Technical Report: Dual-Stream GRU-Conformer Architecture for Brain-to-Text Decoding from Utah Array Recordings

SMU Research Framing - Do Tri Nhan

March 2026

1 Introduction

This report summarizes the experiments conducted on the "High-performance speech neuroprosthesis" dataset. The primary objective was to decode neural signals from the motor cortex into text (phonemes/words) using various deep learning backbones, ranging from Recurrent Neural Networks (RNNs) to advanced Transformers (Conformers/Zipformers) and Large Language Model (LLM) rescoring techniques.

Target of this experiments is for research framing with Dr.Min Lee, reproduce the baseline and propose some method ideas, submit for Interspeech 2026

2 Dataset Overview

The dataset consists of neural recordings during a speech task, focusing on high-accuracy decoding for a speech-to-text Brain-Computer Interface (BCI).

- **Source:** Dryad Repository
- **Baseline Performance:**
  - 9.1% Word Error Rate (WER) with a 50-word vocabulary.
  - 23.8% WER with a 125,000-word vocabulary.
- **Total Samples:** 10,880 total trials.
- **Partitioning:**
  - Train: 8,800 samples (8,709 unique sentences).
  - Test: 880 samples (877 unique sentences).
  - Competition Holdout: 1,200 samples (Unlabeled).

Neural activity was recorded from the **motor cortex** with a bin size of **20 ms**. The input features are extracted from 128 channels across two primary modalities:

The **Competition Dataset** format streamlines training by:

- Extracting only the "Go" period (actual speech).
- Normalizing via `blockIdx` (z-score scaling).
- Consolidating features into a tensor of shape $[Batch, 256, Time]$, where 256 represents 128-ch `tx1` + 128-ch `spikePow`.
- Labeling with 41 unique phoneme IDs (0-40) with 500-token padding.

—

Table 1: Neural Feature Definitions

Feature	Meaning	Details
spikePow	Spike band power	Mean squared voltage, high-pass filtered at 250 Hz, linear regression denoised.
tx1 - tx4	Threshold crossing	Counts of voltage crossings at -3.5, -4.5, -5.5, -6.5 × RMS threshold.

3 Baseline Approach (Hybrid)

The initial reproduction followed the SpeechBCI 2024 pipeline:

1. **Acoustic Model:** RNN (GRU) + Connectionist Temporal Classification (CTC) loss.
2. **First Pass:** Beam Search + 5-gram Language Model (LM).
3. **Second Pass:** N-best hypotheses rescoring using LLMs (OPT 6.7B or Llama).

—

4 Experimental Log and Results

The following table summarizes the history of model iterations ("Submissions") and their performance metrics.

