Mind reading: Predicting transformer activations to create a syllable-based speech neuroprosthesis

Henri GUILLAUME
henri.guillaume@protonmail.com
henri.guillaume@polytechnique.edu

Foreword

This blog post describes the continuation of my work at the Institut de l'Audition. It is not an internship report in the sense that the work described here took place after the end of my contract, however, it is a continuation of my internship's work.

add a cool picture

1 - Introduction

The last decade has seen great improvements in the decoding of neural speech, mainly thanks to numerous advances in machine learning technologies. Syllables make a natural target for this task, as they seem to constitute a base unit of language during speech production [1], [2]. However, the sheer number of unique syllables in most languages makes this approach intractable for machine learning applications [3]. To circumvent this label diversity problem, I suggest a new approach by predicting the continuous state-space embeddings of an audio deep learning model, and apply it to the Brain to text '25 challenge [4].

2 - Dataset

The Brain to text '25 challenge "consists of mapping a variable-length time series of neural activity to text with no ground-truth alignment between the two". Participants are given neural activation recorded while the patient was speaking a sentence, as well as the transcription for that sentence. We therefore have no knowledge of when each word occurs, and at which speed they were spoken. [4]

3 - Approach

It has been shown that deep neural network layers can be predicted from neural activations [5]. We can try to predict these representations, which will in turn be used to generate speech predictions.

3.1 - The target network: Sylber, a syllable-based speech model

Sylber [6] is a transformer-based model obtained by training a HuBERT [7] model via a novel SSL approach called self-segmentation distillation, which forces the model to learn a representation of speech that is quasi-constant within syllables. This gives birth to a salient syllable structure with a coherent state space, clustering phonetically similar syllables closely together. The model allows for reconstruction of the audio via Speech Articulatory Coding (SPARC) [8]. As the dataset only contains textual sentence labels, we artificially obtain the Sylber embeddings by running the model on artificially generated speech using PiperTTS [9].

3.2 - Decoding method: layer prediction and ensembling

We can go beyond simply predicting the output of the model. As both deep neural networks and humans process auditory information hierarchically [10], we can predict activations in shallow layers as well. We can make sense of these by forwarding them through the rest of the transformer stacks.

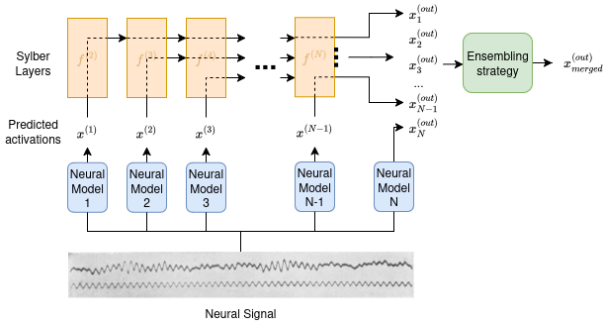

Figure 2: Layer prediction and merging

3.3 - Solving the alignment problem: Optimal Temporal Transport Regression

The nature of the challenge involves not only the neural translation, but also learning an alignment between the target features and the original utterance (which we have no knowledge of). This is usually done using CTC loss [11], [12]. However, since our target are continuous and not categorical this method for learning alignment is not applicable.

3.3.1 - SOTD

Sequences Optimal Transport Distance (SOTD) provides a pseudo-metric over the set of all d-dimensional vector sequences of length at most N [13]. The idea is that the cost of aligning two points of sequences $\{x\}_n$, $\{y\}_m$ depends not only on their temporal position, but also on their similarity via a cost function C. Thus SOTD finds a mass function α over the input sequence that minimizes the transport cost to a distribution β (that we will set as uniform) over the target sequence.

3.3.2 - OTTR

Though SOTD was suggested as a way to outperform CTC on sequence labeling tasks, we can take note of the fact that the definition of SOTD does not require a categorical loss, and can therefore be applied to our continuous state-space alignment problem. We therefore adapt the paper's "OTTC" loss [13] to *optimal temporal regression* (OTTR). We define the loss as:

$$\mathcal{L}_{\text{OTTR}} = -\sum_{i,j=1}^{n,m} \gamma_n^{m,\beta_m}(\alpha[\{x\}_n, W])_{i,j} \cdot \|x_i - y_j\|_2^2$$

Where all notations are the same as the ones in the original paper [13]. The hope is that the network W will learn to put a lot of weight on important features that should be "hard to move".

4 - Results

Bibliography

[1] A. Giraud and D. Poeppel, "Cortical oscillations and speech processing: emerging computational principles and operations," *Nature Neuroscience*, vol. 15, no. 4, pp. 511–517, 2012, doi: 10.1038/nn.3063.

[2] J. Cholin, W. J. M. Levelt, and N. O. Schiller, "Effects of syllable frequency in speech production," *Cognition*, vol. 99, pp. 205–235, 2006, doi: 10.1016/j.cognition.2005.01.009.

[3] Y. M. Oh, "Linguistic Complexity and Information: Quantitative Approaches," 2015. [Online]. Available: http://www.ddl.cnrs.fr/fulltext/Yoonmi/Oh_2015_1.pdf

[4] N. Card *et al.*, "An Accurate and Rapidly Calibrating Speech Neuroprosthesis," *New England Journal of Medicine*, vol. 391, no. 7, pp. 609–618, 2024, doi: 10.1056/nejmoa2314132.

[5] L. Evanson *et al.*, "Emergence of Language in the Developing Brain," *Unpublished Manuscript*, 2025.

[6] C. J. Cho *et al.*, "Sylber: Syllabic Embedding Representation of Speech from Raw Audio," *arXiv preprint arXiv:2410.07168*, 2024, [Online]. Available: https://arxiv.org/abs/2410.07168

[7] 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," *arXiv preprint arXiv:2106.07447*, 2021, [Online]. Available: https://arxiv.org/abs/2106.07447

[8] C. J. Cho, P. Wu, T. S. Prabhune, D. Agarwal, and G. K. Anumanchipalli, "Articulatory Encodec: Vocal Tract Kinematics as a Codec for Speech," *arXiv preprint arXiv:2406.12998*, 2024, [Online]. Available: https://arxiv.org/abs/2406.12998

[9] OHF-Voice, "piper1-gpl: Fast and local neural text-to-speech engine." 2025.

[10] K. M. Rupp, J. L. Hect, E. E. Harford, L. L. Holt, A. S. Ghuman, and T. J. Abel, "A hierarchy of processing complexity and timescales for natural sounds in human auditory cortex," May 2024, doi: 10.1101/2024.05.24.595822.

[11] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proceedings of the 23rd International Conference on Machine Learning (ICML 2006)*, 2006.

[12] F. R. Willett *et al.*, "A high-performance speech neuroprosthesis," *Nature*, vol. 620, no. 7976, pp. 1031–1036, 2023, doi: 10.1038/s41586-023-06377-x.

[13] Y. Kaloga, S. Kumar, P. Motlicek, and I. Kodrasi, "A Differentiable Alignment Framework for Sequence-to-Sequence Modeling via Optimal Transport," *arXiv preprint arXiv:2502.01588*, Feb. 2025.