustrating the steady-state. Figure. 9 illustrates the 1D CNN machine learning configuration. This section describes the training data used to verify the identification accuracy of the proposed method. We prepared a training dataset consisting of eight Japanese words: 1) Yes, 2) No, 3) Start, 4) Stop, 5) Forward, 6) Backward, 7) ON, 8) OFF. Each word was represented by 30 data points, totaling 240 datasets. The electrodes were reattached to the same positions, and a different test dataset was generated with 80 data points (10 data points for each word) across five repetitions. The word recognition accuracy was evaluated in five tests using a confusion matrix to assess the effectiveness of SSR for facial EMG potentials. To mitigate identification failures due to timing variations, the data were augmented by shifting them 160 milliseconds to the left and right. Missing data from these shifts were filled with zero values. Dropout was used to prevent overfitting [11], and Adam was employed as the optimization algorithm [12].

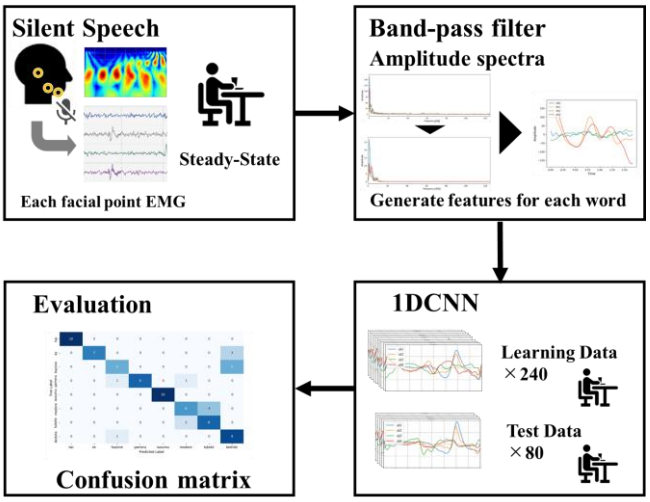

Fig. 8: Flowchart of proposed method in steady-state

Input	input shape=(400,4)
Conv1D	filters=64, k size=24, activation=relu
MaxPooling1D	pooling size=2, stride size=2
Conv1D	filters=64, k size=12, activation=relu
MaxPooling1D	pooling size=2, stride size=2
Conv1D	filters=64, k size=6, activation=relu
MaxPooling1D	pooling size=2, stride size=2
Conv1D	filters=64, k size=3, activation=relu
MaxPooling1D	pooling size=2, stride size=2
Conv1D	filters=64, k size=3, activation=relu
MaxPooling1D	pooling size=2, stride size=2
Dropout	0.7
Flatten	4000
Dense	2000
Dropout	0.5
Dense	8(Numclass)
Output	Accuracy results

Fig. 9:1DCNN training structure

4.3. Result (Steady-state)

Table 1 shows the results of recognition accuracy using the confusion matrix during validation, with an average accuracy of 88.75% over five trials. The highest accuracy achieved was 91.25%. Fig. 10 displays the confusion matrix corresponding to this highest accuracy, confirming that the system can recognize words with high accuracy under steady-state conditions. Next, we explain the proposed method for addressing body motion noise, which is a challenge in SSR.

Table 1: Accuracy Results (Steady-state)

