Interspeech 2024
1-5 September 2024, Kos, Greece

Towards EMG-to-Speech with a Necklace Form Factor

Peter Wu[*1], *Ryan Kaveh*[*1], *Raghav Nautiyal*[1], *Christine Zhang*[1], *Albert Guo*[1], *Anvitha Kachinthaya*[1], *Tavish Mishra*[1], *Bohan Yu*[1], *Alan W Black*[1], *Rikky Muller*[**1], *Gopala Krishna Anumanchipalli*[**1]

[1]University of California, Berkeley

peterw1@berkeley.edu, ryankaveh@berkeley.edu

Abstract

Electrodes for decoding speech from electromyography (EMG) are typically placed on the face, requiring adhesives that are inconvenient and skin-irritating if used regularly. We explore a different device form factor, where dry electrodes are placed around the neck instead. 11-word, multi-speaker voiced EMG classifiers trained on data recorded with this device achieve 92.7% accuracy. Ablation studies reveal the importance of having more than two electrodes on the neck, and phonological analyses reveal similar classification confusions between neck-only and neck-and-face form factors. Finally, speech-EMG correlation experiments demonstrate a linear relationship between many EMG spectrogram frequency bins and self-supervised speech representation dimensions.

Index Terms: electromyography, EMG, EMG-to-speech

1. Introduction

Devices that decode speech from electromyography (EMG) are valuable for assistive communication applications, as they allow users to speak without vocalizing and can be life-changing in helping overcome dysarthria, dysphagia, stutters, and laryngectomies [1, 2, 3, 4, 5, 6, 7]. In single-speaker settings, EMG-to-speech methods have synthesized high-fidelity speech, illustrating the viability of this technology for augmented spoken communication [1, 8, 9]. Current devices measure articulatory signals from electrodes placed on the face [1, 10, 11]. In this paper, we explore a more discreet form factor that only places dry electrodes around a neckband. Thus simplifying system setup, improving subject comfort, and eliminating the use of stigmatizing wet electrode arrays around the ears or face.

Surface EMG devices placed on the neck have shown success in diagnosing medical conditions like dysphagia [12, 13], estimating neck muscle activity [14], and measuring vocal fold vibrations [15]. These techniques have been extended to speech classification using the same neck EMG signals, but high-accuracy speech decoding with wearables remains challenging [16]. Additionally, existing demonstrations of surface EMG typically rely on wet electrodes that require periodic skin preparation to maintain consistent electrode-skin contact. While these wet electrodes can achieve consistent data with minimal electrode-motion-related artifacts, they are cumbersome and limit everyday use cases in public. Significantly more comfortable and easy-to-use 'dry' electrode devices have been demonstrated for arm-based EMG devices [17, 18], but not yet on the neck. While these dry electrodes demonstrate higher electrode-skin impedances and are more susceptible to environmental interference and motion artifacts [19], their improved ease-of-use, lack of hydrogel, and increased longevity make them ideal for wearable speech decoding.

To explore wearable speech decoders, we propose a form factor with reusable dry electrodes and greater spatial coverage compared to traditional EMG neck devices, such as those used for electroglottographs [15]. To this effect, this work presents a wireless, dry-electrode neckband used to capture EMG and speech data from two subjects. The resulting data is used for speech classification and mapping EMG to text and speech. Our high multi-speaker classification performance (Section 4.1) demonstrates an average accuracy of up to 92.7%. We also study the importance of the number of neck electrodes, finding that restricting this to two electrodes, as in electroglottographs [15], noticeably reduces classification accuracy. Phonological confusion analyses in Section 4.2 indicates that classification behavior with neck-only data is similar to that with both neck and face data. Finally, speech-EMG correlation experiments in Section 4.3 indicate that our EMG data can be linearly mapped to speech acoustic representations and that neck electrodes can potentially be used on their own to perform speech decoding.

