emg2speech: synthesizing speech from electromyography using self-supervised speech models

Harshavardhana T. Gowda, Daniel C. Comstock, and Lee M. Miller
University of California, Davis
Correspondence: tgharshavardhana@gmail.com

Abstract

We present a neuromuscular speech interface that translates electromyographic (EMG) signals recorded from orofacial muscles during speech articulation directly into audio. We find that self-supervised speech (S3) representations are strongly linearly related to the electrical power of muscle activity: a simple linear mapping predicts EMG power from S3 representations with a correlation of $r = 0.85$. In addition, EMG power vectors associated with distinct articulatory gestures form structured, separable clusters. Together, these observations suggest that S3 models implicitly encode articulatory mechanisms, as reflected in EMG activity. Leveraging this structure, we map EMG signals into the S3 representation space and synthesize speech, enabling end-to-end EMG-to-speech generation without explicit articulatory modeling or vocoder training. We demonstrate this system with a participant with amyotrophic lateral sclerosis (ALS), converting orofacial EMG recorded while she *silently* articulated speech into audio.

PROJECT PAGE. GITHUB. DATA.

1 Introduction

Neural and neuromuscular interfaces hold significant promise for augmenting human abilities to interact and communicate with the external world. Brain-computer interfaces (BCIs), such as the speech neuroprostheses described in Wairagkar et al. (2025); Metzger et al. (2023); Willett et al. (2023), have shown that individuals with conditions such as anarthria or amyotrophic lateral sclerosis can regain functional speech through invasive neural recordings. While these invasive approaches are well suited for individuals with severe paralysis or complete loss of articulatory control, their widespread deployment is limited by the need for surgical implantation, high cost, and clinical risk. In contrast, we propose a non-invasive speech interface that leverages preserved articulatory muscle activity, enabling a broader range of individuals, including those with laryngectomy, dysarthria, or dysphonia, to regain functional speech without surgical intervention.

In this article, we present a method that leverages self-supervised speech (S3) models to convert electromyographic (EMG) signals collected during speech articulation directly into audio, without explicitly training a vocoder. Our key insight comes from the observation that speech features derived from S3 models can be linearly mapped to the electrical power of muscle action potentials. Because EMG power patterns associated with different articulatory gestures form structured and separable clusters in feature space, these results suggest that S3 models implicitly encode articulatory information, as reflected in EMG activity. This relationship motivates EMG power as an effective intermediate representation for mapping muscle activity to speech features. We exploit this property to design a lightweight EMG-to-audio conversion model that leverages EMG power representations in conjunction with S3 models.

2 Prior work

Converting non-speech signals into audio has been explored across several modalities, including lip-movements-to-speech (Kim et al., 2021; Prajwal et al., 2020), motor-cortex neural signals-to-speech (Wairagkar et al., 2025; Metzger et al., 2023; Littlejohn et al., 2025), and EMG-to-speech (Gaddy and Klein, 2020, 2021). Most existing approaches in these domains (Kim et al., 2021; Prajwal et al., 2020; Wairagkar et al., 2025; Gaddy and Klein, 2020, 2021) assume that the alignment between the input signals (e.g., video or neural activity) and the corresponding audio is known. In contrast, we address a more challenging scenario, similar to Metzger et al. (2023); Littlejohn et al. (2025), where the alignment between neural activity (in our

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16490–16507

case, EMG) and speech is *unknown*. This setting requires the model not only to learn the mapping between EMG activity and audio but also to infer the underlying alignment from an exponential search space.

Work in Littlejohn et al. (2025); Metzger et al. (2023) addresses this alignment-free setting by training an encoder that takes motor-cortex neural signals as input and maps them to discrete HuBERT units (Lakhotia et al., 2021), which are then passed to a pretrained Tacotron (Wang et al., 2017) vocoder following the pipeline in Lakhotia et al. (2021). We adopt a similar high-level pipeline for EMG-to-speech conversion. However, our approach explicitly leverages the *geometric structure* of EMG signals and their relationship to self-supervised (S3) speech representations to design an encoder grounded in articulatory mechanisms.

Despite this progress, prior work faces several practical limitations. For instance, Littlejohn et al. (2025); Metzger et al. (2023) use a small-vocabulary corpus containing only 1,024 words, and in Littlejohn et al. (2025) (where motor-cortex neural activity is converted into speech), each test sentence was presented to the model an average of 6.94 times during training. Moreover, these studies are not fully reproducible due to limited implementation details and the lack of public access to the data used in the experiments. Since non-speech-to-speech conversion typically involves multiple components in an end-to-end pipeline, opaque designs make it difficult to reproduce results and to compare methods fairly. These limitations motivate the need for richer public datasets and reproducible benchmarks for fair evaluation, both of which we provide in this work.

2.1 Our contributions

We make three primary contributions.

First, we open-source one of the largest high quality EMG-to-speech datasets to date, comprising a large-vocabulary corpus of approximately 9 hours of EMG speech data with over 6,800 unique words from a healthy participant, as well as a small-vocabulary corpus of approximately 1 hour of EMG speech data with roughly 300 unique words from a participant with ALS. To the best of our knowledge, these datasets constitute one of the largest and most comprehensive publicly available resources for EMG-to-speech conversion.

Second, building on these datasets, we develop encoder architectures grounded in articulatory mechanisms that exhibit interpretability and operate effectively in low-data regimes, including settings with as little as 40 minutes of training data from an ALS participant. This is particularly important given the practical difficulty of collecting large-scale EMG datasets with current sensing technology. To support learning under limited supervision, we leverage massively pretrained S3 models and use their representations as a structured target space for EMG-to-speech mapping. We further establish, for the first time, a quantitative relationship between EMG signals and S3 representations, and exploit this structure to guide encoder design.

Third, we introduce phoneme-guided decoding for EMG-to-speech synthesis and show that incorporating phonetic structure improves the quality of the generated audio. This is because phonemes are defined by articulatory configurations, and EMG signals recorded from multiple orofacial muscle sites capture phonetic structure more faithfully than HuBERT audio units.

Unlike prior EMG-to-speech benchmarks (Gaddy and Klein, 2020, 2021), our approach does not assume known temporal alignment between EMG and audio during training. This design is motivated by clinically relevant scenarios in which parallel EMG-audio pairs may be unavailable or unreliable, such as laryngectomy (absence of laryngeal voicing) or ALS (degraded acoustic recordings due to bulbar impairment). As a result, the model must learn without frame-level EMG-audio correspondence, which substantially increases the difficulty of the learning problem. Overall, our contributions address fundamental aspects of EMG-to-speech modeling and are simple to implement, yet work well with widely available off-the-shelf pretrained components.

Because our study derives audio from text transcripts rather than using time-aligned EMG-audio pairs, and targets an unaligned EMG-to-speech synthesis setting, there are no existing benchmarks that support direct one-to-one comparisons. Nevertheless, where possible, we compare against representative baselines from recent EMG interface literature, including spectrogram-based feature pipelines from EMG2QWERTY (Sivakumar et al., 2024), to contextualize performance. In appendix C, we additionally provide broader comparisons to prior brain-computer speech interfaces for context, although these are not intended as direct one-to-one comparisons.

3 Data

We use three datasets in this study: (i) a large, general-corpus vocabulary dataset from a healthy participant, denoted DATA $_{\text{GENERAL}}$; (ii) a small, limited-corpus vocabulary dataset from a participant with ALS, denoted DATA $_{\text{ALS}}$; and (iii) a dataset of discrete orofacial gestures underlying speech articulation collected from 12 healthy participants, denoted DATA $_{\text{OROFACIAL GESTURES}}$. Below, we describe each dataset in detail.

3.1 DATA $_{\text{GENERAL}}$

A healthy participant naturally articulates English sentences while the corresponding EMG signals are recorded at 5000 Hz. We record EMG from 31 muscle sites on the neck, chin, jaw, cheek, and lips using monopolar surface electrodes (see figure 5 for electrode placement and appendix A for additional details).

We adapt the language corpus from Willett et al. (2023), who demonstrate a speech brain-computer interface by translating motor-cortex activity into speech. The dataset comprises an English corpus with approximately 6800 unique words and 9660 sentences. Sentences vary in length, and the participant articulated at a normal speaking rate, averaging 115 words per minute. We split the dataset into training, validation, and test sets containing 8500, 760, and 400 sentences, respectively. Sentences in the test set do not appear in the training or validation sets.

The start and end of each sentence are timestamped using mouse clicks. When the participant is ready to begin, they click the mouse to display the sentence on the screen and mark the start time. After articulation is complete, they click again to mark the end time; this second click removes the sentence from the screen. This procedure allows the participant to articulate at their own pace.

3.2 DATA $_{\text{ALS}}$

A participant diagnosed with amyotrophic lateral sclerosis (ALS) silently articulates English sentences (with overt articulatory movements but no audible output) while we record the corresponding EMG signals at 5000 Hz using the same electrode layout as in DATA $_{\text{GENERAL}}$.

We construct a small English corpus comprising approximately 300 unique words and 600 sentences. Sentences vary in length, and the participant articulated at her current comfortable speaking rate, averaging 61 words per minute. We split the dataset into training, validation, and test sets containing 500, 40, and 60 sentences, respectively. Test sentences do not appear in the training or validation sets.

3.3 DATA $_{\text{OROFACIAL GESTURES}}$

Twelve participants perform 13 distinct orofacial movements, with 10 repetitions per movement. The set of movements includes `cheeks: puff out`, `cheeks: suck in`, `jaw: drop down`, `jaw: move backward`, `jaw: move forward`, `jaw: move left`, `jaw: move right`, `lips: pucker`, `lips: smile`, `lips: tuck (as if blotting)`, `tongue: back of lower teeth`, `tongue: back of upper teeth`, and `tongue: roof of the mouth`. These movements are selected to span a broad range of articulatory gestures involved in natural speech production, including lip rounding, jaw positioning, and tongue placement, which are essential for producing different phonemes.

This dataset is recorded using 22 electrodes at a sampling rate of 5000 Hz. Signals are recorded from approximately the same muscle sites as in DATA $_{\text{GENERAL}}$, except that electrodes on the right side of the neck are not used (middle diagram in figure 5). Each gesture is performed for a duration of 1.5 s.

