Prototyping Co-Control Brain–Computer Interfaces Through Brain-to-Text

Anonymous Author(s)
Affiliation
Address
email

Abstract

Co-control between brain–computer interface (BCI) users and intelligent systems requires effective fusion across specialized modules. In Brain-to-Text BCIs, neural decoders (NDs) map neural activity to text token sequences, while language models (LMs) provide compensatory linguistic constraints when ND predictions are uncertain. Integration is typically achieved through probabilistic fusing, yet current systems are poorly calibrated: they encode some notion of confidence in the output distribution but fail to discriminate reliably between correct and incorrect predictions. Through oracle manipulations of the predicted probability distribution, while keeping the same MLE solution, across *over-confident*, *uncertainty-aware*, and *alternative-rich* regimes, we demonstrate that a better calibrated system can substantially improve performance. These results highlight the need for neural decoders to communicate both uncertainty and informative alternatives in order to enable robust multi-module co-control.

1 Introduction

Co-control between a BCI user and an intelligent system enables the execution of complex tasks by leveraging three key design principles: (1) modularization to facilitate specialization of distinct functions, (2) training or fine-tuning on distinct datasets aligned to modularized functionality, and (3) system extensibility as new modules are introduced [1, 2]. For BCI users, who primarily rely on these systems to restore movement or communication, it is critical that their intent is preserved and accurately conveyed through co-control [3–8].

A prominent example system is the Brain-to-Text BCI [9–12], where a ND transcribes continuous neural activity into sequences of sub-word text tokens (e.g., character, phoneme) and benefits from integration with a language model (LM) that prunes invalid words and promotes linguistically probable ones to generate the best hypothesis. These systems incorporate distinct design choices, which include: (1) the ND models the mapping between neural activity and text tokens sequences, while the LM models the sequential statistics of text token sequences; (2) decoders are trained on limited, user-specific neural data, whereas LMs are trained on abundant text corpora; and (3) the decoded text can support additional language-based modules, such as vision–language–action models, that break down abstract commands in the context of a visual scene [13–15]. While (1) and (2) define current systems, (3) highlights an emerging design opportunity: extending Brain-to-Text BCIs beyond transcription into multi-modal co-control frameworks. In such settings, accuracy at the text level is especially important, as even small errors could cascade when passed to downstream modules, distorting meaning or triggering unintended actions. Across all systems, the language model must not "hallucinate" content beyond the user's intent. This raises a central question: how can co-control balance modular specialization with faithful preservation of user intent? In this work, we use Brain-to-Text BCIs as a case study to examine how information shared across modules (ND and LM) shapes reliable co-control.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

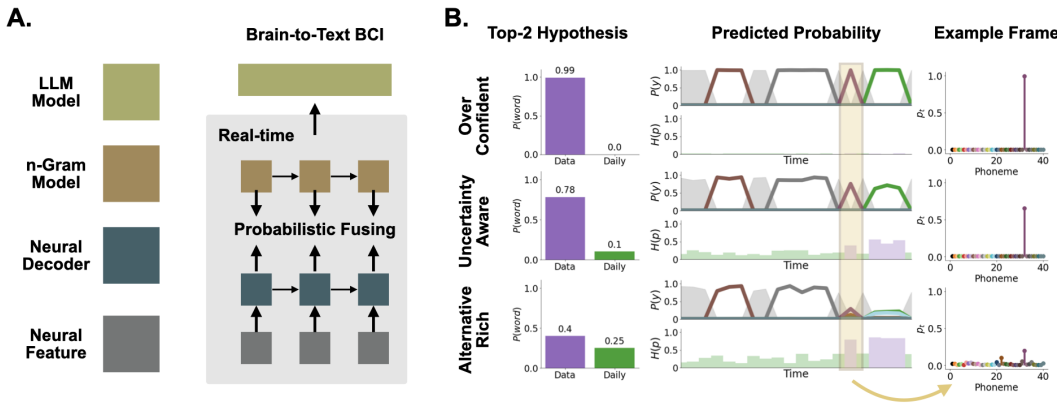

Figure 1: Brain-to-Text BCI with probabilistic fusion between neural decoder and language model. A) Schematic of modules and information flow. B) Illustrative decoder output distribution showing prediction of 'Data' instead of the true word 'Daily', with the error frame highlighted in purple.