Steady-state	Accuracy [%]
Best	91.25
Average	88.75

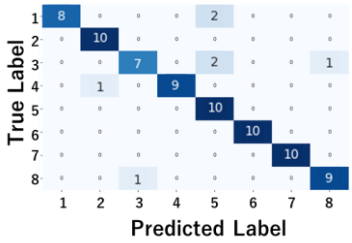

Fig. 10: Best Accuracy

5. ELIMINATION METHOD OF BODY MOTION NOISE

5.1. Effects of body motion noise on signals

In this study, body motion was defined as walking, and electromyogram signals were measured while the subject walked. Vibrations caused by walking interfere with myopotential signals, specifically due to the shaking of the wiring and friction between the electrodes and the skin. Although fixing the device and wiring and applying electrode paste can minimize the effects of body motion, it cannot be entirely eliminated. Fig. 11 illustrates the signals measured at rest and during walking while the same words were silently spoken. The walking speed was not controlled; subjects walked at their normal daily pace. The body movement noise during walking, centered around approximately 5 Hz, overlaps with the frequency band of the EMG and cannot be removed through signal processing methods like bandpass filtering. Consequently, few studies have incorporated gait conditions into measurement methods. Therefore, we propose a method to reduce body motion noise.

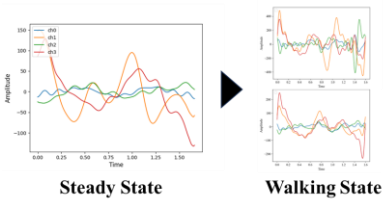

Fig. 11: Effects of body motion noise on signals

5.2. VMD

Variational Mode Decomposition (VMD) [13] is a signal processing technique that decomposes a time-series signal into multiple modes, known as intrinsic mode functions (IMFs). In contrast, Conventional Empirical Mode Decomposition (EMD) [14] is a simpler and computationally less expensive algorithm, as it does not require predefined parameters and performs decomposition directly from the signal. However, EMD can result in overlapping IMFs within the frequency bands, causing mixing between modes. In contrast, VMD requires setting parameters that define the number of modes and the bandwidth of each mode. Although VMD is computationally more expensive due to its optimization-based approach, it is more effective at suppressing noise by producing distinct modes with minimized frequency overlap. The basic structure and features of VMD are described below.

The basic mechanism of VMD is similar to that of EMD. Initially, the input signal is decomposed into multiple IMFs, with each IMF representing a component concentrated in a specific frequency band. The signal $y(t)$ is expressed as the sum of the IMFs. The equation is as follows (2):

$$y(t) = \sum_{j=1}^{J} u_j(t) + r_j(t) \tag{2}$$

This indicates decomposition into J IMFs $u_j(t)$ and residual $r_J(t)$. The decomposition of each IMF is referred to as "sieving." This process recursively identifies the local maxima and minima of the signal, interpolates these extrema to estimate the lower and upper envelopes, and then removes the mean to create a "low-pass" centerline. High-frequency oscillations are separated into "modes." Therefore, the IMF can be expressed by the following equation (3).

$$x(t) = \sum_{k=1}^{K} u_k(t) \tag{3}$$

The IMF is a signal with an AM-FM structure and has the following properties (4).

$$u_k(t) = A_k(t)\cos(\phi_k(t)) \tag{4}$$

$\phi_k(t)$ is the phase of the IMF and is a non-decreasing function. $A_k(t)$ is an envelope that is always positive. Then, both $A_k(t)$ and the instantaneous frequency $\phi'_k(t)$, which is concentrated around the median $\omega_k(t)$, change very slowly compared to phase (5).

$$\begin{aligned} \mathcal{L}(u_k(t), \omega_k, \lambda) = {} & \alpha \sum_{k=1}^{K} \left\| \frac{d}{dt}\left[\left(\delta(t) + \frac{j}{\pi t}\right) * u_k(t)\right] e^{-i\omega_k t} \right\|_2^2 \\ & + \left\| x(t) - \sum_{k=1}^{K} u_k(t) \right\|_2^2 \\ & + \langle \lambda(t), x(t) - \sum_{k=1}^{k} u_p(t) \rangle \end{aligned} \tag{5}$$

Unlike EMD, VMD simultaneously estimates all IMFs and their center frequencies and finds $u_k(t)$ and $\omega_k(t)$ that minimizes the constrained variational problem. To minimize $u_k(t)$ and $\omega_k(t)$, we use the augmented Lagrangian method, as shown in the following equation: Let $\delta_k(t)$ be the Dirac distribution and (*) be the convolution operator. The solution is obtained as a saddle point through a series of iterative

suboptimizations known as the Alternating Direction Method of Multipliers (ADMM) [15].

