A Dual-Stream GRU-Conformer Architecture for Brain-to-Text Decoding from Utah Array Recordings

Anonymous submission to Interspeech 2026

Abstract

Decoding intended speech from intracortical neural signals is critical for restoring communication in individuals with amyotrophic lateral sclerosis (ALS). Existing approaches concatenate multiunit threshold crossings and spike band power into a single stream, conflating two complementary neural modalities. We propose a Dual-Stream GRU-Conformer that encodes each feature independently through parallel bidirectional GRU branches, exchanges cross-modal information via cross-stream attention, and fuses representations through a Conformer encoder. At inference, a triple-LM rescoring pipeline combining a 5-gram language model, Whisper-large-v3, and Qwen2.5-72B-Instruct reranks beam search hypotheses. Evaluated on the Brain-to-Text Benchmark '24, our system achieves 9.38% WER, outperforming the NPTL baseline (9.76%).

Index Terms: human-computer interaction, ulta Array, brain-to-text

1. Introduction

Biosignal-enabled spoken communication has emerged as a compelling interdisciplinary frontier, bridging the domains of biosignal processing and speech technology to restore or augment human communication. Research in this space broadly follows two complementary directions. The first is Silent Speech Interfaces (SSI), which exploits facial and neck muscle activity measured via electromyography (EMG) to decode speech and emotional state even in the complete absence of vocalization, with applications ranging from speech-assistive devices to emotion-aware systems [1, 2]. The second—and the focus of this work—is brain-to-speech decoding, which aims to reconstruct intended speech directly from neural signals recorded from the brain. The motivating application is urgent: neurological conditions such as amyotrophic lateral sclerosis (ALS) and brainstem stroke can deprive individuals of voluntary movement, rendering them locked-in—fully conscious but unable to communicate. Speech brain-computer interfaces (BCIs) offer a pathway to restore communication by decoding what a person intends to say directly from their neural activity and presenting that message as synthesized speech or text [3]. Improving the accuracy and robustness of such systems is therefore not merely a technical challenge but a clinical imperative. Neural signals for brain-to-speech decoding span a broad modality spectrum. Non-invasive approaches such as electroencephalography (EEG), magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI) offer the advantage of accessibility, but suffer from low spatial resolution, susceptibility to noise, and limited vocabulary coverage—typically restricted to closed-vocabulary settings of around 50 words [4]. Indeed, recent work has demonstrated that several prominent EEG-to-text models exhibit performance on purely random noise inputs comparable to their performance on actual EEG, casting doubt on whether these models genuinely learn from brain signals at all [5]. Semi-invasive approaches using electrocorticography (ECoG) provide improved signal fidelity [**?**], while fully invasive methods—including the Utah intracortical microelectrode array (MEA) and, more recently, Neuralink—afford the highest spatial and temporal resolution by recording action potentials directly from individual cortical neurons. This paper focuses on the invasive BCI paradigm using Utah array recordings, which capture multiunit threshold crossings and spike band power from the ventral premotor cortex (area 6V) at a resolution far exceeding non-invasive alternatives. The dataset underpinning our experiments is the openly released corpus accompanying the Brain-to-Text Benchmark '24 [3], which was compiled from a participant with ALS who has lost intelligible speech. The dataset was originally introduced alongside A High-Performance Speech Neuroprosthesis [6], in which intracortical Utah array recordings enabled large-vocabulary decoding with a word error rate of 23.8% across a 125,000-word vocabulary—the first successful demonstration of this scale. The core decoding problem is to map a variable-length time series of neural spike activity to a text sequence, a challenging sequence-to-sequence task that must contend with scarce paired data, high-dimensional noisy inputs, and strong inter-session variability. Existing approaches to this problem predominantly follow a cascade architecture: a recurrent neural network (RNN) first maps neural activity to phoneme probabilities, which a language model (LM) subsequently resolves into words [3]. Competitive entries in the Brain-to-Text '24 challenge further improved upon this paradigm through large-model ensembling, diphone training objectives, and learning rate scheduling. Yet the baseline RNN encoder remains a bottleneck: it processes spike features sequentially and struggles to exploit complementary information encoded in both the spike count and the spike band power simultaneously. To address this, we propose a Dual GRU–Conformer architecture that encodes the two principal neural features—spike count and spike band power—in parallel through independent GRU streams before fusing them with a Conformer [7] encoder that jointly captures local temporal dynamics and long-range contextual dependencies. For decoding, rather than relying on a traditional n-gram language model, we leverage the pre-trained Whisper decoder [8] as a powerful neural language model, enabling more flexible and contextually grounded text generation. Our approach is evaluated on the Brain-to-Text '24 benchmark and demonstrates competitive word error rates, validating the benefit of parallel dual-stream feature encoding for intracortical speech decoding.

