Context-Aware Brain-To-Text Decoding for Speech Brain-Computer Interfaces

Ellie Tanimura
Department of Computer Science
etanim@stanford.edu

1 Introduction

Speech Brain-Computer Interfaces (BCIs) have the potential to restore rapid communication to individuals who cannot physically speak intelligibly by decoding neural activity into text Willett et al. (2023). These systems are typically activated when BCI users think about or attempt to vocalize the words they wish to communicate. This mental activity generates neural spikes in target regions of the brain, which the system decodes into corresponding sentences or phrases. Most systems output several candidate hypotheses, ranked in order of likelihood.

Willett et al. recently demonstrated a groundbreaking speech-to-text inference pipeline that decodes spiking activity into sentences with a 23.8% word error rate (WER) on a 125,000-word vocabulary, which is not yet sufficiently low for everyday use Xiong et al. (2018)He et al. (2019). In their analysis, they observed that phonemes with similar articulations were more likely to be confused by the decoder. They also found that rescoring candidate hypotheses using a large language model (LLM) improved WER by promoting linguistic fluency.

This project explores a complementary rescoring strategy based on conversational context—that is, how well a candidate sentence fits within the dialogue history. Instead of relying solely on linguistic fluency, this method incorporates dialogue history directly into the decoding pipeline to improve contextual coherence and disambiguation.

The core contributions of this project are:

- A novel contextual rescoring approach that integrates prior utterances from the same conversation into hypothesis scoring (this is from my prior work on this project);
- A constrained scoring formulation that balances acoustic, statistical, and neural language model scores;
- A context summarization heuristic that improves computational efficiency with minimal loss in accuracy;
- Preliminary integration into the Brand system for eventual deployment in real-time neural interfaces.

Using these techniques, the system achieved a WER of 10.7%, marking a 12.7% relative improvement over the original decoding pipeline. Summarizing context reduced inference time and memory usage by over 40% with only minor drops in performance. These results highlight the promise of context-aware decoding in next-generation speech BCIs.

1.1 Recent Advancements in Brain-to-Text Systems

The transcription quality of speech BCIs is primarily evaluated using word error rates (WER), which is defined as the number of insertions, deletions, and substitutions required to transform the predicted sentence into the reference sentence, divided by the number of words in the reference. A lower WER indicates better transcription accuracy. Although character error rate (CER) is commonly used in some speech systems, the focus of my evaluation remains on WER due to its stronger alignment with user-facing readability in full-sentence decoding tasks.

Recent research in speech BCIs (brain-computer interfaces) has focused primarily on two challenges: disambiguating acoustically similar phonemes and minimizing transcription latency. This project

addresses the former, building upon the work of Willett et al., whose system uses intracortical electrode arrays to record neural activity. These signals are processed using a five-layer gated recurrent unit (GRU) network, which decodes time-series neural spikes into phoneme sequences by selecting the most probable phoneme at each time step. The resulting phoneme sequences are then passed through a Weighted Finite-State Transducer (WFST), which leverages a trigram language model (LM) to map them into likely word sequences. The decoder outputs an *n*-best list of candidate transcriptions, ranked by their estimated likelihood of accurately reflecting the user's intended speech.

In supplementary work, Willett et al. demonstrated that rescoring these *n*-best hypotheses with a transformer-based language model reduced the WER from 23.8% to 17.4%, a result validated by subsequent studies.

2 Approach

2.1 Incorporating Conversational Context (Prior Work)

With Willett et al.'s decoding pipeline, I used the RNN-generated phoneme probabilities alongside a trigram language model implemented in Kaldi to generate candidate transcriptions. WeNet is employed to streamline the implementation of the real-time language model decoder Willett (2022).

Unlike Willett et al., who output only the top-scoring sentence, I extracted the top-n candidates for each sequence and prepend each one with prior conversational context, defined as the m utterances preceding the target sentence in the Switchboard Dialog Corpus transcript. This augmentation enables the rescoring model to evaluate how well each candidate fits into the ongoing conversation. A transformer-based LLM was then used to rescore the context-enriched hypotheses.

