Predicting Speech Sounds from Speech-Related Muscle Activity Using Electromyography

Evan Holbrook
Dpt. of Biological Engineering
Massachusetts Institute of Technology
Cambridge, USA
evanholb@mit.edu

Daniel Rodriguez Rodriguez
Master in Design Engineering
Harvard University
Cambridge, USA
drodriguezrodriguez@mde.harvard.edu

Hessan Sedaghat
Master in Design Engineering
Harvard University
Cambridge, USA
hsedaghat@mde.harvard.edu

Julius Stein
Master in Design Engineering
Harvard University
Cambridge, USA
jstein@mde.harvard.edu

***Abstract*—This study explores the development of a model to classify electrical activity in speech-related muscles to one of five possible phonemes. This constitutes groundwork to build a Glottis-Computer Interface (GCI) aimed at restoring speech capabilities in individuals who have undergone laryngectomy or suffer from vocal cord disorders. Data was obtained with surface electromyography (sEMG) from six muscles in the throat and mouth area, and a random forest classifier was trained to predict phonemes. The experimental results demonstrated a phoneme classification accuracy of 86.7%, highlighting the system's effectiveness in interpreting muscle signals. Future work will focus on improving sensor technology, expanding the dataset, and developing a wearable device to enhance usability and accuracy, making this technology accessible to a broader range of users.**

Keywords: **Glottis-Computer-Interface, Brain-Computer Interface, EMG, Speech Restoration, Laryngectomy, Vocal Cord Disorders, Machine Learning, Assistive Technology.**

I. Introduction

Laryngectomy is a surgical procedure that removes the larynx (or voice box) as a treatment for laryngeal cancer and severe laryngeal tuberculosis, resulting in a permanent loss of natural voice production, although the patients' ability to move speech-related muscles remains [1]. As of 2013, over 60000 people in the United States alone had undergone a laryngectomy [2]. Existing communication aids, such as electrolarynx devices, try to replace the vibration previously given by the vocal cords, but often lack the naturalness and ease of use desired by patients, making speech sound mechanical and less intelligible [3]. These limitations underscore the need for innovative solutions that can provide more natural and efficient communication means.

The advent of brain-computer interfaces (BCIs) has opened new avenues for assisting individuals with speech production impairments, Recent advances in surface electromyography (sEMG) technology, which measures muscle electrical activity through electrodes placed on the skin that cover the target muscles, offer a promising alternative that is also non-invasive. Literature shows that capturing the unique activation patterns of the glottis and mouth muscles during speech production makes it possible to predict intended speech from the act of silently mouthing words (Silent Speech Recognition, or SSR) [4, 5]. Current technology also allows us to predict the patterns of stress and intonation that humans can give to sentences during speech, also known as prosody [6].

Recent work is focused on predicting speech at the word or sentence level, which has a fundamental scalability problem. For instance, a study on prosody used a corpus of 2500 English words to train its model, far below the 490,000 entries in Merriam-Webster's dictionary as of May 2024 [7], a number that excludes plural forms, verbal variations, and other words too modern or too niche to appear in said dictionary. If we factor in languages other

than English, word-level prediction seems to be an unscalable approach to SSR. In contrast, English has 44 phonemes we combine to create those words. Efforts to predict phonemes in silent speech, coupled with the current work on prosody, could bring us closer to a system for a language-agnostic speech production assistive technology.

This study investigates the potential of a Glottis-Computer Interface (GCI) to predict phonemes. Our approach leverages machine learning, specifically a random forest classifier, to analyze sEMG data of isolated phoneme enunciations by individuals with a voice box. This study aims to bridge the gap between current assistive communication devices and the natural speech experience desired by users.

II. METHODS

A. Participants

Seven participants were selected based on the following inclusion criteria to ensure the collection of reliable and consistent data:

- Age: 18 years or older.
- Language Proficiency: Fluent in English.
- Health Status: No existing speech production impairments and healthy skin in the mouth and throat areas.

The recruitment process aimed to eliminate variability due to pre-existing conditions, ensuring that the sEMG data accurately reflected the muscle activity associated with phoneme production.