4 Methods

4.1 Electromyography (EMG)

EMG signals are collected by a set of sensors $\mathcal{V}$ and represented as functions of time t. A sequence of EMG signals E corresponding to articulated speech, associated with an audio signal A and phonemic content L, is represented as $E = \{\mathbf{f}_v(t)\}_{\forall v \in \mathcal{V}}$. Here, $\mathbf{f}_v(t)$ denotes the EMG signal captured at sensor node v as a function of time. The audio signal A encodes both phonemic (lexical) content and expressive aspects of speech such as volume, pitch, prosody, and intonation, while L represents only the phonemic content—a sequence of phonemes. For example, the phonemic content L of the word <FRIDAY> is denoted by the phoneme sequence <F-R-IY-D-AY>.

EMG covariance matrices: for an EMG signal $E_{\mathcal{V}\times\tau}$ collected from $\mathcal{V}$ sensor nodes over a duration of τ samples, we construct a symmetric positive definite (SPD) covariance matrix $\mathcal{E}_{\mathcal{V}\times\mathcal{V}} = \epsilon EE^\top$, where ϵ is a scaling factor. We denote the diagonal of $\mathcal{E}$ as $\mathbb{D}(\mathcal{E})$ and its lower triangular part

as $\lfloor \mathcal{E} \rfloor$. The vector $\mathbb{D}(\mathcal{E})$ represents the muscle action potential power at each electrode $v \in \mathcal{V}$ during the interval τ, while the off-diagonal elements capture the pairwise cross-channel covariance, reflecting the spatial co-activation structure across electrodes. A vectorized representation of $\mathcal{E}$ is denoted as $\mathtt{vec}(\mathcal{E})$, a column vector of dimension $\mathcal{V}^2$.

The geodesic distance between two SPD matrices $\mathcal{E}_1$ and $\mathcal{E}_2$ is the same as the distance between their corresponding Cholesky matrices $\mathcal{L}_1$ and $\mathcal{L}_2$ (Lin, 2019) and is calculated as

$$d(\mathcal{L}_1, \mathcal{L}_2) = \{||\lfloor \mathcal{L}_1 \rfloor - \lfloor \mathcal{L}_2 \rfloor||_F^2 + ||\log \mathbb{D}(\mathcal{L}_1) - \log \mathbb{D}(\mathcal{L}_2)||_F^2\}^{1/2}, \quad (1)$$

where $||\cdot||_F$ denotes the Frobenius norm. Here, $\mathcal{L}_1$ and $\mathcal{L}_2$ are the *Cholesky factors* of the SPD matrices $\mathcal{E}_1$ and $\mathcal{E}_2$, i.e., lower triangular matrices such that $\mathcal{E} = \mathcal{L}\mathcal{L}^\top$.

EMG spectrograms: for an EMG signal $E_{\mathcal{V}\times\tau}$ collected from $\mathcal{V}$ sensor nodes at a sampling frequency f_s, we compute the short-time Fourier transform (STFT) over successive time windows to obtain a power spectrogram representation $\mathcal{S}_{\mathcal{V}\times F\times\tau'} = \left|\mathrm{STFT}(E_{\mathcal{V}\times\tau})\right|^2$, where F denotes the number of frequency bins and τ' the number of time frames. Each slice $\mathcal{S}^{(t)}_{\mathcal{V}\times F}$ captures the frequency-domain energy distribution of EMG activity across $\mathcal{V}$ electrodes at time frame t. To reduce the spectral granularity, we bin the frequency axis into B frequency bands to obtain $\mathcal{B}$. In practice, we use either five bands B_1 = `[80, 125] Hz`, B_2 = `[125, 250] Hz`, B_3 = `[250, 375] Hz`, B_4 = `[375, 687.5] Hz`, and B_5 = `[687.5, 1000] Hz`, following Kaifosh et al. (2025), or 31 linearly spaced frequency bands between `80 and 1000 Hz`. A vectorized representation of $\mathcal{B}$ is denoted as $\mathtt{vec}(\mathcal{B})$, a column vector of dimension $\mathcal{V}B$.

4.2 Audio (A)

Audio spectrograms: for a speech waveform $a(t)$ sampled at frequency f_s, we compute a mel-scaled power spectrogram using a Hann-windowed short-time Fourier transform (STFT), followed by projection onto a mel filterbank with B mel bands. Specifically, we first obtain the power spectrogram $\mathcal{M}_{F\times\tau'} = |\mathrm{STFT}(a(t))|^2$, where F denotes the number of frequency bins and τ' the number of time frames. This spectrogram is then projected onto a mel filterbank W_{mel} spanning the frequency range $[f_{\min}, f_{\max}]$, yielding

$$\mathcal{A}_{B\times\tau'}(b,t) = \sum_f W_{\mathrm{mel}}(b,f)\,\mathcal{M}_{f,t},$$

where $b \in \{1,\ldots,B\}$ indexes the mel bands. Each vector $\mathcal{A}^{(t)}_B$ encodes the mel-band power distribution of the speech signal at frame t, emphasizing perceptually relevant frequency regions. We use B = 80 mel bands, $f_{\min} = 20$ Hz, and $f_{\max} = f_s/2$. We denote the column vector of an audio spectrogram by $\mathcal{A}$ throughout the article.

Audio features from S3 models: for a speech waveform $a(t)$, we extract self-supervised speech (S3) representations by passing the signal through a pretrained model $\mathcal{S}$, yielding $\mathcal{H} = \mathcal{S}(a(t))$. The model $\mathcal{S}$ can be instantiated as WAV2VEC 2.0 (Baevski et al., 2020), HUBERT (Hsu et al., 2021), or WAVLM (Chen et al., 2022). We denote the column vector of S3 audio representations by $\mathcal{H}$ throughout the article.

4.3 Sequence-to-sequence models

We construct sequences of $\mathtt{vec}(\mathcal{E})$, $\mathtt{vec}(\mathcal{B})$, $\mathcal{A}$, and $\mathcal{H}$, emitted every 20 ms and use a context length of 25 ms. For temporal relation modeling, we employ a causal time depth separable convolutional network (TDS), as described below.

We adapt the TDS model originally designed for EMG-based keyboard typing in Sivakumar et al. (2024) with minor modifications. The model relies exclusively on local temporal context, with a 1 *s* causal receptive field. To improve robustness to spatial variability in electrode activity, the architecture incorporates a *shift-tolerant* module consisting of a linear layer followed by a ReLU activation. This module is applied to electrode channel shifts of -1, 0, and $+1$ positions, and the resulting outputs are averaged. The concatenated outputs from the shift-tolerant module are then fed into the TDS network for temporal modeling.

5 Results

5.1 $\mathcal{E}$ and $\mathbb{D}(\mathcal{E})$ encode articulatory information

Here we test whether covariance-based EMG features preserve discriminative structure related to articulation. We evaluate this on

DATA OROFACIAL GESTURES, where each trial is an orofacial movement recorded from 22 electrodes over 1.5 s. Each trial is represented by an EMG signal matrix $E \in \mathbb{R}^{22\times7500}$. We summarize each trial with a symmetric positive definite (SPD) covariance matrix $\mathcal{E} \in \mathbb{R}^{22\times22}$, and additionally consider its diagonal $\mathbb{D}(\mathcal{E}) \in \mathbb{R}^{22}$, which represents per-channel EMG power.

The vectors $\mathbb{D}(\mathcal{E})$ corresponding to different orofacial gestures naturally form distinct clusters, as shown in figure 1. We further quantify this separability using the unsupervised k-medoids clustering algorithm (Kaufman and Rousseeuw, 1990), achieving an accuracy of 61.41% using $\mathbb{D}(\mathcal{E})$ (averaged across 12 subjects). When using the full covariance matrix $\mathcal{E}$ with the geodesic distance defined in equation 1, the k-medoids accuracy increases to 73.7%; both results are well above the random-chance level of 7.69%.

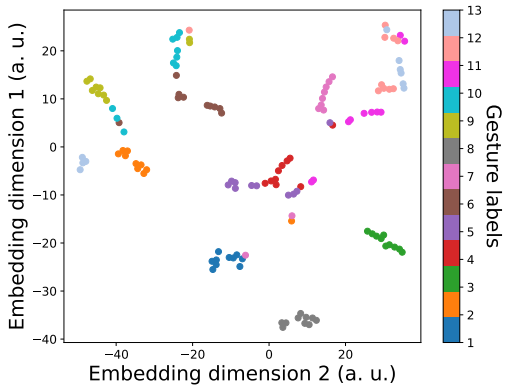

Figure 1: Different orofacial gestures are naturally separable using covariance-based EMG features. t-SNE visualization of diagonal vectors $\mathbb{D}(\mathcal{E})$ for 13 orofacial movements from a single subject. Embedding is color-coded by articulatory gesture type (*a.u.* = arbitrary units).

These results demonstrate that both $\mathcal{E}$ and $\mathbb{D}(\mathcal{E})$ encode discriminative articulatory information. While $\mathbb{D}(\mathcal{E})$ alone is sufficient to distinguish among different orofacial movements, incorporating the full covariance structure in $\mathcal{E}$ improves decoding accuracy.

Note that other widely used EMG features, such as log-spectrograms (Sivakumar et al., 2024) or rectified time-domain signals (Halliday and Farmer, 2010), are not as straightforward to probe for this type of global structure. When raw EMG signals $E \in \mathbb{R}^{22\times7500}$ are featurized using spectrograms or rectified signals, the temporal dimension may be reduced in granularity but is not collapsed into a single frame. In contrast, covariance-based representations aggregate temporal information into a single fixed-dimensional feature, yielding $\mathbb{D}(\mathcal{E}) \in \mathbb{R}^{22}$ or $\mathcal{E} \in \mathbb{R}^{22\times22}$. We analyze $\mathbb{D}(\mathcal{E})$ using Euclidean distance, while $\mathcal{E}$ is compared using the metric defined in equation 1. Consequently, commonly used time-frequency or time-domain features do not yield a directly comparable fixed-dimensional representation that captures articulatory structure in the same way.

