TOWARDS BCI SPEECH DECODING AS MACHINE TRANSLATION (MT)

KAMIL A. GRAJSKI

28 FEB 2025

ABSTRACT

Speech decoding in recent Human BCI studies reflect elements of speech recognition frameworks. That is, neural activity is processed to estimate time-series of phonemes and text generation search space is managed with NLP.

Here, it is hypothesized that machine translation (MT) may provide a more powerful set of signal processing and language modeling methods. A data-driven unsupervised method (VQ-VAE) may be applied to “learn” a discrete representation of the spatio-temporal electrode array neural data.

Here, we download open-source data from the Willett, et al. (2023) *Nature* study. We demonstrate that VQ-VAE readily converges on low-dimensional discrete representations of high-dimensional electrode array data.

The next steps are to leverage these learned discrete representations to fine-tune an LLM as well as to explore video processing-inspired VQ-VAE.

INTRODUCTION

- Willet, et al. (2023) published a study of a speech neuroprosthesis comprised of electrode arrays implanted in human BA6. Speech decoding methods featured prescriptive use of phonemes, RNN networks, and NLP to manage the phoneme-to-text search space.
- Here, it is hypothesized that a machine translation (MT) framework may be effective. The input "language" is a discrete representation of neural activity. Output (text) is by a fine-tuned LLM.
- A purely data-driven unsupervised method (VQ-VAE) is applied to an *open source* data set from Willet, et al. to obtain a discrete representation of neural activity.
- We show convergence of VQ-VAE in an initial case: the electrode array time slices treated as independent 2D "image" samples. This provides a baseline for more complex 3D VQVAE and ViT.

METHODS

Data set: 13 sessions; 200-300 trials per session; 100-500 20 msec frames of pre-processed electrode array per trial. ~1.2M samples.

Electrode array comprised of 4 sets of 8x8 electrodes. Here, only the Area 6 (ventral) 8x8 array is used.

Here, the ETL, Training, Testing, Validation, and Visualization was coded from scratch in Python and PyTorch on AWS (GPU) with VS Code and GitHub Co-Pilot. VQ-VAE Class from DeepMind (modified).