B. Materials

MyoWare 2.0 Muscle Sensors (Advancer Technologies, LLC; Raleigh, NC, USA) were chosen to acquire sEMG data for their high sensitivity and reliability in detecting muscle activations (see Figure 1). The sensor has a triangular shape, with two leads for the target muscles and one reference lead for a non-target muscle or bone that will not show activity during the protocol task. The sensors' specifications are as follows:

- Voltage Input: Minimum = +2.27V, Typical = +3.3V or +5V, Maximum = +5.47V
- Input Bias Current: 250 pA, maximum 1 nA
- Input Impedance: 800 kΩ
- Common Mode Rejection Ratio (CMRR): 140 dB
- Filters:
  - High-pass Filter: Active 1st order, cutoff frequency (fc) = 20.8 Hz, -20 dB
  - Low-pass Filter: Active 1st order, fc = 498.4 Hz, -20 dB
- Rectification Method: Full-wave
- Envelope Detection: Linear, Passive 1st order, fc = 3.6 Hz, -20 dB

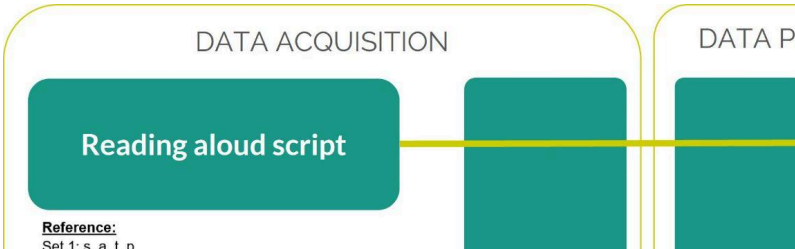
Fig. 1. MyoWare 2.0 Muscle Sensor

The analog output of the MyoWare 2.0 Muscle Sensors were converted to digital with an Arduino microcontroller and a shield previously programmed with code written in Arduino IDE. Arduino IDE and CoolTerm were the pieces of software used to visualize and acquire the signals in Windows computers. Data analysis was performed on Google Colab, which runs on Python.

C. Protocol

The overall protocol diagram is shown in Figure 2. Due to issues encountered during data processing, the protocol was revised. The revised protocol is shown in Figure 3.

Participants were seated comfortably–in a position where none of the targeted muscles were under strain–in a noise-controlled environment. After cleaning the skin with



alcohol wipes, six MyoWare Muscle Sensors were positioned on three key throat and mouth locations on the left side and its mirror counterparts on the right to capture the muscle activity associated with speech production for six total positions (see locations in a test subject in Figure 4):

1. The first sensor pair was positioned laterally on the upper jaw along the masseter muscle with a reference electrode positioned at the back of the neck.
2. The second sensor pair was positioned under the jawline in the midpoint between the chin and the end of the jaw along the mandibular muscles, with measuring electrodes placed horizontally.
3. The third and final pair was positioned vertically along the platysma cervicalis, with reference electrodes placed on the clavicle.