2. Related Work

2.1. Intracortical Speech Decoding: The Cascade Paradigm

The dominant paradigm for decoding speech from Utah intracortical microelectrode array recordings follows a two-stage cascade architecture. This approach was established by Willett et al. [6], who demonstrated the first large-vocabulary brain-to-text system: a five-layer RNN maps binned multiunit threshold crossings and spike band power to per-timestep phoneme logits via a CTC loss, after which a 5-gram language model performs beam search to generate sentence hypotheses, and an OPT large language model rescores the candidates. This system achieved a word error rate of 9.1% on a 50-word closed vocabulary and 23.8% on a 125,000-word open vocabulary at a decoding throughput of 62 words per minute—the first demonstration of large-vocabulary continuous speech decoding from intracortical signals. The publicly released code and dataset subsequently became the foundation for the Brain-to-Text Benchmark '24 [3].

2.2. Brain-to-Text Benchmark '24: Competition Entries

The Brain-to-Text Benchmark '24 [3] formalized the evaluation of neural speech decoding algorithms on a held-out test set of 1,200 sentences, measured by word error rate. Four competitive entries each improved upon the PyTorch baseline (WER 9.76%) by June 2024, and their approaches illuminate the key directions in the field.

The **NPTL baseline** [6] uses a unidirectional RNN encoder with a 5-gram language model and OPT rescoring, yielding approximately 9.46% WER after training optimizations such as learning rate scheduling and a diphone-based training objective.

Stanford LISA (3rd place, 8.93% WER) [9] introduced LLM Integrated Scoring Adjustment (LISA) as part of the MONA LISA framework, originally developed for EMG-based silent speech interfaces. Applied to neural decoding, LISA uses a fine-tuned GPT-3.5 to rerank an ensemble of 5-gram beam search candidates, effectively replacing fixed n-gram scoring with a neural language model at the hypothesis selection stage.

TeamCyber / CIBR Okubo Lab (2nd place, 8.26% WER) [10] adopted a bidirectional GRU as the neural encoder in place of the unidirectional RNN baseline, combined with model ensembling and Llama-2-7B for hypothesis selection. The use of a bidirectional recurrent architecture—while precluding real-time deployment—demonstrated that increased contextual access in the encoder meaningfully improves phoneme decoding accuracy.

DCoND-LIFT (1st place, 5.77% WER) [11], submitted by the Shlizerman Lab, achieved the strongest published result on the benchmark. Their Divide-and-Conquer Neural Decoder (DCoND) replaces single-phoneme targets with *diphones*—pairwise phoneme transitions yielding approximately 1,600 classes—and marginalizes over the preceding phoneme to recover per-phoneme probabilities. This context-aware formulation reduced the phoneme error rate from 16.62% to 15.34%. Combined with an ensembling strategy that aggregates predictions from multiple independent decoders and passes both decoded phonemes and transcription candidates to a fine-tuned GPT-3.5, DCoND-LIFT reached 5.77% WER, a 41% relative reduction over the baseline.

UCLA NECL (4th place, 5.68% WER) [3] employed a causal Transformer encoder paired with Llama-3 8B as the language model decoder. While no accompanying publication is available at the time of this submission, the result indicates that causal Transformer architectures combined with large open-source LLMs can be competitive with RNN-based systems when adequately optimized.

