INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# Hybrid Silent Speech Interface Through Fusion of Electroencephalography and Electromyography

*Huiyan Li*[1], *Mingyi Wang*[1], *Han Gao*[1], *Shuo Zhao*[1], *Guang Li*[1*] , *You Wang*[1]

[1]State Key Laboratory of Industrial Control Technology, Institute of Cyber Systems and Control, Zhejiang University, China

huiyanli@zju.edu.cn, wangmingyi@zju.edu.cn, gao_han@zju.edu.cn, zhaoshuo@zju.edu.cn, guangli@zju.edu.cn, king_wy@zju.edu.cn

## Abstract

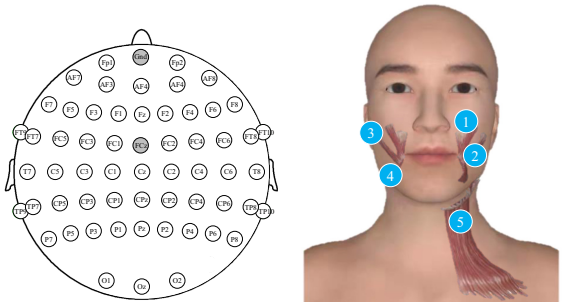
Silent Speech Interface (SSI) can enable interaction in a new and natural way based on no-audible biosignals generated by the human body. Electroencephalography (EEG) or surface electromyography (sEMG) generated during speech production can be utilized to decode silent speech. However, obtaining complementary information from EEG and sEMG is still challenging. This paper presents a hybrid SSI based on the converter between bimodal electrophysiological signals and audio signals. EEG and sEMG are fused through two sequence-to-sequence models, and multi-task losses are applied to achieve complementarity between speech intention and muscle activity in silent speech. The feasibility of the proposed fusion method is validated in the silent speech dataset, and an average objective character error rate (CER) of 7.22% among eight speakers is obtained. The experimental results show that our bimodal-based hybrid SSI facilitates the conversion of electrophysiological signals to audio.

**Index Terms**: Silent Speech Interface (SSI), electroencephalography (EEG), surface electromyography (EMG), bimodal fusion

## 1. Introduction

Silent Speech Interface (SSI) is a system that acquires the speech-related physiological signals from the human speech production process without audible acoustic signal [1]. Recognition and speech synthesis algorithms are applied in SSI for decoding the intended message [2]. SSI offers a new communication method that can be used in noisy environments and privacy scenarios [3]. In addition, SSI can also be applied as a clinical application for people who have undergone a laryngectomy and provides assistive devices to restore oral communication [4].

Electroencephalography (EEG) [5] and surface electromyography (sEMG) [4] are the common physiological measurement methods to capture speech intention and muscle activity in silent speech. EEG characterizes the neural processing of speech production while sEMG records neuromuscular information from the brain to the speech-related articulators during muscle fiber contraction [3]. Single-modal physiological signals first succeeded in recognizing isolated words and continuous sentences [6, 7]. Then deep learning models, such as Long-Short Time Memory (LSTM) and Transformer, were introduced in SSI to convert single-mode electrophysiological signals into speech [4, 8, 9]. In 2022, Gaddy et al. obtained a word error rate of 42% on the English single-speaker silent speech dataset [10]. The results of those papers show that there is room for methodological improvement in reconstructing speech from silent speech. Previous studies have shown that in tasks such as motion and emotion recognition[11, 12], fusing information from different modalities can improve classification accuracy and reliability [11]. However, only a few studies have focused on signal fusion in SSI[13, 14]. Hueber et al. [13] fused ultrasound and optical video for continuous speech recognition. The fusion of EMG and EEG was currently limited to the recognition of phonemes [15]. In order to integrate brain activity signals with muscle activity signals, decision-level fusion is considered effective since it contains information about the different phases of neural activity in vocalization.




(a) *EEG* (b) *sEMG*

Figure 1: *Electrode positions of EEG and sEMG. Gnd is the ground electrode and FCz is the reference electrode.*

In this paper, we propose a novel mapping method by fusing EEG and EMG signals to implement bimodal-based hybrid SSI[1]. The method demonstrates the effectiveness of fusion methods in reconstructing speech from physiological signals. The EEG and sEMG signals are trained with sequence-to-sequence (seq2seq) models separately to obtain complementary information at the decision level. Moreover, the alignment information is optimized by the decision fusion method, and a multi-task strategy is applied to improve the performance of silent speech. We conduct experiments to demonstrate that bimodal fusion physiological signal in SSI is feasible. Besides, our proposed decision-level fusion method is more effective than single-modal and other fusion types.

