Improving speech brain-computer interface with conversation context

Stanford CS224N Custom Project

Brian Lee
Department of Computer Science
Stanford University
bjlee25@stanford.edu

Allison Tee
Department of Mathematics
Stanford University
ateecup@stanford.edu

Abstract

For people with paralysis, the inability to communicate is one of the most debilitating aspects of their condition. Speech brain-computer interfaces (BCIs) have the potential to help people overcome this problem by translating neural patterns during speaking attempts into sentences. We build on a pipeline consisting of an RNN to decode neural data, an n-gram language model (LM) to output a list of possible word sequences ordered by likeliness, and transformer large language model (LLM) to choose the most probable sequence. We implemented conversational context into the LLM, where the model was given additional contextual information to improve the word error rate (WER) and performed a structured hyperparameter search. We found that performance does not vary much across different context lengths, but the optimal context window was 1,000 characters, giving a final WER of 14.0%, an improvement over the original (no context) WER of 16.7%. We could only find context for 170 out of the 600 test phrases, and our WER on the sentences with context is 10.6%. Additionally, we use OpenAI's ChatGPT to evaluate the sentences directly, but it was unsuccessful but can provide interpretable results. We experimented with different OPT model sizes but found that the second-largest model (6.7B parameters) yielded the best results by a slim margin.

1 Key Information to include

- Mentor: Chaofei Fan
- External Collaborators: No
- Sharing project: No

2 Introduction

Speech brain-computer interfaces are a promising technology to enable people with paralysis to communicate by decoding incoherent speech attempts. The most advanced speech BCI currently available was created by the Neural Prosthetics Translational Laboratory (NPTL) here at Stanford (Willett et al. (2023)). Our project enhances the decoding algorithm of the paper's SoTA speech neuroprosthesis by incorporating conversational context.

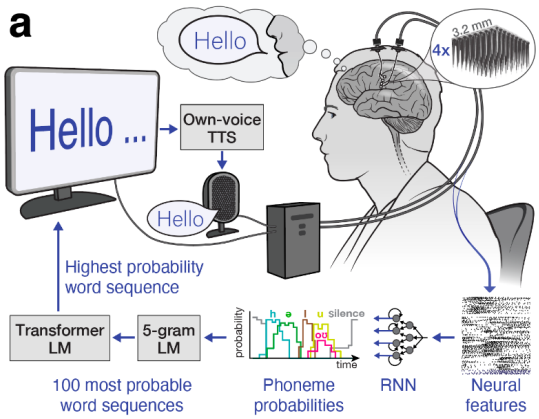

Figure 1: The translation pipeline of the original BCI. The neural data is decoded by an RNN, which outputs phoneme probabilities, which is then fed into an n-gram LM. This n-gram LM then outputs a list of possible word sequences, of which the transformer model chooses the one that is the most likely.

By taking intracortical electrode measurements, researchers at the NPTL looked to decode a subject's intended speech through brain signals, through an intermediary step of phoneme prediction (phonemes being single-syllable utterances that can sometimes be made, even by those with paralysis).

The key contributions of the study include using an RNN and language models for decoding phonemes from neural signals and words from phonemes; vastly increased accuracy, speed, and vocabulary size of the word decoder model over previous state-of-the-art (SOTA) models; and demonstrating that the ventral premotor cortex (area 6v) well-represents speech articulators even for people suffering years of paralysis.

The RNN decoder translated the neural spiking data into a set of probabilities indicating the intended phoneme, the output of which was then fed into a language model. The language model was an n-gram model created with Kaldi, and the output was a list of probable sentences. This list was then given to a transformer model to find the most likely sentences. The error rate reported in this study (23.8%) is around two times smaller as the previous SOTA speech BCI, and it manages to perform well on a large vocabulary as well, which is a first for this type of model. Furthermore, the speed at which it decodes is quite fast, over 3 times as fast as the previous SOTA method. These results show promise for allowing individuals with paralysis to have natural-level conversations.

Here, we outline a modified pipeline that achieves a WER as low as 10.6% over a large vocabulary of 125,000 words with a less expensive 3-gram LM with a context length of 1,000 characters. Our improvements led to the optimal score weighting of the LLM as opposed to the 3-gram LM to be 1 (compared to the original study's weight of 0.5), suggesting that the LLM with context improvements alone is significantly better than the 3-gram at determining sentence scoring. The main components of our work include:

- Implementing context-based sentence rankings by matching each sentence to its larger conversation in the dataset (if available).
- Optimizing context window length and other hyperparameters to score sentences, along with the parameter size of OPT models.
- Exploring and analyzing GPT models with prompt engineering as an alternative method to determine the best sentences.