2 Related Work

A central challenge in training high-performance Brain-to-Text neural decoders lies in mapping continuous neural activity into symbolic text tokens states [10, 11]. This task is complicated by noise and variability arising from recording artifacts, spike-train stochasticity, temporal variability, and individual differences in movement style [16–18]. Automatic speech recognition (ASR) faces an analogous challenge [19]: although the input signal differs, both domains must contend with noisy, variable data and produce symbolic text outputs. A shared feature is that model predictions ultimately take the form of probability distributions over sub-word text tokens, creating opportunities to transfer insights from ASR to Brain-to-Text BCIs.

ASR research has extensively explored language model (LM) integration. Conservative approaches use probabilistic fusing with beam search via weighted finite-state transducers (WFSTs) to generate top-K hypotheses [20, 21]. More recent methods employ large LMs to rescore the tok-K hypotheses from WSFTs [22, 23] or, more aggressively, to generate transcriptions directly from the top-K hypotheses [24–26]. A recurring lesson is that top-K hypotheses often contain informative alternatives that can improve final transcription quality [24, 27, 23]. However, substantial uncertainty remains unresolved without additional support from acoustic information. Directly fusing acoustic features into LLMs is challenging due to modality gaps. As a result, a promising direction is to leverage model uncertainty to dynamically weight contributions from different modules, thereby balancing modalities and improving integration [24, 28]. In this paper, we advance these ideas to Brain-to-Text BCIs, showing how neural decoder uncertainty and alternative hypotheses could support more reliable fusion between the decoder and language model as a prototype for co-control.

3 Method

Brain-to-Text Neural Decoder. Results reported in this paper are based on a gated recurrent unit (GRU) trained with the Connectionist Temporal Classification (CTC) [29] loss on two publicly available speech-BCI datasets (Participants T12 [10] and T15 [11]). Hyperparameters were set to the empirically optimal values reported in the original studies. The decoder transcribes multichannel neural recordings $\mathbf{X}_{1:T} \in \mathbb{R}^{D \times T}$, into a target phoneme sequence $\mathbf{y}_{1:N} \in \mathcal{V}^N$ over phoneme tokens $\mathcal{V}$ (details in A). Training maximizes the sequence probability $\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{X}_{1:T})$ with frame-wise predictions $s_{1:T}$ collapsed into $\mathbf{y}$ by compressing consecutive repeats, so that each contiguous segment of a token v in time is mapped to a single symbol in the output sequence.

Probabilistic Fusion with n-gram LM via WFSTs. State-of-the-art Brain-to-Text BCIs integrate neural decoder outputs with external language models through WFSTs [20] composed of three transducers: the token graph V, which encodes decoder output probabilities as transition costs; the lexicon L, which maps token sequences to words; and the grammar G, which encodes an n-gram language model as transition costs. The overall $\text{score}(\mathbf{y}) = \log p_{\text{ND}}(\mathbf{y} \mid \mathbf{X}_{1:T}) + \beta \log p_{\text{LM}}(\mathbf{y})$ where β controls the LM contribution (Appendix A.1 for detail).

Analysis and Simulation. Sequence-level analyses were performed on the top-K hypotheses ranked by score$(\mathbf{y})$, while frame-level analyses used the most probable alignment $\hat{s}_{1:T}$. To simulate oracle decoders with identical maximum-likelihood outputs but different distributions, we manipulated the frame-level probability vector $\mathbf{p}_t = [p_t^{(1)}, p_t^{(2)}, \dots]$. Three regimes were considered: (i) an **over-confident**: where all mass is concentrated on the predicted class; (ii) an **uncertainty-aware**: where the predicted class retains its calibrated probability and residual mass is spread uniformly across others; and (iii) an **alternative-rich**: where residual mass is concentrated on plausible confusions while improbable classes are suppressed. The exact manipulation is provided in Appendix C.

4 Experimental Results

4.1 Lower-Ranked Hypotheses Contain Better Candidates

For real-time speech BCI decoding, the system ideally produces a single top-1 hypothesis that reflects the best estimate of the neural decoder, while integrating linguistic constraints through probabilistic fusing when the neural decoder is uncertain. Although post-hoc rescoring or generative error correction based on the top-K hypotheses is possible, such approaches introduce additional delay and may require user confirmation, potentially interrupting the natural flow of speech production.

Here, we analyze the correctness of the top-K hypotheses to assess whether informative alternatives are present beyond the top-1 output. Lower-ranked hypotheses correspond to sequences with reduced scores score$(\mathbf{y}) = \log p_{\text{ND}}(\mathbf{y} \mid \mathbf{X}_{1:T}) + \beta \log p_{\text{LM}}(\mathbf{y})$ with $\beta = 0.7$ for subject T12 and $\beta = 0.3$ for subject T15. Lower-ranked hypotheses may arise either from low-probability alternatives suggested by the ND or from higher probability ones down-weighted by the LM. Importantly, even tokens assigned low probability close to chance level by the ND at time t can still surface in the top-K list if strongly supported by the language model. Using $\min$ WER and Word-Recall (see Appendix B.1), we found that more accurate candidates frequently appeared among lower-ranked hypotheses, with substantial performance gaps observed between the top-1, top-10, and even top-100 outputs. This indicates that while neural decoders may generate informative alternatives, they are assigned such low probabilities that they rarely meaningfully impact real-time decoding.