Fig. 4. EMG sensor setup on test subject

Fig. 2. Original study design diagram

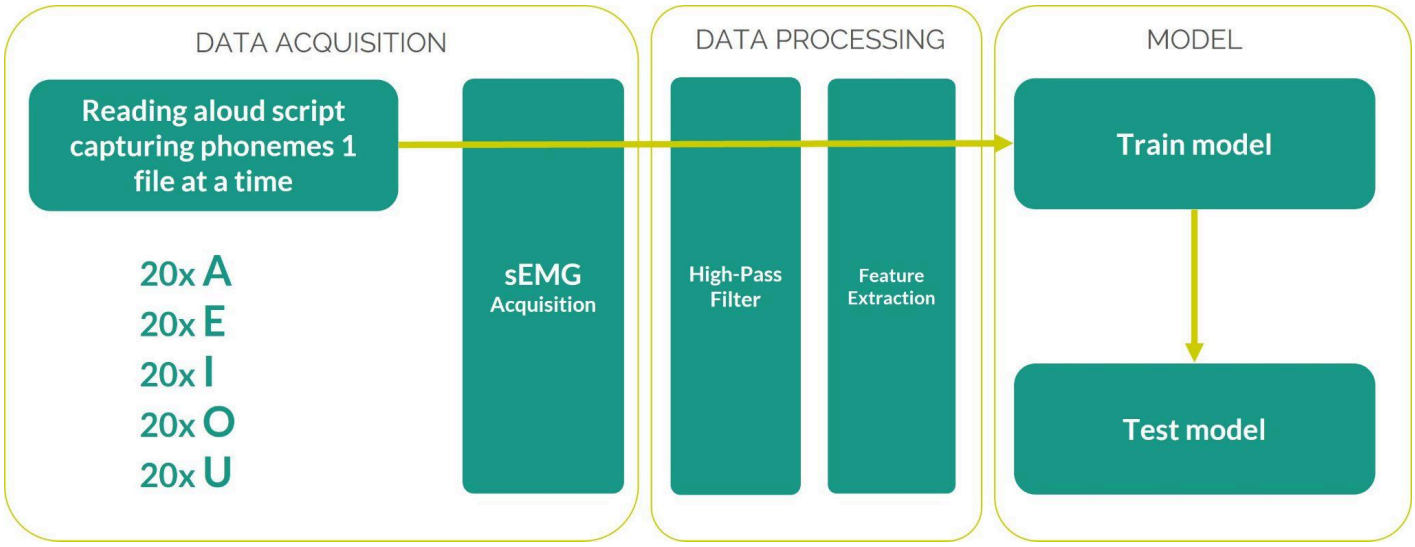


Fig. 3. Revised study design diagram

Each participant was instructed to reproduce the 44 phonemes in the English language. where a woman teaches how to pronounce the phoneme and a noun with that phoneme to make it easier for the audience. Once the participant demonstrated they could replicate the phoneme, a single 6-channel sEMG recording of the participant saying the phoneme 10 times was collected. The sampling frequency of the Analog-to-Digital converter was approximately 440 Hz.

In the revised protocol, each participant was instructed to replicate specific phonemes 20 times each in separate sEMG recordings. Given the significant increase in time to acquire data, this protocol only focused on five phonemes:

- [æ] as in the ‘a’ in ‘apple’.
- [e] as in the ‘e’ in ‘elephant’.
- [ɪ] as in the ‘i’ in ‘igloo’.
- [ɒ] as in the ‘o’ in ‘octopus’ pronounced in American English.
- [ʌ] as in the ‘u’ in ‘umbrella’.

D. Data Processing

The rectified envelope signals coming from the MyoWare 2.0 Muscle Sensors were used over the raw signals to make signal processing more efficient. To remove the effects from the DC offset, a high-pass filter with a cutoff frequency of 0.1 Hz and an order of 5 was applied to all signals.

TABLE I. Features extracted from phoneme sEMG recordings

Feature	**Domain**	**Formula/Definition**
Mean Absolute Value (MAV)	Time	$\mathrm{MAV} = \frac{1}{N}\sum_{i=1}^{N}\lvert x_i\rvert$
Root Mean Square (RMS)	Time	$\mathrm{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}$
Number of Zero Crossings (ZC)	Time	$ZC = \mathrm{count}\left((x_i \cdot x_{i-1}) < 0\right)$
Waveform Length (WL)	Time	$\mathrm{WL} = \sum_{i=1}^{N-1}\lvert x_{i+1} - x_i\rvert$
Number of Slope Sign Changes (SSC)	Time	$SSC = \mathrm{count}\left(\left(\frac{x_i - x_{i-1}}{x_{i-1} - x_{i-2}}\right) < 0\right)$
Spectral Entropy (SE)	Frequency	$\mathrm{SE} = -\sum_{f=1}^{F} P(f)\log_2(P(f))$
		where P(f) is the normalized power spectral density.
Median Frequency (MF)	Frequency	Frequency below which 50% of the power in the power spectrum is located.
Peak Frequency (PF)	Frequency	$\mathrm{PF} = \arg\max_f P(f)$ where P(f) is the power spectral density.
Total Power (TP)	Frequency	$\mathrm{TP} = \sum_{i=1}^{N} P(f_i)$ where P(f) is the power spectral density.