3 Related Work

Willett et. al's paper is a proof of concept for using neural activity to decode attempted speech movements with a large vocabulary, so the results are far from complete and clinically viable. The error rate is still too high for fluent communication compared to a 4-5% error rate for speech-to-text systems. The authors state that further work should aim to reduce decoder train time and adapt to

neural activity changes. The hardware for measuring electrical signals in the cortex is in development, and it is unclear how well the results generalize to other people with paralysis because the speech BCI was only tested on one English-speaking person with ALS. The paper represents a significant achievement in the field of speech BCIs, being the first to decode neural spiking data and attempted speech of unconstrained sentences using a large vocabulary. The word decoding speed of this technology is significantly faster than that of other speech BCIs and alternative communication technologies, such as those based on eye tracking and hand movement (Räihä and Ovaska (2012); Willett et al. (2021)). A more recent study has reached significantly lower WERs of 2.66% with improvements such as doubling the number of electrodes recording activity in the ventral precentral gyrus, language model improvements, and fine-tuning the decoder (Card et al. (2023)). A number of successful models that enhance speech BCIs for this use case improve the RNN in the pipeline, but this is not the focus of our project.

NPTL's speech BCI builds on foundational research that decodes individual letters and a more limited selection of words from neural signals but does not allow for conversation in the normal sense. This work uses electrode arrays to achieve significantly lower word error rates and faster decoding speeds compared to previous speech BCIs. Such technological advancements not only have practical implications for improving communication aids for individuals with speech impairments but also contribute to the broader understanding of the neural representation of speech. This research could potentially help developments in NLP tasks that require real-time processing of complex datasets, by demonstrating how complex speech patterns can be interpreted from neural data, helping expand the field's approach to both speech synthesis and recognition technologies.

4 Approach

- **Code:** We began by using the code provided by Willett et al.'s paper. It implemented the RNN, the n-gram language models, and a rescoring algorithm for LLMs. As we were mostly working on the LLM part of the model, we modified the inputs to those rescoring functions. Furthermore, we modified the rescoring functions themselves to accommodate contextual integration and hypothesis modification.
- **Models:** The primary LLM that we are working with is OPT 6.7B, an open-source pretrained transformer model. The original paper uses a 5-gram language model to predict possible next words in a sequence, but because it requires nearly 240 GB of memory to run, we use the computationally cheaper 3-gram language model. After calculating the top n-gram predictions, the original paper's transformer LLM refines these predictions by calculating the log probability of each token in a hypothesis given the tokens before it and summing them up to calculate the sentence score (see equation 1).
- **Adding context:** Every sentence in the dataset is sampled from a larger conversation in the Switchboard Dialog Act Corpus (Stolcke et al. (2000)). We wrote a function to search through this corpus to find each ground truth sentence in the training set and append the relevant conversation text before it to a list of contexts. When we calculate the log probabilities of tokens with OPT, we also conditioned on the most recent tokens of their respective context.
- **Hyperparameter Search:** Willett et al.'s original implementation calculates the sentence score with a combination of scores from an n-gram language model and the OPT model. We performed hyperparameter grid searches before and after incorporating context over three hyperparameters: the acoustic scale (α), the length penalty, and LLM weight (β). The acoustic scale and β are the weights given to the RNN's and the n-gram's log probabilities in equation 1. The length penalty scales the number of characters the model outputs and applies it as a penalty to the score.

$$score(s) = \alpha * log(P_{RNN}(s)) + (1 - \beta) * log(P_{ngram}(s)) + \beta * log(P_{opt}(s))^{1} \quad (1)$$

  Additionally, we tested different context window sizes.
- **LLM Playground:** We asked GPT 3.5 Turbo and GPT 4 to pick the most likely sentence from lists of ranked, probable sentences given conversation context to test if it could perform

[1]This equation is slightly modified from the original scoring equation, as it switches around the complementary weights for the n-gram and OPT log probabilities. This is done to be consistent with the code. Note that in the code, β is also labelled as alpha, and α is labelled as acoustic_scale.

better than OPT 6.7B. We tested various OPT model sizes (1.3B, 2.7B, 6.7B, and 13B). Due to computational considerations, our exploration in this area is limited.