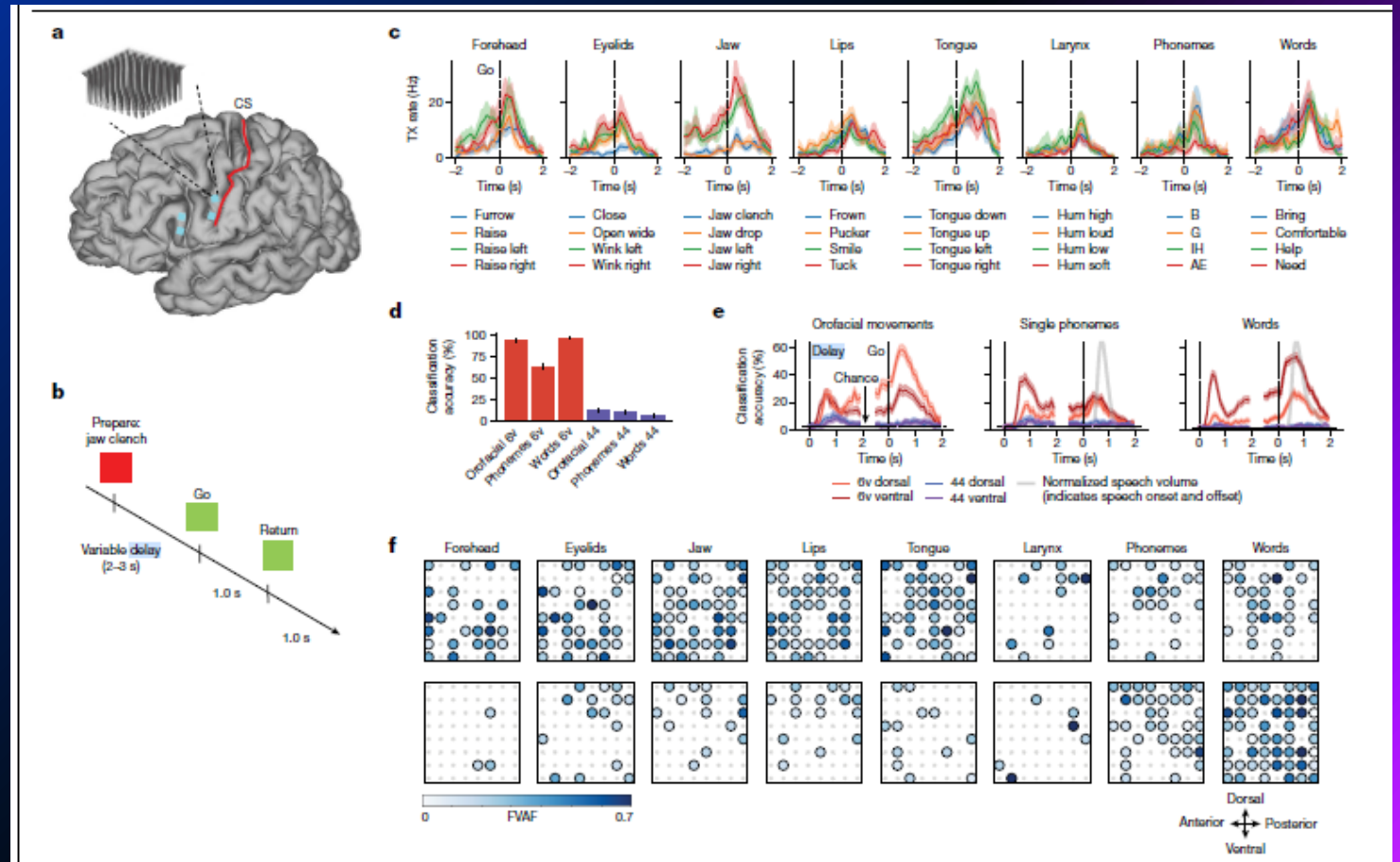


Fig. 1 | Neural representation of orofacial movement and attempted speech. **a**, Microelectrode array locations (cyan squares) are shown on top of MRI-derived brain anatomy (CS, central sulcus). **b**, Neural tuning to orofacial movements, phonemes and words was evaluated in an instructed delay task. **c**, Example responses of an electrode in area 6v that was tuned to a variety of speech articulator motions, phonemes and words. Each line shows the mean threshold crossing (TX) rate across all trials of a single condition ($n = 20$ trials for orofacial movements and words, $n = 16$ for phonemes). Shaded regions show 95% confidence intervals (CIs). Neural activity was denoised by convolving with a Gaussian smoothing kernel (80 ms s.d.). **d**, Bar heights denote the classification accuracy of a naive Bayes decoder applied to 1 s of neural population activity from area 6v (red bars) or area 44 (purple bars) across all movement conditions (33 orofacial movements, 39 phonemes, 50 words). Black lines denote 95% CIs. **e**, Red and blue lines represent classification accuracy across time for each of the four arrays and three types of movement. Classification was performed with a 100 ms window of neural population activity for each time point. Shaded regions show 95% CIs. Grey lines denote normalized speech volume for phonemes and words (indicating speech onset and offset). **f**, Tuning heatmaps for both arrays in area 6v, for each movement category. Circles are drawn if binned firing rates on that electrode were significantly different across the given set of conditions ($P < 1 \times 10^{-5}$ assessed with one-way analysis of variance; bin width, 800 ms). Shading indicates the fraction of variance accounted for (FVAF) by across-condition differences in mean firing rate.

From Willett, et al. (2023)