A wave-snipping algorithm was built to obtain the 10 phonemes from each file containing the 10 enunciations of the corresponding phoneme. The algorithm computed the mean and standard deviation of 45 random samples within the first second of the signal, which were then used to determine activation regions in the signal. An activation in any of the 6 channels in the sEMG recordings meant overall activation. The difficult application of this algorithm for the first four subjects prompted the switch to the revised protocol, which focused on getting pre-sliced phonemes.

Five time-domain features and four frequency-domain features were extracted from each of the 6 channels in every phoneme sEMG recording, which adds up to a total of 54 features per single phoneme. The features and the way they were calculated are shown in Table 1.

E. Model Training and Evaluation

A random forest classifier was employed to train the model using the extracted features. The dataset was randomly split into a training set (80%) and an evaluation set (20%) using the random state 42. Hyperparameter optimization was performed using grid search to achieve the best performance. The hyperparameters tried were:

- Number of estimators: 100, 200, 300
- Maximum features: auto, sqrt

- Maximum depth: None, 10, 20, 30
- Bootstrap: True, False

Overall accuracy and individual phoneme classification rates were calculated, and a confusion matrix was generated to visualize the classification accuracy for each phoneme. Due to the study design revision, the primary focus of the model was the individual phonemes [æ], [e], [ɪ], [ɒ] and [ʌ]. Detailed precision and recall metrics were computed for each phoneme to understand the classifier's strengths and weaknesses in distinguishing between similar-sounding phonemes.

III. RESULTS

A total of 5 female participants and 2 male participants finished the study. Their mean age was 27 years old and they did not present speech production impediments. From these 7 participants, 2 male participants and 1 female participant went through the revised protocol. 296 phoneme files were included in the model, and 4 corrupted phoneme files were discarded.

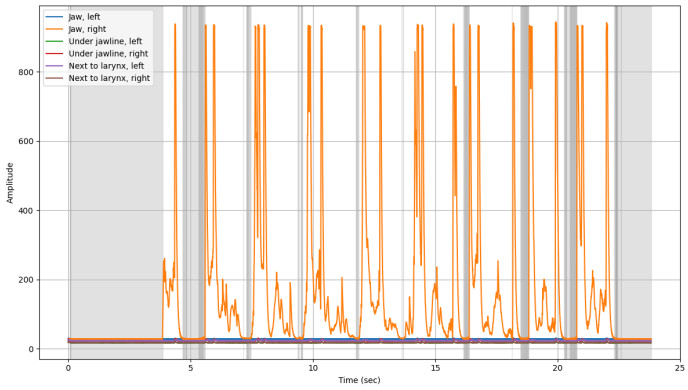


Fig. 5. EMG recording of ten enunciations of phoneme [s] from test subject 1.

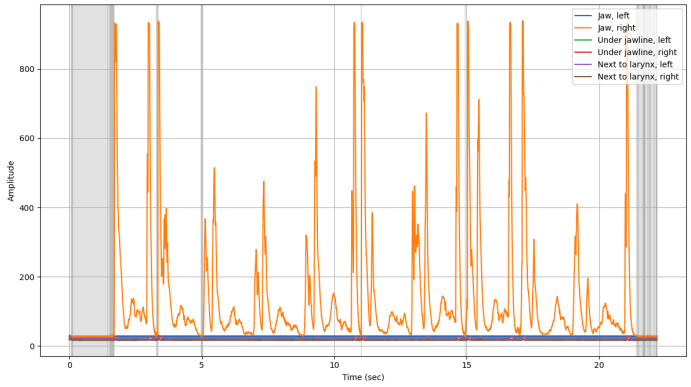


Fig. 6. EMG recording of ten enunciations of phoneme [w] from test subject 1.

