Silent Speech Recognition using Electromyography Signals

Darshan Jain and Amitangshu Pal
Computer Science and Engineering, Indian Institute of Technology Kanpur, Kanpur, Uttar Pradesh, India
E-mail:{darshanjain,amitangshu}@cse.iitk.ac.in

***Abstract*—This paper presents a novel autoregressive approach for sentence-level silent speech recognition (SSR) using surface electromyography (sEMG) signals. We propose an attention-enhanced Sequence-to-Sequence (Seq2Seq) model trained on synthetically augmented multi-word EMG sequences built from a fixed vocabulary. Unlike conventional Connectionist Temporal Classification (CTC) based models, our approach supports flexible decoding of word sequences and demonstrates a Word Error Rate (WER) of 9.3% on real continuous sEMG data. We further introduce a pipeline for high-quality data acquisition, preprocessing, and augmentation using OpenBCI Cyton hardware. This work provides a promising step toward practical, low-resource silent speech interfaces.**

***Index Terms*—Silent speech recognition, surface electromyography, sequence-to-sequence model**

I. Introduction

Silent speech recognition (SSR) seeks to decode speech-related information without relying on acoustic signals, offering transformative potential for individuals with speech impairments and for applications demanding secure or noise-robust communication. Traditional assistive devices often depend on residual vocalization or cumbersome interfaces, limiting their usability for those with severe motor impairments and in noise-sensitive environments such as military operations or space missions. Surface electromyography (sEMG) based SSR, which captures facial and articulatory muscle activity through non-invasive electrodes, has emerged as a leading solution owing to its portability, affordability, and resilience to acoustic interference.

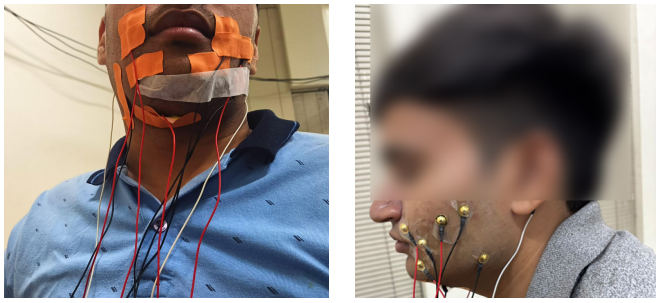
Prior literature: Early SSR efforts have exploited diverse sensing modalities beyond sEMG. In [1], [2], the authors have used ultrasound tongue imaging combined with optical lip video, to enable phoneme decoding via Hidden Markov Models (HMMs), however, the solution requires bulky probes and intensive processing. In reference [3], [4], electromagnetic articulography (EMA) sensors are placed inside the mouth that track articulatory movements, which show moderate accuracy at the cost of intrusive hardware and rigid calibration. In reference [5], [6], electroencephalography (EEG) is used to decode imagined speech directly from cortical signals, however, the solution has low spatial resolution, high artifact susceptibility, and limited vocabularies. Vision-based lipreading leveraging deep convolutional and transformer networks are proposed in [7], [8], to predict viseme sequences, but the solution remains vulnerable to lighting changes, occlusions, and viseme ambiguities (e.g., /p/ vs. /b/). Collectively, these non-EMG methods have laid foundational insights into silent speech decoding but encounter practical barriers that steered the community toward sEMG.

Within the sEMG domain, initial classification-based approaches have addressed isolated units and small vocabularies. The authors in [9] have distinguished five Japanese vowels using perioral sEMG with three electrodes and basic statistical features. Subsequent work [10] has employed Linear Discriminant Analysis (LDA) and k-Nearest Neighbors (k-NN) for command recognition, achieving reasonable performance for limited command sets. Multi-stream HMMs are used in [11], by combining time and frequency domain features, yet all these methods have relied on handcrafted features and rigid model assumptions.

The transition to continuous speech recognition has leveraged phoneme-level modeling and acoustic-inspired frameworks. The authors in [12] have applied HMMs to decode phoneme sequences, moving toward large-vocabulary continuous SSR. Authors in [13] have adapted Gaussian Mixture Model–HMM pipelines from acoustic Automatic Speech Recognition (ASR) system to sEMG, recognizing over 100 words with moderate word error rates. The authors in [14] have further scaled vocabularies to thousands of words by integrating extensive preprocessing routines, but at the expense of substantial training data and complexity.

Recently, deep learning techniques are used heavily to build end-to-end and hybrid architectures, that outperform traditional models. In [15], the authors have replaced GMMs with Deep Neural Networks (DNNs) in hybrid HMM systems, improving continuous SSR accuracy. Authors in [16] have combined Convolutional Neural Networks (CNNs) with Bidirectional Long Short–Term Memory (BiLSTM) layers trained via Connectionist Temporal Classification (CTC), achieving 14.8% word error rate on a 20-word vocabulary. However, fixed-vocabulary constraints and CTC's inflexibility to variable-length sequences limit the scalability of such approaches. Recent advances leverage transformer architectures and data augmentation to model long-range dependencies and address data scarcity. In [17], the authors have proposed a self-attention encoder–decoder on high-density sEMG, reaching a 5.14% character error rate yet relying on 128 channels and massive corpora. Authors in [18] have demonstrated a practical

This research was supported by the IIT Kanpur Initiation Grant IITK/CS/2021164, and Research-I Foundation of the Department of CSE at IIT Kanpur



(a) (b)

Fig. 1: Gold cup electrodes placed over face with two white wires placed behind each ear representing bias and reference.