## 2. Method

In this section, we first introduce the acquisition and preprocessing of experimental data. Then, we detail the pipeline

*Corresponding author

[1]Audio samples can be found at https://stone-wave.github.io/.



10.21437/Interspeech.2023-805

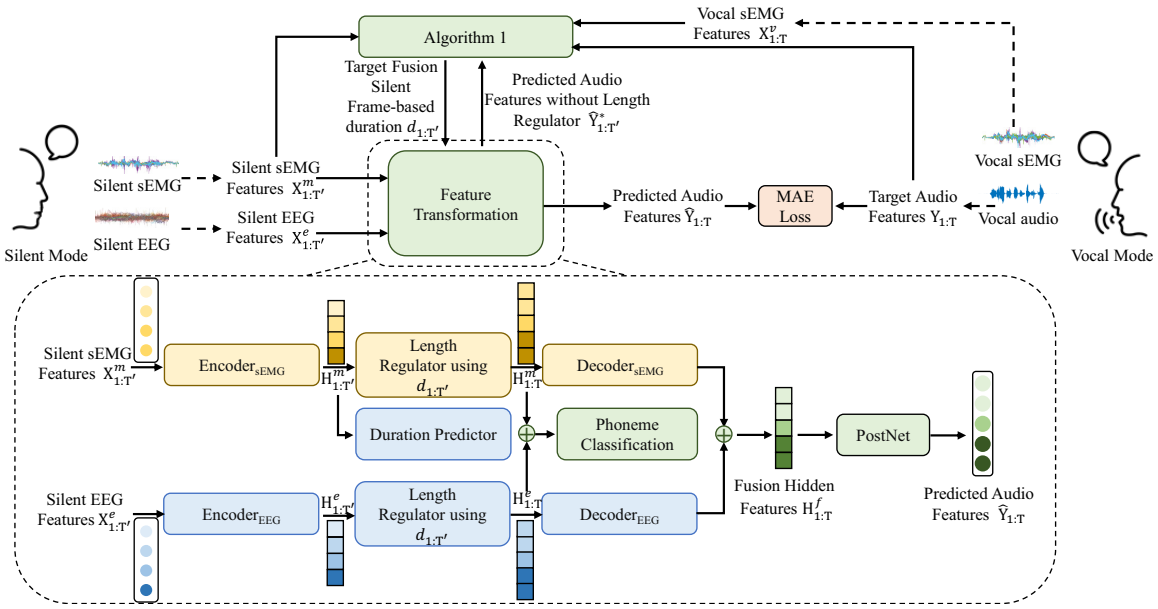



Figure 2: *The pipeline of hybrid SSI through the fusion of EEG and sEMG.*

of hybrid SSI through EEG and sEMG.

### 2.1. Data Acquisition

Simultaneous EEG and EMG signals are recorded using a 64-channel Brain Products actiCHamp Plus amplifier at a sampling rate of 1 kHz. 57 channels of EEG electrode positions are shown in Figure 1(a). 5 channels of EMG positions are shown in Figure 1(b). In addition, audio signals are recorded from a desktop USB microphone at a sampling rate of 16 kHz. Audio and electrophysiological signals are synchronized using arriving time stamps. Eight subjects, two of whom are female, participate in the experiment. All subjects are healthy and native Mandarin speakers, aged between 22 and 28 years, with normal vision and oral communication skills. During the data acquisition, the subjects need to press the start button, then read the sentences displayed on the computer screen without sound or aloud, and press the end button when they have finished reading. The sentences displayed on the screen are from the Mandarin corpus AISHELL3 [16]. Each subject has at least 33.21 minutes and 520 sentences of silent speech. The corpora contain 1170 words and 913 characters. Subjects are required to repeat the text five times silently which is articulated but without sound and once audibly. The vocal electrophysiological signals are only involved in training and not in testing and validating. The silent dataset for each subject is divided into a training set, a validation set, and a testing set in a ratio of 8:1:1. One repeat time of the sentence is selected as the testing or validation set, while the rest are performed as the training set.