The competition summary [3] identified two key findings. First, ensembling multiple independent decoders followed by fine-tuned LLM merging was the single highest-impact improvement across all top entries. Second, and perhaps surprisingly, direct architectural substitution of the RNN with Transformers or deep state-space models did not yield consistent improvements at the encoder stage, suggesting that RNNs may retain advantages for neural spike sequence modeling at current data scales.

2.3. Feature Representation in Intracortical Decoding

The Willett et al. dataset provides two complementary neural features at each 10 ms time bin: **multiunit threshold crossings** (`tx1`), a proxy for spike count, and **spike band power** (`spikePow`), a measure of high-frequency local field potential energy. Prior work consistently concatenates these two features before feeding them into the encoder [6, 11, 10], treating them as a single unified representation. While this simplifies the architecture, it conflates two signals that may encode speech kinematics at different temporal scales and with different noise characteristics: threshold crossings capture sharp transient spiking activity, whereas spike band power reflects a smoother, more sustained envelope of neural activity. Prior studies on intracortical motor decoding [12] have noted that these two features are complementary rather than redundant, motivating a more deliberate treatment of each modality during encoding.

2.4. Language Models for Neural Speech Decoding

The role of language models in neural speech decoding has evolved substantially from n-gram rescoring to full neural LM integration. The original pipeline [6] uses a 5-gram language model for beam search followed by OPT for rescoring. Benster et al. [9] and Li et al. [11] both demonstrated that replacing or augmenting n-gram models with fine-tuned LLMs provides substantial gains, with a recurring insight being that neural LMs can leverage both the decoded phoneme sequence and the textual n-gram candidates simultaneously, correcting systematic errors introduced at the beam search stage [3].

An alternative to LLM rescoring is to use a pre-trained speech model's decoder directly as the language model. Whisper [8], trained on 680,000 hours of transcribed speech with a sequence-to-sequence Transformer architecture, offers a decoder whose conditional distributions capture both linguistic and acoustic priors over word sequences. Several recent works in non-invasive brain-to-speech decoding have exploited the Whisper decoder as a plug-in language model [13], suggesting that its internally learned priors generalize beyond acoustic inputs. Compared to a fine-tuned GPT-3.5, the Whisper decoder requires no separate fine-tuning stage, is differentiable end-to-end, and can be jointly optimized with the neural encoder—properties that motivate its adoption in our proposed system.

3. Proposed Method

3.1. System Overview

Our system follows the cascade paradigm established in prior work [6]: a neural encoder maps intracortical spike sequences to per-timestep phoneme logits via CTC, which are then decoded

using a language model pipeline. We introduce modifications at both stages: a dual-stream GRU-Conformer encoder that processes the two neural feature modalities separately before fusing them, and a triple-LM rescoring pipeline that combines n-gram, Whisper, and LLaMA scores at inference time.

3.2. Input Preprocessing

Each input trial is represented as a matrix $\mathbf{X} \in \mathbb{R}^{T \times 256}$, where T is the number of 10 ms time bins and 256 is the number of electrode channels. The first 128 channels correspond to multiunit threshold crossings (`tx1`) and the remaining 128 channels correspond to spike band power (`spikePow`). Following the baseline, a Gaussian smoothing kernel with width $\sigma = 2$ bins is applied along the time axis prior to any further processing.

To compensate for inter-session non-stationarities, a per-day affine calibration layer is applied to the smoothed signal:

$$\tilde{\mathbf{x}}_t = \phi\left(\mathbf{x}_t \mathbf{W}_d + \mathbf{b}_d\right), \quad (1)$$

where $\mathbf{W}_d \in \mathbb{R}^{256 \times 256}$ and $\mathbf{b}_d \in \mathbb{R}^{256}$ are trainable parameters specific to recording day d, initialized to the identity and zero respectively, and $\phi(\cdot)$ denotes the Softsign activation function. After calibration, the signal is split along the channel axis into two streams: $\mathbf{X}^{\text{tx1}} \in \mathbb{R}^{T \times 128}$ and $\mathbf{X}^{\text{sp}} \in \mathbb{R}^{T \times 128}$.