2. Wireless EMG Neckband

The wireless recording setup was designed to explore easy-to-use, discreet, and comfortable wearables for speech decoding. As a result, the system consists of a dry electrode EMG neckband and a wireless recording module (Fig. 1a and b). In addition, three wet Ag/AgCl electrodes are included on the face to simultaneously record a benchmark for the neck recorded EMG. Emphasis was placed on modularity, and system characteristics such as electrode count, location, and chemistry can be easily augmented due to our 3D printing-based fabrication process and recording hardware that can record electrophysiological signals from up to 64 electrodes.

2.1. Electrodes

The dry electrode necklace comprises 10 equally spaced dry electrodes placed all around an elastic neckband (Fig. 1b). The dry electrodes are 3D printed and electroless gold-plated to achieve an inert, biocompatible surface [20]. This fabrication process results in dry electrodes with a high effective surface area that can be used without any hydrogel application or skin abrasion, greatly improving user comfort while also reducing the risk of skin irritation [21]. Lastly, this fabrication process can easily be repeated or augmented to result in arbitrarily shaped electrodes that conform to the jaw or other features in future studies.

The electrodes were arranged to comfortably fit the average neck circumference of men (40.1 cm +/- 3.05 cm) and women (34.8 cm +/- 2.79 cm) [22]. As a result, the final neck elec-

[*]These authors contributed equally to this work. [**]Equal advising.

10.21437/Interspeech.2024-1568

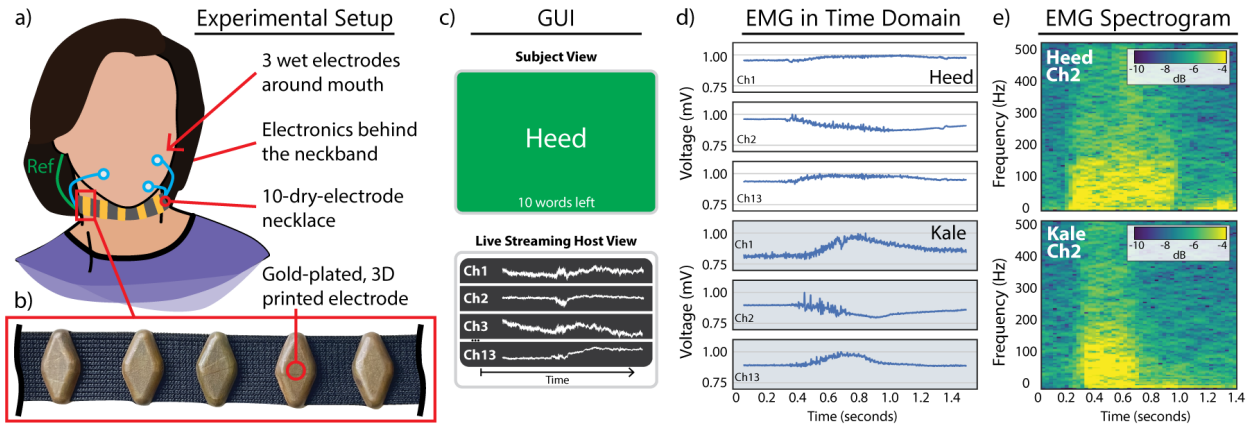

Figure 1: *(a) Experimental setup including a dry electrode neckband, baseline monitoring face electrodes, wet reference electrode behind the right ear, and neckworn electronics behind the head. (b) Partial photograph of 3D printed, gold plated neck electrodes. (c) Sample renders of the experiment GUI's subject and host views. Subject view displays a teleprompter while raw EMG data is live plotted on the host view. (d) Raw sample EMG from a single utterance of the words 'Heed' and 'Kale'. (d) Sample EMG time-frequency spectrograms (see section 3.2) from a single utterance of the words 'Heed' and 'Kale'.*

trodes have a cross-sectional surface area of 2 cm^2 and a pitch of 3 cm. These electrodes achieve an average 50 Hz electrode-skin impedance (ESI) and phase of 55.1 $k\Omega s$ and -95° at 50 Hz (N = 10) which, while potentially higher than the ESI of wet electrodes, is well within the input parameters of the system's neural recording frontends. The electrodes exhibit a mean electrode DC offset of -13.3 mV with a standard deviation of 14.1 mV (N = 600). After fabrication, the electrodes are clipped into an elastic, velcro neckband. Each electrode is soldered to a 36 AWG jumper cable that is threaded through the elastic band to minimize wire-motion-related artifacts.