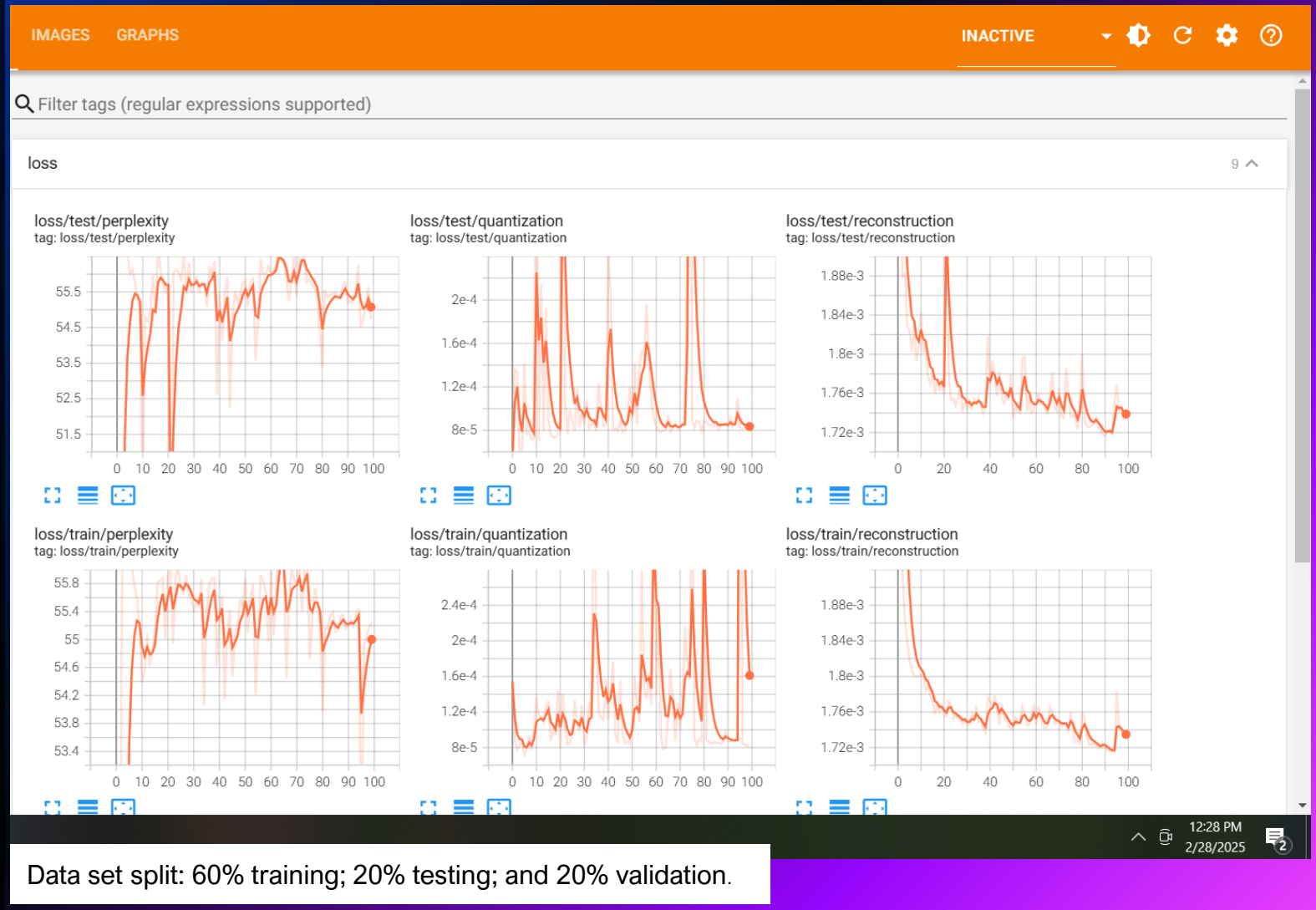
RESULTS
TRAIN & TEST

Tensor Board results (*right*).

VQ-VAE Perplexity, Quantization, and Reconstruction Loss Values vs Training Epoch. (*left to right*)

Training Set (*bottom*) and independent Testing Set (*top*). Independent Validation Set (*not shown*).

Results indicate convergence. Hyperparameters to explore include VQ-VAE complexity, learning parameters, etc.



Data set split: 60% training; 20% testing; and 20% validation.

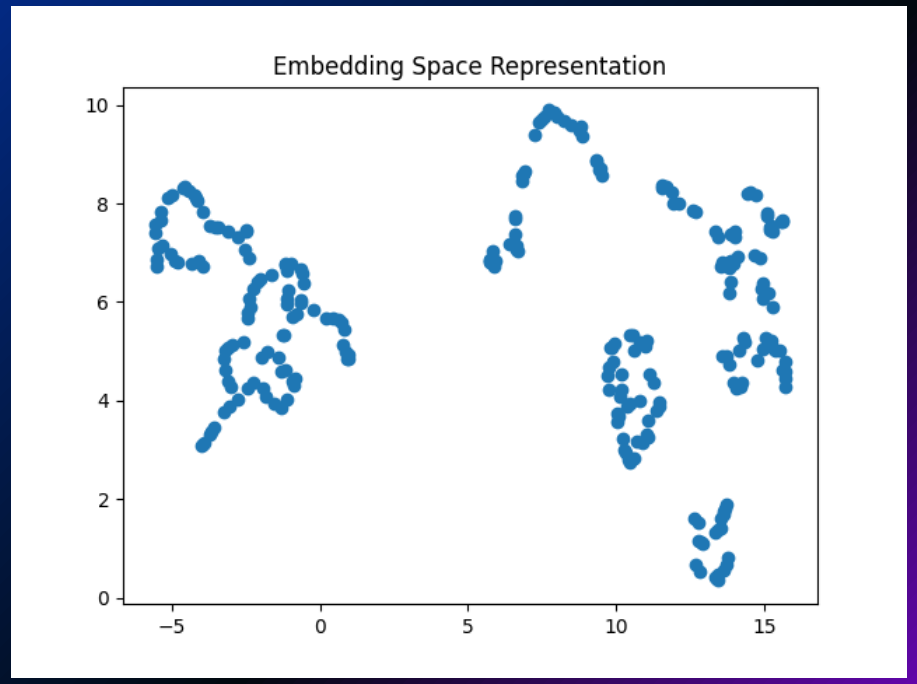
RESULTS EMBEDDING (VQ)