8-channel system with gated recurrent unit (GRU) backbones and augmentation techniques, achieving up to 97.2% accuracy in single-subject tests for Mandarin phrases. Nevertheless, most state-of-the-art sEMG-SSR solutions still depend on high-density arrays, large-scale multimodal datasets, and complex training schemes, constraining real-world adoption.

Our contributions: The key challenges across these modalities and methods are (a) fixed sentence templates, or in case of limited/fixed vocabularies, addition of further set of words, (b) reliance on handcrafted features or generative model assumptions, (c) solutions like CTC and HMM struggle with variable-length outputs, and (d) requirement of a comprehensive dataset, i.e. in order to collect the EMG data, a significant amount of time is required, which is tiring for the subject and also requires constant observation on signal acquisition setup. Our approach addresses these limitations by combining low-density 8-channel sEMG hardware, a robust preprocessing pipeline, synthetic data augmentation (i.e. cross-fading, time warping, noise injection, baseline-drift simulation etc.), and an attention-based sequence-to-sequence architecture trained on a moderately sized English corpus. Through extensive experiments, we show that the proposed framework achieves high-fidelity sentence-level SSR with a 9.3% word error rate on real, unconstrained sentences, demonstrating the feasibility of lightweight, accessible, and extensible SSR systems. We made our experimental data public for community use[1].

Paper organization: The paper is organized as follows. Our experimental details, along with data collection, cleaning and the proposed solution, are discussed in Section II. Experimental results and performance evaluations are summarized in Section III. The paper is concluded in Section IV.

II. Proposed Solution

Hardware setup: To develop a silent speech recognition (SSR) system that can decode full sentences from facial electromyographic (sEMG) signals, a reliable and well-structured setup is created. This section describes the experimental setup, the process used to prepare the data, and the attention-based sequence-to-sequence model that together make up the core of the system.

[1]https://github.com/DarshanJain295/Silent-Speech-recognition-dataset

The OpenBCI Cyton Board [19] is used for recording the multi-channel sEMG data. The Cyton board features 8 analog input channels, each sampled at 250 Hz, with gold cup electrodes filled with Ten20 conductive gel. These electrodes are placed over targeted facial muscles as shown in Fig. 1. These locations are selected based on both anatomical relevance and evidence from previous studies [14], [20] that have highlighted effective zones for SSR. Specifically, we place the elctrodes around the orbicularis oris below the lower lip, the submental region under the chin, bilateral sternocleidomastoid muscles on the neck, the zygomaticus major and risorius near the cheeks, and the masseter region along the jawline. This spatially distributed setup captures diverse muscle activations involved in silent articulation. Prior research [2], [21] supports the effectiveness of these placements in decoding articulatory gestures during non-vocal speech. In addition to these, a common reference electrode is placed behind the left ear, while a driven bias electrode is placed behind the right ear, to help reduce the common-mode noise and enhance signal stability.

Dataset curation: To construct a focused vocabulary, we start with the TIMIT English speech corpus [22]. After removing punctuation and converting all words to lowercase, we analyze word frequencies across all sentences. By iteratively removing sentences containing the least frequent words, we reduce the dataset until only 22 high-frequency, everyday English words remain. Each resulting sentence is limited to 4–5 words and uses distinct combinations from this vocabulary:

> “all”, “anyone”, “anything”, “are”, “back”, “can”, “come”, “could”, “did”, “doing”, “find”, “help”, “here”, “in”, “now”, “out”, “please”, “right”, “stop”, “that”, “tomorrow”, “you”

For data collection, sEMG signals are first recorded for each of the 22 selected words. Each word is silently mouthed at least 20 times by the subject, who remains seated in a quiet environment and refrains from vocalization. Once the subject is comfortable, data for longer utterances is collected by recording 34 distinct real sentences, each composed of four to five words drawn from the curated vocabulary. The recorded sentences are shown in Table I. These sentences serve as the test set. For training, synthetic sentence sequences are generated by concatenating isolated word recordings, ensuring a robust alignment between training and real-world testing conditions.