5.2 $\mathcal{H}$ can linearly map to $\mathbb{D}(\mathcal{E})$

We test whether there exists a linear mapping defined by a weight matrix W and bias b such that $\mathbb{D}(\mathcal{E}) \approx W\mathcal{H} + b$ with a high correlation[1].

We use the training set described in section 3 DATA GENERAL to learn this mapping and evaluate it on the corresponding test set. We report the Pearson correlation between the predicted sequences $\mathbb{D}(\mathcal{E}')$ and the ground-truth $\mathbb{D}(\mathcal{E})$ on the test set. The representations $\mathcal{H}$ are extracted using HUBERT (Hsu et al., 2021), WAV2VEC 2.0 (Baevski et al., 2020), and WAVLM (Chen et al., 2022). We evaluate BASE models with latent space dimension of 768 and 12 transformer layers, LARGE models with latent space dimension of 1024 and 24 transformer layers, and FINE-TUNED (FT) models that have been tuned for automatic speech recognition (ASR).

Correlation coefficients (r) across models and layers are shown in figure 2. We find that a simple linear model can predict $\mathbb{D}(\mathcal{E})$ from $\mathcal{H}$ with a correlation as high as $r = 0.85$. The layer-wise trends across different models partially mirror the observations reported in (Cho et al., 2023, 2024) for electromagnetic articulography (EMA), where two local peaks were consistently observed across models. In our case, we observe two local peaks for WAV2VEC 2.0 models but only a single dominant peak for HUBERT and WAVLM models. A sharp decline in correlation emerges in the upper layers of fine-tuned models, reflecting the growing influence of task-specific objectives. This effect

[1]We actually aim to probe whether $\mathcal{H}$ (768-1024 dimensions) can map to $\mathrm{vec}(\mathcal{E})$ (961 dimensions). However, the resulting $\sim 10^6$-parameter linear transformation would be severely ill-posed . To make this analysis tractable, we use $\mathbb{D}(\mathcal{E})$ as a proxy because it provides a compact, well-conditioned, and physically meaningful representation grounded in articulatory mechanisms, making it well suited for linear probing. Importantly, this substitution is justified because both $\mathcal{E}$ and $\mathbb{D}(\mathcal{E})$ encode structured articulatory information, and the latter serves as a low-dimensional surrogate for the former, as shown in section 5.1.

is especially pronounced for WAV2VEC 2.0 compared to HUBERT and WAVLM.

Notably, for the HUBERT-BASE model, the peak correlation at layer 6 aligns with the layer previously identified as optimal for discrete speech resynthesis and spoken language modeling (Lakhotia et al., 2021). While prior work established this empirical result, the mechanistic basis for this peak remained unclear. Our analysis provides a principled interpretation: layer 6 exhibits the strongest linear predictive power for $\mathbb{D}(\mathcal{E})$, which encodes structured and discriminative articulatory information (i.e., different articulatory gestures such as tongue and jaw positions naturally form separable clusters). This tight alignment between articulatory structure and model representations offers a direct explanation for why layer 6 is particularly effective for downstream speech resynthesis and language modeling. In short, the layer that best captures articulatory mechanisms is also the one that yields the strongest downstream performance, providing convergent evidence for its functional role.

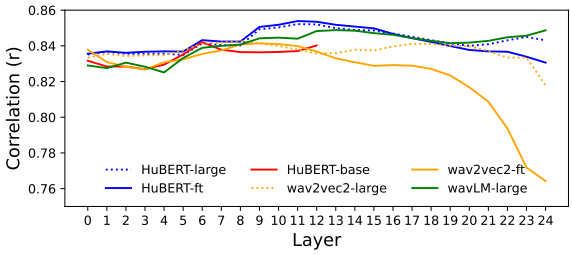

Figure 2: Layer-wise correlation (r) between $\mathbb{D}(\mathcal{E}')$ and $\mathbb{D}(\mathcal{E})$ across different self-supervised speech models. A simple linear mapping is used to predict $\mathbb{D}(\mathcal{E}')$ from $\mathcal{H}$.

We also examine whether a similar linear mapping exists between EMG spectrogram features ($\mathrm{vec}(\mathcal{B})$) and $\mathcal{H}$. Frequency bands of $\mathcal{B}$ are obtained using five frequency bins, as described in section 4. However, the resulting correlation coefficients are substantially lower, with a maximum correlation of approximately $r = 0.57$ (figure 3). For comparison, we also compute correlations for linear mapping between $\mathcal{A}$ and $\mathbb{D}(\mathcal{E})$ ($r = 0.61$), which is considerably lower than the correlation between $\mathcal{H}$ and $\mathbb{D}(\mathcal{E})$.

The above observations indicate that among the different EMG feature representations considered, $\mathbb{D}(\mathcal{E})$ exhibits the strongest linear alignment with the self-supervised speech feature space $\mathcal{H}$. This strong correspondence suggests that $\mathbb{D}(\mathcal{E})$ (consequently, $\mathrm{vec}(\mathcal{E})$) and $\mathcal{H}$ encode highly compatible representations, making them particularly well suited for EMG-to-audio learning. In contrast, EMG spectrogram features ($\mathcal{B}$) and their alignment with ($\mathcal{H}$) yield notably weaker correlations.

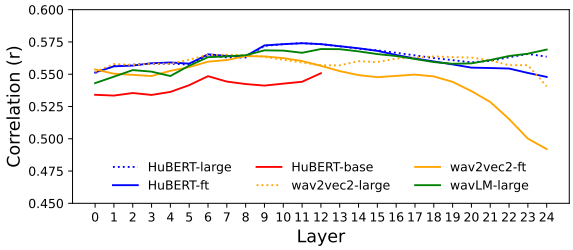

Figure 3: Layer-wise correlation (r) between $\mathcal{B}'$ and $\mathcal{B}$ across different self-supervised speech models. A simple linear mapping is used to predict $\mathcal{B}'$ from $\mathcal{H}$.

5.3 emg2speech synthesis

As shown earlier, the following relationship holds:

$$\mathcal{H} \xrightarrow{\texttt{linear mapping}} \mathbb{D}(\mathcal{E}) \xrightarrow{\texttt{gesture-specific clustering}} \text{OROFACIAL MOVEMENTS.}$$

The existence of a simple linear mapping from $\mathcal{H}$ to $\mathbb{D}(\mathcal{E})$ is informative because it indicates that the self-supervised representations $\mathcal{H}$ encode articulatory structure consistent with underlying muscle activations. This forward mapping is well posed: $\mathcal{H}$ has moderate dimensionality (768–1024), whereas $\mathbb{D}(\mathcal{E})$ is low dimensional (31), allowing the mapping to be estimated stably using linear regression.

By contrast, the inverse direction $\mathbb{D}(\mathcal{E}) \rightarrow \mathcal{H}$ is intrinsically underdetermined and not uniquely invertible in the linear setting, since multiple high-dimensional speech representations can correspond to the same low-dimensional articulatory configuration. This challenge is further compounded when temporal alignment between EMG and audio is unknown. Nevertheless, the existence of a reliable forward mapping suggests that recovering $\mathcal{H}$ from EMG is a feasible learning problem when using an appropriate nonlinear mapping [2].

Motivated by this observation, we consider alignment-free prediction of $\mathcal{H}$-derived representations from EMG features ($\mathrm{vec}(\mathcal{E})$, $\mathbb{D}(\mathcal{E})$, or $\mathrm{vec}(\mathcal{B})$). Because the inverse linear mapping is ill posed, we model it using a nonlinear sequence-to-sequence architecture that can capture tempo-

[2]For linear probing, we use low-dimensional versions of both covariance and spectrogram representations: $\mathbb{D}(\mathcal{E})$ and a 5-bin spectrogram $\mathcal{B}$. For speech synthesis, we instead use full-resolution features ($\mathrm{vec}(\mathcal{E})$ and a 31-bin $\mathcal{B}$) to preserve fine-grained cross-channel and spectral structure.

ral and contextual dependencies present in $\mathcal{H}$. Concretely, EMG features are provided as input to a TDS convolutional network (section 4.3) that predicts discrete units associated with $\mathcal{H}$. We use the 100-unit discretization from layer 6 of HuBERT-BASE (Lakhotia et al., 2021), denoted $\texttt{dis}(\mathcal{H})_{\text{HuBERT}}$. Training is performed with the connectionist temporal classification (CTC) loss (Graves et al., 2006), which enables learning without explicit frame-level alignment between EMG sequences and $\texttt{dis}(\mathcal{H})_{\text{HuBERT}}$. Finally, the predicted $\texttt{dis}(\mathcal{H})_{\text{HuBERT}}$ sequence is converted to audio using a pretrained Tacotron vocoder (Wang et al., 2017). The overall architecture is illustrated in figure 4.

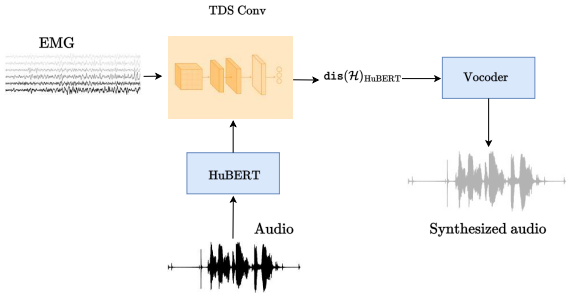

Figure 4: Multivariate EMG signals are converted into $\texttt{vec}(\mathcal{E})$, $\mathbb{D}(\mathcal{E})$, or $\mathcal{B}$, and then passed through a TDS CONV block to predict $\texttt{dis}(\mathcal{H})_{\text{HuBERT}}$, which are subsequently fed into a vocoder to synthesize audio. Frozen neural network components are shown in blue, and trainable components are shown in orange.

5.3.1 Results for DATA GENERAL

We use Google text-to-speech (gTTS) to synthesize audio from the corresponding text transcripts. From this synthesized audio, we extract the discrete HuBERT units $\texttt{dis}(\mathcal{H})_{\text{HuBERT}}$[3]. We report $\texttt{dis}(\mathcal{H})_{\text{HuBERT}}$ decoding results in table 1.