The neck electrodes are used in conjunction with three wet, Ag/AgCl electrodes around the subject's lips to track lip, palate, and jaw movements [1]. Two electrodes are placed above and below the left side of the subject's lips, while one electrode is placed on the right side of the lips. These wet electrodes are used to provide a comparison benchmark for decoding tasks described in section 3.3 and 4. All neck and face electrodes are used to perform differential measurements with the shared reference wet electrode placed on the subject's right mastoid. Each differential channel is connected to a wireless recording module worn behind the neck.

2.2. Wireless Recording Module

EMG was recorded using an existing compact recording platform, known as the miniature, wireless, artifact-free neuromodulation device (WANDmini) [23]. WANDmini, originally built for implanted neural recording, has been adapted for wearable applications and deployed in multiple electrophysiological studies [24, 25]. Due to being small (2.5 $\times$ 2.5 cm^2) and lightweight (3.8 g), WANDmini can discreetly fit on the back of the neckband and be comfortably worn for hours. WANDmini records and digitizes EMG signals with a custom neuromodulation IC [26] (NMIC, Cortera Neurotechnologies, Inc.). The recorded data is then processed and packetized by an onboard FPGA SoC (166 MHz ARM Cortex M3 processor - SmartFusion2 M2S060T, Microsemi). The packetized data is then transmitted to the base station by a 2.4 GHz BLE radio (nRF51822, Nordic Semiconductor) that is also used to configure WANDmini. When powered by a 3.7 V, 300 mAh lithium polymer (LiPo) battery, the neckband and WANDmini can operate and

Table 1: *WANDmini System Electrical Specifications*

Maximum Recording Channels	64
Recording Channels Used	13
Reference Location	Right Mastoid
Input Range	100 mVpp
ADC Resolution	15 bits
ADC Sample Rate	1 kSps
Noise Floor	$70nV/\sqrt{Hz}$
Wireless Data Rate	2 Mbps
WANDmini Power	46 mW
Weight (w/o battery)	3.8 g
Battery Life	44 Hours

digitize 64 channels of data for roughly 24 hours. Pertinent system specifications are described in Table 1.

Packetized EMG signals are received by a wireless base station connected to a laptop running a Python graphical user interface (GUI) that not only provides real-time data visualization but also provides cues and teleprompting for subjects (Fig. 1c). In addition to data visualization and teleprompting, the GUI also records audio of the subject vocalizing the cued utterance. All audio and EMG are synchronized with trigger signals sent by the GUI to the laptop's microphone and WANDmini. During each experiment, the GUI consists of two main components: a screen displaying real-time EMG signals from the EMG device, and a teleprompter that presents words sequentially from a given utterance list. This setup allows the subject to vocalize the words shown while a host monitors the recorded EMG and audio data. The teleprompter's pacing can be adjusted to accommodate different speakers, ensuring that the speech produced during data collection is natural.