### 2.2. Pre-processing and Feature Extraction

The recorded EEG and sEMG signals are analyzed with EEGLAB [17]. 0.5 $\sim$ 128 Hz bandpass filter is applied to the EEG signals, and 1 $\sim$ 400 Hz is applied to the sEMG signals. In addition, ADJUST [18] is used to remove blink artifacts from the EEG. Both electrophysiological signals and audio signals are windowed with a 64 ms Hanning window and 16 ms hop length. The extraction process of sEMG features and the audio features, i.e., mel spectrograms, is followed [9]. The time domain features of EEG are extracted in the same with sEMG. In addition, three energy of Wavelet decomposition coefficients is extracted from EEG for each channel [19]. Finally, the EMG feature dimension is 195 ($5 \times 39$), the EEG feature dimension is 513 ($57 \times 9$), and the mel spectrograms dimension is 80, while 5 and 57 are the channels for sEMG and EEG respectively.

### 2.3. Model Structure

This paper presents a hybrid SSI model that fuses brain activity and facial neuromuscular motor information to synthesize audio signals. The pipeline of the proposed hybrid SSI model is shown in Figure 2. It converts the EEG signals with the sEMG signals in silent mode into audio features in vocal mode. This is the first attempt to convert silent signals into audio by fusing EEG and sEMG signals. In this fusion task, the silent EEG feature is defined as $\mathrm{X}^{\mathrm{e}}_{1:\mathrm{T}'}$, with $\mathrm{X}^{\mathrm{m}}_{1:\mathrm{T}'}$ of the silent sEMG features, where $\mathrm{T}'$ is the number of frames of silent features. Mel spectrograms are defined as $\mathrm{Y}_{1:\mathrm{T}}$, and vocal sEMG features acquired synchronously are defined as $\mathrm{X}^{\mathrm{v}}_{1:\mathrm{T}}$, where $\mathrm{T}$ is the number of frames of vocal features.

$\mathrm{X}^{\mathrm{e}}_{1:\mathrm{T}'}$ and $\mathrm{X}^{\mathrm{m}}_{1:\mathrm{T}'}$ are fed into two separate seq2seq backbones with the same structure. The results of these two backbones are then combined before PostNet to obtain the predicted mel spectrograms. Besides, a silent frame-based duration predictor and a phoneme classification module are introduced to enhance speech reconstruction performance. Finally, a pretrained vocoder generates the waveform with the converted predicted mel spectrograms.

The seq2seq model combines an encoder, a length regulator, and a decoder. This structure has been demonstrated effective in a single-modal sEMG-to-speech task [9]. The encoder used in this paper is an optimized Conformer [20] structure, removing the Macaron structure and using the activation function called scaled exponential linear unit (SeLU) [21] in the convolutional module. This new structure aims to explore the correlation between audio and physiological signals in silent speech