2.2 Constrained Hypothesis Scoring

To better balance contributions from different components of the above pipeline, I modified the original scoring formulation. Willett et al. combined LLM, trigram, and acoustic model scores via an unconstrained linear combination:

$$S_{\text{total}} = \alpha \cdot S_i + (1 - \alpha) \cdot \text{oldLMScores} + \beta \cdot \text{acousticScores}.$$

Previously, I used this scoring function to evaluate my system. I propose a normalized, convex combination:

$$S_{\text{total}} = \alpha \cdot S_i + \gamma \cdot \text{oldLMScores} + (1 - \alpha - \gamma) \cdot \text{acousticScores},$$

where $\alpha + \gamma \leq 1$, ensuring interpretability and improved stability during tuning. All scores are normalized to operate on comparable scales prior to combination.

Here, α controls the influence of the LLM, while γ weights the trigram model. The acoustic model's contribution is implicitly determined as $1 - \alpha - \gamma$, ensuring that all components remain proportionally balanced. This formulation offers key advantages. First, it guarantees that the scoring function remains a convex combination, preventing any single component from disproportionately influencing rankings. Without this constraint, the weights can become arbitrarily large, leading to instability. Second, it regularizes the model, discouraging redundancy between the correlated acoustic and language model scores. By ensuring that an increase in one term's influence is offset by a decrease in another, the system avoids overfitting to any particular score component.

Additionally, the constrained scheme improves hyperparameter stability. In an unconstrained setting, small weight adjustments can dramatically alter rankings, making optimization unpredictable. Here, changes in α and γ inherently rebalance contributions, leading to smoother tuning and better generalization.

This approach assumes comparable scales among the three scoring components. If one operates on a different magnitude (e.g., acoustic scores), normalization is required to preserve meaningful weight distributions. With proper scaling, this formulation offers a principled, stable, and interpretable method for rescoring hypotheses while maintaining flexibility for contextual integration.

2.3 Context Summarization

Incorporating conversational context improves decoding accuracy, but full-context inference with large language models (LLMs) like OPT-6.7B introduces significant compute challenges. Long dialogue histories—often exceeding 500 words—inflate memory usage and inference time, limiting scalability in real-world systems. To address this, I developed a context summarization heuristic designed to retain semantically relevant information while drastically reducing context length.

To reduce computational overhead from long conversational histories, I implemented a context summarization heuristic. The heuristic operates as follows:

1. Remove stopwords, modal verbs, and generic verbs from the dialogue history to extract semantically meaningful keywords;

2. Retain only sentences containing at least one keyword;

3. From this filtered list, select the N most recent relevant sentences as the summary context.

This approach enables a significant reduction in input length to the LLM, allowing more scalable real-time inference.

2.4 System Integration

To support live neural interface testing, I am integrating the context-aware rescoring pipeline into the Brand system Ali et al. (2023), a software stack used by BrainGate for neural signal decoding and communication. The integration was designed to support real-time operation using inter-process communication via Redis. Although real-time participant testing is still pending, the infrastructure is in place for future deployment and evaluation.

3 Results

3.1 Constrained Hypothesis Scoring

To systematically evaluate the impact of incorporating conversational context with the new rescoring function, I calculated the WER for each candidate transcription in the *n*-best list against the corresponding ground truth. The results are displayed with a heatmap matrix, where each row corresponds to a ground truth sentence and each column represents a candidate transcription from the *n*-best list. The highest-ranked rescored hypotheses appear on the leftmost side. Dark red regions represent higher error values, indicating significant deviations from the ground truth, while dark blue regions signify lower error values, suggesting a closer match.

The heatmaps in Figure 1 revealed two primary failure modes: cases where the correct transcription was entirely absent from the candidate set, indicating fundamental decoding limitations; and cases where the correct transcription was present but improperly ranked, suggesting inefficiencies in the ranking mechanism. This distinction is crucial, as the former suggests a need for improved candidate generation, while the latter highlights the potential for better rescoring strategies.

Table 1 illustrates the second case, where the correct transcription ("it might be kind of scary") was misranked due to confusion with phonetically similar phrases such as "story" and "tech". After applying our conversational context-based rescoring, the model correctly promoted the ground-truth sentence to the top position, demonstrating the value of context for disambiguation.

In contrast, Table 2 shows an example where the correct transcription ("they roll up") was never generated by the decoder. In such situations, rescoring cannot recover the target, underscoring the need for improved acoustic modeling and richer candidate generation upstream of the rescoring pipeline.