Both streams are then temporally compressed using a sliding window unfold operation with kernel size $K = 32$ bins and stride $S = 4$ bins, yielding strided representations of dimension $128K$ at each output time step. This produces an effective frame rate of approximately 25 Hz and introduces local temporal context into each frame.

3.3. Dual-Stream GRU Encoder

Rather than concatenating the two feature modalities before encoding, we process each stream independently with a separate bidirectional GRU:

$$\mathbf{H}^{\text{tx1}} = \text{BiGRU}_{\text{tx1}}(\mathbf{X}^{\text{tx1}}), \quad (2)$$

$$\mathbf{H}^{\text{sp}} = \text{BiGRU}_{\text{sp}}(\mathbf{X}^{\text{sp}}), \quad (3)$$

where each BiGRU has 5 layers and a hidden size of 512 per direction (1024 after concatenating forward and backward states), matching the total parameter budget of the single-stream baseline. Weight matrices are initialized using orthogonal initialization for recurrent connections and Xavier uniform for input projections. Layer normalization is applied to the output of each GRU to stabilize training.

3.4. Cross-Stream Attention

After the GRU stage, we introduce a bidirectional cross-stream attention module to allow each modality to attend to information in the other. Concretely, given the normalized GRU outputs $\hat{\mathbf{H}}^{\text{tx1}}$ and $\hat{\mathbf{H}}^{\text{sp}}$, we compute:

$$\tilde{\mathbf{H}}^{\text{tx1}} = \hat{\mathbf{H}}^{\text{tx1}} + \text{MHA}\left(\hat{\mathbf{H}}^{\text{tx1}}, \hat{\mathbf{H}}^{\text{sp}}, \hat{\mathbf{H}}^{\text{sp}}\right), \quad (4)$$

$$\tilde{\mathbf{H}}^{\text{sp}} = \hat{\mathbf{H}}^{\text{sp}} + \text{MHA}\left(\hat{\mathbf{H}}^{\text{sp}}, \hat{\mathbf{H}}^{\text{tx1}}, \hat{\mathbf{H}}^{\text{tx1}}\right), \quad (5)$$

where $\text{MHA}(Q, K, V)$ denotes multi-head attention with 8 heads and layer normalization applied to the query and key-value inputs before attention. The residual formulation ensures that each stream retains its own representation while being enriched by complementary information from the other.

3.5. Conformer Encoder

The cross-attended streams are concatenated and projected to a common dimension $D_c = 512$ via a linear layer followed by layer normalization:

$$\mathbf{Z} = \text{LayerNorm}\left(\left[\tilde{\mathbf{H}}^{\text{tx1}}; \tilde{\mathbf{H}}^{\text{sp}}\right] \mathbf{W}_{\text{proj}}\right). \quad (6)$$

The fused representation $\mathbf{Z}$ is then passed through $N = 2$ Conformer blocks [7]. Each Conformer block follows the macaron structure:

$$\mathbf{z}' = \text{FFN}_{0.5} \rightarrow \text{MHSA} \rightarrow \text{DepthConv} \rightarrow \text{FFN}_{0.5} \rightarrow \text{LayerNorm}, \quad (7)$$

where the two feed-forward sub-layers each contribute with a half-step residual weight of 0.5, the multi-head self-attention (MHSA) has 8 heads, and the depthwise convolutional module uses a kernel size of 15 with a gated linear unit (GLU) activation. The convolutional module is designed to capture local phoneme-boundary dynamics that complement the global context modeled by MHSA.

The output of the Conformer stack is projected to $C+1$ logits via a linear layer, where $C = 40$ is the number of phoneme classes and the additional class corresponds to the CTC blank token.