2.3. Experimental Setup and Data Collection

To verify the neckband system performance, electrophysiological and speech measurements were performed on two subjects. At the start of each experiment, the subject would arrive and don the neckband. The two center electrodes of the neckband were aligned around the subject's Adam's apple and then tightened so each electrode was making contact with the skin. After the neck electrodes were placed around the entire neck, the wet Ag/AgCl electrodes were placed around the subject's lips. Once all the

electrodes were connected to the recording frontend, WAND-mini would be stowed in a 3D-printed enclosure affixed to the back of the neckband. The subject would then sit in front of the laptop, approximately 3 feet away from the microphone, and vocalize the utterances displayed on the screen (1c). In total, 11 words were displayed 10 times each (each instance of a word is referred to as an utterance). The GUI first shows 'wait' for three seconds, then the specified word for three seconds, and a final 'wait' for three seconds. As a result, each recording totals 9 seconds but is sliced to the 1.5 seconds around the actual utterance. This user study was approved by UC Berkeley's Institutional Review Board (CPHS protocol ID: 2018-09-11395).

3. Computational Methods

3.1. Dataset

Our complete dataset is composed of 13 channels of EMG (sampled at 1000 Hz) and voiced acoustics (sampled at 44100 Hz) for 11 words. These 13 channels comprise 10 neck channels and 3 face channels. This dataset was further segmented into a 10-channel dataset (consisting only of the dry neck electrodes), and a 13-channel dataset (consisting of all electrodes).

Our utterance list consists of the following words, listed with their IPA transcriptions: heed [hid], had [hæd], hood [hʊd], tail [tʰeɪl], kale [kʰeɪl], doe [doʊ], goat [goʊt], aba [aba], ada [ada], aga [aga], and aka [akʰa]. In this set of utterances, we have minimal pairs that differ only in the vowel ([hid]/[hæd]/[hʊd]), (near-)minimal pairs that differ only in plosive place of articulation ([aba]/[ada]/[aga] and [doʊ]/[goʊt] and [tʰeɪl]/[kʰeɪl]), and minimal pairs that differ in voicing and aspiration of a plosive ([aga]/[akʰa]). Minimal pairs in our data allow us to isolate certain aspects of speech while keeping other factors constant, which helps us assess the performance of our model on each of these aspects.

Each word is recorded 10 times by two native male English speakers (arbitrarily fixed as Speaker 1 and Speaker 2). The pertinent 1.5 seconds recorded for each utterance are kept, yielding 5 minutes and 30 seconds of data in total. Since EMG and voiced speech audio are recorded at the same time (Section 2.3), we treat these two modalities as temporally aligned.

3.2. EMG Representations

Raw EMG in the time-domain can be readily featurized for classification tasks. Inspired by openSMILE, we compute the following statistics for each EMG channel: max, min, range, max position, min position, arith-mean, quad-mean, std, var, kurtosis, skewness, 25-percentile, 75-percentile, number of peaks, mean peak amplitude, mean abs slope, rise time, fall time, zcr, mcr [27]. We concatenate the vectors from each dimension into a single vector, giving us a feature vector for each EMG utterance. Sample raw EMG data of a single utterance of the words 'Heed' and 'Kale' are shown in Figure 1d.

While the above time-domain and statistical features are adequate for classification, they are also more noise susceptible and contain fewer dimensions relative to frequency-domain representations of EMG. A time-frequency spectrogram, on the other hand, can be easily filtered and readily compared to acoustic speech representations. As a result, spectrograms are computer for every utterance's channels using consecutive Fourier transforms. The number of samples per segment was set to 100, while the overlap between segments was set to 50. Additionally, the number of Fast Fourier Transform (FFT) points per segment was set to 128. These parameters allowed for a high quality spectrogram to be generated in order to better enable data analysis. Figure 1e depicts two spectrograms extracted from channel 2 of the raw EMG shown in Figure 1d.