5 Experiments

5.1 Data

We use the “competitionData" version of the data collected by NPTL that has over 12,100 sentences (sourced from the Switchboard corpus) spoken by a person with paralysis and the corresponding neural spiking activity recorded in speech-related cortex areas (NPTL). It contains a training, test, and validation split, where each partition has both sentences and hypotheses for use in training. We also utilize the Switchboard corpus consisting of 5-minute phone calls for a total of 122,646 utterances to find the corresponding conversation context.

5.2 Evaluation method

The study evaluates the model's decoding performance using word error rate, which is also the metric we adopt. The word error rate is defined as the number of word substitutions, insertions, or deletions needed to make the decoded and true sentences match. The original paper achieved an 11.8% offline word error rate on a vocabulary with 125,000 words and 5-gram language model, but the more comparable figure is a 15.4% WER with the 3-gram LM.

5.3 Experimental details

All of our training was done on a Linux VM on Google Cloud. The hardware specifications of the VM were a Nvidia T4 GPU and 16 vCPUS (for a total of 32 GB of memory). As our focus was on the language model side of the decoder, we imported the neural decoding RNN as-is from the paper while editing the language model portion. For our baseline, we combined the results of a pretrained 3-gram language model with a custom version of a pretrained OPT transformer. Aside from hyperparameter search, the OPT transformer used the Switchboard contextual information as mentioned above, and we modified and wrote additional code to incorporate this. The evaluation time was approximately 20-40 minutes per run.

Hyperparameter	Range	Step Size
acoustic scale (α)	$(0.2, 1)$	0.1
LLM weight (β)	$(0, 1)$	0.1
length penalty	$(-3, 2)$	<1
context window (chars)	$(0, 2, 000)$	variable

Figure 2: Hyperparameters tested in an initial grid search. Following this, we conducted searches with smaller ranges.

5.4 Results

The most successful model is OPT 6.7B with a context window of 500 characters, achieving a 14.0% WER on the test dataset (10.6% WER on sentences with context). Our point of comparison is the 3-gram + transformer model baseline, which achieves a WER of 15.4%. Using the optimal hyperparameters of $\alpha = 0.5$, length penalty of 0.4, and $\beta = 1$ decreased the WER by 1-2 %.

GPT 3.5-Turbo performed poorly at choosing the best sentence from the n best outputs of the LM with context. The model often did not follow instructions, claiming that there wasn't enough information provided to make a choice, and wrote a function to simply choose the highest score, even when prompted to incorporate other metrics like grammatical correctness and contextual information. However, GPT 4-Turbo performed much better. It evaluated the sentences on not only scores, but grammatical correctness, syntax, and contextual information.

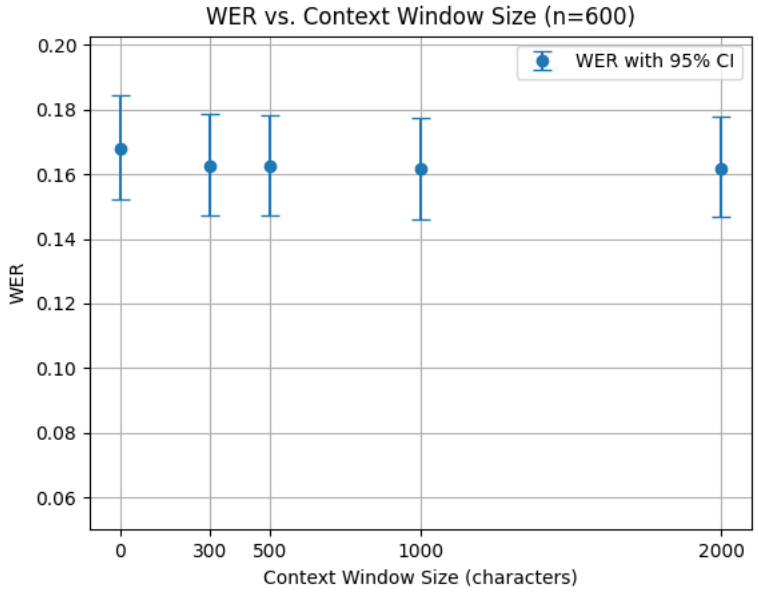

Figure 3: Graph of WER on the test set $(n = 600)$ across a range of context window sizes for the 3-gram with OPT-6.7B with default hyperparameters.