The reason for switching to the revised protocol can be observed in Figures 5 and 6. Figure 5 shows the results of a successful wave snipping for phoneme [s], with gray areas representing parts of the EMG recording ruled out. Figure 6 shows an unsuccessful attempt, as seen by the unsnipped regions of the signal. Figures 7 to 11 show pre-sliced EMG recordings for phonemes [æ], [e], [ɪ], [ɒ] and [ʌ] for illustrative purposes.

The random forest classifier achieved an overall accuracy of 86.7%, indicating a respectable level of precision in predicting the correct phonemes. The confusion matrix is shown in Table 1. The classifier exhibited the highest accuracy for phoneme [æ] and performed worst on phoneme [ʌ], which was often misclassified as phoneme [ɒ]. This indicates a need for further refinement in feature extraction or model complexity to better differentiate these phonemes. Table II reports precision, recall, and f-1 scores for each of the phonemes.

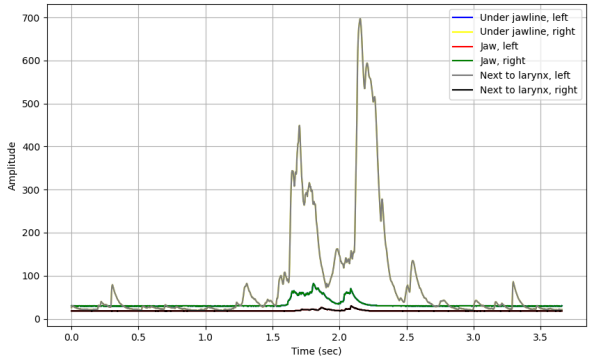


Fig. 7. EMG recording of one enunciation of phoneme [æ] from test subject 6.

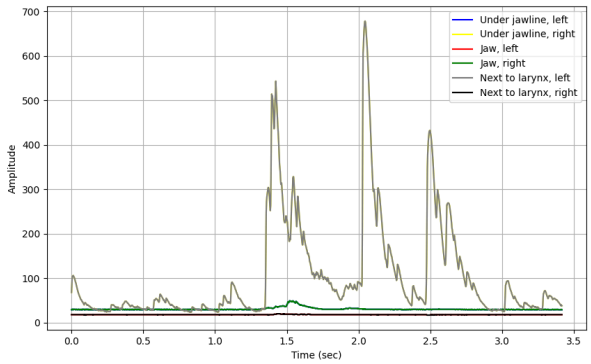


Fig. 8. EMG recording of one enunciation of phoneme [e] from test subject 6.

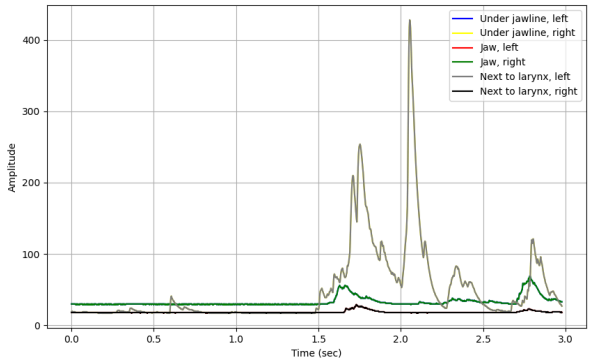


Fig. 9. EMG recording of one enunciation of phoneme [ɪ] from test subject 6.

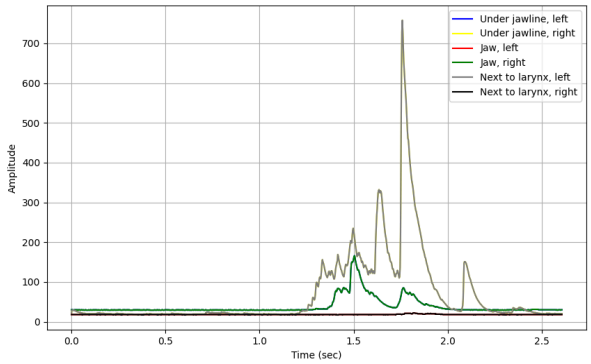


Fig. 10. EMG recording of one enunciation of phoneme [ɒ] from test subject 6.