**Algorithm 1:** Fusion Silent Frame-based Duration Extraction in hybrid SSI

**Output:** The set of fusion silent frame-based duration sequence $\boldsymbol{d}_{1:\mathrm{T}'}$

Given a fixed alignment weight $\lambda_{\mathrm{align}}$;
Initialize accumulated cost matrix $\mathbf{D} \in \mathbb{R}^{\mathrm{T}' \times \mathrm{T}}$ ;
**Function** *get_align(*$\mathbf{C}$*)* **:**
  **for** $j = 1$ to $\mathrm{T}$ **do**
    **for** $i = 1$ to $\mathrm{T}'$ **do**
      $\mathbf{D}(i,j) = \mathbf{C}(i,j) + \min\{\mathbf{D}(i-1,j), \mathbf{D}(i,j-1), \mathbf{D}(i-1,j-1)\}$ ;
    **end**
    $align[j] = \arg\min_i \mathbf{D}(i,j)$ ;
  **end**
**end**
**for** *each sample in features set* **do**
  $\mathbf{C}(i,j) = \|\mathrm{X}^{\mathrm{m}}_{1:\mathrm{T}'}[\mathrm{i}] - \mathrm{X}^{\mathrm{y}}_{1:\mathrm{T}}[\mathrm{j}]\|$ ;
  $align_{1:\mathrm{T}} = get\_align(\mathbf{C})$;
  Calculate the initial silent frame-based duration $\boldsymbol{d}_{1:\mathrm{T}'}$ using Eq. 1;
**end**
**while** *Train* **do**
  Train five epochs ;
  **for** *each sample in features set* **do**
    Hidden features after the $\mathrm{Encoder}_{\mathrm{EEG}}$ $\mathrm{H}^{\mathrm{e}}_{1:\mathrm{T}'} \leftarrow \textbf{Encoder}_{\mathrm{EEG}}(\mathrm{X}^{\mathrm{e}}_{1:\mathrm{T}'})$ ;
    Hidden features after the $\mathrm{Encoder}_{\mathrm{sEMG}}$ $\mathrm{H}^{\mathrm{m}}_{1:\mathrm{T}'} \leftarrow \textbf{Encoder}_{\mathrm{sEMG}}(\mathrm{X}^{\mathrm{m}}_{1:\mathrm{T}'})$ ;
    Predicted mel spectrograms without length regulator $\hat{\mathrm{Y}}^{*}_{1:\mathrm{T}'} \leftarrow \textbf{PostNet}(\textbf{Decoder}_{\mathrm{EEG}}(\mathrm{H}^{\mathrm{e}}_{1:\mathrm{T}'}) + \textbf{Decoder}_{\mathrm{sEMG}}(\mathrm{H}^{\mathrm{m}}_{1:\mathrm{T}'})$ ;
    $\mathbf{C}(i,j) = \|\mathrm{X}^{\mathrm{m}}_{1:\mathrm{T}'}[\mathrm{i}] - \mathrm{X}^{\mathrm{y}}_{1:\mathrm{T}}[\mathrm{j}]\| + \lambda_{\mathrm{align}}\|\hat{\mathrm{Y}}^{*}_{1:\mathrm{T}'}[\mathrm{i}] - \mathrm{Y}_{1:\mathrm{T}}[\mathrm{j}]\|$ ;
    $align_{1:\mathrm{T}} = get\_align(\mathbf{C})$;
    Calculate target fusion silent frame-based duration $\boldsymbol{d}_{1:\mathrm{T}'}$ using Eq. 1 ;
  **end**
  Update $\boldsymbol{d}_{1:\mathrm{T}'}$;
**end**
**return** $\boldsymbol{d}_{1:\mathrm{T}'}$

production and accelerate the convergence of the network during training. $\mathrm{H}^{\mathrm{e}}_{1:\mathrm{T}'}$ and $\mathrm{H}^{\mathrm{m}}_{1:\mathrm{T}'}$ are the hidden features after the $\mathrm{Encoder}_{\mathrm{EEG}}$ and $\mathrm{Encoder}_{\mathrm{sEMG}}$. The length regulator is designed to solve the problem of frame mismatch between the electrophysiological features in silent mode and mel spectrograms in vocal mode [9, 22], solving the mismatch between the speech modes. Based on the length regulator, $\mathrm{H}^{\mathrm{e}}_{1:\mathrm{T}'}$ and $\mathrm{H}^{\mathrm{m}}_{1:\mathrm{T}'}$ can be regulated into $\mathrm{H}^{\mathrm{e}}_{1:\mathrm{T}}$ and $\mathrm{H}^{\mathrm{e}}_{1:\mathrm{T}}$ using fusion silent frame-based duration $d_{1:\mathrm{T}'}$. We can match the speech rates of different patterns. $d$ utilized in the length regulator is computed by Algorithm 1 and Eq. 1 in the training stage. The duration predictor is trained using $\mathrm{H}^{\mathrm{m}}_{1:\mathrm{T}'}$, which aims to consider that sEMG is closer to the final production of speech than EEG. The fusion silent frame-based duration sequence is obtained from the pre-trained duration predictor in the inference stage. Finally, $\mathrm{H}^{\mathrm{e}}_{1:\mathrm{T}}$ and $\mathrm{H}^{\mathrm{m}}_{1:\mathrm{T}}$ are summed to train the phoneme classification.

$$d_{1:\mathrm{T}'}[i] = \sum_{j=1}^{\mathrm{T}} \mathbb{I}(align_{1:\mathrm{T}}[j] == i) \tag{1}$$

where $\mathbb{I}$ is an indicator function.