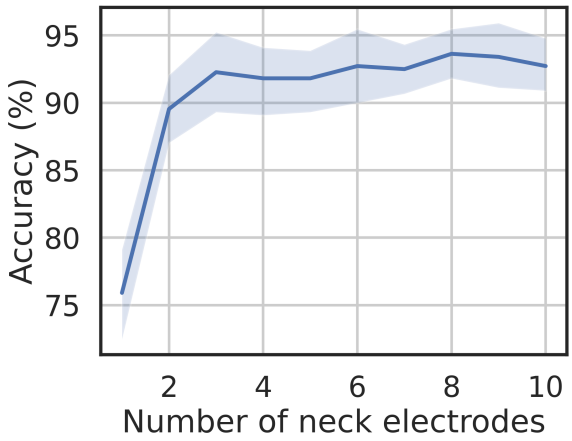

Figure 2: *Classification accuracy for different numbers of neck electrodes. Solid lines are means and opaque regions are 95% confidence intervals.*

3.3. Self-Supervised Acoustic Representations

Self-supervised acoustic speech representations encode waveforms into a dense vector space suitable for speech production and perception [28, 29]. Given the versatility of mapping to and from these representations, we compare our EMG data with these features to study the feasibility of mapping EMG signals to acoustic ones. We choose WavLM as the self-supervised speech representation in this paper due to its success in speech benchmarks and articulatory tasks [28, 29, 30, 31]. WavLM is a state-of-the-art pre-trained Transformer model that extends HuBERT, a masked speech prediction method, by adding a denoising objective [32, 33]. This model also employs a gated relative position bias to better utilize sequence ordering and is trained on a larger and more diverse dataset across different scenarios and languages [28]. Given the generalizability of WavLM, we hypothesize that features extracted by this model may be robust enough to map to EMG data. Since WavLM accepts 16000 Hz waveform inputs, we downsample our speech audio from 44100 Hz. We extract features from the tenth layer of WavLM, given the articulatory properties around that model depth [30]. This yields 1024-dimensional vectors at each time step, where time steps have a sampling rate of 50 Hz.

4. Results

4.1. EMG Classification

To check our EMG data quality, we first classify words from EMG with a random forest classifier with max depth of 32. Here, our inputs are the EMG statistics vectors described in Section 3.2. On 10 random 80%-20% train-test splits of our two-speaker dataset, we achieve a mean accuracy of 93.9% with a 95% confidence interval of [92.7%, 95.0%] using all 13 electrodes. With only the 10 neck electrodes, we achieve a mean accuracy of 92.7% with a 95% confidence interval of [90.9%, 94.8%]. Accuracies in both cases are much higher than chance, which is around 9% given that each of our 11 words has the same number of utterances. This suggests that our EMG data and the neck-only subset both contain useful linguistic content.

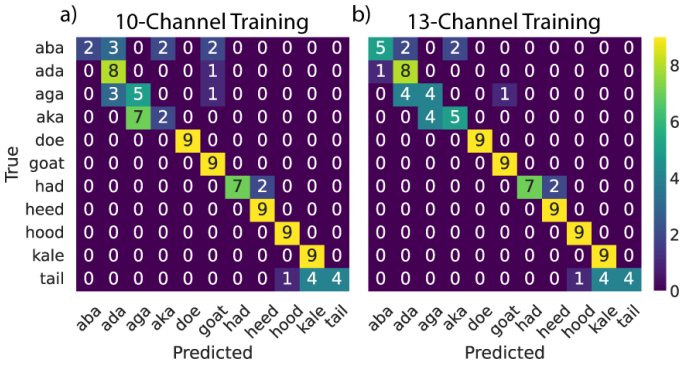

Figure 3: *Confusion matrices using model trained on (a) the 10 neck channels and (b) all 13 channels.*

We also check the importance of the number of neck electrodes for EMG decoding. With the aforementioned train-test splits, we calculate the word classification accuracy given different numbers of neck electrodes as input. Accuracy means and 95% confidence intervals are plotted on Figure 2. Accuracy noticeably improves after adding the second and third electrode, and continues improving a bit up to adding the eighth electrode. This suggests that language decoding from EMG can improve when the device has more than two electrodes, such as in the typical electroglottograph setup[15].