Manual event markers are then inserted via the OpenBCI GUI for precise alignment; one marker is placed before articulation and one after, ensuring accurate temporal segmentation of the signals corresponding to words. This approach contrasts sharply with earlier methods [13], that relied on Montreal Forced Alignment, Dynamic Time Warping and canonical correlation analysis (CCA) to infer word boundaries, a method susceptible to cascading alignment errors due to lack of visual ground truth. Notably, all the recordings are preserved, even noisy samples. Additional repetitions are recorded only if quality concerns arise. The sentence-level recordings captured

TABLE I: Sentences used for our experiments

1.	anyone can come in	2.	can anyone help out	3.	can you come back tomorrow	4.	can you come here tomorrow
5.	can you come right now	6.	can you find out	7.	can you find out	8.	can you please stop that
9.	are you all right	10.	are you doing anything	11.	can anyone help you	12.	come back here now
13.	come back right now	14.	come here now please	15.	come here right now	16.	come out right now
17.	could you come here	18.	could you come tomorrow	19.	could you stop please	20.	did anyone help you
21.	did you find anything	22.	please come back tomorrow	23.	please come in now	24.	please come out here
25.	please come right now	26.	please stop doing that	27.	stop doing that please	28.	stop that right now
29.	you are all right	30.	you can all help	31.	you can come in	32.	you can come tomorrow
33.	you can stop now			34.	you come back here		

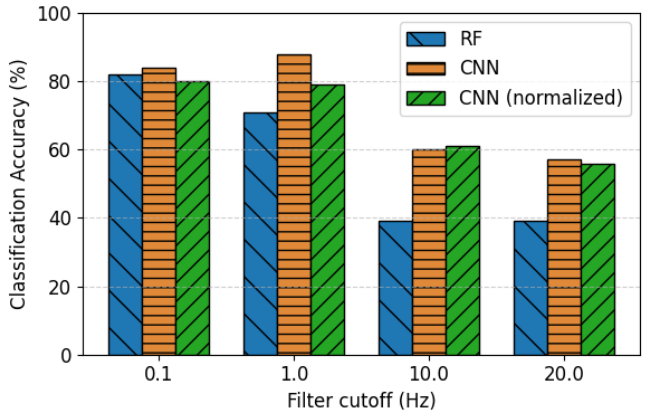


Fig. 2: Comparison of classification accuracy for different high-pass filter cutoff frequencies with and without normalization.

natural timing and brief pauses, closely reflecting realistic silent speaking behavior.

This raw sEMG data taken from the setup is passed through a multi-stage preprocessing pipeline, designed to enhance the signal fidelity for downstream classification and sequence modeling. A key priority during preprocessing is the removal of artifacts such as DC offsets, and baseline drift, which can obscure the subtle temporal dynamics of facial muscle activity. These components are mitigated using high-pass filtering, a common strategy to stabilize EMG baseline [23], [24]. However, determining the appropriate cutoff frequency requires careful consideration due to the unique spectral characteristics of facial EMG, which often contain informative low-frequency components absent in non-facial EMG.

Signal processing: To empirically identify the optimal filter configuration, four high-pass cutoff values of 0.1 Hz, 1 Hz, 10 Hz, and 20 Hz are tested. This process involves training word-level classifiers on filtered datasets to quantify the impact of each cutoff on recognition performance. Classification experiments are conducted using both a Random Forest (RF) and a Convolutional Neural Network (CNN), and CNN followed by normalization. Intuitively, the normalization step plays a critical role in addressing the amplitude variability arising from minor differences in electrode placement, skin conductivity, and inter-session muscle tension. However, performance comparisons from Fig. 2 demonstrate that normalization shows comparable performance with traditional CNN, with no major impact in lower cutoff frequencies.

As shown in Fig. 2, the 0.1 Hz high-pass filter yielded the highest collective classification accuracy across both RF and CNN. This finding suggests that preserving low-frequency signal components is vital when working with facial EMG, as they carry subtle but discriminative muscle activity essential for distinguishing silent speech gestures. Consequently, a 0.1 Hz cutoff is selected for all subsequent experiments, balancing effective noise suppression with the retention of meaningful low-frequency variations.

While collecting data it is observed that sometimes the subject sneezed or didn't mouthed the samples properly, In order to avoid such noisy samples for generating training data we need to remove them. To identify the most consistent and reliable samples for each word, a Multi-DTW (Dynamic Time Warping) analysis [25] is performed. This method calculates pairwise temporal alignment distances between all samples of a given word. From this, the 17 exemplars having the smallest average DTW distance to the rest are selected empirically. These central exemplars represent the most temporally consistent and least noisy ones, effectively forming a core set for data augmentation and modeling. From the 17 exemplars for all words, 14 are randomly designated for training (hereafter referred to as training exemplars) and the remaining 3 for kept for validation, with this split fixed across all experiments to maintain consistency.

Fig. 3 presents the normalized sEMG waveforms for all exemplars of the words *all*, *anyone*, *anything*, and *are* across the eight recording channels. The overlaid signals illustrate that there is substantial overlap among exemplars of the same word. At the same time, distinct activation profiles across channels highlight the discriminative information available for classification. However, this also explains some of the confusions observed between some words. For instance, *anything* and *anyone* display nearly identical waveform shapes in several channels, reflecting overlapping muscle activations that make them harder to distinguish.