We provide $\texttt{vec}(\mathcal{E})$, $\mathbb{D}(\mathcal{E})$, or $\texttt{vec}(\mathcal{B})$ as input to the TDS network, and train it to predict the corresponding $\texttt{dis}(\mathcal{H})_{\text{HuBERT}}$ unit sequence using the CTC loss. For example, for the sentence $_{\text{T-START}}$ < IT WAS PAID FOR >$_{\text{T-END}}$ with target HuBERT units 71-12-71-12-4-12-4-40-93-86-13-58-32-1-99-..., the model learns a mapping from the EMG feature sequence to the target unit sequence. During inference, the model outputs a probability distribution over all 100 $\texttt{dis}(\mathcal{H})_{\text{HuBERT}}$ units at each time step, and we decode the most likely unit sequence using *greedy search*. For instance, the decoded sequence may be 71-12-57-4-54-40-93-86-13-58-16-14-76-6-36-.... We compute the unit error rate (UER) as the Levenshtein distance between the target and predicted $\texttt{dis}(\mathcal{H})_{\text{HuBERT}}$ sequences, normalized by the target sequence length (see appendix D for ablation studies on training data size).

[3]For both DATA $_{\text{GENERAL}}$ and DATA $_{\text{ALS}}$, we do not use subject-recorded audio for EMG-to-speech synthesis. The healthy participant vocalized the sentences during recording, whereas the ALS participant articulated them silently. In both cases, we rely only on transcript-based gTTS audio to derive $\texttt{dis}(\mathcal{H})_{\text{HuBERT}}$. Subject-recorded audio from DATA $_{\text{GENERAL}}$ is used only for probing linearity in the previous section. This design is motivated by clinical scenarios in which parallel EMG and audio recordings may be unavailable. Note that the gTTS audio is not temporally aligned with the EMG, which makes the translation problem more challenging than settings with paired, time-synchronized supervision.

Table 1: Unit error rate (UER) for different EMG feature representations when predicting $\texttt{dis}(\mathcal{H})_{\text{HuBERT}}$ units on DATA $_{\text{GENERAL}}$. The dataset and preprocessing details are described in section 3. Lower UER is better. Values are averaged over 5 random seeds.

MODEL INPUT	UER (% ↓)	INPUT FEATURE DIMENSION
$\texttt{vec}(\mathcal{B})$	62.71 ± 0.50	961
$\mathbb{D}(\mathcal{E})$	62.16 ± 0.50	31
$\texttt{vec}(\mathcal{E})$	**56.08 ± 0.91**	961

We also present the results of phoneme-level decoding in table 2. For the sentence $_{\text{T-START}}$ <IT WAS PAID FOR>$_{\text{T-END}}$ with the corresponding phonemic transcription IH-T $_{\text{SPACE}}$ W-AA-Z $_{\text{SPACE}}$ P-EY-D $_{\text{SPACE}}$ F-AO-R, the TDS model is trained to learn the mapping from $\texttt{vec}(\mathcal{E})$, $\mathbb{D}(\mathcal{E})$, or $\texttt{vec}(\mathcal{B})$ to phoneme sequences using the CTC loss. During inference, the model outputs probabilities for all 40 English phonemes at each time step, and the predictions are decoded using *greedy search*. For example, the decoded output might be IH-T $_{\text{SPACE}}$ W-AA-Z $_{\text{SPACE}}$ P-EY-T $_{\text{SPACE}}$ F-AO-R. We compute the phoneme error rate (PER) as the Levenshtein distance between the target and decoded phoneme sequences, normalized by the length of the target sequence.

Table 2: Phoneme error rate (PER) for different EMG feature representations when predicting phonemes on DATA $_{\text{GENERAL}}$. The dataset and preprocessing details are described in section 3. Lower PER is better. Values are averaged over 5 random seeds.

MODEL INPUT	PER (% ↓)	INPUT FEATURE DIMENSION
$\texttt{vec}(\mathcal{B})$	44.40 ± 2.28	961
$\mathbb{D}(\mathcal{E})$	44.40 ± 1.51	31
$\texttt{vec}(\mathcal{E})$	**32.78 ± 0.66**	961

As shown in tables 1 and 2, $\texttt{vec}(\mathcal{E})$ outperforms $\texttt{vec}(\mathcal{B})$. $\mathcal{B}$ was computed using 31 linearly spaced frequency bins, and for any given time frame, both $\texttt{vec}(\mathcal{E})$ and $\texttt{vec}(\mathcal{B})$ have 961 dimensions. Notably, even $\mathbb{D}(\mathcal{E})$, which has only 31 dimensions (i.e., $31\times$ fewer dimensions than $\texttt{vec}(\mathcal{B})$), performs on par with $\texttt{vec}(\mathcal{B})$. This finding is consistent with the linear mapping results shown in figures 2 and 3.

PHONEME GUIDED DECODING OF $\texttt{dis}(\mathcal{H})_{\text{HUBERT}}$: as shown in table 2, phoneme sequences can be decoded more accurately than $\texttt{dis}(\mathcal{H})_{\text{HUBERT}}$ unit sequences. Motivated by this observation, we train the TDS convolutional encoder in figure 4 with two prediction heads: one for phonemes, producing framewise posteriors $P(\text{PHONEME} \mid \text{EMG})$, and one for $\texttt{dis}(\mathcal{H})_{\text{HUBERT}}$ units, producing framewise posteriors $P(\texttt{dis}(\mathcal{H})_{\text{HUBERT}} \mid \text{EMG})$. The model is optimized with CTC losses for both outputs, together with an additional consistency loss that encourages phoneme-consistent $\texttt{dis}(\mathcal{H})_{\text{HUBERT}}$ predictions. Specifically, we use a precomputed lookup table $P(\text{PHONEME} \mid u)$, where $u \in \mathcal{U}$ and $\mathcal{U} = \texttt{dis}(\mathcal{H})_{\text{HUBERT}}$ denotes the discrete HuBERT unit set, to transform the unit posterior at each frame t into a phoneme distribution by marginalizing over units:

$$\tilde{P}(\text{PHONEME} \mid \text{EMG})_t = \sum_{u \in \mathcal{U}} P(\text{PHONEME} \mid u)\, P(u \mid \text{EMG})_t.$$

We then minimize a cross-entropy loss between $\tilde{P}(\text{PHONEME} \mid \text{EMG})_t$ and the phoneme-head posterior $P(\text{PHONEME} \mid \text{EMG})_t$ (see appendix B for more details). At inference time, we decode only $\texttt{dis}(\mathcal{H})_{\text{HUBERT}}$ units from the unit head using greedy decoding. Resulting improvement is shown in table 3.

Table 3: Effect of phoneme-guided training on HUBERT unit decoding with $\texttt{vec}(\mathcal{E})$ as input on DATA GENERAL. Values are averaged over 5 random seeds. The improvement is statistically significant ($p < 10^{-6}$).

TRAINING OBJECTIVE	UER (% ↓)
Unit CTC only	56.08 ± 0.91
Phoneme-guided decoding	**51.81 ± 0.62**

Furthermore, three human raters (see appendix D.3) listened to all 400 synthesized audio samples in the test set (generated from $\texttt{dis}(\mathcal{H})_{\text{HUBERT}}$ units obtained via phoneme-guided decoding with $\texttt{vec}(\mathcal{E})$ as input) and transcribed them. We compute the word error rate (WER) as the Levenshtein distance between each rater's transcription and the ground-truth transcript, normalized by the length of the ground-truth transcript. We report the resulting WERs in table 5.

5.3.2 Results for DATA ALS

We follow the same preprocessing, model architecture, and training procedure used for DATA GENERAL. We report unit decoding performance in table 4. To assess end-to-end intelligibility, we synthesize speech from the decoded units and measure word error rate (WER) using human transcriptions (on all 60 synthesized audios in the test set); the resulting WER is reported in table 5.

Table 4: Unit error rate (UER) for different EMG feature representations when predicting $\texttt{dis}(\mathcal{H})_{\text{HUBERT}}$ units on DATA ALS. Dataset and preprocessing details are described in section 3. Lower UER is better. Values are averaged over 5 random seeds.

MODEL INPUT	UER (% ↓)
$\texttt{vec}(\mathcal{B})$	59.18 ± 2.81
$\mathbb{D}(\mathcal{E})$	54.27 ± 0.56
$\texttt{vec}(\mathcal{E})$	52.29 ± 0.91
$\texttt{vec}(\mathcal{E})$ with phoneme-guided decoding	**46.98 ± 1.11**

Table 5: WER as rated by human transcribers.

HUMAN TRANSCRIBER	DATA GENERAL WER (% ↓)	DATA ALS WER (% ↓)
1	65.09	52.69
2	59.32	50.97
3	63.23	48.81
Average	**62.55 ± 2.40**	**50.82 ± 1.59**

We contextualize these WERs relative to prior brain-computer interface studies in appendix C. Please see the following links for demonstrations: Audio. We also provide the ground-truth transcripts and test-set transcripts from three human raters: Transcriptions.

6 Conclusion

We present methods and datasets that convert orofacial EMG signals directly into speech, and we demonstrate the system with both a healthy participant and a participant diagnosed with ALS.

7 Limitations

This work primarily focuses on the technical aspects of EMG-to-speech modeling, including characterizing the structure of orofacial EMG signals, quantifying their relationship to self-supervised speech representations, and designing encoder architectures grounded in articulatory mechanisms. Our clinical demonstration uses approximately one hour of data from a single participant with ALS. As a result, we do not yet characterize how performance evolves under day-to-day distribution shifts in EMG signals (e.g., changes in electrode placement, skin impedance, fatigue, or disease progression). Consequently, we also do not evaluate whether this non-stationarity is more or less challenging than the distribution shifts observed in invasive neural interfaces (Wairagkar et al., 2025; Willett et al., 2023).

In addition, we do not demonstrate sustained, long-term performance of this non-invasive neuroprosthesis. In contrast, prior work on invasive neuroprostheses has reported stability over extended periods in related decoding settings, including brain-to-text (Fan et al., 2023) and cursor-based brain-computer interfaces (Wilson et al., 2025).

Finally, we do not explore whether large-scale pretrained EMG-to-speech models can improve decoding performance. For example, in related work on EMG-based keyboard typing (EMG2QWERTY) (Sivakumar et al., 2024), pretraining on data from 100 individuals improved accuracy after fine-tuning to new individuals, although zero-shot performance remained limited. Speech may be more challenging than discrete key typing, and future work should investigate how to build and effectively leverage large-scale pretrained models for EMG-to-speech translation.