Table 1: min WER and word-recall across Top-K hypotheses for two speechBCI datasets

	Participant T12		**Participant T15**	
	min WER ↓	Word-Recall ↓	min WER ↓	Word-Recall ↓
Top 1	0.204	0.197	0.059	0.056
Top 10	0.122	0.109	0.018	0.016
Top 100	0.079	0.062	0.009	0.007

4.2 Model Predictions Approximate Poorly Calibrated Systems

To better characterize the uncertainty and alternatives proposed by the neural decoder, we examined its frame-wise predictions, since the intrinsic objective of decoding is to recognize the phoneme token v over time. Specifically, we analyzed the Viterbi alignment path $\hat{s}_{1:T} = \arg\max_{s_{1:T}} p(s_{1:T} \mid \mathbf{X}_{1:T})$where, in practice, CTC-trained models often associate the maximum-likelihood sequence $\hat{y}$ with a unique high-probability alignment path $\hat{s}_{1\,T}$ [29, 30]. This alignment provides an estimate of the temporal boundaries of each token $v \in \mathcal{V}$ during the attempted speech.

We first evaluated whether frame-level uncertainty, measured by normalized entropy (0 indicating high confidence and 1 indicating high uncertainty; see example in Figure 1), is predictive of correctness. As shown in Figure 2, entropy values remain consistently close to 0 across frames, with no significant distinction between correct and incorrect predictions. This indicates that the neural decoder produces highly confident outputs regardless of accuracy.

We next investigated whether the reference token $s^\star$, the symbol from the reference transcription aligned to a given frame, occurs among the top candidates for each frame, and what margin separates the model's predicted output $\hat{s}$ from the reference token $s^\star$, a factor that directly affects the search path in LM integration [31]. We found that $\hat{s}$ is consistently associated with high probability. In incorrect frames, large probability margins remain between $\hat{s}$ and $s^\star$, even though $s^\star$ frequently appears within the top-10 candidates among the 41 phoneme tokens. This suggests that, even when

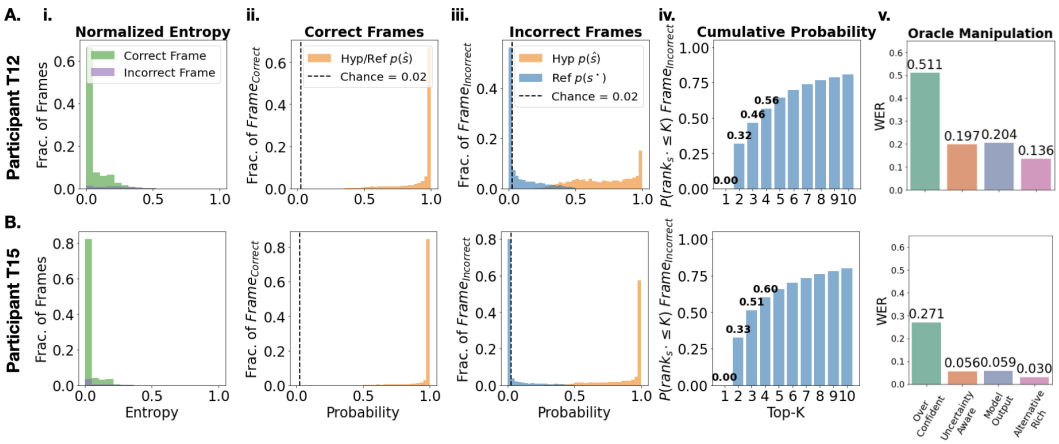

Figure 2: Frame-level analysis of model-predicted output distributions for participant T12 in A) and T15 in B). i. Normalized entropy across frames (range 0–1). ii–iii. Distribution of the predicted $p(\hat{s})$ and reference $p(s^\star)$ phoneme probability for frames with ii) correct and iii) incorrect predictions. iv. Cumulative probability for $s^\star$ appearing within the top-K in incorrect frame. v. WER for oracle manipulation of neural decoder output.

correct alternatives are present in the neural decoder output, they are often assigned significantly lower probabilities, relegating them to lower-ranked search paths during fusion.