Data augmentation: To create synthetic sentences, a sequence of words (typically 4 to 7 words long) is randomly selected. For each word, a exemplar from the training exemplars is chosen. These selected exemplars are then joined together to simulate continuous speech. However, simply joining them would create abrupt transitions at the boundaries between words. To solve this problem, a smoothing technique is used where the last 50 milliseconds of one word and the first 50 milliseconds of the next word are overlapped and gradually faded into each other. This overlap helps reduce abrupt jumps in the signal, making the transitions between words smoother and more natural, similar to how words are articulated in fluent speech, even in silent articulation.

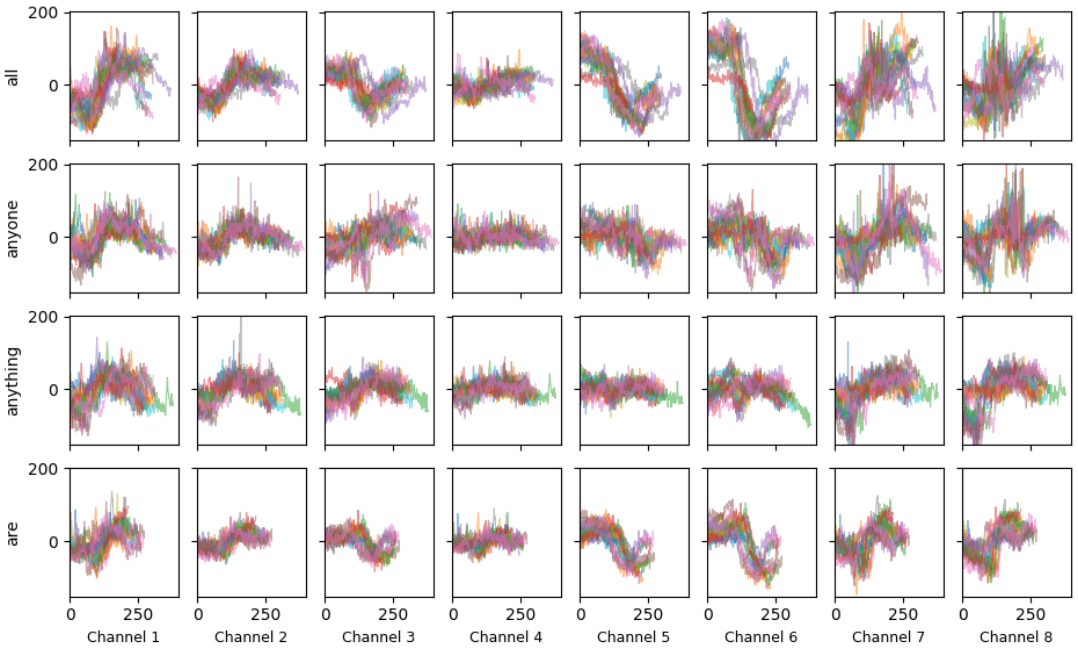


Fig. 3: Normalized sEMG waveforms for the exemplars of "all," "anyone," "anything," and "are," across eight channels.

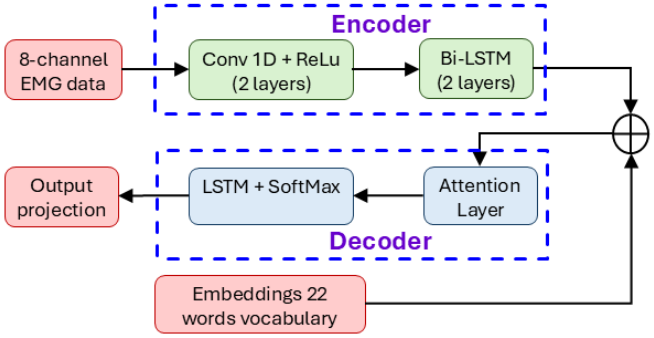


Fig. 4: Architecture of the proposed sequence-to-sequence model.

This synthesis process is further enhanced to make it more robust. First, Gaussian noise is added (SNR between 20–30 dB) to imitate minor variations. Next, we extend or contract the samples by upto 10%, in order to stimulate variations in how fast or slow a person speaks. Finally, a slow, wave-like drift is added to the signals to imitate small shifts in the electrode baseline. All these steps are introduced randomly for every training point. These augmentations introduce natural variations into the data while still preserving the original word structure, helping the model learn to generalize better in real-world scenarios.

To handle the vast combinatorial space without incurring heavy storage demands, a runtime data generator is implemented. This generator produces augmented sentence batches on the fly during training having 6000 sentences, dynamically sampling and modifying word exemplars with each epoch. Final evaluation is conducted on 34 real-sentence recordings to assess generalization beyond the synthetic training domain.

Proposed model: The proposed model follows a sequence-to-sequence architecture enhanced with an attention mechanism, inspired by frameworks used in neural machine translation and speech recognition systems [26], [27]. As shown in Fig. 4, the encoder consists of a convolutional front-end that extracts the local temporal and spatial features from the sEMG signals, followed by a bidirectional Long Short-Term Memory (BiLSTM) network to model long-range dependencies and capture the global contextual information. The decoder is an autoregressive unidirectional LSTM that generates output words sequentially. At each step, it conditions on the previously predicted word, its current hidden state, and an attention vector computed over the encoder outputs. This attention mechanism enables the decoder to focus selectively on relevant portions of the input sequence, allowing the model to handle variability in word duration and coarticulation effects [28], [29].