We are actively addressing these limitations through ongoing longitudinal studies and by expanding data collection to build larger EMG-to-speech corpora from individuals with diverse clinical etiologies, including ALS and laryngectomy.

8 Ethical considerations

Research was conducted in accordance with the principles embodied in the Declaration of Helsinki and in accordance with the University of California, Davis Institutional Review Board (IRB) Administration protocol 2078695-1. All participants provided written informed consent. All participants also provided consent for publication of deidentified data. Volunteers of any gender and from all racial and ethnic groups were eligible to participate. Participants were required to be at least 18 years old, able to understand spoken and written English, and able to follow task instructions. Participants had no skin conditions or wounds at electrode placement sites and were excluded if they had uncorrected vision problems. Children, individuals unable to provide informed consent, and prisoners were not included in the experiments. All participants were compensated in accordance with IRB protocols.

The participant with ALS was first diagnosed in 2019 and has non-familial ALS with spasticity.

Acknowledgments

This work was supported by awards to Lee M. Miller from: Accenture, through the Accenture Labs Digital Experiences group; CITRIS and the Banatao Institute at the University of California; the University of California Davis School of Medicine (Cultivating Team Science Award); the University of California Davis Academic Senate; a UC Davis Science Translation and Innovative Research (STAIR) Grant; and the Child Family Fund for the Center for Mind and Brain.

Harshavardhana T. Gowda is supported by Neuralstorm Fellowship, NSF NRT Award No. 2152260 and Ellis Fund administered by the University of California, Davis.

We thank Karen Dhillon and Craig M. McDonald's Neuromuscular Research Lab at UC Davis for valuable guidance and participant recruitment.

Conflict of interest

H. T. Gowda and L. M. Miller are inventors on intellectual property related to *silent* speech owned by the Regents of University of California, not presently licensed.

Author contributions

- Harshavardhana T. Gowda: Conceptualization, mathematical formulation, method development, data analysis, experimental design, data collection software development, data collection, and manuscript preparation.
- Daniel C. Comstock: Data collection and manuscript review.
- Lee M. Miller: Conceptualization, funding,

participant recruitment, and manuscript review.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Tyler Benster, Guy Wilson, Reshef Elisha, Francis R Willett, and Shaul Druckmann. 2024. A cross-modal approach to silent speech with llm-enhanced recognition. *arXiv preprint arXiv:2403.05583*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Cheol Jun Cho, Abdelrahman Mohamed, Alan W Black, and Gopala K Anumanchipalli. 2024. Self-supervised models of speech infer universal articulatory kinematics. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12061–12065. IEEE.

Cheol Jun Cho, Peter Wu, Abdelrahman Mohamed, and Gopala K Anumanchipalli. 2023. Evidence of vocal tract articulation in self-supervised learning of speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Lorenz Diener, Gerrit Felsch, Miguel Angrick, and Tanja Schultz. 2018. Session-independent array-based emg-to-speech conversion using convolutional neural networks. In *Speech Communication; 13th ITG-Symposium*, pages 1–5.

Chaofei Fan, Nick Hahn, Foram Kamdar, Donald Avansino, Guy Wilson, Leigh Hochberg, Krishna V Shenoy, Jaimie Henderson, and Francis Willett. 2023. Plug-and-play stability for intracortical brain-computer interfaces: a one-year demonstration of seamless brain-to-text communication. *Advances in neural information processing systems*, 36:42258–42270.

David Gaddy and Dan Klein. 2020. Digital voicing of silent speech. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5521–5530.

David Gaddy and Dan Klein. 2021. 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)*, pages 175–181.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

D. M. Halliday and S. F. Farmer. 2010. On the need for rectification of surface emg. *Journal of Neurophysiology*, 103(6):3547.

Awni Hannun, Ann Lee, Qiantong Xu, and Ronan Collobert. 2019. Sequence-to-sequence speech recognition with time-depth separable convolutions.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Matthias Janke and Lorenz Diener. 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.

Szu-Chen Jou, Tanja Schultz, Matthias Walliczek, Florian Kraft, and Alex Waibel. 2006. Towards continuous speech recognition using surface electromyography. In *Ninth International Conference on Spoken Language Processing*.

Patrick Kaifosh, Thomas R. Reardon, and CTRL labs at Reality Labs. 2025. A generic non-invasive neuromotor interface for human-computer interaction. *Nature*, 645:702–711.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Arnav Kapur, Utkarsh Sarawgi, Eric Wadkins, Matthew Wu, Nora Hollenstein, and Pattie Maes. 2020. Non-invasive silent speech recognition in multiple sclerosis with dysphonia. In *Machine Learning for Health Workshop*, pages 25–38. PMLR.

Leonard Kaufman and Peter J. Rousseeuw. 1990. Partitioning around medoids (program pam). In *Wiley Series in Probability and Statistics*, pages 68–125. John Wiley & Sons, Inc., Hoboken, NJ, USA. Retrieved 2021-06-13.

Minsu Kim, Joanna Hong, and Yong Man Ro. 2021. Lip to speech synthesis with visual context attentional gan. *Advances in Neural Information Processing Systems*, 34:2758–2770.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Zhenhua Lin. 2019. Riemannian geometry of symmetric positive definite matrices via cholesky decomposition. *SIAM Journal on Matrix Analysis and Applications*, 40(4):1353–1370.

Kaylo T Littlejohn, Cheol Jun Cho, Jessie R Liu, Alexander B Silva, Bohan Yu, Vanessa R Anderson, Cady M Kurtz-Miott, Samantha Brosler, Anshul P Kashyap, Irina P Hallinan, and 1 others. 2025. A streaming brain-to-voice neuroprosthesis to restore naturalistic communication. *Nature neuroscience*, pages 1–11.

Geoffrey S Meltzner, James T Heaton, Yunbin Deng, Gianluca De Luca, Serge H Roy, and Joshua C Kline. 2018. Development of semg sensors and algorithms for silent speech recognition. *Journal of neural engineering*, 15(4):046031.

Sean L Metzger, Kaylo T Littlejohn, Alexander B Silva, David A Moses, Margaret P Seaton, Ran Wang, Maximilian E Dougherty, Jessie R Liu, Peter Wu, Michael A Berger, and 1 others. 2023. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976):1037–1046.

M. A. Mines, B. F. Hanson, and J. E. Shoup. 1978. Frequency of occurrence of phonemes in conversational english. *Language and Speech*, 21(3):221–241.

Mehryar Mohri, Fernando C. N. Pereira, and Michael Riley. 2008. Speech recognition with weighted finite-state transducers. In *Handbook on Speech Processing and Speech Communication, Part E: Speech recognition*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2020. Learning individual speaking styles for accurate lip to speech synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13796–13805.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Viswanath Sivakumar, Jeffrey Seely, Alan Du, Sean R Bittner, Adam Berenzweig, Anuoluwapo Bolarinwa, Alexandre Gramfort, and Michael I Mandel. 2024. emg2qwerty: A large dataset with baselines for touch typing using surface electromyography. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Arthur R. Toth, Michael Wand, and Tanja Schultz. 2009. Synthesizing speech from electromyography using voice transformation techniques. In *Interspeech 2009*, pages 652–655.

Maitreyee Wairagkar, Nicholas S Card, Tyler Singer-Clark, Xianda Hou, Carrina Iacobacci, Lee M Miller, Leigh R Hochberg, David M Brandman, and Sergey D Stavisky. 2025. An instantaneous voice-synthesis neuroprosthesis. *Nature*, pages 1–8.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards end-to-end speech synthesis.

Francis R Willett, Erin M Kunz, Chaofei Fan, Donald T Avansino, Guy H Wilson, Eun Young Choi, Foram Kamdar, Matthew F Glasser, Leigh R Hochberg, Shaul Druckmann, and 1 others. 2023. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036.

Guy H Wilson, Elias A Stein, Foram Kamdar, Donald T Avansino, Tsam Kiu Pun, Ronnie Gross, Tommy Hosman, Tyler Singer-Clark, Anastasia Kapitonava, Leigh R Hochberg, and 1 others. 2025. Long-term unsupervised recalibration of cursor-based intracortical brain–computer interfaces using a hidden markov model. *Nature Biomedical Engineering*, pages 1–19.

Maria K Wolters, Karl B Isaac, and Steve Renals. 2010. Evaluating speech synthesis intelligibility using amazon mechanical turk. In *Proc. 7th Speech Synthesis Workshop (SSW7)*, pages 136–141.

A Experimental details

We collect EMG signals from 31 sites on the neck, chin, jaw, cheek, and lips using monopolar electrodes. An ACTICHAMP PLUS amplifier and associated active electrodes from BRAINVISION (Brain Vision) are used to record EMG signals at 5000 Hz. To ensure proper contact between the skin surface and electrodes, we use SUPERVISC, a high-viscosity electrolyte gel from EASYCAP (Easycap). We develop a software suite in a PYTHON environment to provide visual cues to participants and to collate and store timestamped data. For time synchronization, we use Lab Streaming Layer (LSL). See figure 5 for electrode placement. In addition to the 31 data electrodes, we also use a GROUND electrode (marked as GND) and a REFERENCE electrode (marked as 32). The GROUND electrode is placed

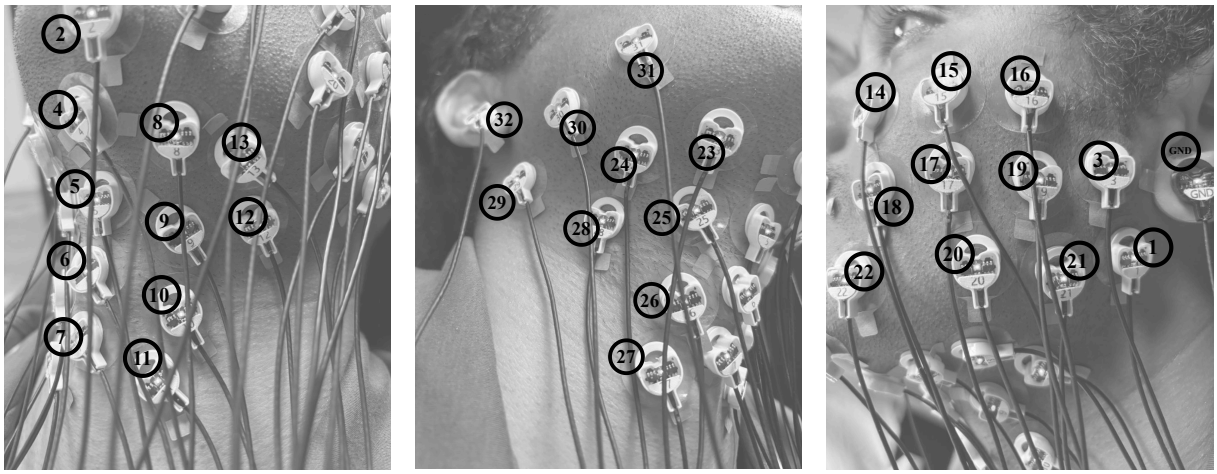

Figure 5: LEFT: electrode placement on the left side of the neck. MIDDLE: electrode placement on the right side of the neck. RIGHT: electrode placement on the left cheek.

on the left earlobe and the REFERENCE electrode is placed on the right earlobe.

