Brain-to-Text

Stanford CS224N Custom Project

Ethan Trepka
Department of Computer Science
Stanford University
trepka@stanford.edu

Abstract

Brain-computer interfaces could enable people who have lost the ability to speak to communicate efficiently by decoding intended speech from neural activity. Current state-of-the-art models utilize recurrent neural networks to map neural activity to sequences of phonemes and the viterbi algorithm to map sequences of phonemes to text with prior probabilities given by a large language model. I hypothesized that these systems could be improved by (a) identifying pretrained language models that well-describe the distribution of text used in day-to-day speech, (b) utilizing a transformer instead of an RNN for mapping neural activity to phonemes. Surprisingly, I found that a simple trigram language model trained on the switchboard corpus outperformed a pretrained large language model in capturing the statistical properties of spoken language. Unfortunately, replacing the RNN with a variety of transformer architectures did not improve the performance of the neural activity to phoneme system. Taken together, these results suggest that a principled selection of a language model prior has the potential to improve communication neuroprostheses.

1 Key information to include

- Mentor: Kaylee Burns
- External Collaborators: None
- Sharing project: No

2 Introduction

Brainstem stroke and ALS can cause people to lose the ability to control small face and mouth muscles, rendering them unable to speak. Existing assistive technologies leverage preserved motor functions to decode language from large hand movements and eye movements to enable individuals to control a computer and generate speech. However, these technologies are limited to typing speeds of less than 20 words per minute. Luckily, even when fine-grained facial motor function is lost, the neural activity planning and supporting those motor functions in the brain may be maintained. Thus, one way to enable more efficient communication could be to decode intended sentences from neural activity recorded in speech-related areas in the brain.

3 Related Work

A wide range of approaches have recently been proposed for communication neuroprostheses, from decoding imagined handwritten letters from motor cortical areas to decoding intended phonemes or complete words from speech premotor areas.

In 2021, Moses and colleagues demonstrated a speech neuroprosthesis that involved training a word classifier on neural activity in sensorimotor cortex (Moses et al., 2021). However word classification approaches are generally limited to relatively small vocabularies due to the constraints of collecting training datasets with large vocabularies and limitations in the separability of neural activity sampled at a limited number of sites in the brain.

More recently, a number of studies have proposed decoding individual characters or phonemes from neural activity to enable flexible communication with large vocabularies. In 2021, Willet and colleagues developed a brain-computer interface by decoding characters from neural activity in motor cortex while participants imagined writing each character. An RNN was used to map neural activity to character probabilities which were then combined to predict words and sentences using the viterbi algorithm with a language model prior. The authors demonstrated typing speeds of 90 characters per minute with greater than 90% accuracy with this system (Willett et al., 2021). In 2023, Willet and colleagues developed a speech neuroprosthesis by decoding intended speech from neural activity in a premotor area and Broca's area. Using an RNN to decode phonemes in conjunction with a language model, they demonstrated a 9 percent error rate on a 50 word vocabulary decoding task and 23 percent error rate on a 125,000 word vocabulary decoding task. Excitingly, they also released all of the data they collected in this experiment which consisted of recordings of neural activity from premotor and speech areas recorded while a participant attempted to speak 12,100 sentences. (Willett et al., 2023)

4 Approach

Here, I have examined two potential strategies to improve communication neuroprostheses via improving the language model prior and the neural activity to phoneme model. I have tested these strategies using the aforementioned, publicly available dataset (Willett et al., 2023).

What pretrained language model should be used as the 'prior' for mapping a sequence of phonemes to a sequence of words for a communication neuroprostheses? I hypothesize that language models that better represent the distribution of text used in day-to-day speech will outperform those that represent the distribution of text in news articles or text on the internet. I will train simple n-gram models on a variety of different text corpuses and compute their perplexity on sentences in the brain-to-text validation set (see Table 1). These sentences are intended to represent things that a patient may want to communicate and thus should capture the distribution of text used in day-to-day speech. I will also compare a variety of pretrained LLMs using this approach. I predict that models trained on a large speech specific corpus will outperform models trained on more general text databases or a smaller speech corpus. The language model with lowest perplexity on communication-relevant datasets should be the optimal model for the viterbi algorithm prior.