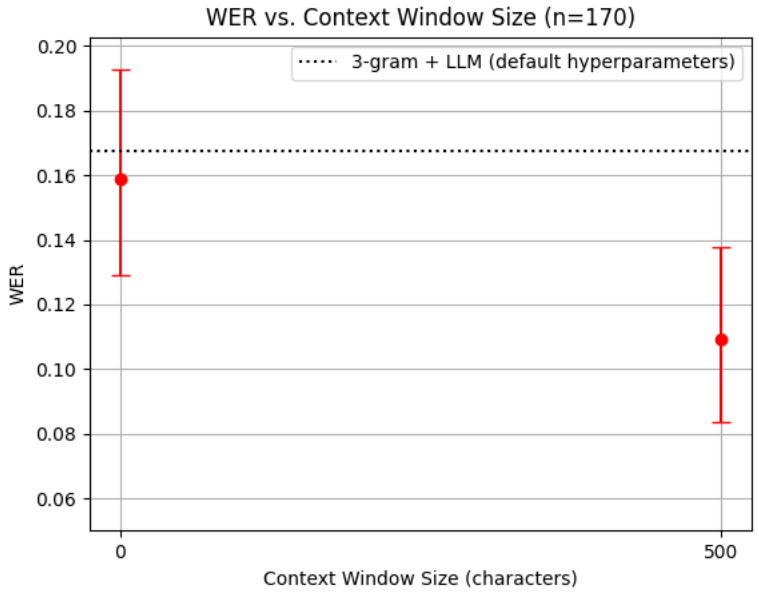

Figure 4: Graph of the improvement on the sentences with context of the model with context and optimized hyperparameters with respect to the vanilla 3-gram + OPT-6.7B.

6 Analysis

Our results for the sentences with context are notably better than the baseline model. Our experiments show that model performance does not vary significantly with context size, as a context window of 300 characters results in a WER of 10.9%– only 0.3% greater than our best result. Combined with the drawback of larger context lengths taking significantly more time to evaluate the test dataset (15 minutes for 300 characters and over 30 minutes for 1,000 characters), incorporating a few sentences

of context is a time-efficient solution that still results in significant WER improvement. This is likely because the most relevant context for any sentence is typically right before it. Note that the conversation context we use includes dialog from multiple parties, which suggests the use of a microphone in a real-life setting. However, because we don't need much context to result in an improvement, just keeping track of what the speech BCI subject says throughout the conversation should be sufficient.

Our method has limited improvement on the entire test dataset because we could only find context for 170 out of the 600 total sentences. This may be because of the vernacular nature of the corpus. For some training points, the sample is drawn from the beginning of the conversation and doesn't result in a context-improved data point. While this is not extremely common, it happened a nonzero number of times throughout training, which could have lowered the efficacy of this method. This is not something we can fix, as the chosen samples are all we have to work with.

The Switchboard dataset also includes filler words (e.g. "um," "like") that introduce noise in the log probability calculations. This is expected in conversations, so it isn't a major concern. An interesting result that we discovered is that the optimal weight for the OPT model's log probability when determining sentence score is 1, which suggests that the context-improved LLM is much better at scoring sentences than the 3-gram LM. The model is likely bottlenecked by the 3-gram LM, which could be improved to a 5-gram LM.

The original LLM used (OPT 6.7B) is more accurate than OPT 1.3B, 2.7B, and 13B by a slight margin. We also reason that the OpenAI GPT models are too costly to employ in a large-scale, daily setting. There are some benefits in terms of transparency, as when asked to explain its rationale for choosing sentences (without additional context), GPT 4 uses grammar, syntax, and context-based reasoning:

```
Ground Truth:  "you know it was just in jest"
GPT 4's Output:  "you knew it was just in jest"
Rationale:  " "you knew it was just in jest" (makes sense in a playful or
non-serious context)
"you knew it was just in last" (grammatically incorrect)
"you knew it was just in lace" (unlikely unless discussing something
specific to fabric)
"you knew it was just in dash" (incorrect but could be a typo for something
else)..."
```

While GPT 3.5 was clearly lacking in many ways, GPT 4 showed initial promising results. This was most obvious in the way that it dealt with contextual information. For instance, for one word sequence, the word sequences whose correct transcription was "just way in the back" were passed to GPT 4, with and without context. Without context, GPT determined the most likely sequence to be "the way in the back." However, given context (which was the string "sounds lovely course in the spring we get the bluebonnets and Indian paintbrush h Oh the wild flowers are beautiful I think I was thinking about throwing some for in the back just where the kids"), it actually returned "just way in the back," the correct string.