Before signal acquisition, participants are briefed on the experimental protocol and seated comfortably in a chair. Sentence start and end times are timestamped using mouse clicks. When a participant is ready to articulate a sentence, they click the mouse to prompt the sentence to appear on the screen. Once articulation is complete, they click again to indicate the end, which causes the sentence to disappear. This allows participants to articulate at their own pace.

The data collection environment is carefully controlled to reduce AC electrical interference. EMG signals undergo minimal preprocessing. The signal from the REFERENCE channel (electrode 32) is subtracted from all other channels. The resulting signals are bandpass filtered using a third-order Butterworth filter between 80 and 1000 Hz and segmented according to sentence start and end times based on synchronized timestamps. The segmented sentences are subsequently z-normalized along the time dimension for each channel.

The electrodes are positioned over regions that overlay muscle groups involved in speech articulation, providing coverage of key articulators such as the tongue, jaw, lips, and larynx. Electrode locations 19, 21, 3, and 1 approximately overlie the *hyoglossus*, *palatoglossus*, and *styloglossus* muscles. These muscles are located in the lower cheek region and play a vital role in tongue movement. They are also consistently recruited across a wide range of articulatory gestures. Muscles in the upper and posterior cheek regions include the *masseter* and *temporalis*, which control jaw motion, and the *zygomaticus*, which is involved in upper lip elevation. These muscles correspond approximately to electrode regions around nodes 22, 18, 17, and 15 in figure 5. Electrodes located beneath the jaw capture activity from muscles involved in tongue protrusion and jaw-tongue coordination, such as the *genioglossus* near electrodes 8, 9, 23, and 25, as well as the *digastric*. Additionally, electrodes near the laryngeal region, including nodes 6, 7, 10, 11, 26, and 27, reflect activity from muscles that modulate laryngeal and hyoid position, such as the *sternohyoid*, *stylohyoid*, and *digastric*. These muscles contribute to pitch control, vowel shaping, and jaw movement.

B Detailed explanation: phoneme guided decoding of $\texttt{dis}(\mathcal{H})_{\text{HUBERT}}$

We train a bidirectional gated recurrent unit (GRU) model to map $\texttt{dis}(\mathcal{H})_{\text{HUBERT}}$ unit sequences to phoneme sequences using a CTC objective on DATA GENERAL (train-val-test split as described in section 3). After CTC decoding, the model achieves $\text{PER} = 0\%$ (phoneme error rate) on the corresponding test split. Using this trained model, we estimate a unit-to-phoneme conditional table by aggregating framewise posteriors: for each HuBERT unit u, we collect the predicted phoneme distribution $P_{\text{GRU}}(\text{PHONEME}_t \mid u)$ at every frame t where $u_t = u$, and average across all such frames, i.e.,

$$P(\text{PHONEME} \mid u) = \frac{1}{N_u} \sum_{t:\, u_t = u} P_{\text{GRU}}(\text{PHONEME}_t = \text{PHONEME} \mid u)\,,$$

where $N_u = |\{t : u_t = u\}|$ denotes the total number of frames in the dataset for which the HuBERT unit equals u. We remove the CTC blank symbol from $P(\cdot \mid u)$ and renormalize. This yields $P(\text{PHONEME} \mid \texttt{dis}(\mathcal{H})_{\text{HuBERT}})$, which we use as a fixed probabilistic mapping in our consistency regularization.

For phoneme-guided decoding of $\texttt{dis}(\mathcal{H})_{\text{HuBERT}}$, we train the TDS convolutional encoder in figure 4 with two heads: one for phonemes, producing framewise posteriors $P(\text{PHONEME} \mid \text{EMG})$, and one for $\texttt{dis}(\mathcal{H})_{\text{HuBERT}}$ units, producing framewise posteriors $P(u \mid \text{EMG})$, where $u \in \mathcal{U}$ and $\mathcal{U} = \texttt{dis}(\mathcal{H})_{\text{HuBERT}}$ denotes the discrete HuBERT unit set. The model is optimized with CTC losses for both outputs. Additionally, we impose a consistency loss by using the precomputed lookup table $P(\text{PHONEME} \mid u)$ to transform the unit posterior at each frame t into a phoneme distribution via marginalization over units:

$$\tilde{P}(\text{PHONEME} \mid \text{EMG})_t = \sum_{u \in \mathcal{U}} P(\text{PHONEME} \mid u) \cdot P(u \mid \text{EMG})_t.$$

We then minimize a cross-entropy loss between $\tilde{P}(\text{PHONEME} \mid \text{EMG})_t$ and the phoneme-head posterior $P(\text{PHONEME} \mid \text{EMG})_t$:

$$\mathcal{L}_{\text{cons}} = -\sum_t \sum_{\text{PHONEME}} P(\text{PHONEME} \mid \text{EMG})_t \cdot \log \tilde{P}(\text{PHONEME} \mid \text{EMG})_t.$$

The total training objective is a weighted sum of the two CTC losses and the proposed consistency term:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{unit}} \mathcal{L}_{\text{CTC}}^{\text{unit}} + \lambda_{\text{phone}} \mathcal{L}_{\text{CTC}}^{\text{phone}} + \lambda_{\text{cons}} \mathcal{L}_{\text{cons}}.$$

In our experiments, we use $\lambda_{\text{unit}} = 0.8$, $\lambda_{\text{phone}} = 0.1$, and $\lambda_{\text{cons}} = 0.1$.

We further probe the structure of $P(\text{PHONEME} \mid \texttt{dis}(\mathcal{H})_{\text{HuBERT}})$. In figure 6, we visualize, for each HuBERT unit, the phoneme with the highest conditional probability (i.e., $\texttt{argmax}_{\text{PHONEME}} P(\text{PHONEME} \mid u)$). Multiple HuBERT units may map to the same most-probable phoneme p. In such cases, for each phoneme p, we order the corresponding units by increasing entropy of $P(\cdot \mid u)$; lower entropy indicates a sharper, more confident association between the unit and phoneme p. We find that alveolar consonants (e.g., `T`, `S`, `N`) and `A` and `I` based vowels (e.g., `AA`, `AY`, `AH`, `IY`, `IH`) have many HuBERT units mapping to them.

This is consistent with prior analyses showing that these phones are among the most frequently occurring in conversational English and that their phonetic realizations have different manifestations depending on coarticulatory context (Mines et al., 1978). These observations suggest that our lookup dictionary $P(\text{PHONEME} \mid \texttt{dis}(\mathcal{H})_{\text{HuBERT}})$ captures meaningful phonetic structure and is grounded in known articulatory regularities of speech.

C Detailed literature review

Here, we review prior work on speech neural and neuromuscular interfaces and contextualize our results relative to state-of-the-art methods. A substantial body of research (Jou et al., 2006; Kapur et al., 2020; Meltzner et al., 2018; Toth et al., 2009; Janke and Diener, 2017; Diener et al., 2018; Littlejohn et al., 2025) has laid the groundwork for EMG-based speech interfaces. Among the earliest studies, Jou et al. (2006) demonstrate EMG-to-speech conversion on a small corpus of 50 sentences. Kapur et al. (2020) use a corpus of 15 sentences and, rather than performing phoneme-level decoding, formulate the task as a 15-way classification problem. Meltzner et al. (2018) study EMG-to-text recognition for isolated words, phrases drawn from a ∼200-word vocabulary, and continuous sentences using a custom grammar-based recognition model over a set of 1200 scripted phrases. Toth et al. (2009) present EMG-to-speech conversion on a corpus of 500 sentences. Janke and Diener (2017) demonstrate EMG-to-speech conversion using up to two hours of data and 2000 utterances.

Overall, these studies rely on private datasets and task-specific pipelines, and they typically evaluate on small, constrained corpora. In addition, the works do not release full implementations (e.g., code repositories) or sufficient methodological details to enable direct reproducibility. As a result, it is difficult to directly compare performance across systems, and all the above results do not establish generalization to open-vocabulary English settings.