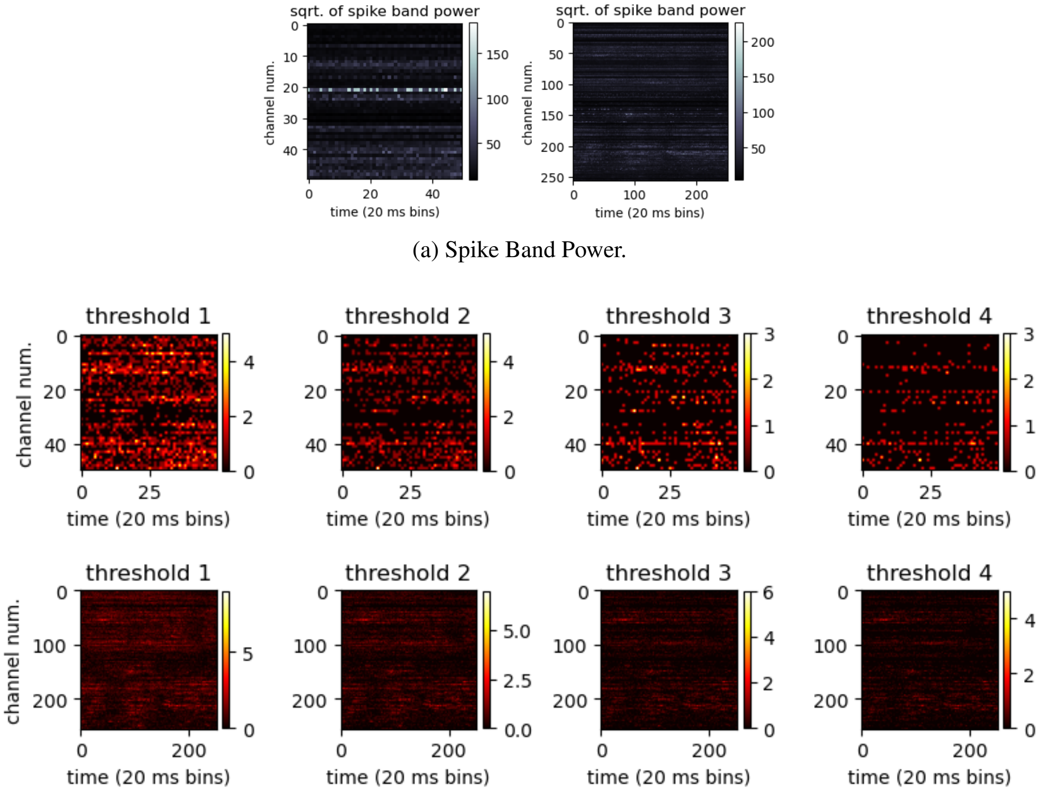
What sequence-to-sequence model should be used to map neural activity to phonemes? I hypothesize that a transformer architecture will outperform RNN-based approaches for decoding sequences of phonemes from neural activity. I will test this approach by replacing the state-of-the-art GRU model presented in (Willett et al., 2023) with a variety of different transformer architectures.

5 Experiments

5.1 Data

The dataset consists of neural activity recorded while a participant intended to speak 12,000 sentences (Willett et al., 2023). Neural activity was recorded with two electrode arrays, each with 128 extracellular electrodes, implanted in the brain of the participant. These extracellular electrodes capture voltage fluctuations outside of neurons at 256 points in the brain. Large changes in measured voltage typically correspond to a neuron firing an action potential nearby. However, interpreting the recorded voltage signal directly is not straightforward (the "spike sorting" problem), and thus for real-time applications such as brain-computer interfaces, simpler features are typically extracted from the voltage trace. These features include counting the number of threshold crossing events (which roughly corresponds to the number of spikes in nearby neurons) as well as the 'spike band power'. Spike band power is used to capture spiking activity in cases where there is low signal-to-noise ratio and threshold crossings perform poorly (Nason et al., 2020). Figure 1a shows spike-band power in a subset of channels and time bins in a trial on the left and in all channels and time bins in a single trial on the right. Spike-band power is very high in channel number 20 in the left plot, suggesting there may be many neurons firing nearby. Figure 1b shows threshold crossing events in a subset of channels and time bins in a trial (top) and in channels and time bins in a single trial (bottom). A count of four threshold crossing events roughly indicates that one or two nearby neurons fired a total of four spikes in that time period, but exactly what threshold should be used is not clear and is likely different for different electrode channels. As a result, the authors filter the data with four different thresholds, all of which could be used simultaneously as features. Here, I use only spike band power (128 features) and threshold 1 crossings (128 features) from one of the two electrode arrays (256 dimensional input total).

There are 8880 training trials and 880 validation trials in the dataset. Neural data on each trial is thus a 256 x trial length array. The language data on each trial is a single sentence which is preprocessed as a sequence of phonemes. Examples of sentences in the training dataset are shown in Table 1.



(a) Spike Band Power.

(b) Threshold crossing events (threshold 4 > 3 > 2 > 1)

Figure 1: Example neural data from one trial of train set.