The problem with GPT-4 was not the results, but rather the difficulty of getting standardized results. Even when told to only output the best sentence, or to only output the index of the best sentence, it would try to format the output in a weird way. A sample of what these outputs were like are shown below:

```
1.  Selected sequence:  The best word sequence is:  "it's really hard to
find something that works"
2.  Selected sequence:  "and i love the tying" - LM Score:
-41.400733947753906, AC Score:  -35.67578125
3.  Selected sequence:  The best word sequence is:  "i do have a friend
that run" - LM Score:  -38.738162994384766, AC Score:  -39.4306640625
```

As you can see, the three different outputs had three different formats. This made it impossible to compute a WER without using methods like function-calling, and we could only do qualitative analyses as done above.

7 Conclusion

In this paper, we apply methods to modify the speech BCI pipeline by Willett et. al (WER 23.8 %) that achieves a WER as low as 10.6% over a large vocabulary of 125,000 words with a 3-gram LM and context length of 1,000 characters. Our improvements led to the optimal score weighting to give to the LLM as opposed to the 3-gram LM is 1 (compared to the original study's weight of 0.5), suggesting that the LLM with context improvements alone is best at determining sentence scoring. Using our implementation of context-based sentence rankings by matching each sentence to its larger conversation in the dataset, we found that various conversational context windows from 300 to 2,000 characters have minimal impact on the performance of the model, with the 300-character window achieving 10.9% on the subset of sentences with context. We found that the optimal sentence scoring system does not consider the 3-gram predictions at all and weights the improved LLM's log probabilities twice as much as the RNN's log probabilities (though note that the initial n-best sentence selection provided by the 3-gram is used in the rest of the pipeline). The performance of the various OPT models are within 1% WER of each other. Our limited exploration of GPT models with prompt engineering to find the best sentences did not yield substantial results due to the hallucinating and sometimes capricious nature of LLMs.

However, because initial qualitative results using GPT 4 seem promising, we would like to explore ways to quantify the outputs of GPT 4 in a way that would allow us to analyze its effectiveness when compared to traditional LLM pipelines. Furthermore, we would like to see how effective context could be when decoding a longer string; by dividing the decoding process into smaller steps, and treating the previously decoded sequence as the context, we may be able to improve the accuracy of the next chunk of decoding. With more resources, we believe that using a 5-gram language model with context will result in additional improvements in WER.

8 Ethics Statement

One of the possible ethical concerns that arise with this project is that it may make it such that the intended users of this device and model become potential victims of fraud. As the device is meant to interpret words from an individual who has difficulty communicating (often unable to communicate at all), individuals may look to take advantage of them by making it appear as though the person with paralysis has consented to something even when they haven't by using this technology. This ethical concern is largely mitigated by the fact that the barrier to actually set up this BCI is quite extensive; as an invasive BCI that requires electrodes to be surgically implanted into the brain by a team of surgeons, the chances of a malicious individual taking advantage of the model is quite small. However, this risk can be further mitigated by making the model's software locally installed on the BCI, and disconnected from the internet as to prevent any hacking or tampering with the model. This could stop people from manipulating either the weights in a way that would make the model give inaccurate decodings or from directly manipulating the output itself. In general, it is also important to ensure safety and informed consent in implantable BCI research.

Another potential concern is in the field of criminal justice: what if a court wants to use this technology to "read someone's mind" to determine guilt? First of all, since the model is decoding speech intent (which is distinct from conscious thought), the model should not say anything unless it is the express intent of the individual. However, even if it managed to have these capabilities, a countermeasure we could take is to clearly advertise this aspect of the product, making it clear that it can only decode conscious efforts to speak as opposite to some hidden guilt like a polygraph claims to do.

References

Nicholas S Card, Maitreyee Wairagkar, Carrina Iacobacci, Xianda Hou, Tyler Singer-Clark, Francis R Willett, Erin M Kunz, Chaofei Fan, Maryam Vahdati Nia, Darrel R Deo, et al. 2023. An accurate and rapidly calibrating speech neuroprosthesis. *medRxiv*.

Neural Prosthetics Translational Lab (NPTL). 2023. Brain-to-text benchmark '24. EvalAI.

Kari-Jouko Räihä and Saila Ovaska. 2012. An exploratory study of eye typing fundamentals: dwell time, text entry rate, errors, and workload. In *Proceedings of the SIGCHI Conference on Human*

Factors in Computing Systems, CHI '12, page 3001–3010, New York, NY, USA. Association for Computing Machinery.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

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, et al. 2023. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036.