A reproducible benchmark for open-vocabulary EMG-to-speech conversion was introduced by Gaddy and Klein (2020, 2021). However, these works rely on time-aligned EMG-audio pairs for

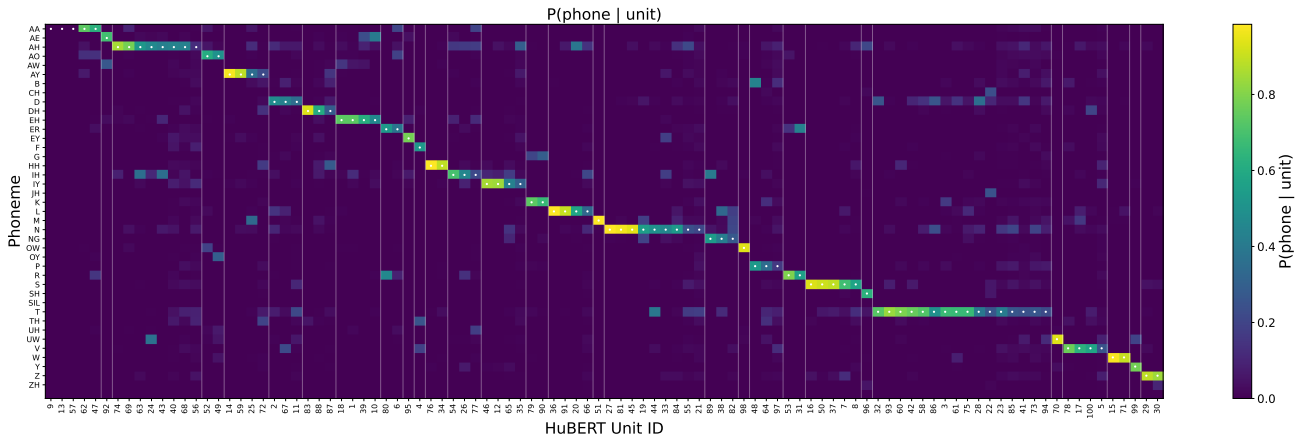

Figure 6: For each HuBERT unit, the phoneme with the highest conditional probability (i.e., $\texttt{argmax}_{\text{PHONEME}} P(\text{PHONEME} \mid u)$) is shown. Multiple HuBERT units may map to the same most-probable phoneme p. In such cases, for each phoneme p, we order the corresponding units from left to right by increasing entropy of $P(\cdot \mid u)$.

training. Building on Gaddy and Klein (2021), Benster et al. (2024) propose an approach that leverages an audio-only corpus in addition to paired EMG-audio data. While effective in the benchmark setting, such methods can be difficult to deploy in clinical scenarios where parallel EMG-audio recordings may be unavailable or unreliable. On the large-vocabulary corpus, Gaddy and Klein (2020) report a word error rate (WER) of 68%, and Gaddy and Klein (2021) reduce this to 42%. Benster et al. (2024) report WER $< 10\%$ on the Gaddy and Klein (2020) dataset by using a large language model (LLM) to post-correct the intermediate EMG-to-phoneme output. However, their system no longer supports streaming synthesis, and the evaluation does not fully characterize potential information leakage through the LLM (e.g., memorization or exposure to overlapping text distributions). Consequently, their results are not directly comparable to strictly streaming EMG-to-speech systems. Littlejohn et al. (2025) report a WER of 74% on the Gaddy and Klein (2020) dataset using a CNN+RNN transducer model; however, their train-test splits and implementation details are not publicly available, which prevents direct comparison. In our setting, we address a harder learning problem by not assuming time-aligned EMG-audio pairs during training, and we report a WER of 62% on an open-vocabulary corpus. We emphasize that these WER values should not be compared one-to-one across studies, since the data collection setup, training targets and alignment assumptions, problem formulation, and evaluation methodology differ substantially. We report these results to provide context relative to prior EMG-based speech interfaces.

To address these limitations, we build on widely available pretrained speech models and vocoders, but adapt them to the EMG setting through principled, articulatorily motivated design choices. We characterize the structure of orofacial EMG signals and introduce methods that are straightforward to implement yet carefully tailored for this problem, yielding substantial gains. For example, we represent EMG signals using covariance matrices and propose phoneme-guided decoding of speech units (section 5.3.1), which exploits phonetic structure to improve the fidelity of synthesized speech. Our encoder design and phoneme-guided decoding are grounded in articulatory mechanisms and established regularities of speech production.

The WERs reported in this work fall broadly within the range observed for invasive speech neural interfaces. For example, Wairagkar et al. (2025) report a median WER of 43% using 256 intracortical electrodes, trained on approximately 8300 cued sentences spanning about 38 hours of data; the large number of hours for a comparable number of sentences reflects the substantially slower speaking rate in that study relative to ours. Similarly, Metzger et al. (2023) report a WER of 54% using 253 electrodes with a 1024-word vocabulary. In our setting, we achieve a WER of 62% on a large-vocabulary corpus. These error rates should be interpreted in the context of neural speech interfaces, where the sensing modality, signal-to-noise ratio,

data size, and evaluation setup differ substantially from conventional audio ASR (automatic speech recognition) benchmarks.

D Additional technical details

Here, we provide a detailed description of the model architectures, training procedures, participant instructions, and ablation studies used throughout this work, with the goal of making our methodology clear, reproducible, and easy to interpret.

D.1 Effect of training data size

We study how training-set size affects the unit error rate (UER) when decoding $\texttt{dis}(\mathcal{H})_{\text{HuBERT}}$. Using DATA GENERAL, we train the model with between 2000 and 8000 sentences, while keeping the validation and test splits fixed as in section 3. Over this range, we observe an approximately linear improvement in UER with increasing training data. This differs from the power-law behavior commonly reported for large-scale from-scratch training (Kaplan et al., 2020). Although we cannot draw strong conclusions from a single-participant study, these results suggest that, when leveraging frozen pretrained speech representations, performance may remain data-limited and continue to benefit predictably from additional EMG training data, yielding lower UER (and consequently WER) as the dataset size scales.

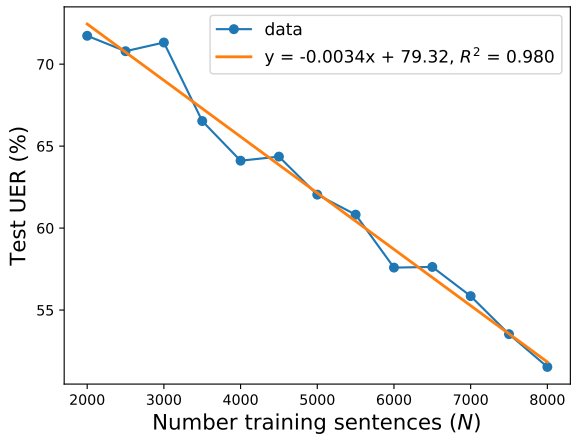

Figure 7: Test UER as a function of the number of training sentences. Over the evaluated range (2000-8000 sentences), UER decreases approximately linearly with increasing training data, indicating consistent gains from additional EMG data. UER is computed by decoding $\texttt{dis}(\mathcal{H})_{\text{HuBERT}}$ from a model trained with $\texttt{vec}(\mathcal{E})$ as input; during training, we use phoneme-guided decoding.

D.2 Losses

As shown in figure 8, all loss terms decrease smoothly during the early stages of training, indicating stable optimization. Validation losses begin to increase after approximately 25 epochs, suggesting the onset of overfitting. The consistency loss $\mathcal{L}_{\text{cons}}$ and the phoneme-level CTC loss $\mathcal{L}_{\text{CTC}}^{\text{phone}}$ decrease more rapidly than the unit-level loss $\mathcal{L}_{\text{CTC}}^{\text{unit}}$, consistent with their role as auxiliary objectives that regularize training and encourage better alignment for unit prediction. Together with figure 7, which shows that UER decreases approximately linearly with training set size over the evaluated range and has not yet saturated, and the overfitting observed in figure 8, these results suggest that the model remains data-limited and can continue to benefit from additional training data.

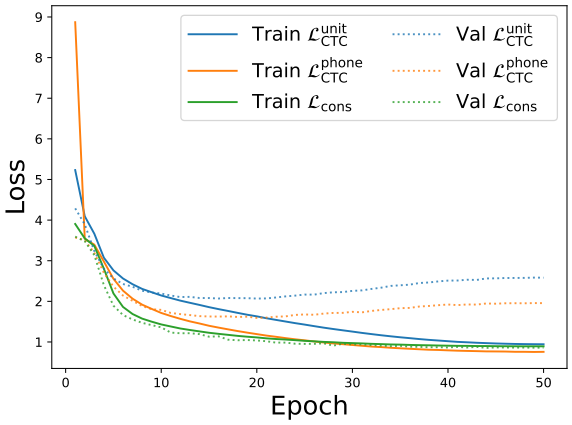

Figure 8: Training and validation losses versus epoch. The causal TDS convolutional model in section 4.3 is trained for 50 epochs with $\texttt{vec}(\mathcal{E})$ as input, using $\mathcal{L}_{\text{CTC}}^{\text{unit}}$, $\mathcal{L}_{\text{CTC}}^{\text{phone}}$, and $\mathcal{L}_{\text{cons}}$.

D.3 Transcriptions

Human evaluation of synthesized speech is commonly used to assess the intelligibility and overall quality of generated audio (Wolters et al., 2010). For both DATA GENERAL and DATA ALS, we ask human raters to listen to the synthesized audio and transcribe each utterance in English. Raters are not restricted to a predefined vocabulary and may write any English words for both corpora (even though DATA ALS contains only about 300 unique words, raters are not informed of this constraint). In this sense, our evaluation targets open-vocabulary recognition and is less constrained than evaluations such as Metzger et al. (2023) and Littlejohn et al. (2025), which use fixed vocabularies of 1,024 words.

In figure 9, we summarize the distribution of WER and PER across three human transcribers for all evaluated sentences (460 sentences in total across DATA GENERAL and DATA ALS). The raters exhibit similar central tendency and spread, indicating strong agreement. To compute PER, we phonemize each rater's transcription and compare it against the ground-truth phonemized sentence.

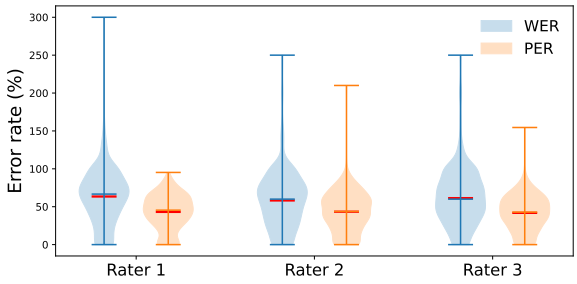

Figure 9: Distributions of WER and PER across three human transcribers for 460 sentences. Means are shown in red.

Across raters, the mean PER is 42.79%, which is lower than the mean WER of 61.02%. This suggests that, even when transcribed words are incorrect, the errors are often phonetically plausible.