Ultimately, the sum of the decomposed modes reconstructs the original signal. The user can select only the necessary modes that do not contain noise and reconstruct the signal. In this study, we applied VMD to remove body motion noise, with the principal diagram of this process illustrated in Fig. 12.

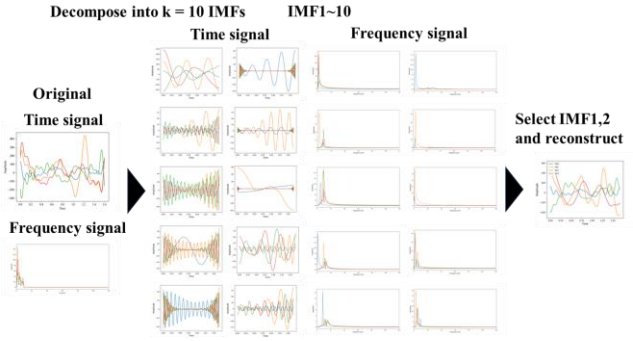

Fig. 12: Principal diagram of the VMD

In this study, the number of modes (IMFs) to be decomposed in the VMD was set to four; the decision to decompose into four modes is based on the principle that increasing the number of decomposition modes reduces the number of features per time signal. As the number of modes is increased, the decomposition becomes more detailed, and it is difficult to identify which mode contains the signal with the required characteristics. Increasing the number of mode decompositions is effective when the signal sampling time is long. However, the signal sampling time per word in this study is as short as 1.6 seconds. Referring to previous studies with equal signal sampling times [6], we decomposed the signal into four modes.

The criterion for the center frequency at which the signal is decomposed is that a smaller IMF number represents a lower frequency and a larger IMF number represents a higher frequency. IMF3 and IMF4 are high frequency portions of the original signal and inevitably contain signals accompanied by body motion. Furthermore, IMF1 is a low-frequency signal that is close to the average of the upper and lower envelopes and suppresses fluctuations caused by walking. In a previous study [6], only IMF1 was selected to reconstruct the signal. However, when the signal is reconstructed using only IMF1, not only body movement noise but also necessary myopotential frequency components are removed, resulting in smaller features. Therefore, the signal differs significantly from that of the steady-state model, and words cannot be correctly identified. In other words, IMF2 contains the necessary myopotential signals. IMF2 also has the effect of suppressing intermediate fluctuations in the signal. Therefore, selecting IMF1 and 2 may be effective in minimizing the effects of body vibration while maintaining the characteristics of the measured signal. This finding was confirmed by empirical experiments in which multiple modes were selected; parameter setting through empirical experiments is important for effective noise reduction using VMD. Fig. 13 shows the EMG potential features with IMF1 or IMF1 and 2 selected after decomposition into four modes.

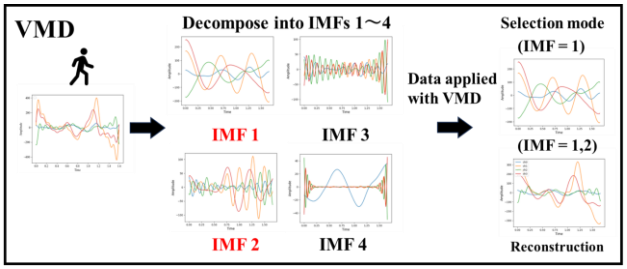

Fig. 13: VMD (Decomposition 4, IMF1, IMF1 and 2)

5.3. SSR-system (Walking-state)

Figure. 14 presents a flowchart illustrating the walking state. Two patterns were prepared to evaluate the accuracy improvement for walking using the proposed method in this study: 1) training data consist of steady-state data, and test data consist of walking-state data; 2) 3) training data consist of steady-state data, and test data consist of walking-state data with VMD applied.