The proposed model is trained to optimize three loss functions simultaneously: a speech synthesis loss, which predicted mel spectrograms before and after the PostNet as [23]; a duration loss and a phoneme classification loss as auxiliary tasks to aid convergence. Besides, this study utilizes Parallel WaveGAN (PWG) [24] as the vocoder in hybrid SSI.

# 3. Results

In this section, we evaluate the proposed method in hybrid SSI. We first introduce the experimental setting. Then, we evaluate the proposed model with the objective and human metrics. Furthermore, we conduct a comparison study on the fusion types. Finally, we provide more insights into the brain region study from the aforementioned results.

## 3.1. Experimental Settings

The implementation of the proposed hybrid SSI model is based on the open-source ESPnet toolkit [25]. The number of encoder and decoder blocks is 6, and the number of attention heads is 4. The attention dimension of feature transformation is set to 384, and the kernel size of the convolutional module in the optimized Conformer is 7. Besides, the size of the phoneme vocabulary is 139, including consonants and toned vowels. The batch size is eight signals, and the epoch is 160. $\lambda_{\mathrm{align}}$ is set to 10 as [8]. The feature transformation module's learning rate schedule and the other hyper-parameters are consistent as [9]. PWG is pre-trained by audio signals from all speakers with the implementation[2].

The character error rate (CER) using automatic speech recognition (ASR) evaluation tool is the objective evaluation criterion in our paper. CER is obtained by computing the ground truth text and the text recognized from an ASR called Citrinet [26][3]. The lower CER suggests that the context of predicted audio signals are close to the silent speech. As a reference, the average CER of the ground truth audio signal for the testing set is 0.85%.

The CER of the validation set is calculated for each epoch during training. The parameter of the epoch with the lowest CER is chosen as the best model for testing. EEG and the sEMG features from the testing set are input to the best parameter model to generate the predicted audio as the testing result for each speaker. Besides, for the objective quality evaluation of the predicted audio in the testing set, mel cepstral distortion (MCD) [27] is also introduced. The lower MCD indicates a better similarity of the predicted audio signals with the ground truth.

## 3.2. Model Performance

The objective evaluation of the results of the proposed fusion method on the hybrid SSI task is shown in Table 1. ↑ (↓) means higher (lower) is better. The second row shows the CER results obtained by the ASR toolkit for the predicted speech obtained from the testing set. Our proposed method obtains an average

[2]https://github.com/kan-bayashi/ParallelWaveGAN

[3]https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_zh_citrinet_512

Table 1: *The objective evaluation of the proposed model*

| | $Spk_1$ | $Spk_2$ | $Spk_3$ | $Spk_4$ | $Spk_5$ | $Spk_6$ | $Spk_7$ | $Spk_8$ | Average | Std |
|---|---|---|---|---|---|---|---|---|---|---|
| CER(%)↓ | **1.00** | 9.63 | 5.42 | 11.3 | 5.93 | 12.39 | 9.14 | 2.94 | 7.22 | 3.79 |
| MCD(dB)↓ | **2.76** | 3.09 | 2.95 | 3.55 | 4.03 | 3.10 | 3.05 | 2.89 | 3.18 | 0.39 |

Table 2: *The human evaluation of the proposed model*

| | $Spk_1$ | $Spk_2$ | $Spk_3$ | $Spk_4$ | $Spk_5$ | $Spk_6$ | $Spk_7$ | $Spk_8$ | Average | Std |
|---|---|---|---|---|---|---|---|---|---|---|
| CER(%)↓ | **1.48±1.05** | 7.51±4.42 | 1.48±1.87 | 4.72±4.85 | 2.50±3.95 | 1.48±1.87 | 14.56±0.92 | 4.59±3.42 | 6.89 | 5.22 |
| MOS↑ | 4.33±0.57 | 3.88±0.59 | **4.63±0.56** | 3.96±0.84 | 4.42±0.61 | 2.96±0.99 | 3.83±0.74 | 4.13±0.83 | 3.92 | 4.82 |

Table 3: *Comparison results of the fusion type study*

| Fusion Type | CER(%) ↓ | MCD(db)↓ |
|---|---|---|
| Only EEG | 38.85±14.98 | 4.55±0.77 |
| Only sEMG | 10.52±5.24 | 3.10±0.31 |
| Feature Concatenate | 13.14±6.74 | 3.26±0.31 |
| Concatenate after Encoder | 8.34±4.47 | 3.14±0.41 |
| Add after Encoder | 9.50±4.41 | **3.05±0.23** |
| Concatenate after Decoder | 9.90±7.19 | 3.30±0.49 |
| Proposed | **7.22±3.79** | 3.18±0.39 |