In addition, we evaluate perceptual quality using mean opinion score (MOS). Specifically, we randomly sample 25 sentences from the 400-sentence test set of DATA GENERAL and 10 sentences from the 60-sentence test set of DATA ALS, yielding 35 sentences in total. We then ask 10 native English speakers to rate the quality of the synthesized audio on a 1-to-5 MOS scale, resulting in 350 individual ratings. We restrict MOS evaluation to this subset because three human listeners had already listened to and transcribed all 460 test sentences for the WER/PER evaluation shown in figure 9, and collecting MOS ratings for the full test set from a larger pool of raters was not practically feasible.

Figure 10 shows the distribution of MOS ratings across the 10 raters. The mean MOS across all ratings is 3.79. This value should be interpreted in the context of our generation setting. In conventional text-to-speech or voice conversion evaluations, the input typically contains clean linguistic content, and MOS primarily reflects the naturalness and perceptual quality of the generated speech. In our setting, by contrast, speech is synthesized directly from EMG, where the underlying linguistic representation is itself noisy and may already be corrupted prior to waveform generation. In addition, we use a pretrained vocoder (Lakhotia et al., 2021) that was trained on natural speech rather than on linguistically corrupted intermediate representations. Consequently, the synthesized outputs can remain acoustically speech-like and perceptually plausible even when the recovered linguistic content is inaccurate. This helps explain, at least in part, why MOS remains moderately high despite high WER: in our setting, MOS reflects the perceptual plausibility of the synthesized waveform, whereas WER reflects the fidelity of the recovered linguistic content.

We open-source all 460 synthesized sentences from the test set, together with the ground-truth audio for all 9,660 sentences from the healthy participant. Ground-truth audio for the participant with ALS is unavailable because she articulated the sentences *silently*. These audio samples are available on the PROJECT PAGE. To provide a reference point for future work, we transcribed the healthy participant's ground-truth audio using WHISPER (Radford et al., 2023), obtaining a WER of 14.25% on the training-validation split and 12.66% on the test set. These values should be treated as baseline references for future improvements.

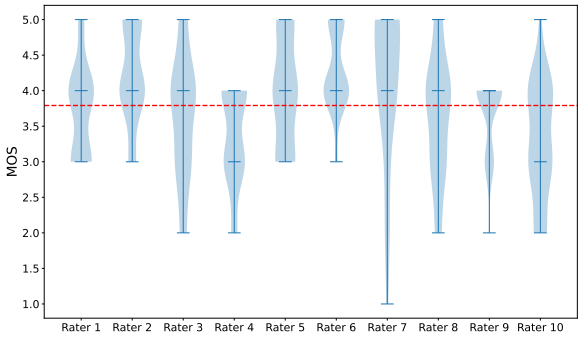

Figure 10: Distribution of mean opinion score (MOS) ratings across 10 raters for 35 synthesized sentences. Mean MOS across raters is shown in red.

D.4 emg2text

Furthermore, we train the TDS convolutional model to decode phonemes on all 460 sentences (following the procedure in section 4.3) and then map the predicted phoneme sequences to words using a weighted finite-state transducer (WFST) decoder[4].

For phoneme-to-word decoding, we use the LibriSpeech-100 transcripts (Panayotov et al.,

[4]We use the WFST decoding implementation provided by ICEFALL (github.com/k2-fsa/icefall).

2015), which contain roughly 38000 sentences and 35000 unique words. From these transcripts, we build a pronunciation lexicon WFST, L, that maps phoneme sequences to words. We also train a 4-gram language model with KenLM (Heafield, 2011) and convert it into a grammar WFST, G. Finally, we construct the CTC topology WFST, H, which encodes the allowable label sequences under the CTC criterion.

We compose these components into a decoding graph (Mohri et al., 2008), $HLG = H \circ L \circ G$, which integrates the CTC constraints (H), lexicon mapping (L), and language model probabilities (G). At inference time, we perform beam search over HLG with beam width 50 and compute WER as the normalized Levenshtein distance between the reference and decoded word sequences. This procedure yields a WER of 51.17% and a PER of 38.19%. The language model is trained only on LibriSpeech-100 transcripts; sentences from our train-validation-test splits are not included. We summarize this approach, denoted EMG2TEXT with direct EMG2SPEECH in table 6.

Table 6: PER and WER for EMG2TEXT and EMG2SPEECH.

TRAINING METHOD	PER (% ↓)	WER (% ↓)
EMG2SPEECH	42.79	61.02
EMG2TEXT	38.19	51.17

Overall, EMG2TEXT achieves lower PER and WER than direct EMG2SPEECH. However, direct EMG-to-speech generation remains important for neural prostheses because it can enable a more fluid, natural interaction (e.g., without requiring an explicit intermediate text interface). We therefore view improving direct EMG-to-speech as an important direction for future work.

D.5 Model architecture and implementation details

Input representation. Let $\mathbf{x} \in \mathbb{R}^{T \times N \times \texttt{dim}}$ denote the input, where N is the batch size, T is the number of time steps, and $\texttt{dim}$ denotes the dimensionality of the input EMG representation, namely $\texttt{vec}(\mathcal{E})$, $\mathbb{D}(\mathcal{E})$, or $\texttt{vec}(\mathcal{B})$.

Channel normalization. Before the multi-layer perceptron (MLP), we apply channel-wise batch normalization to the input. For $\texttt{vec}(\mathcal{E})$ and $\texttt{vec}(\mathcal{B})$, we reshape $\mathbf{x}$ to $\mathbb{R}^{T \times N \times \mathcal{V} \times \mathcal{V}}$ and $\mathbb{R}^{T \times N \times \mathcal{V} \times B}$, respectively, and apply 2D batch normalization with $\mathcal{V}$ as the channel dimension. For $\mathbb{D}(\mathcal{E})$, we apply 1D batch normalization directly to $\mathbf{x} \in \mathbb{R}^{T \times N \times \mathcal{V}}$. In all cases, normalization statistics are computed independently for each channel $\mathcal{V}$ across the batch and temporal dimensions, and additionally across the final feature dimension when present. In our case, $\mathcal{V} = 31$ and $B = 31$.

Spatially robust MLP. After channel normalization, we apply an MLP frontend designed to improve robustness to spatial variability across electrodes. Specifically, we construct an ensemble of views by circularly shifting the input along the channel dimension with offsets $o \in \mathcal{O}$, where $\mathcal{O} = \{-1, 0, 1\}$. For matrix-valued inputs, i.e., $\texttt{vec}(\mathcal{E})$ or $\texttt{vec}(\mathcal{B})$, of shape $\mathbb{R}^{T \times N \times \mathcal{V} \times \mathcal{F}}$, where $\mathcal{F} = \mathcal{V}$ for $\texttt{vec}(\mathcal{E})$ and $\mathcal{F} = B$ for $\texttt{vec}(\mathcal{B})$, each shifted view is flattened across the last two dimensions and passed through a shared MLP. For $\mathbb{D}(\mathcal{E}) \in \mathbb{R}^{T \times N \times \mathcal{V}}$, we analogously shift along the channel dimension and apply the same shared MLP directly. The resulting embeddings are aggregated across shifts using mean pooling:

$$\mathbf{h} = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} \mathrm{MLP}(\phi(\mathrm{shift}(\mathbf{x}, o))),$$
$$\mathcal{O} = \{-1, 0, 1\},$$

where $\phi(\cdot)$ denotes flattening for matrix-valued inputs and the identity map for $\mathbb{D}(\mathcal{E})$. This yields $\mathbf{h} \in \mathbb{R}^{T \times N \times H}$, where H is the output dimensionality of the final MLP layer. In our case, $H = 384$.

TDS convolutional encoder. The output of the spatially robust MLP, denoted by $\mathbf{h} \in \mathbb{R}^{T \times N \times H}$, is passed to a temporal encoder based on the time-depth separable (TDS) architecture of Hannun et al. (2019). For a given TDS block, let $\mathbf{u} \in \mathbb{R}^{T \times N \times H}$ denote the block input. We reshape $\mathbf{u}$ to $\mathbb{R}^{N \times K \times W \times T}$, where $H = KW$. A causal convolution with kernel tensor $\Theta \in \mathbb{R}^{K \times K \times 1 \times k}$ and kernel size $1 \times k$ is then applied along the temporal dimension, with $k = 14$ and replicate padding of length $k - 1$ on the left:

$$\tilde{z}_{n,k,w,t} = \sum_{i=0}^{k-1} \sum_{k'=1}^{K} \Theta_{k,k',0,i}\, \tilde{u}_{n,k',w,t-i},$$
$$\Theta \in \mathbb{R}^{K \times K \times 1 \times k}.$$

The convolution output is passed through a ReLU nonlinearity, reshaped back to $\mathbb{R}^{T \times N \times H}$, added to the block input through a residual connection, and

normalized with LayerNorm:

$$\mathbf{z}_{\text{conv}} = \text{LayerNorm}(\text{ReLU}(\text{Conv}(\mathbf{u})) + \mathbf{u}),$$
$$\mathbf{z}_{\text{conv}} \in \mathbb{R}^{T \times N \times H}.$$

The subsequent fully connected block applies two linear layers with a ReLU nonlinearity in between, followed by a residual connection and LayerNorm:

$$\mathbf{z}_{\text{fc}} = \text{LayerNorm}(\text{FC}_2(\text{ReLU}(\text{FC}_1(\mathbf{z}_{\text{conv}}))) + \mathbf{z}_{\text{conv}}),$$
$$\mathbf{z}_{\text{fc}} \in \mathbb{R}^{T \times N \times H}.$$

We stack four such TDS blocks with channel configuration [24, 24, 24, 24]. Since the MLP output dimensionality is fixed to $H = 384$, each block uses width $W = 384/24 = 16$. All convolutions are causal.

Dual output heads. Following the encoder, a shared bottleneck maps H-dimensional representations to a 512-dimensional space. Two independent linear heads then produce log-softmax outputs: a unit head `Linear`(512, 101) over 100 HUBERT units plus a CTC blank symbol (index 100), and a phone head `Linear`(512, 41) over 40 phonemes plus a CTC blank symbol (index 40).

Optimization. We use AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.98$, weight decay = 10^{-4}, learning rate $= 3 \times 10^{-4}$) with a linear warm-up over 5 epochs, starting from $0.1\times$ the base learning rate, followed by cosine annealing to a minimum of 10^{-6} over the remaining 45 epochs. Models are trained for 50 epochs with per-epoch temporal jitter resampling of the training set.