4. Experiments

4.1. Dataset

The dataset consists of 12,100 sentences of intended speech recorded from a single participant with late-stage Amyotrophic Lateral Sclerosis (ALS), with neural activity captured via 256 electrodes in speech-related motor cortex regions. The data is split into 8,800 training and 880 test samples, evaluated using Word Error Rate (WER) on 1,200 held-out sentences.

The input is constructed by concatenating two neural signal features: spike counts (`tx1`), representing threshold-crossing firing events across 128 channels, and spike band power (`spikePow`), capturing signal energy in the $\sim$250–5000 Hz range across another 128 channels. This yields an input tensor of shape $[B,\ 256,\ T]$, with each time bin corresponding to 20 ms. Target labels are phoneme ID sequences drawn from a vocabulary of 41 phonemes (IDs 0–40), padded to a fixed length of 500 tokens.

4.2. Data Augmentation

We apply a four-stage augmentation pipeline during training to improve generalisation and reduce overfitting to the relatively small number of recorded trials. First, additive noise is injected by superimposing white noise and a per-channel constant offset onto the input, simulating common sources of recording variability such as thermal noise and DC drift. Second, Temporal CutMix [14] pastes a contiguous temporal segment from one sample into another, preventing the model from relying on absolute temporal position within a trial. Third, Input Mixup [15] linearly interpolates pairs of training samples with a mixing coefficient drawn from a $\text{Beta}(0.3, 0.3)$ distribution, encouraging smoother decision boundaries in neural feature space. Fourth, SpecAugment [16] masks two random time spans (up to 20 frames each) and two random channel spans (up to 40 channels each), forcing the model to decode speech from partial observations and improving robustness to electrode drop-out.

4.3. Optimisation

The model is optimised with AdamW [17] (β_1=0.9, β_2=0.98, ε=10^{-8}, weight decay = 10^{-4}). The learning rate follows a cosine decay schedule preceded by a 5,000-step linear warmup from zero to a peak of 3×10^{-4}, then decays to 1% of the peak value by the end of training. The warmup phase allows the model parameters — particularly the randomly initialised calibration matrices — to reach a stable operating regime before the learning rate begins to decay.

4.4. Triple-LM Rescoring

At inference, a **triple-LM rescoring** scheme is applied to rerank the 100-best hypotheses produced by beam search under a Kaldi 5-gram language model. Each hypothesis receives a composite score that linearly combines four terms: the CTC log-likelihood from the acoustic model, the 5-gram LM score, the log-probability assigned by Whisper-large-v3 [18] acting as a speech-domain language model (using only its text decoder, with a silent audio input), and the log-perplexity under Qwen2.5-72B-Instruct [19] quantised to 4-bit NF4 and distributed across two 40 GB GPUs. The four weights (α=0.5, β_1=0.2, β_2=0.4, β_3=0.4) are tuned on the held-out validation set.

Incorporating both Whisper and the large instruction-tuned LLM as rescoring components allows the pipeline to leverage complementary sources of linguistic knowledge: Whisper provides strong priors over phonetically plausible word sequences learned from large-scale speech data, while the LLM enforces broader grammatical and semantic coherence. Optionally, a second-stage **instruction correction** pass prompts the LLM to repair systematic phoneme-level transcription errors (e.g., converting *"thee"* to *"the"* or *"wun"* to *"one"*) while preserving the utterance length and overall meaning, providing a final layer of post-hoc linguistic normalisation.

4.5. Results

Table 1 reports WER on the held-out test set. Our proposed dual-stream GRU-Conformer encoder achieves a WER of 9.38%, outperforming the NPTL PyTorch Baseline (5-gram + OPT-6B) which scores 9.76%.