Sentences
We got pretty lucky on that.
I guess that’s about all I have to say.
Live without dessert for the most part.
Don’t buy them that way.
I live in a farming community.

Table 1: Example sentences from five trials of train set.

5.2 Evaluation method

Perplexity was used to evaluate how well language models captured the distribution of spoken text in the Brain2Text dataset. The perplexity of a sentence of length t words $x_1, x_2, ..., x_t$ is defined as: $P(x_1, x_2, x_3, ..., x_t) = exp(-1/t \sum_i^t logp_\theta(x_i|x_<i))$.

I report median perplexity across all sentences in the validation dataset here which consists of 880 senetences similar to those in Table 1.

Connectionist temporal classification (CTC) loss was used to evaluate the neural activity to phoneme models and was computed using PyTorch. CTC loss is ideal for this scenario because the output of the model is a continuous, unsegmented time series and thus there may be many appropriate alignments of the output phoneme predictions to the target sequence of phonemes.

I report the CTC loss averaged across each sentence (phoneme sequences) in the validation dataset here.

5.3 Experimental details

For the experiments on the language model prior, n-gram models were implemented with custom code, and pre-trained GPT-2 and OPT350M models were obtained from HuggingFace. Train and validation sentences were preprocessed by removing punctuation.

The most important aspect of this experiment was ensuring a fair comparison between between n-gram models and different LLMs, which is challenging given that they have different tokenizers/vocabularies. To address this, I enforced an identical vocabularly for all models. I defined the vocabulary as the set of all words in the Brain-to-Text train and validation set that occur more than once. For computing perplexity with pre-trained LLMs, I computed logprobabilities using softmax over only the subset of words in the vocabularly.

My experiments on the neural activity to phoneme decoder were built off of code provided by Willett et al. (2023). Thus, I used identical parameters to the paper for training the GRU-state-of-the art model (additional details and implementation provided here).

I then replaced the GRU with my own version of a transformer model based on a combination of this code and the code from assignment 4. All transformers had 8-heads and an embedding dimension of 512 (neural data from two adjacent time points concatenated together). All other parameters were identical to those in assignment 4. I varied the number of blocks (2 or 3), type of attention (self-attention vs causal self attention), and dropout probability (0.1 or 0.3) in three different experiments as shown in the figure.

Batch size was 64 trials and training progressed until CTC loss on the test set plateaued.

5.4 Results

A trigram model trained on the switchboard corpus achieved the lowest perplexity on the brain-to-text validation set, lower than both pretrained LLMs, such as GPT-2 and OPT350M, and other n-gram models trained on a non-speech corpus (Reuters) or a small speech corpus (brain-to-text train set) (Table 2). I was surprised by the performance of the trigram model trained on the switchboard corpus, although this is consistent with my prediction that language models trained on a speech corpus will better represent the distribution of language used in day-to-day communication.

The three transformer architectures I tested did not outperform the state-of-the art GRU model which achieved a CTC loss of 0.82. The closest transformer model was the model with 3-layers, causal attention, and a dropout probability of 0.1 which achieved a CTC loss of 1.26. I may have stopped this experiment too prematurely though and this could potentially have further decreased with additional training. This was somewhat surprising to me, although, I think theses results could be explained by inadequate hyperparameter tuning. I still think there likely exists a transformer model that would outperform this GRU baseline, I just was not able to find it here.

Language Model	Train Corpus	B2T Validation Set Perplexity
Unigram	B2T Train Sentences	383.82
Bigram	B2T Train Sentences	219.99
Trigram	B2T Train Sentences	455.96
Unigram	Switchboard Corpus	392.81
Bigram	Switchboard Corpus	**116.54**
Trigram	Switchboard Corpus	**57.40**
Unigram	Reuters Corpus	1778.46
Bigram	Reuters Corpus	1295.53
Trigram	Reuters Corpus	3178.61
GPT-2	N/A	**168.46**
OPT350M	N/A	**141.38**

Table 2: Median perplexity of language models measured on sentences in the brain-to-text validation set. Language models include unigram, bigram, and trigram models trained on the brain-to-text training sentences, on text transcribed from a speech corpus (switchboard), and written text from a news outlet (Reuters), along with pretrained large-language models, GPT-2 and OPT350M.