4.2. Phonological Confusion

To analyze our EMG data phonologically, we generate confusion matrices for EMG classification (Fig. 3). We use the same random forest classifier, utterance set, statistical EMG input vector representations, and target labels as in Sec. 4.1. In order to emphasize model confusion, we report results for 1-shot classification. Specifically, we train on all of one speaker's data and 1 random utterance for each word from the other speaker, and test on the remaining data from the second speaker. We generate two confusion matrices, one for each first-second speaker assignments, and add together the matrices element-wise for conciseness. This process was performed with the full 13-channel dataset (Fig. 3a) and with the 10-channel dataset (Fig. 3b).

As is can be seen from the confusion matrices in Figure 3, our model can generally predict the correct vowel based on the EMG data, but it has more trouble with identifying specific consonants. In particular, we observe that the model trained on all 13 channels primarily confuses plosives that either: (1) differ in place of articulation but match in voicing i.e. [d]/[b] and [k]/[t], or (2) differ in voicing but match in place of articulation i.e. [k]/[g]. In addition, the model may also be weaker at differentiating between front vowels as opposed to other vowels as the model confuses had and heed but can distinguish between these words and hood. This may be because of the placement of electrodes on the neck given that back vowels cause more muscles to be engaged near the neck while front vowels do not.

We note that there is some improvement in model performance after adding three extra wet electrodes to the face. The model trained on only neck electrode channels confuses [goʊt] and [aba] while the model trained on all 13 channels does not. The 3 extra electrodes near the lips may help the model identify the labial consonant [b] in [aba]. In addition, for the ground truth word [ada], the 13-channel model predicts [aba] while the 10-channel model predicts [goʊt]. The facial electrodes likely also help the model distinguish between vowels and consonants since the 13-channel model is able to identify a word-initial vowel as opposed to the 10-channel model, which confuses them. Generally, the model confusability is similar for the 13- and 10-channel settings, suggesting that our necklace form factor may capture enough information to decode speech.

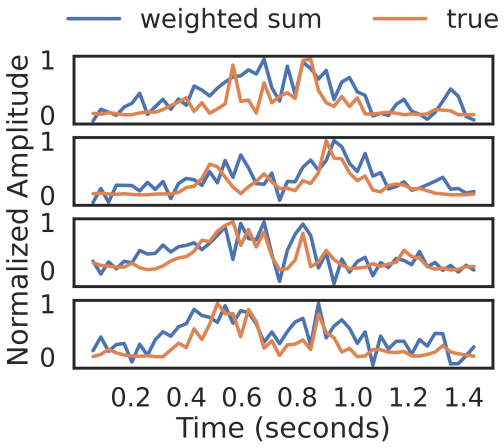

Figure 4: *Weighted sum of self-supervised speech features match EMG spectrogram frequency bins. Here, we plot 1 EMG channel of a "kale" utterance for bins 90-94 Hz, 102-105 Hz, 238-242 Hz, and 348-352 Hz (Top-to-bottom).*

4.3. Speech-EMG Correlation

Through a speech-EMG correlation experiment, we observe that self-supervised speech features and time-frequency representations of EMG correlate noticeably. For all utterances in our dataset, we encode waveforms into WavLM representations [28] (Section 3.3) and EMG into spectrograms (Section 3.2). For our EMG representation, we flatten the 10 EMG channels and 129 spectrogram frequencies into a 1290 vector, yielding 1290 values varying over time. We linearly interpolate the WavLM features to match the length of the EMG features. Then, we train a linear regression model on all of the data to map the WavLM vector at each time step to the respective flattened EMG vector. In other words, we approximate each flattened EMG dimension with a weighted sum of WavLM dimensions. Out of the 1290 EMG dimensions, 33.1% of them have a mean Pearson correlation coefficient of at least 0.5 with a linear combination of WavLM dimensions, with examples visualized in Figure 4. 33.1% is noticeably higher than the 0.0% that occurs when we replace WavLM elements with numbers randomly uniformly sampled from [0,1]. This suggests that our EMG data contains useful speech acoustic information.