To assess whether more effective fusion with the language model could be achieved while keeping the same MLE solution but varying the output distribution, we simulated alternative forms of the decoder's predicted distribution: *over-confident*, *uncertainty-aware*, and *alternative-rich*, and compared them against the model output. We found that probabilistic fusing with the WFST substantially outperforms the over-confident regime, as the system can exploit lower-probability alternatives when they are available. In contrast, when comparing the model's original outputs against the uncertainty-aware and alternative-rich regimes, the neural decoder most closely resembled the uncertainty-aware system. Its distributions were sharply biased toward the predicted token, with minimal probability mass assigned to plausible alternatives. As a result, the language model could not reliably detect and correct errors from the decoder probabilities, and instead tended to reinforce erroneous words.

5 Conclusions and Perspectives

In this work, we examined how probabilistic fusing between neural decoders and language models shapes performance in speech BCIs. Our analysis revealed that effective fusion is hindered by two critical limitations in the estimated probabilities that are the output of existing decoders. First, the decoder showed limited discriminability between correct and incorrect frames, producing overconfident predictions in both cases. Second, even when correct alternatives were present in incorrect frames, they were assigned very low probability and separated by large margins from the top-1 prediction, making them unlikely to surface during fusion. Oracle simulations demonstrated that a better calibrated neural decoder—one that reliably communicates both uncertainty and structured alternatives—would enable more effective probabilistic fusing, allowing language models to provide complementary corrections when uncertainty is high.

Beyond the specific case study of Brain-to-Text BCIs, our findings highlight a broader perspective: BCIs could be viewed as integrated systems of interacting modules, rather than as pipelines forming an end-to-end decoder. Probabilistic fusing exemplifies a fusing paradigm for co-control, where modules trained on different modalities or tasks collaborate to resolve ambiguity. Future BCIs may incorporate additional modules that extend beyond neural decoding. A key challenge will be learning how these modules communicate rich and informative signals, not just essentially final, confident outputs, to achieve robust collaboration. By identifying this limitation in current neural decoders and proposing a path forward, our work points toward a new direction for BCI research: moving beyond output-only training objectives to develop decoders that explicitly convey uncertainty and alternatives. Such decoders would lay the foundation for co-controlled systems that integrate heterogeneous modules, enabling flexible and reliable, faithful communication in next-generation neuroprostheses.

References

[1] Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, et al. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*, 2024.

[2] Bang Liu, Xinfeng Li, Jiayi Zhang, Jinlin Wang, Tanjin He, Sirui Hong, Hongzhang Liu, Shaokun Zhang, Kaitao Song, Kunlun Zhu, et al. Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems. *arXiv preprint arXiv:2504.01990*, 2025.

[3] MH Farhadi, Ali Rabiee, Sima Ghafoori, Anna Cetera, Wei Xu, and Reza Abiri. Human-centered shared autonomy for motor planning, learning, and control applications. *arXiv preprint arXiv:2506.16044*, 2025.

[4] Johannes Y Lee, Sangjoon Lee, Abhishek Mishra, Xu Yan, Brandon McMahan, Brent Gaisford, Charles Kobashigawa, Mike Qu, Chang Xie, and Jonathan C Kao. Brain–computer interface control with artificial intelligence copilots. *Nature Machine Intelligence*, pages 1–14, 2025.

[5] Brandon McMahan, Zhenghao Mark Peng, Bolei Zhou, and Jonathan Kao. Shared autonomy with ida: interventional diffusion assistance. *Advances in Neural Information Processing Systems*, 37:128330–128354, 2024.

[6] Siddharth Reddy, Anca D. Dragan, and Sergey Levine. Shared autonomy via deep reinforcement learning. In *Robotics: Science and Systems (RSS)*, 2018. Presented at RSS 2018.

[7] Takuma Yoneda, Luzhe Sun, Ge Yang, Bradly C. Stadie, and Matthew R. Walter. To the noise and back: Diffusion for shared autonomy. In *Robotics: Science and Systems (RSS)*, 2023.

[8] Weihao Tan, David Koleczek, Siddhant Pradhan, Nicholas Perello, Vivek Chettiar, Vishal Rohra, Aaslesha Rajaram, Soundararajan Srinivasan, HM Sajjad Hossain, and Yash Chandak. On optimizing interventions in shared autonomy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5341–5349, 2022.

[9] Francis R Willett et al. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858):249–254, 2021.

[10] Francis R Willett et al. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023.

[11] Nicholas S Card et al. An accurate and rapidly calibrating speech neuroprosthesis. *New England Journal of Medicine*, 391(7):609–618, 2024.