In addition to the attention-based sequence-to-sequence model, we also train a baseline model using a CNN+BiLSTM encoder combined with a CTC loss function [30]. This loss function has been widely adopted in prior sEMG-based silent speech recognition studies [7], [13], [16], [17], particularly for its ability to align variable-length input sequences with target label sequences without requiring explicit word boundaries during training. The CNN layers in this model extract local spatiotemporal features from the raw sEMG signals, while the stacked BiLSTM layers capture long-range temporal dependencies. The CTC loss function then enables direct mapping of the input sequence to the target word sequence by learning to collapse repeated labels and ignore blank tokens. This baseline provides a direct comparison for evaluating the benefits of attention mechanisms and autoregressive decoding in the proposed model.

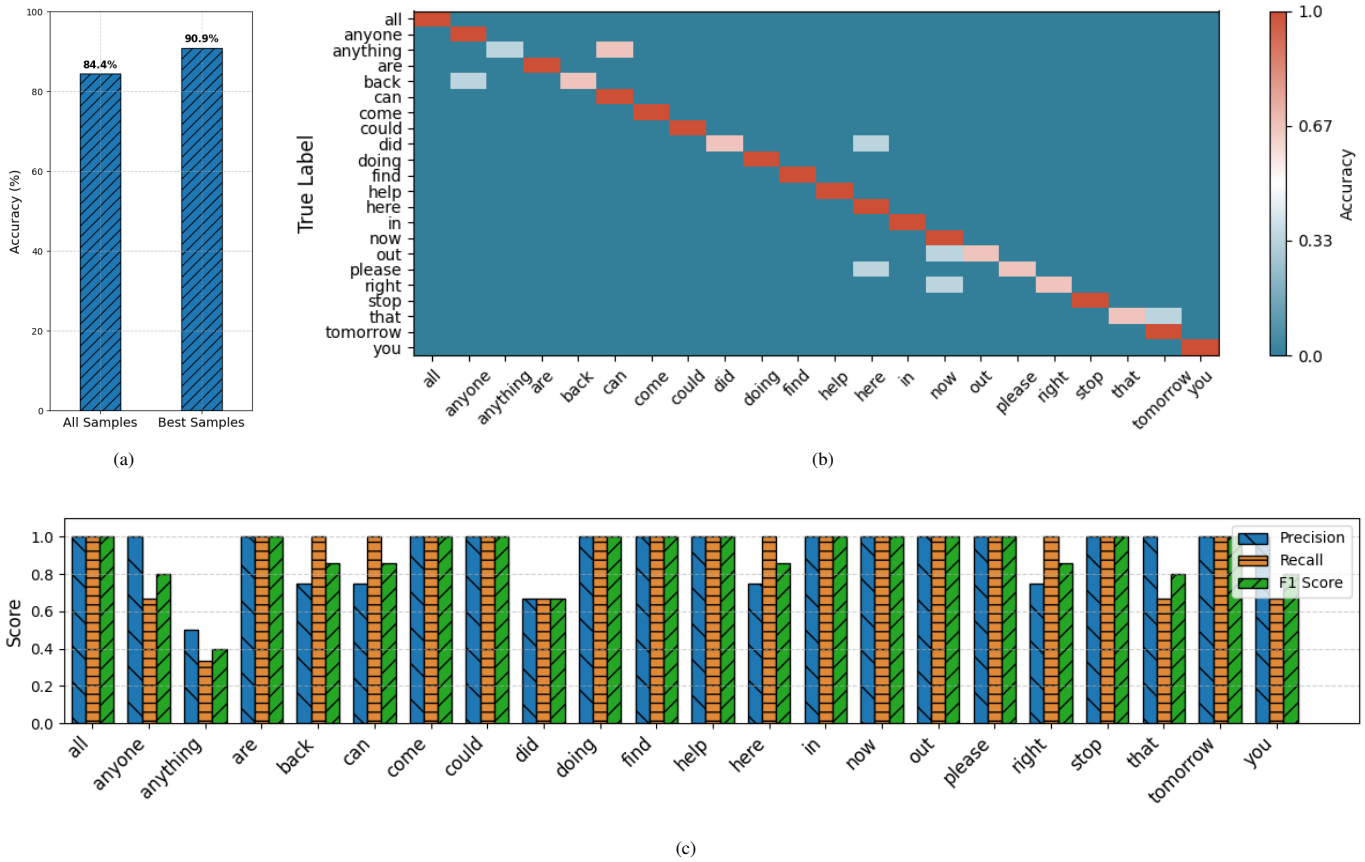


Fig. 5: (a) Word-level classification accuracy, and (b) the confusion matrix for the CNN classifier. (c) Precision, recall and F1 scores per word.