5. Conclusion

This work presents a discreet EMG neckband with reusable dry electrodes capable of performing speech classification and analysis. Ablation studies with our device indicate that the neck electrodes can achieve a high classification accuracy on their own (92.7%) which is similar to classification accuracies achieved with both neck and face electrodes (93.9%). Additionally, speech-EMG correlation experiments reveal that our device can record useful speech information for further speech decoding work. Moving forward we will collect sentence-length utterances from a larger set of speakers to explore wearable EMG-to-speech synthesis through the use of a necklace form factor without any face electrodes.

6. References

[1] D. Gaddy and D. Klein, "An improved model for voicing silent speech," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 175–181.

[2] J. S. Galego, O. V. Casas, D. Rossato, A. Simões, and A. Balbinot, "Surface electromyography and electroencephalography processing in dysarthric patients for verbal commands or speaking intention characterization," *Measurement*, vol. 175, p. 109147, 2021.

[3] C. E. Stepp, J. T. Heaton, R. G. Rolland, and R. E. Hillman, "Neck and face surface electromyography for prosthetic voice control after total laryngectomy," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 17, no. 2, p. 146–155, Apr. 2009. [Online]. Available: http://dx.doi.org/10.1109/TNSRE.2009.2017805

[4] F. Wang and E. M.-L. Yiu, "Surface electromyographic activity of the suprahyoid and sternocleidomastoid muscles in pitch and loudness control," *Frontiers in Physiology*, vol. 14, May 2023.

[5] A. D. Chan, K. Englehart, B. Hudgins, and D. F. Lovely, "Myoelectric signals to augment speech recognition," *Medical and Biological Engineering and Computing*, vol. 39, pp. 500–504, 2001.

[6] G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy, and J. C. Kline, "Silent speech recognition as an alternative communication device for persons with laryngectomy," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 12, pp. 2386–2398, 2017.

[7] S.-C. Jou, L. Maier-Hein, T. Schultz, and A. Waibel, "Articulatory feature classification using surface electromyography," in *2006 IEEE international conference on acoustics speech and signal processing proceedings*, vol. 1. IEEE, 2006, pp. I–I.

[8] M. Lyu, C. Xiong, and Q. Zhang, "Electromyography based chinese voice command recognition," in *2014 IEEE International Conference on Information and Automation (ICIA)*. IEEE, Jul. 2014.

[9] J. Wu, Y. Zhang, L. Xie, Y. Yan, X. Zhang, S. Liu, X. An, E. Yin, and D. Ming, "A novel silent speech recognition approach based on parallel inception convolutional neural network and mel frequency spectral coefficient," *Frontiers in Neurorobotics*, vol. 16, Sep. 2022.

[10] K. Scheck and T. Schultz, "Multi-speaker speech synthesis from electromyographic signals by soft speech unit prediction," in *ICASSP*, 2023.

[11] A. Kapur, S. Kapur, and P. Maes, "Alterego: A personalized wearable silent speech interface," in *23rd International conference on intelligent user interfaces*, 2018, pp. 43–53.

[12] M. Vaiman and E. Eviatar, "Surface electromyography as a screening method for evaluation of dysphagia and odynophagia," *Head & face medicine*, vol. 5, no. 1, pp. 1–11, 2009.

[13] V. Gupta, N. P. Reddy, and E. P. Canilang, "Surface emg measurements at the throat during dry and wet swallowing," *Dysphagia*, vol. 11, pp. 173–179, 1996.

[14] C. M. Sommerich, S. M. Joines, V. Hermans, and S. D. Moon, "Use of surface electromyography to estimate neck muscle activity," *Journal of Electromyography and Kinesiology*, vol. 10, no. 6, pp. 377–398, 2000.