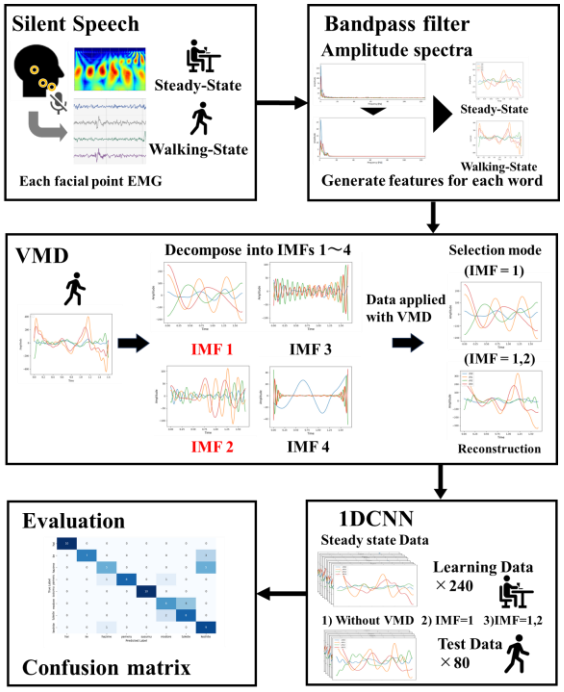

Fig. 14: Flowchart of proposed method in walking state

5.4. Result (Walking state)

Table 2 displays the walking accuracy results. Fig. 15, Fig. 16 and Fig. 17 show the confusion matrices with and without the proposed method, respectively. The proposed method with decomposition 4 and IMF1and2 selection modes reduced the number of errors in word identification and improved accuracy from 48.75% to 73.25%. This indicates that the proposed method can identify words with approximately 75% accuracy using myopotentials, even while walking. The optimal parameter settings of the VMD improved the accuracy over previous studies [3]. However, this method has certain limitations. The frequency band of the EMG signal overlaps with body motion noise, and there is a limit to body motion noise removal in signal processing. The main source of body motion noise is friction between the electrode and the skin, necessitating improved electrode

fixation and miniaturization of the sensor device. Additionally, user tuning plays a crucial role. Training data were developed through trial and error, and the model included voice data of all variations. Therefore, it is important to clarify the minimum tuning time required for this study. Furthermore, to enhance the system's practicality, the vocabulary must be expanded beyond the current eight words. To increase the vocabulary, it is necessary to take into account the more subtle differences in EMG features of each word. Therefore, tongue movement should be taken into account and an additional electrode should be added under the jaw.

Table 2: Accuracy Results for (walking-state)

Walking-state	Accuracy [%]
Without VMD	48.75
VMD (IMF=1)	53.75
VMD (IMF=1,2)	73.25

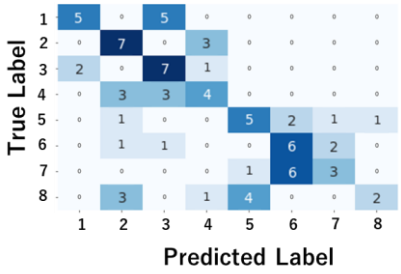

Fig. 15: Without VMD

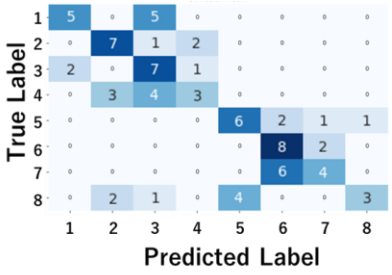

Fig. 16: VMD (IMF=1)

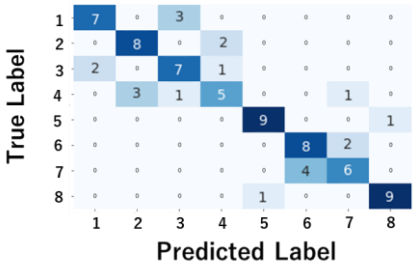

Fig. 17: VMD (IMF=1,2)