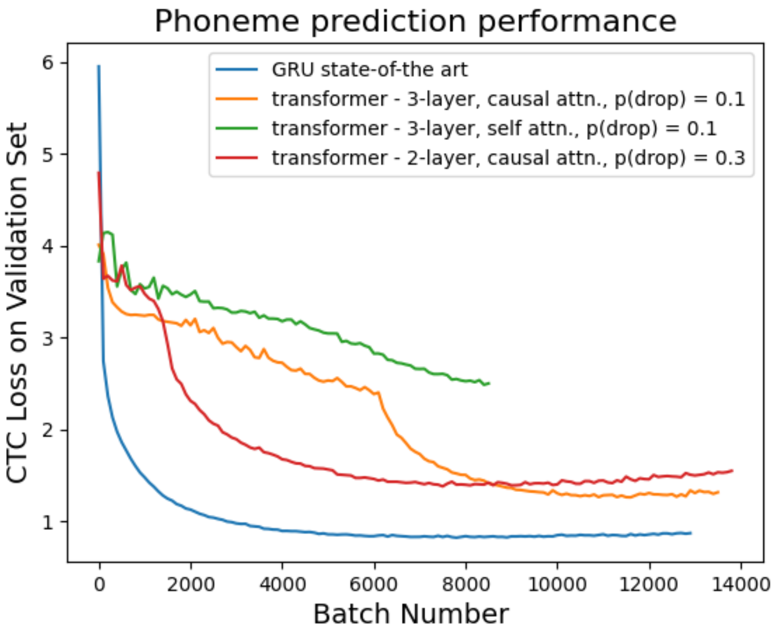


Figure 2: Performance of various transformers on phoneme prediction task compared to a state-of-the art GRU model.

6 Analysis

Why does the trigram model trained on the Switchboard Corpus outperform GPT-2 on this task? To get at this question, I examined example validation sentences, some where GPT-2 outperformed the trigram model and some where the trigram model outperformed GPT-2 (Table 3). The first trend that seemed to pop out to me was that GPT-2 seems to have lower perplexity than the trigram model on longer and more complex sentences, such as "his voice was nearly drowned out by the crowd" and higher perplexity than the trigram model on sentence fragments, such as "A little more flexibility". This is reasonable because the language model can use a much longer context window to predict the probability of words later in long sentences. The second trend that popped out to me was that GPT-2

Table 3: Perplexity of different validation sentences

Validation Sentences	**GPT2 Perplexity**	**Trigram Perplexity**
A little more flexibility.	53.87	**29.39**
They even looked halfway decent.	318.76	**45.46**
They are not sure of themselves.	104.56	**33.47**
If you're ever looking for that.	201.61	**28.71**
A food processor.	**728.54**	2202.79
Two sides of the same coin.	**29.15**	919.77
There is no timetable for a decision.	**21.61**	66.68
His voice was nearly drowned out by the crowd.	**27.1**	98.81

seemed to perform worse with more "speechy" phrases like "if you're every looking for that" and "they even looked halfway decent" that would probably not appear very frequently in written text.

7 Conclusion

The main finding of this project is that a simple language model trained on a speech corpus may be able to outperform more complex pretrained LLMs in representing the distribution of simple sentences used in day-to-day speech. This suggests that rather than using pretrained LLMs out of the box, finetuning on speech specific datasets may be an important step to improve performance of a speech neuroprosthesis.

Although this conclusion is relatively straightforward, it has, to the best of my knowledge, not yet been implement in speech BCI systems. Willett et al. (2023) used either ngram models trained on openwebtext2 or a pretrained llm such as gpt-2 without additional finetuning.

I was not able to find evidence here that a transformer could improve performance on the neural activity to phoneme mapping task. This result, however, may simply be the result of poor architecture and/or hyperparameter choices.

competition.

References

David A. Moses, Sean L. Metzger, Jessie R. Liu, Gopala K. Anumanchipalli, Joseph G. Makin, Pengfei F. Sun, Josh Chartier, Maximilian E. Dougherty, Patricia M. Liu, Gary M. Abrams, Adelyn Tu-Chan, Karunesh Ganguly, and Edward F. Chang. 2021. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227. PMID: 34260835.

Samuel R. Nason, Alex K. Vaskov, Matthew S. Willsey, Elissa J. Welle, Hyochan An, Philip P. Vu, Autumn J. Bullard, Chrono S. Nu, Jonathan C. Kao, Krishna V. Shenoy, Taekwang Jang, Hun-Seok Kim, David Blaauw, Parag G. Patil, and Cynthia A. Chestek. 2020. A low-power band of neuronal spiking activity dominated by local single units improves the performance of brain–machine interfaces. *Nature Biomedical Engineering*, 4(10):973–983.

Francis R. Willett, Donald T. Avansino, Leigh R. Hochberg, Jaimie M. Henderson, and Krishna V. Shenoy. 2021. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858):249–254.

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, Krishna V. Shenoy, and Jaimie M. Henderson. 2023. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036.