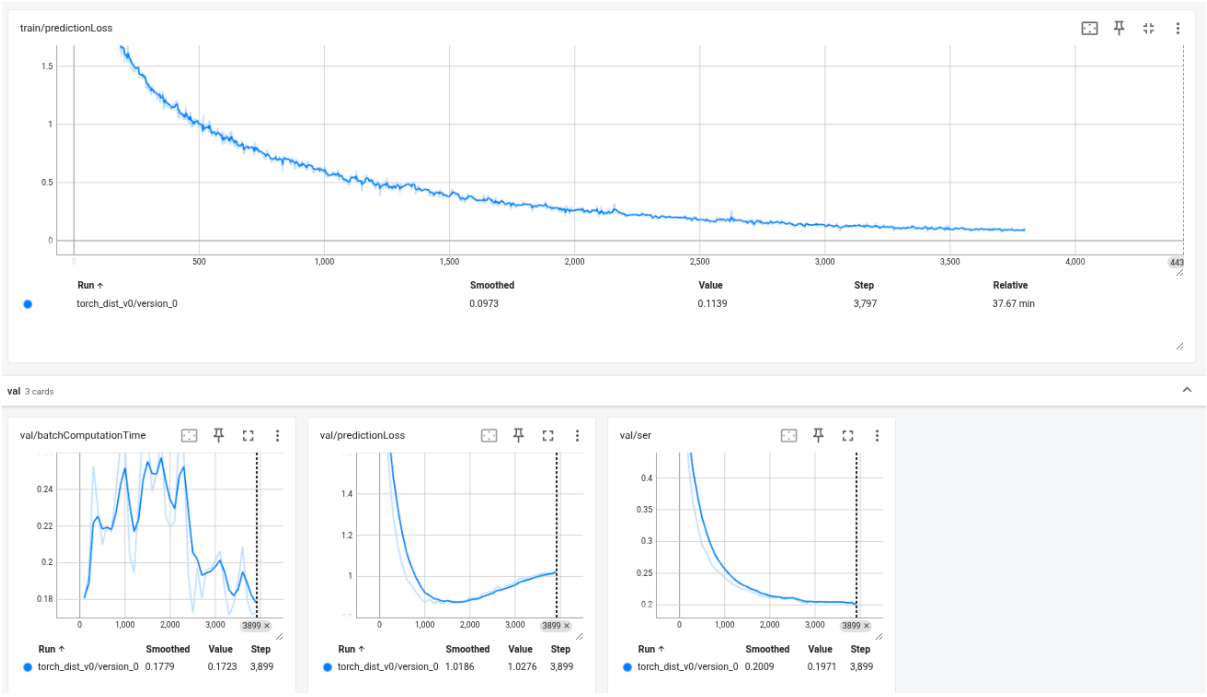

Figure 1: Training interface

ID	Model/Backbone	Approach	Result/WER
Sub 1	Baseline Text	Leaderboard submission	15%
Sub 3	GRU + Llama Rescore	Reproduced 2024 Baseline	9.76%
Sub 4	Zipformer	Heavy backbone experiment	17%
Sub 5	Conformer (Deep)	170M params	Overfit
Sub 6a-d	Conformer (Scaled)	Reduced params (33M), added layers	Slow convergence
Sub 7a	SwiGLU + RoPE + MixUp	Advanced Regularization	Instability
Sub 7b	Parameter-Limited	Params $\leq N \times 3000$ (24M)	CER 25% (Stalled)
Sub 8	Dual-Stream Cross-Attention	tx1/spikePow attention fusion	Failed
Sub 8a/b	Dual-Stream GRU	Pipeline parity with architecture swap	9.38%
Sub 14	WhisperBrain	Whisper decoder as LM	Failed (> 200%)
Sub 16	Dual Conformer CTC	Gaussian smoothing + LayerDrop	Failed
Sub 17	Enhanced Dual-Stream	Cross-stream attention + AdamW	Progressing...

Rank	Participant team	WordErrorRate (↑)	Last submission at
1	Brainaudio (Lightbeam Unidirectional)	4.98	7 days ago
2	Brain-to-Text (BIT)	5.10	5 months ago
3	DCoND-LIFT (DCoND-LIFT)	5.77	2 years ago
4	ZJU_BBIC (BSF)	6.92	4 months ago
5	neurips_13184 (neurips_13184_tformer_bidirect)	7.98	6 months ago
6	Linderman_Lab (Modified NPTL Pytorch RNN)	8.00	1 year ago
7	TeamCyber	8.26	2 years ago
8	PIRL	8.31	2 years ago
9	Stanford Silent Speech (10 x (RNN + 5-gram) + LISA)	8.93	2 years ago
10	MetaMason (Hybrid with LLM)	9.38	13 days ago
11	NPTL Pytorch Baseline (5gram + OPT6B)	9.76	2 years ago