III. Performance Evaluation

To evaluate how well our attention-based model can decode silent speech from facial sEMG signals, we conduct experiments at both the word and sentence levels. The results show that combining proper signal preprocessing, careful selection of training examples, and sentence-level data augmentation leads to accurate and generalizable decoding of continuous EMG input.

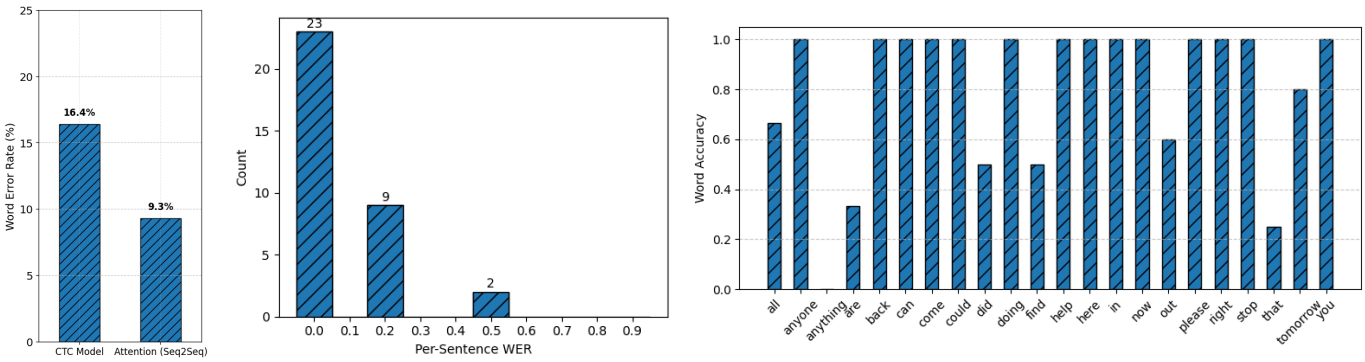
Word level classification accuracy: Before training the full sequence model, we first assess how exemplar selection impacts isolated word classification performance. Specifically, we compare two training sets: one containing all samples available (more than 20 per word), and another is exemplars selected using Multi-DTW [25]. In both cases, the dataset is split into 80% training and 20% testing. As shown in Fig. 5(a), the CNN classifier achieves word classification accuracy of above 80%; specifically, when trained on the training exemplars, reaching an accuracy of 90.9%, compared to 84.4% using all samples. This result supports our hypothesis that Multi-DTW effectively filters out noisy or misaligned recordings, improving model performance. It also suggests that the proposed solution extracts robust and discriminative features from EMG signals.

The confusion matrix, shown in Fig. 5(b), further illustrates the model's ability to distinguish between classes. Each cell represents the classification accuracy for a word, aggregated over three test samples per class. Nearly all words are classified correctly, with strong diagonal dominance indicating high specificity for most words. However, a small number of misclassifications occur, mainly among phonetically similar words. For example, some samples of "anything," "that," and "right" are occasionally confused with other classes. This is likely due to similarities in sEMG signal patterns for these words, which may challenge even experienced human annotators.

To provide a more nuanced perspective on classification performance across the vocabulary, Fig. 5(c) presents precision, recall, and F1 scores for each word class. As observed from the figure, most words achieve high scores (i.e. above 0.8), reflecting both consistent true positive rates and minimal false positives or negatives. Some classes exhibit lower scores, especially those that also show confusion in Fig. 5(b), demonstrating the difficulty in distinguishing certain word pairs with highly overlapping muscle activation patterns.

Sentence-level decoding accuracy: For sentence-level decoding, we train our Seq2Seq model on synthetic sentence samples built from the training exemplars per word, while validation is done on the validation exemplars. Testing is conducted on real sentences, each consisting of 4–5 words from the 22-word vocabulary.

Fig. 6(a) shows the accuracy of our model. From this figure, we can observe that our proposed model achieves a Word Error



(a) (b) (c)

Fig. 6: (a) Sentence-level WER comparison between CTC and Seq2Seq model on real recordings. (b) Distribution of per-sentence WER for the Seq2Seq model, along with (c) the per-word recognition accuracy.

TABLE II: Predictions from Seq2Seq model. The words in "blue" show the misclassified ones.

Reference	Hypothesis	Reference	Hypothesis
1. can you help out	can you help out	2. are you doing anything	are you doing can
3. come out right now	come are right now	4. please come out here	please come out here
5. can you please stop that	can you please stop right	6. you can all help	you can all help
7. please come right now	please come right now	8. please come in now	please come in now
9. stop doing that please	stop doing that please	10. you are all right	you all now right
11. could you come tomorrow	could you come tomorrow	12. could you stop please	could you stop please
13. can anyone help out	can anyone help now	14. come back here now	come back here now
15. can anyone help you	can anyone help you	16. anyone can come in	anyone can come in
17. are you all right	all you all right	18. could you come here	could you come here
19. come here right now	come here right now	20. you can come in	you can come in
21. come back right now	come back right now	22. can you come back tomorrow	can you come back tomorrow
23. did anyone help you	back anyone help you	24. come here now please	come here now please
25. please come back tomorrow	please come back back tomorrow	26. please stop doing that	please stop doing right
27. you can stop now	you can stop now	28. did you find anything	did you did back
29. can you come right now	can you come right now	30. you come back here	you come back here
31. stop that right now	stop right right now	32. can you come here tomorrow	can you come here tomorrow
33. can you find out	can you find out	34. you can come tomorrow	you can come tomorrow