6. Conclusion

In this paper, we proposed a method to reduce body motion noise and improve the accuracy of SSR using facial surface EMG. While our results demonstrate that the proposed method enhances accuracy during walking, we acknowledge that signal processing alone has limitations in eliminating body movement noise. In this study, the sensor device was fixed to a cap for measurement. Although expensive, small wireless electrode devices weighing 7g-12g are available among sensor devices. To improve the practicality and accuracy, we plan to address this issue in the future using two approaches: fabrication of a small, wearable sensor device with minimal cost (simple appearance that does not bother the human eye, electrodes fixed without seals) and development of advanced signal processing technology. These combined efforts aim to enable reliable SSR regardless of the user's environment and usage conditions.

Acknowledgements

This work was supported by KEIRIN JKA (2024M-381) and the Institute of Science and Engineering of Chuo University.

References

[1] Laxmi Pandey, Ahmed Sabbir Arif: "LipType: A Silent Speech Recognizer Augmented with an Independent Repair Model", Conference on Human Factors in Computing , Article No.: 1, pp.1-19(2021)

[2] Naoki Kimura, Tan Gemicioglu, Jon Womack,Yuhui Zhao, Thad Starner, Abdelkareem Bedri, Richard Li, Alex Olwal, Jun Rekimoto: "Mobile, Hands-free, Silent Speech Texting Using SilentSpeller", Conference on Human Factors in Computing, No.178, pp.1-5(2021)

[3] Arnav Kapur, Shreyas Kapur, Pattie Maes: "AlterEgo: A Personalized Wearable Silent Speech Interface", International Conference on Intelligent User Interfaces, pp 43-53(2018)

[4] David Gaddy, Dan Klein: "Digital Voicing of Silent Speech", Conference on Empirical Methods in Natural Language Processing, pp.5521–5530(2020)

[5] Arnav Kapur, Utkarsh Sarawgi, Eric Wadkins, Matthew Wu, Nora Hollenstein, Pattie Maes: "Non-Invasive Silent Speech Recognition in Multiple Sclerosis with Dysphonia", Proceedings of the Machine Learning for Health NeurIPS Workshop, pp.116:25-38(2020)

[6] H. Ikeda, T. Ohhira and H. Hashimoto: "Classification of silent speech words considering walking using VMD applied facial EMG", the 9th International Symposium on Affective Science and Engineering (ISASE2023), C000013, March 8 2023.

[7] "OpenBCI — Home", https://openbci.com/

[8] Hung-Yuan Chung: "The review of applications and measurements in facial electromyography", Journal of Medical and Biological Engineering, 25 pp.15-20(2005)

[9] D.T. Mewett, H. Nazeran, K.J. Reynolds: "Removing power line noise from recorded EMG", 2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, (2001)

[10] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu: "WaveNet: A Generative Model for Raw Audio", arXiv:1609.03499, (2016)

[11] Yarin Gal, Zoubin Ghahramani: "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning", ICML, (2016)

[12] Diederik P. Kingma, Jimmy Lei Ba: "ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION", International Conference for Learning Representations, (2015)

[13] Shihan Ma, Bo Lv, Chuang Lin, Xinjun Sheng, Xiangyang Zhu: "EMG Signal Filtering Based on Variational Mode Decomposition and Sub-Band Thresholding", IEEE Journal of Biomedical and Health Informatics, vol. 25, pp.47-58(2021)

[14] I. Daubechies, J. Lu, and H.-T. Wu: "Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool", Appl. Computat. Harmon. Anal, vol. 30, No.2, pp.243-261(2011)

[15] Konstantin Dragomiretskiy, Dominique Zosso: "Variational Mode Decomposition", IEEE TRANSACTIONS ON SIGNAL PROCESSING, vol. 62, No.3(2014)