Table 1: *Word Error Rate (%) on the held-out test set.*

Method	WER (%)
NPTL PyTorch Baseline (5-gram + OPT-6B)	9.76
Dual-stream GRU-Conformer (ours)	**9.38**

5. Conclusion

In this paper, we presented a Dual-Stream GRU-Conformer system for intracortical speech decoding that addresses a key limitation of prior work: the indiscriminate concatenation of two complementary neural feature modalities. By encoding multiunit threshold crossings and spike band power through independent bidirectional GRU branches before exchanging information via cross-stream attention and fusing representations in a shared Conformer encoder, our architecture allows each modality to develop a specialized representation while benefiting from cross-modal context. This design is complemented at inference by a triple-LM rescoring pipeline that leverages a 5-gram language model alongside Whisper-large-v3 and Qwen2.5-72B-Instruct, providing both speech-domain and broad linguistic priors for hypothesis reranking. On the Brain-to-Text Benchmark '24, our system achieves a WER of 9.38%, improving over the NPTL baseline of 9.76%.

6. References

[1] J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. M. M. Donas, J. L. Perez-Cordoba, and A. M. Gomez, "Silent speech interfaces for speech restoration: A review," *IEEE Access*, vol. 8, pp. 177 995–178 021, 2020.

[2] J. S. Brumberg, A. Nieto-Castanon, P. R. Kennedy, and F. H. Guenther, "Brain-computer interfaces for speech communication," *Speech Communication*, vol. 52, no. 4, pp. 367–379, 2010.

[3] F. R. Willett, J. Li, T. Le, C. Fan, M. Chen, E. Shlizerman, Y. Chen, X. Zheng, T. Singer-Clark, T. Benster, H. D. Lee, M. Kounga, E. K. Buchanan, D. Zoltowski, S. W. Linderman, and J. M. Henderson, "Brain-to-text benchmark '24: Lessons learned," 2024. [Online]. Available: https://arxiv.org/abs/2412.17227

[4] H. Liu, H. Luo, J. Zhou, and L. Dai, "EEG2TEXT: Open vocabulary EEG-to-text decoding with EEG pre-training and multi-view transformer," 2024. [Online]. Available: https://arxiv.org/abs/2405.02165

[5] H. Jo, Y. Yang, J. Han, Y. Duan, H. Xiong, and W. H. Lee, "Are EEG-to-text models working?" 2024. [Online]. Available: https://arxiv.org/abs/2405.06459

[6] F. R. Willett, E. M. Kunz, C. Fan, D. T. Avansino, G. H. Wilson, E. Y. Choi, F. Kamdar, M. F. Glasser, L. R. Hochberg, S. Druckmann, K. V. Shenoy, and J. M. Henderson, "A high-performance speech neuroprosthesis," *Nature*, vol. 620, no. 7976, pp. 1031–1036, 2023.

[7] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[9] T. Benster, G. Wilson, R. Elisha, F. R. Willett, and S. Druckmann, "A cross-modal approach to silent speech with LLM-enhanced recognition," 2024. [Online]. Available: https://arxiv.org/abs/2403.05583

[10] Y. Chen, X. Zheng, and T. S. Okubo, "TeamCyber: Second-place solution for the brain-to-text benchmark '24 (bidirectional GRU + Llama-2 ensembling)," GitHub repository. https://github.com/CIBR-Okubo-Lab/speechBCI_2024, 2024, submitted to the Brain-to-Text Benchmark '24 competition. Result reported in [3].

[11] J. Li, T. Le, C. Fan, M. Chen, and E. Shlizerman, "Brain-to-text decoding with context-aware neural representations and large language models," 2024. [Online]. Available: https://arxiv.org/abs/2411.10657

[12] F. R. Willett, D. T. Avansino, L. R. Hochberg, J. M. Henderson, and K. V. Shenoy, "High-performance brain-to-text communication via handwriting," *Nature*, vol. 593, no. 7858, pp. 249–254, 2021.

[13] A. Défossez, C. Caucheteux, J. Rapin, O. Kabeli, and J.-R. King, "Decoding speech perception from non-invasive brain recordings," *Nature Machine Intelligence*, vol. 5, no. 10, pp. 1097–1107, 2023.

[14] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Training strategy that makes use of sample mixing and its cutout regularization in a single framework," in *Proc. ICCV*, 2019, pp. 6023–6032.

[15] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. ICLR*, 2018.

[16] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[17] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[18] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023, pp. 28 492–28 518.

[19] Qwen Team, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2025.