Table 4: *Results of the brain region study*

| Brain Region | Number of Channels | CER(%)↓ | MCD(dB)↓ |
|---|---|---|---|
| w/o Frontal Lobe | 37 | 7.34±4.58 | 4.01±0.35 |
| w/o Central Sulcus | 38 | 9.46±7.89 | 4.00±0.31 |
| w/o Parietal Lobe | 42 | 7.15±5.41 | 4.05±0.43 |
| w/o Occipital Lobe | 54 | **6.61±4.04** | 4.02±0.41 |
| w/o Temporal Lobe | 47 | 7.57±4.73 | 3.18±0.39 |
| Full | 57 | 7.22±3.79 | **3.18±0.39** |

CER of 7.22% with a standard deviation of 3.79% on all subjects. The results of silent speech reconstruction vary between speakers, with $Spk_1$ obtaining the best result of 1.00% while $Spk_6$ have the worst CER of 12.39%. The third row shows the MCD obtained between the predicted audio signals and the ground truth. The speech quality evaluation results differ from the accuracy, but the best quality results are also reflected in $Spk_1$.

12 native Chinese listeners ($20 \sim 28$ years old) are recruited to evaluate the proposed method's results. These listeners have all passed a pretest of normal hearing function. We provide 5 randomly selected speech samples from the predicted speech results of each subject's testing set and for a total of 40 speech samples. Listeners are required to transcribe the speech in a quiet environment with headphones. In addition, Mean Opinion Scores (MOS) are employed as the quality evaluation of the predicted speech, with evaluation scores ranging from 0.5 to 5.0 with an interval of 0.5 points. Higher MOS scores represent higher voice quality.

Table 2 lists the results of the human evaluations, where $\pm$ indicates the standard deviation of the metrics across 12 listeners. The predicted speech obtained an average CER of 6.89% and a MOS of 3.92 among human listeners. Due to the variability in subjective ratings between listeners, all subsequent subsections of this paper are objective evaluations.

### 3.3. Comparison with Other Fusion Types

To illustrate the method's superiority in this paper, we compared the single-modal method with the fusion method at different locations. As shown in the second and third rows of Table 3, compared to the single-modal method using only sEMG, the method proposed in this paper achieves a 3.30% reduction in CER. Compared to the single-modal method using only EEG, the method proposed in this paper achieves a 31.63% reduction in CER and a 1.37 reduction in MCD. These results demonstrate that the fusion method proposed in this paper successfully obtains bimodal information in hybrid SSI. The fourth row in Table 3 shows the results of the direct concatenation of sEMG with EEG features. The fifth and sixth rows show the results of concatenating or summing hidden representations obtained after $\text{Encoder}_{\text{sEMG}}$ and $\text{Encoder}_{\text{EEG}}$, respectively. The seventh line shows the results of the positional concatenate consistent with the methods of this paper. Compared to these methods, the fusion method proposed in this paper outperforms others. Fusing in the middle of the model or the decision level achieves a lower CER and improves the speech quality compared to the direct concatenate of features. Direct concatenation may ignore the delay between brain activity and neuromuscular motor information in speech production.

### 3.4. Brain Region Study

To analyze the effect of different brain regions on the hybrid SSI, we compare the objective results of removing a single brain region from the EEG, as shown in Table 4, w/o is without. Comparing the results of removing a single brain region show that the Occipital lobe has a negative effect on the hybrid SSI. Conversely, the Central sulcus has a more significant effect on the results because it contains motor information [28].

## 4. Conclusions

In this paper, we propose a bimodal fusion-based SSI and investigate the feasibility of converting fused EEG and sEMG signals into audio signals. We demonstrate experimentally that the bimodal fusion model performs better than the single-modal in the hybrid SSI task. The method proposed in this paper leads to complementary information that can be obtained from EEG and sEMG to improve the accuracy of silent speech decoding. It also provides a feasible direction for related SSI research.

## 5. Acknowledgements

This work is supported by the Zhejiang University Global Partnership Fund and the Fundamental Research Funds for the Central Universities 226-2022-00086.