Rate (WER) of 9.3%, showing strong generalization from synthetic to real data. In contrast, a baseline CNN+BiLSTM model trained using the CTC loss function achieves a higher WER of 16.4%, highlighting the limitations of CTC-based decoding in handling coarticulation and variable word durations. Fig. 6(b) shows the distribution of sentence-level WER across all test samples. Most sentences have near-perfect recognition (23 out of 34 are predicted perfectly), with the few falling below 20% WER. Only a couple of sentences have 50% WER. These results suggest that the model effectively generalizes to the real sentence data despite being trained exclusively on synthetic sequences.

A per-class accuracy analysis is also conducted across all 22 vocabulary words for all the occurrences of those words in test sentences. The result is shown in Fig. 6(c). From this figure, we can observe that most of the words achieve over 90% recognition accuracy, though certain words like "anything" and "anyone" show slightly lower scores due to acoustic and muscular similarity during articulation. This suggests the need for additional training diversity or tailored augmentation for confusable word pairs.

Qualitative analysis: To better understand how the model performs beyond numbers, Table II shows a few examples of real test sentences alongside their predicted outputs. In most cases, the model is able to generate sentences that are grammatically correct and meaningfully aligned with the original input. This suggests that the model not only recognizes individual words, but also understands how they fit together in natural language. When mistakes happen, they are usually minor, often involving similar looking words like "are" instead of "out" or "right" instead of "that." These kind of words tend to produce very similar facial muscle movements during silent articulation, making them hard to distinguish. Still, the predicted sentences are generally easy to understand, and the errors do not significantly affect the overall meaning.

What stands out in these examples is that the model isn't just memorizing patterns. This is because of the attention mechanism, that learns to focus on the most relevant parts of the input while generating each word. This helps it deal with real-world challenges like signal noise, variable word lengths, and natural coarticulation. Considering that it is trained only on synthetic data, its ability to perform well on real, continuous speech is a strong sign of how well the system generalizes.

Latent space structure: To analyze the model's internal representations, t-SNE is applied to the encoder output vectors of all training exemplars. The resulting 2D projection shown in Fig. 7, reveals clear clustering by word identity, indicating that the model learns meaningful word-specific encodings from

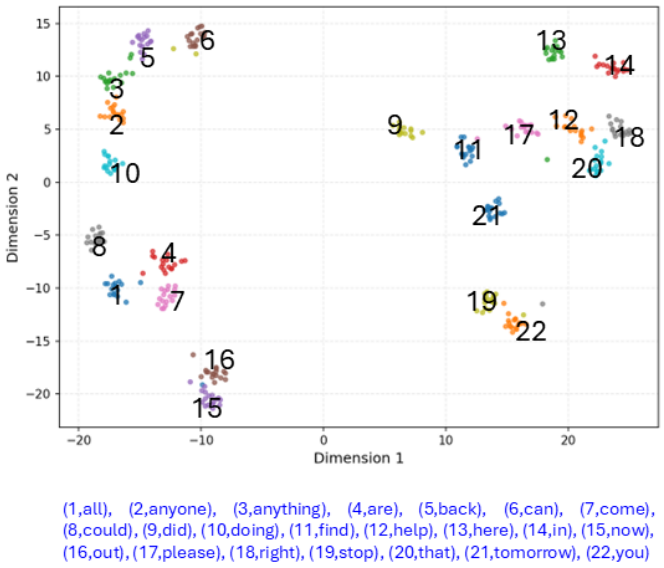


Fig. 7: t-SNE visualization of the encoder outputs for exemplar for all the words. The (x, y) in blue text denotes the number used to show each cluster, and the corresponding word respectively.

the raw EMG input. Close proximity between some clusters, such as "anything" and "anyone," correlates with observed confusion during decoding.

IV. Conclusion

This paper introduces a robust and scalable framework for silent speech recognition solution using surface electromyography signals and an attention-based Sequence-to-Sequence model. Unlike traditional methods based on Connectionist Temporal Classification (CTC), the proposed autoregressive architecture allows for fine-grained alignment and decoding of multi-word silent utterances. To address variability in articulation and timing, dynamic sentence-level augmentation is included to simulate realistic muscle activation patterns. This approach results in a WER of 9.3% on natural sEMG recordings, and outperforms the CNN+BiLSTM+CTC baseline.

Beyond accuracy, the attention mechanism enables the model to learn context-aware and interpretable word-level representations, as validated by latent space analysis and per-word performance metrics. Notably, although the model is trained solely on synthetically composed sEMG sentences, it generalizes well to real-world EMG inputs, highlighting the effectiveness of the proposed augmentation strategy. These findings demonstrate that combining attention-based architectures with carefully designed synthetic training data can yield high-fidelity SSR systems, offering a promising foundation for non-acoustic communication interfaces in assistive, secure, or silent environments.