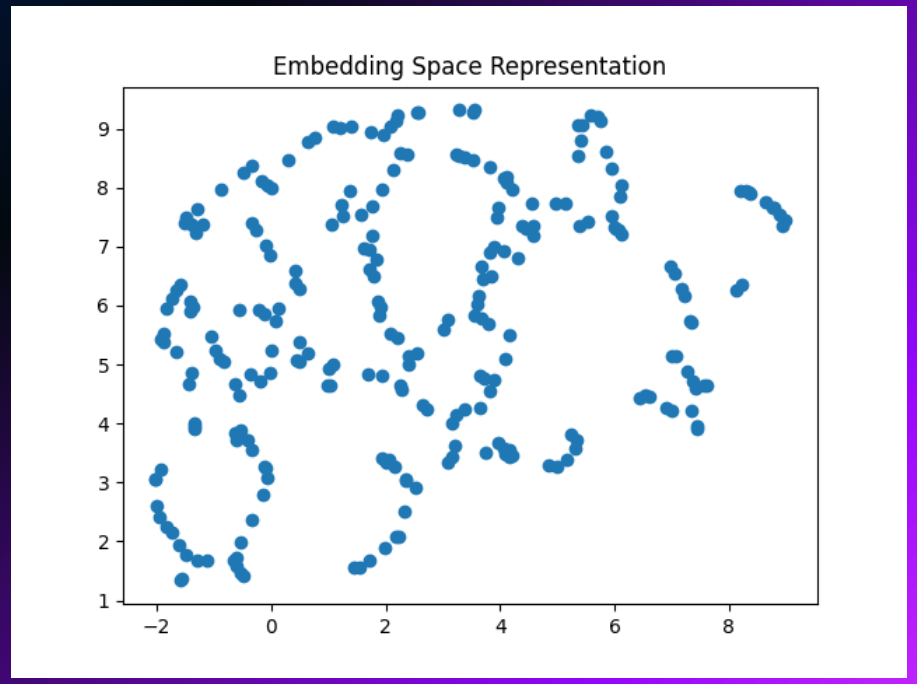
Embedding space projected to 2D (*right*). Each point corresponds to a vector in the embedding space (VQ “codebook”).

Before training (*bottom*), the “points” are spread out.

Following training (*top*), the “points” distribute non-uniformly.

Results indicate convergence. Hyperparameters to explore include embedding space dimension, number of embedding vectors, etc.





DISCUSSION

- It is hypothesized that MT offers a useful framework for processing neural speech data. As such, it is necessary to represent the input “language” as a time series of discrete symbols. It is preferred to adopt a data-driven unsupervised approach such as VQ-VAE and its variants.
- Here, we’ve shown in a simplistic way – treating each electrode array time frame as an independent image – that the unsupervised approach does indeed converge towards a discrete representation. But this is just the first step.
- Towards demonstrating MT, the next step is to fine-tune an LLM on such inputs.
- Further steps include multi-time step processing of the electrode array data in the time range 100-1000 msec with methods such as 3D Convolution, 3D VQ-VAE, and Visual Transformers (ViT).
- Finally, it is hypothesized that these techniques can be readily adapted to ECoG data.

REFERENCES

- Dosovitskiy, et al., (2021). An image is worth 16x16 words: transformers for image recognition at scale. https://arxiv.org/abs/2010.11929.
- Metzger, et al. (2023). A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*. 620:1037-1046.
  - [Note: I have applied for and am access to the raw data is pending approvals.]
- Van den Oord, et al. (2018). Neural discrete representation learning. https://arxiv.org/abs/1711.00937v2.
- Walker, et al. (2021). Predicting video with VQVAE. https://arxiv.org/abs/2103.01950
- Willett, et al. (2023). A high-performance neural speech prosthesis. *Nature*. 620:1031-1036.
- https://datadryad.org/dataset/doi:10.5061/dryad.x69p8czpq

GLOSSARY

1. ECoG – Electro-corticogram
2. MT – Machine Translation
3. ViT – Vision Transformer
4. VQ – Vector Quantize
5. VQ-VAE – Vector Quantized Variational Encoder

THANK YOU

Kamil A Grajski, PhD

408-966-8204

kgrajski@nurosci.com

www.nurosci.com