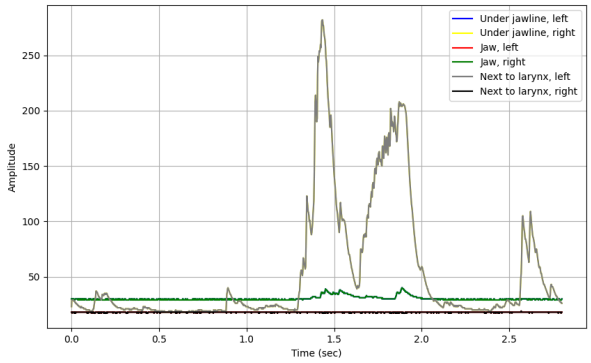


Fig. 11. EMG recording of one enunciation of phoneme [ʌ] from test subject 6.

TABLE II. Confusion matrix for random forest classifier tested on evaluation set. The diagonal of the matrix corresponds to phonemes correctly classified, the rest are misclassifications.

	[æ]	**[e]**	**[ɪ]**	**[ɒ]**	**[ʌ]**
[æ]	11	0	0	0	0
[e]	1	10	0	0	0
[ɪ]	0	0	6	1	0
[ɒ]	0	0	2	11	0
[ʌ]	0	0	0	4	14

TABLE III. Classification report for random forest classifier tested on evaluation set.

	Precision	**Recall**	**f1-score**	**Support**

[æ]	0.92	1.00	0.96	11
[e]	1.00	0.91	0.95	11
[ɪ]	0.75	0.86	0.8	7
[ɒ]	0.69	0.85	0.76	13
[ʌ]	1.00	0.78	0.88	18

IV. DISCUSSION

Our study hypothesized that the unique activation patterns of the glottis muscles, captured via sEMG, could be used to predict phonemes accurately, thus enabling the reconstruction of natural speech for individuals with vocal cord disabilities in the future. The results substantiated our hypothesis, as the random forest classifier achieved an overall phoneme classification accuracy of 86.7%. These findings indicate that sEMG data can effectively capture the nuanced differences in muscle activations associated with different speech sounds.

The findings align closely with recent advancements in the field of electromyography for speech synthesis. For instance, studies such as Janke and Diener (2017) and Chan et al. (2021) [5,6] have demonstrated the feasibility of using facial sEMG data to generate speech. However, our research extends these findings by focusing specifically on phoneme prediction, which has advantages over word or sentence prediction.

If further progress is made to enable live phoneme prediction from muscle activations, our design would allow for capturing accent, intonation, and song in addition to speech reconstruction. By reducing the input space to 44 phonemes instead of all words, this approach dramatically increases the density of samples in comparison to an equal number of word recordings.

A. Limitations

While our study achieved high accuracy, several limitations need to be addressed. Firstly, variability in sensor placement can significantly affect data quality. Future studies should focus on developing more robust methods for sensor placement to minimize this variability. This study was also limited by only recruiting individuals without speech impairments. Including participants with diverse speech and health conditions could help generalize our findings. Lastly, our small sample size makes our random forest classifier vulnerable to overfitting. More participants should be recruited to develop a more robust model.

The current study strategy, while it could potentially lead to language-agnostic speech production assistive technology, would not be accent-agnostic. Unlike phonemes, phones provide a way to refer to a specific sound or speech gesture. For instance, the phoneme [p] includes the allophones [p] and [p^h], the latter being an aspirated ‘p’ sound. Switching from phoneme to phone prediction, although more complex, could make our technology truly predict unique sounds and eventually replicate the accent the individual naturally possesses or is trying to emulate. Thus, individuals without a voice box would retain their own accent when speaking a language other than their mother tongue, and accent assimilation would occur organically with use, as it happens in individuals with a healthy voice box.

