Leveraging deep state-space models for silent speech decoding

G. Wilson[1*†], R. Elisha[1†], T. Benster[1†], Y. Lee[1†], K. Shenoy[1‡], J. Henderson[1‡], Z. Bao[1‡], S. Druckmann[1‡]
[1]Stanford University, Stanford, USA; [†]co-lead, [‡]co-senior
* 318 Campus Drive West, W100-A Clark Center, Stanford, CA 94305. E-mail: ghwilson@stanford.edu

Introduction: Silent speech interfaces (SSIs) are a promising approach for people with communication disorders. Noninvasive studies have focused heavily on surface electromyography (sEMG) for obtaining electrical signals from orofacial muscles but often rely on handcrafted input features and smaller machine learning pipelines [1, 2]. In this work, we use grids of flexible, hydrogel-based sEMG electrodes to obtain high signal-to-noise ratio (SNR) recordings during silent speech production, and a recent class of machine learning models developed for long time-series inference [3] to decode these recordings.

Methods: We applied grids of hydrogel-based sEMG sensors to a volunteer's face as they silently uttered a dataset of fifty words in an instructed delay task. Simultaneous activity from 48 sEMG channels across the arrays were recorded at 30 KHz. Signals were digitally filtered offline using a 5th order Butterworth 1 Hz highpass filter and downsampled to 1 KHz.

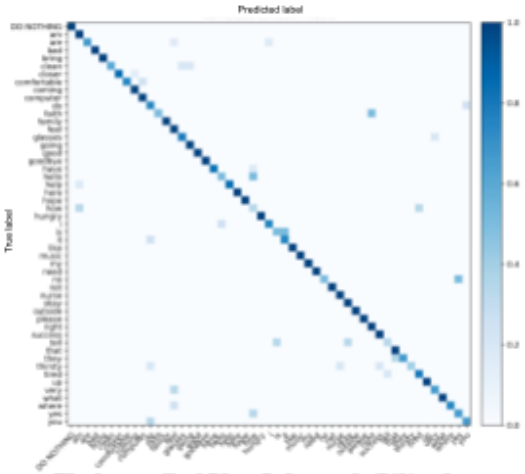

Fig. 1: normalized S4 confusion matrix (84% acc)

Results: Using a ConvMLP model, we obtained a 33% absolute improvement in fifty-way classification over a logistic regression with manually-derived features in. Furthermore, a deep learning model specifically designed to incorporate long timescales (‘S4’) improved accuracy by an additional 8%, with an overall accuracy of 84%. We demonstrated increasing accuracy with increasing channel count and dataset size.

Discussion: S4’s performance highlights the potential for deep state-space models in SSIs and, more broadly, extracting nonlinear features from raw EMG activity. Hyperparameter sweeps indicate that higher performance is likely achievable with increased channel counts and training data.

Significance: Our results indicate that deep state-space models are well-suited for sEMG-based silent speech decoding. Increasing performance with higher channel count and more data suggests room for future improvements by leveraging higher-density arrays and continuous speech corpuses.

References:

[1] Wang, Y. et al. “All-weather, natural silent speech recognition via machine-learning-assisted tattoo-like electronics”. Flexible Electronics, (2021).
[2] Meltzner, G. S. et al. “Development of sEMG sensors and algorithms for silent speech recognition.” *J. Neural Eng*. (2018).
[3] Gu, A.. et al. “Efficiently Modeling Long Sequences with Structured State Spaces.” *arxiv* (2021).