Figure 2: Leaderboard results

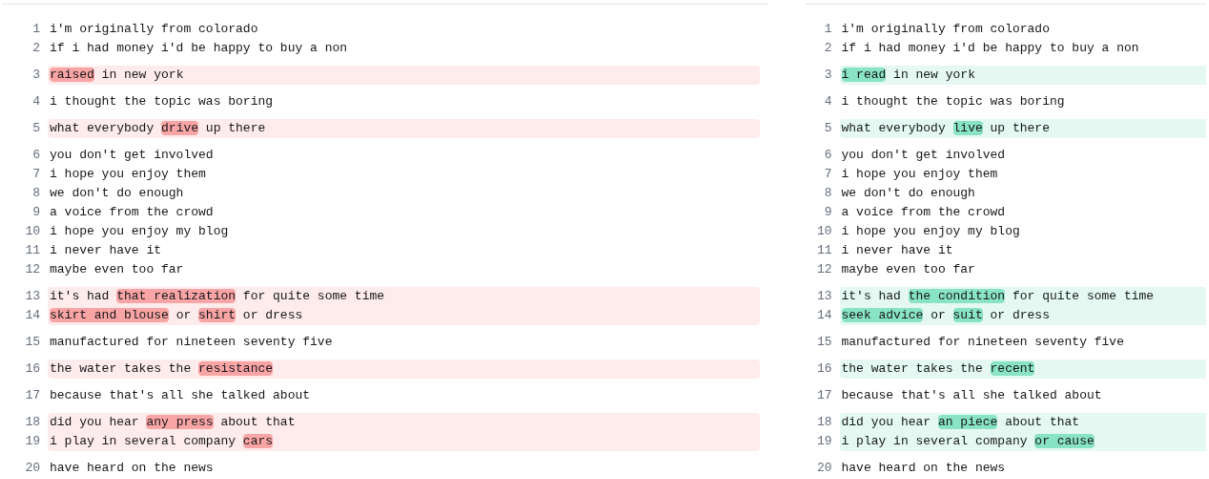

Figure 3: Investigation dashboard

—

5 Evaluation Strategies

We implemented four distinct evaluation scripts to maximize the utility of Large Language Models in the decoding process:

1. **Standard Eval (`03_eval_v2.py`):** 5-gram LM + OPT-6.7B rescoring.
2. **Whisper LM (`03_eval_v2_whisper_lm.py`):** Uses the Whisper decoder as a speech-domain language model by feeding silent audio to the encoder.
3. **Strong LLM (`03_eval_v2_llm_strong.py`):** Two-stage process using Llama-3.1-8B for perplexity scoring followed by **Instruction Correction** to fix common phoneme confusion.
4. **Triple-LM (`03_eval_v2_combined_40g.py`):** Ensemble of 5-gram, Whisper, and Llama-70B. Requires 2 × 40GB GPUs.

6 Future Work and Backlog

- Implement **Test-Time Augmentation (TTA)** with white noise averaging.
- Refactor model loading for WandB integration.
- Investigate Phoneme-to-Phoneme mapping codecs.
- Optimize "Closed 50 Vocabulary" prediction using constrained beam search.

7 Project Backlog and Future Directions

To further optimize the Word Error Rate (WER) and address current bottlenecks such as overfitting and stability, the following tasks and research ideas have been identified:

7.1 Current Backlog

- ☐ **Data Infrastructure:** Synchronize the processed dataset to the Basepod environment and conduct a thorough review of all module-specific README files.

☐ **Evaluation Optimization:** Investigate the closed 50-vocabulary prediction task and implement a constrained beam search for this subset.

☐ **Codebase Refactoring:** Decouple the model loading logic from the training loop and integrate **Weights & Biases (WB)** for real-time experiment tracking and artifact versioning.

☐ **Regularization and Stability:**

- Mitigate overfitting via aggressive data augmentation and model scaling.
- Implement **Test-Time Augmentation (TTA)**: Instead of a single forward pass, execute N passes with injected white noise and average the logits in log-probability space (potential 5-10% WER reduction without retraining).

☐ **LLM Integration:** In-depth investigation of LLM-based rescoring methods and leveraging Large Language Models for final transcript correction.

☐ **Hyperparameter Tuning:** Systematically increase epoch counts and parameter scales; explore pre-trained neural weights and codec mapping.

☐ **Ablation Analysis:** Document why standard ASR models (e.g., Whisper) are ill-suited for raw neural-to-text tasks without specialized encoders.

Table 2: Proposed Research Directions

Focus Area	**Proposed Hypothesis / Implementation**
Feature Decoupling	Separate `tx1` (sparse spike counts) and `spikePow` (continuous energy) into parallel encoders to learn distinct temporal dynamics before fusion.
Constrained Decoding	Apply a 50-word fixed vocabulary constraint during the beam search phase to benchmark against the baseline's restricted set performance.
Backbone Evolution	Transition from the current GRU-based architecture to a **Conformer** backbone while maintaining the stability of the hybrid CTC pipeline.
Ensemble Methods	Evaluate cross-model ensembling combining multiple LLM rescorers (e.g., Llama + OPT) to stabilize the N-best hypothesis selection.
Augmentation Logic	Formulate a theoretical justification for each configuration parameter and augmentation choice to be included in the final manuscript.

8 Discussion: Why Whisper ASR fails?

Preliminary results from Submission 14/15 indicate that using a frozen Whisper decoder yields a WER $> 200\%$. We hypothesize that:

1. The distribution of latent neural features significantly differs from the speech-derived Mel-spectrograms Whisper was trained on.
2. The limited 8,800 samples are insufficient to "re-align" the Whisper decoder to the new neural modality without catastrophic forgetting or extreme overfitting.