B. Future Research Directions

To enhance the capabilities of the GCI system, future research should focus on expanding the phoneme set. For example, including consonants and complex phonetic combinations will improve the system's versatility and applicability. Further, optimizing the system for real-time speech synthesis is essential for practical applications. This involves not only improving the computational efficiency of the model but also ensuring the stability and reliability of the sEMG data acquisition. Lastly, model overfitting should be assessed using a brand new set of participants and evaluating whether our model achieves similar accuracy in phoneme prediction.

While conducting the present study, research involving a novel technique using a sensing-actuating system based on

soft magnetoelasticity was published [8]. Instead of measuring the electrical activity of speech-related muscles, the movement of the muscles is translated into electrical signals by the system. Given that the bioimpedance of human skin diminishes the power of the electrical signals detected by the surface electromyography electrodes, using another data source for speech prediction seems a valid alternative pathway. The technique was tested at the sentence level in 5 sentences and should be further explored to be applicable in real situations.

Because small signals are generated in the glottis when speech is thought of, our approach could further enable an alternative to a brain-computer interface by detecting the glottal signals as a proxy for our thoughts. This level of detection would require either embedded electrodes or a magnetoelastic system as described above given the magnitude of the signals produced.

V. CONCLUSION

This study successfully demonstrated the potential of using surface electromyography (sEMG) data from the glottis muscles to accurately predict phonemes, achieving a classification accuracy of 86.7%. This validates the hypothesis that unique muscle activation patterns can be utilized to reconstruct natural speech for individuals who have undergone laryngectomy or suffer from severe vocal cord disorders. The high accuracy of the Glottis-Computer Interface (GCI) underscores its potential to significantly enhance communication aids, providing a more natural and effective solution for speech restoration. The high accuracy of phoneme classification suggests that sEMG-based systems could greatly improve the naturalness and intelligibility of speech synthesis devices for individuals who have undergone laryngectomy. The development of a wearable device is crucial for practical application, promising to revolutionize assistive communication devices and significantly improve the quality of life for individuals with speech impairments.

Beyond medical applications, this technology holds the potential to enable silent speech communication in noisy environments or for individuals seeking discreet communication methods.

REFERENCES

1. Ceachir, O., Hainarosie, R., & Zainea, V. (2014). Total laryngectomy - past, present, future. Maedica (Bucur), 9(2), 210-216. Retrieved from https://pubmed.ncbi.nlm.nih.gov/25705281/

2. Cleveland Clinic. (n.d.). Laryngectomy. Cleveland Clinic. Retrieved May 2, 2024, from https://my.clevelandclinic.org/health/treatments/24072-laryngectomy

3. Kaye, R., Tang, C. G., & Sinclair, C. F. (2017). The electrolarynx: Voice restoration after total laryngectomy. Medical Devices: Evidence and Research, 10, 133–140. Retrieved from https://doi.org/10.2147/mder.s133225

4. S. Ma et al. (2019). Silent Speech Recognition Based on Surface Electromyography. 2019 Chinese Automation Congress (CAC), Hangzhou, China, 4497-4501. Retrieved from https://doi.org/10.1109/TIM.2023.3244849

5. Janke, M., & Diener, L. (2017). EMG-to-Speech: Direct Generation of Speech From Facial Electromyographic Signals. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(12), 2375-2385. Retrieved from https://doi.org/10.1109/TASLP.2017.2738568

6. Chan, M. D., Shiwani, B., Roy, S. H., Heaton, J. T., Meltzner, G. S., Contessa, P., De Luca, G., Patel, R., & Kline, J. C. (2021). Surface electromyography–based recognition, synthesis, and perception of prosodic subvocal speech. Journal of Speech, Language, and Hearing Research, 64(6S), 2134–2153. Retrieved from https://doi.org/10.1044/2021_jslhr-20-00257

7. Merriam-Webster, I. (n.d.). Subscription plans and pricing. Merriam-Webster unabridged. https://unabridged.merriam-webster.com/subscriber/register/p1

8. Che, Z., Wan, X., Xu, J., et al. (2024). Speaking without vocal folds using a

machine-learning-assisted wearable sensing-actuation system. Nat Commun, 15, 1873. Retrieved from https://doi.org/10.1038/s41467-024-45915-7