Table 1: Example of phonetic confusion before applying conversational context. The correct hypothesis was present but incorrectly ranked. After rescoring with context, the correct sentence was ranked first.

Rank	Candidate Hypothesis
1	it might be kind of story
2	it might be kind of tech
3	it might be kind of tacky
4	it might be kind of turkey
5	**it might be kind of scary**
6	it might be kind of techie

The ground truth sentence was "it might be kind of scary." Before rescoring, phonetically similar alternatives such as "story" and "tech" were ranked above the correct phrase "scary". After integrating conversational context, the correct hypothesis was promoted to the top rank, illustrating the disambiguation power of context-aware rescoring.

Table 2: Example where the ground-truth transcription was missing from the top-5 candidate list.

Rank	Candidate Hypothesis
1	they're all up
2	well up
3	they're well up
4	they were up
5	they fill up

The correct transcription ("they roll up") did not appear among the top-5 candidates. Since conversational context cannot recover a sentence that was never proposed, this type of failure highlights the limitations of the initial candidate generation process rather than the rescoring step.

In most cases, after applying incorporating conversational context and applying the constrained scoring function, the highest-ranked candidate transcriptions tend to have the lowest word error rates. This pattern is clearly visible in the heatmap, where top-ranked hypotheses are associated with minimal errors.

Quantitatively, incorporating conversational context with the optimized scoring strategy resulted in a WER of 10.7%, representing a substantial improvement over both the original baseline and previous rescoring approaches, including context-aware rescoring with the original rescoring function.

In summary, the following table summarizes the performance of various rescoring strategies, highlighting the incremental improvements achieved through context integration and scoring optimization:

Table 3: Word Error Rate (WER) Across Different Rescoring Strategies

System Variant	WER (%)
Baseline Willett et al. system	23.8
OPT-6.7B rescoring without context	16.5
OPT-6.7B + conversational context	11.4
OPT-6.7B + context + optimized scoring	**10.7**

These results demonstrate several key trends. Rescoring with a transformer-based language model alone yields a significant WER reduction of 5.5% compared to the baseline (23.8% $\rightarrow$ 16.5%). Incorporating conversational context further reduces WER by 4.7%, underscoring the value of dialogue

history in disambiguating candidate transcriptions. Applying the optimized, convex-constrained scoring strategy brings an additional 0.7% absolute improvement, resulting in a total WER reduction of 12.7% relative to the original system.

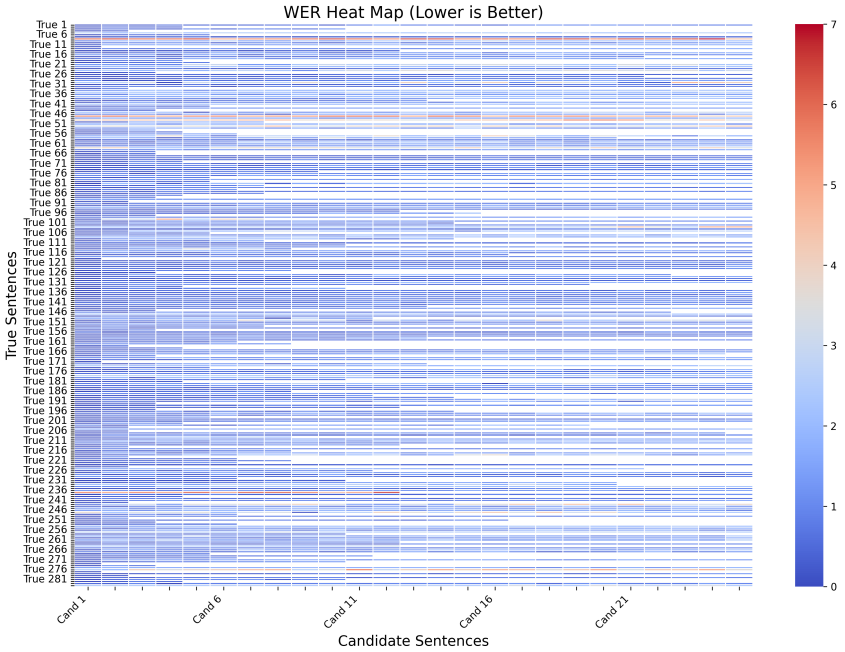

Figure 1: Heatmap visualization of WER for all candidate transcriptions

3.2 Context Summarization

As shown in Table 4, context summarization yields substantial reductions in memory usage and inference time. A 200-word summary maintains a relatively low WER of 11.2% while halving memory consumption. Increasing to 500 words further improves WER to 11.0%, while still using approximately 40% less memory than the full-context baseline.