# 6. References


[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[2] J. A. G. López, A. G. Alanís, J. M. Martín-Doñas, J. L. Pérez-Córdoba, and A. M. Gomez, "Silent speech interfaces for speech restoration: A review," *IEEE Access*, vol. 8, pp. 177 995–178 021, 2020.

[3] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-based spoken communication: A survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2257–2271, 2017.

[4] M. Janke and L. Diener, "Emg-to-speech: Direct generation of speech from facial electromyographic signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2375–2385, 2017.

[5] G. Krishna, C. Tran, Y. Han, M. Carnahan, and A. H. Tewfik, "Speech synthesis using EEG," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1235–1238.

[6] C. Jorgensen, D. D. Lee, and S. Agabont, "Sub auditory speech recognition based on emg signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, 2003, pp. 3128–3133.

[7] S.-C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards continuous speech recognition using surface electromyography," in *Proc. INTERSPEECH 2006 – Annual Conference of the International Speech Communication Association*, 2006.

[8] D. Gaddy and D. Klein, "Digital voicing of silent speech," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 5521–5530.

[9] H. Li, H. Lin, Y. Wang, H. Wang, M. Zhang, H. Gao, Q. Ai, Z. Luo, and G. Li, "Sequence-to-sequence voice reconstruction for silent speech in a tonal language," *Brain Sciences*, vol. 12, no. 7, p. 818, 2022.

[10] D. Gaddy, D. Klein, C. Zong, F. Xia, W. Li, and R. Navigli, "An improved model for voicing silent speech," in *Proceedings of the 59th Conference of the Association for Computational Linguistics (ACL)*, 2021, pp. 175–181.

[11] J. Tryon and A. L. Trejos, "Classification of task weight during dynamic motion using EEG–EMG fusion," *IEEE Sensors Journal*, vol. 21, no. 4, pp. 5012–5021, 2021.

[12] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal transformer fusion for continuous emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3507–3511.

[13] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Communication*, vol. 52, pp. 288–300, 2010.

[14] E. Tatulli and T. Hueber, "Feature extraction using multimodal convolutional neural networks for visual speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2971–2975.

[15] K. Chuysud and Y. Punsawad, "Hybrid EEG-fEMG based Human-Machine Interface for communication and control applications," in *2019 16$^{th}$ International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2019, pp. 1–5.

[16] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A multi-speaker mandarin TTS corpus and the baselines," *arXiv preprint arXiv:2010.11567*, 2020.

[17] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[18] A. Mognon, J. Jovicich, L. Bruzzone, and M. Buiatti, "Adjust: An automatic eeg artifact detector based on the joint use of spatial and temporal features," *Psychophysiology*, vol. 48, no. 2, pp. 229–240, 2011.

[19] C. A. Teixeira, B. Direito, H. Feldwisch-Drentrup, M. Valderrama, R. P. Costa, C. Alvarado-Rojas, S. Nikolopoulos, M. Le Van Quyen, J. Timmer, B. Schelter, and A. Dourado, "EPILAB: A software package for studies on the prediction of epileptic seizures," *Journal of Neuroscience Methods*, vol. 200, no. 2, pp. 257–271, 2011.

[20] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. INTERSPEECH 2020 – 21 $^{st}$ Annual Conference of the International Speech Communication Association*, 2020, pp. 5036–5040.

[21] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2017, pp. 971–980.

[22] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019, pp. 3165–3174.

[23] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[24] R. Yamamoto, E. Song, and J. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203.

[25] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: end-to-end speech processing toolkit," in *Proc. INTERSPEECH 2018 – 19 $^{th}$ Annual Conference of the International Speech Communication Association*, 2018, pp. 2207–2211.

[26] S. Majumdar, J. Balam, O. Hrinchuk, V. Lavrukhin, V. Noroozi, and B. Ginsburg, "Citrinet: Closing the gap between non-autoregressive and autoregressive end-to-end models for automatic speech recognition," *arXiv preprint arXiv:2104.01721*, 2021.

[27] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1, 1993, pp. 125–128.

[28] H. Li and F. Chen, "Classify imaginary Mandarin tones with cortical EEG signals," in *Proc. INTERSPEECH 2020 – 21$^{st}$ Annual Conference of the International Speech Communication Association*, 2020, pp. 4896–4900.