References

[1] T. Hueber *et al.*, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Commun.*, vol. 52, no. 4, pp. 288–300, 2010.

[2] J. Cai *et al.*, "A visual speech recognition system for an ultrasound-based silent speech interface," in *ICPhS*, 2011, pp. 384–387.

[3] P. Heracleous *et al.*, "A pilot study on augmented speech communication based on electro-magnetic articulography," *Pattern Recognition Letters*, vol. 32, no. 8, pp. 1119–1125, 2011.

[4] B. Cao *et al.*, "Magtrack: A wearable tongue motion tracking system for silent speech interfaces," *Journal of Speech, Language, and Hearing Research*, vol. 66, pp. 3206–3221, 2023.

[5] M. Wester, "Unspoken speech - speech recognition based on electroencephalography," https://www.researchgate.net/publication/36453500_Unspoken_Speech_-_Speech_Recognition_based_on_Electroencephalography, 2006.

[6] M. Angrick *et al.*, "Speech synthesis from ecog using densely connected 3d convolutional neural networks," *J Neural Eng*, vol. 16, no. 3, 2019.

[7] T. Afouras *et al.*, "Deep lip reading: A comparison of models and an online application," in *Interspeech*, 2018, pp. 3514–3518.

[8] M. Kim *et al.*, "Lip to speech synthesis with visual context attentional gan," in *NeurIPS*, 2021.

[9] N. Sugie *et al.*, "A speech prosthesis employing a speech synthesizer-vowel discrimination from perioral muscle activities and vowel production," *IEEE Transactions on Biomedical Engineering*, vol. BME-32, no. 7, pp. 485–490, 1985.

[10] C. Jorgensen *et al.*, "Sub auditory speech recognition based on emg signals," in *IJCNN*, 2003, pp. 3128–3133.

[11] H. Manabe *et al.*, "Multi-stream hmm for emg-based speech recognition," in *IEEE EMBS*, 2004, pp. 4389–4392.

[12] T. Schultz *et al.*, "Modeling coarticulation in emg-based continuous speech recognition," *Speech Commun.*, vol. 52, no. 4, pp. 341–353, 2010.

[13] D. Gaddy *et al.*, "Digital voicing of silent speech," 2020. [Online]. Available: https://arxiv.org/abs/2010.02960

[14] G. S. Meltzner *et al.*, "Development of semg sensors and algorithms for silent speech recognition," *Journal of Neural Engineering*, vol. 15, no. 4, p. 046031, 2018.

[15] M. Wand *et al.*, "Deep neural network frontend for continuous emg-based speech recognition," in *Interspeech*, 2016, pp. 3032–3036.

[16] L. Xie *et al.*, "Neural chinese silent speech recognition with facial electromyography," *Speech Commun.*, vol. 171, p. 103230, 2025.

[17] R. Song *et al.*, "Decoding silent speech from high-density surface electromyographic data using transformer," *Biomed. Signal Process. Control.*, vol. 80, 2023.

[18] B. Huang *et al.*, "Design and implementation of a silent speech recognition system based on semg signals: A neural network approach," *Biomed. Signal Process. Control.*, vol. 92, p. 106052, 2024.

[19] "OpenBCI, Cyton Biosensing Board (8-Channel)," 2024. [Online]. Available: https://www.openbci.com/products/cyton-biosensing-board-8-channel

[20] X. Tan *et al.*, "Extracting spatial muscle activation patterns in facial and neck muscles for silent speech recognition using high-density semg," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–13, 2023.

[21] M. Zhu *et al.*, "Towards optimizing electrode configurations for silent speech recognition based on high-density surface electromyography," *Journal of Neural Engineering*, vol. 18, no. 1, p. 016005, 2021.

[22] J. S. Garofolo *et al.*, "TIMIT : acoustic-phonetic continuous speech corpus." 1993. [Online]. Available: http://www.worldcat.org/isbn/1585630195

[23] A. van Boxtel, "Optimal signal bandwidth for the recording of surface emg activity of facial, jaw, oral, and neck muscles," *Psychophysiology*, vol. 38, no. 1, pp. 22–34, 2001.

[24] C. J. De Luca *et al.*, "Filtering the surface emg signal: Movement artifact and baseline noise contamination," *Journal of Biomechanics*, vol. 43, no. 8, pp. 1573–1579, 2010.

[25] M. Müller, *Information Retrieval for Music and Motion*. Berlin, Heidelberg: Springer-Verlag, 2007.

[26] D. Bahdanau *et al.*, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2015.

[27] W. Chan *et al.*, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *IEEE ICASSP*, 2016, pp. 4960–4964.

[28] T. Luong *et al.*, "Effective approaches to attention-based neural machine translation," in *EMNLP*, 2015, pp. 1412–1421.

[29] A. Vaswani *et al.*, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[30] A. Graves *et al.*, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006, pp. 369–376.