[15] F. Lecluse, M. Brocaar, and J. Verschuure, "The electroglottography and its relation to glottal activity," *Folia Phoniatrica et Logopaedica*, vol. 27, no. 3, pp. 215–224, 1975.

[16] P. Chen, L. Chen, and X. Mao, "Content classification with electroglottograph," in *Journal of Physics: Conference Series*, vol. 1544, no. 1. IOP Publishing, 2020, p. 012191.

[17] P. Laferriere, E. D. Lemaire, and A. D. C. Chan, "Surface electromyographic signals using dry electrodes," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 10, p. 3259–3268, Oct. 2011.

[18] A. C. Myers, H. Huang, and Y. Zhu, "Wearable silver nanowire dry electrodes for electrophysiological sensing," *RSC Advances*, vol. 5, no. 15, p. 11627–11632, 2015.

[19] Y. M. Chi, T. P. Jung, and G. Cauwenberghs, "Dry-contact and noncontact biopotential electrodes: Methodological review," *IEEE Reviews in Biomedical Engineering*, vol. 3, pp. 106–119, 2010.

[20] R. Kaveh, N. Tetreault, K. Gopalan, J. Maravilla, M. Lustig, R. Muller, and A. C. Arias, "Rapid and scalable fabrication of low impedance, 3d dry electrodes for physiological sensing," *Advanced Materials Technologies*, p. 2200342, 5 2022.

[21] S. Stjerna, P. Alatalo, J. Mäki, and S. Vanhatalo, "Evaluation of an easy, standardized and clinically practical method (sureprep) for the preparation of electrode–skin contact in neurophysiological recordings," *Physiological Measurement*, vol. 31, no. 7, p. 889, may 2010.

[22] J. Kornej, H. Lin, L. Trinquart, C. R. Jackson, D. Ko, E. J. Benjamin, and S. R. Preis, "Neck circumference and risk of incident atrial fibrillation in the framingham heart study," *Journal of the American Heart Association*, vol. 11, no. 4, p. e022340, 2022.

[23] A. Zhou, S. R. Santacruz, B. C. Johnson, G. Alexandrov, A. Moin, F. L. Burghardt, J. M. Rabaey, J. M. Carmena, and R. Muller, "A wireless and artefact-free 128-channel neuromodulation device for closed-loop stimulation and recording in non-human primates," *Nature Biomed. Eng.*, vol. 3, no. January, pp. 15 – 26, 2019.

[24] R. Kaveh, J. Doong, A. Zhou, C. Schwendeman, K. Gopalan, F. L. Burghardt, A. C. Arias, M. M. Maharbiz, and R. Muller, "Wireless user-generic ear eeg," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, pp. 727–737, 8 2020.

[25] A. Moin, A. Zhou, A. Rahimi, A. Menon, S. Benatti, G. Alexandrov, S. Tamakloe, J. Ting, N. Yamamoto, Y. Khan, F. Burghardt, L. Benini, A. C. Arias, and J. M. Rabaey, "A wearable biosensing system with in-sensor adaptive machine learning for hand gesture recognition," *Nature Electronics*, vol. 4, pp. 54–63, 2021.

[26] B. C. Johnson, S. Gambini, I. Izyumin, A. Moin, A. Zhou, G. Alexandrov, S. R. Santacruz, J. M. Rabaey, J. M. Carmena, and R. Muller, "An implantable 700μW 64-channel neuromodulation IC for simultaneous recording and stimulation with rapid artifact recovery," *IEEE Symp. VLSI Circuits*, pp. C48–C49, 2017.

[27] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[28] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, M. Zeng, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2021.

[29] S. wen Yang *et al.*, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Interspeech*, 2021.

[30] C. J. Cho *et al.*, "Evidence of vocal tract articulation in self-supervised learning of speech," in *ICASSP*, 2023.

[31] P. Wu *et al.*, "Speaker-independent acoustic-to-articulatory speech inversion," in *ICASSP*, 2023.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[33] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.