Table 4: Performance trade-offs using context summarization

Experiment	**WER (%)**	**Peak Memory Util (MB)**	**Avg. Rescore Time (sec)**
Full context (500–600 words)	10.7	14890.22	5.37
Summarized context (100 words)	11.7	–	–
Summarized context (200 words)	11.2	7778.90	2.23
Summarized context (500 words)	11.0	9010.29	3.21

These results demonstrate that intelligent context filtering can enable real-time BCI decoding pipelines to scale more efficiently, without significant degradation in transcription accuracy.

4 Challenges, Limitations, and Future Work

While context-aware rescoring approach shows significant improvements in transcription accuracy, several challenges emerged throughout development. One of the primary difficulties was managing the trade-off between model accuracy and real-time performance. Large language models like OPT-6.7B, though powerful, introduce considerable inference latency, which is a barrier to deployment in live BCI systems. Tuning hyperparameters such as α and γ in the constrained scoring formulation also proved computationally expensive, requiring multiple full decoding passes over the dataset. Additionally, I found that integrating conversational context was not always straightforward—spoken

dialogues often include disfluencies, non-sequiturs, or shifts in topic that challenge the model's ability to track coherence over long spans.

Another limitation of this work is that evaluation was restricted to offline decoding using the Switchboard Dialog Corpus. While Switchboard offers realistic conversation data, it may not fully reflect the speech patterns or vocabulary of real BCI users, especially in medical or assistive contexts. Furthermore, the summarization heuristic, though effective at reducing input size, is based on handcrafted rules that may discard useful context. More sophisticated summarization techniques or retrieval-based context selection might yield better results. I also observed that transformer-based language models occasionally ranked semantically inconsistent hypotheses highly, suggesting that their fluency bias could sometimes override semantic alignment with the prior context.

Looking forward, several directions hold promise for improving the system. One key goal is to deploy the pipeline in a live BCI setting via the Brand software stack and evaluate it using real neural signals.

5 Conclusion

This project presents a context-aware decoding pipeline for speech BCIs that improves transcription accuracy by incorporating conversational history into hypothesis rescoring. Building upon the decoding architecture introduced by Willett et al., I demonstrate that adding contextual information and using a constrained scoring formulation substantially reduces WER. I introduce a summarization heuristic that enables real-time inference with large language models by reducing memory and compute load without significant accuracy degradation.

The final system achieves a 10.7% WER—a 12.7% improvement over the original baseline—highlighting the importance of contextual awareness in neural decoding. These results represent a meaningful step toward practical deployment of speech BCIs in real-world, interactive settings. Future work will focus on optimizing runtime performance, incorporating user-specific adaptations, and validating the pipeline with live neural recordings.

References

Yahia H. Ali, Kevin Bodkin, Mattia Rigotti-Thompson, Kushant Patel, Nicholas S. Card, Bareesh Bhaduri, Samuel R. Nason-Tomaszewski, Domenick M. Mifsud, Xianda Hou, Claire Nicolas, Shane Allcroft, Leigh R. Hochberg, Nicholas Au Yong, Sergey D. Stavisky, Lee E. Miller, David M. Brandman, and Chethan Pandarinath. 2023. Brand: A platform for closed-loop experiments with deep network models. *bioRxiv*.

Yanzhang He, Tara N. Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, Qiao Liang, Deepti Bhatia, Yuan Shangguan, Bo Li, Golan Pundak, Khe Chai Sim, Tom Bagby, Shuo-yiin Chang, Kanishka Rao, and Alexander Gruenstein. 2019. Streaming end-to-end speech recognition for mobile devices. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6381–6385.

Frank Willett. 2022. speechbci: A software repository for speech brain-computer interface research. `https://github.com/fwillett/speechBCI`. Accessed: 2024-05-20.

Frank R. Willett, Emily M. Kunz, C. Fan, et al. 2023. A high-performance speech neuroprosthesis. *Nature*, 620:1031–1036.

W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke. 2018. The microsoft 2017 conversational speech recognition system. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5934–5938.