[12] Justin J Jude et al. An intuitive, bimanual, high-throughput qwerty touch typing neuroprosthesis for people with tetraplegia. *medRxiv preprint*, April 2025.

[13] An-Chieh Cheng et al. Navila: Legged robot vision–language–action model for navigation. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

[14] Brianna Zitkovich et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning (CoRL)*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR, 2023.

[15] Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on robot learning*, pages 287–318. PMLR, 2023.

[16] John P Cunningham et al. Dimensionality reduction for large-scale neural recordings. *Nature neuroscience*, 17(11):1500–1509, 2014.

[17] Alex H Williams and Scott W Linderman. Statistical neuroscience in the single trial limit. *Current opinion in neurobiology*, 70:193–205, 2021.

[18] Alex H Williams et al. Discovering precise temporal patterns in large-scale neural recordings through robust and interpretable time warping. *Neuron*, 105(2):246–259, 2020.

[19] Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. How might we create better benchmarks for speech recognition? In *Proceedings of the 1st workshop on benchmarking: Past, present and future*, pages 22–34, 2021.

[20] Mehryar Mohri, Fernando Pereira, and Michael Riley. Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16(1):69–88, 2002.

[21] Mehryar Mohri and Michael Riley. Speech recognition with weighted finite-state transducers. In *Springer Handbook of Speech Processing*, pages 559–584. Springer, 2008.

[22] Anjuli Kannan, Yonghui Wu, Patrick Nguyen, Tara N Sainath, Zhijeng Chen, and Rohit Prabhavalkar. An analysis of incorporating an external language model into a sequence-to-sequence model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5828. IEEE, 2018.

[23] Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Eng-Siong Chng. Hyporadise: An open baseline for generative speech recognition with large language models. *Advances in Neural Information Processing Systems*, 36:31665–31688, 2023.

[24] Chen Chen, Ruizhe Li, Yuchen Hu, Sabato Marco Siniscalchi, Pin-Yu Chen, Eng Siong Chng, and Chao-Han Huck Yang. It's never too late: Fusing acoustic information into large language models for automatic speech recognition. In *International Conference on Learning Representations (ICLR)*, 2024. Poster, ICLR 2024; arXiv:2402.05457.

[25] Ehsan Variani, Tongzhou Chen, James Apfel, Bhuvana Ramabhadran, Seungji Lee, and Pedro Moreno. Neural oracle search on n-best hypotheses. In *ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7824–7828, 2020.

[26] Yuchen Hu, Chen Chen, Chao-Han Huck Yang, Ruizhe Li, Chao Zhang, Pin-Yu Chen, and EnSiong Chng. Large language models are efficient learners of noise-robust speech recognition. In *International Conference on Learning Representations (ICLR)*, 2024. Accepted as Spotlight (top 5%) at ICLR 2024.

[27] Mingda Li, Weitong Ruan, Xinyue Liu, Luca Soldaini, Wael Hamza, and Chengwei Su. Improving spoken language understanding by exploiting asr n-best hypotheses. *arXiv preprint arXiv:2001.05284*, 2020.

[28] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pages 41753–41769. PMLR, 2023.

[29] Alex Graves. Connectionist temporal classification. In *Supervised sequence labelling with recurrent neural networks*, pages 61–93. Springer, 2012.

[30] Ruizhe Huang, Xiaohui Zhang, Zhaoheng Ni, Li Sun, Moto Hira, Jeff Hwang, Vimal Manohar, Vineel Pratap, Matthew Wiesner, Shinji Watanabe, et al. Less peaky and more accurate ctc forced alignment by label priors. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11831–11835. IEEE, 2024.

[31] Frank Wessel, Ralf Schluter, Klaus Macherey, and Hermann Ney. Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on speech and audio processing*, 9(3):288–298, 2002.

[32] Hongzhu Li and Weiqiang Wang. Reinterpreting ctc training as iterative fitting. *Pattern Recognition*, 105:107392, 2020.

Supplementary Material

A Brain-to-Text Brain Machine Interface

Let $\mathbf{X}_{1:T} \in \mathbb{R}^{D \times T}$ denote the multichannel neural recordings, where D indexes the dimension of the neural feature, and T is the length of the input sequence, sliding windows of neural activity during attempted speech. Let $\mathbf{y}_{1:N} \in \mathcal{V}^N$ denote the target phoneme sequence of length N, where $\mathcal{V}$ is the phoneme vocabulary augmented with the silence token SIL, representing the silence between word production, and the blank symbol ϵ, representing the padding ignored when converting frame-wise prediction to sequence.

The neural decoder M trained with connectionist temporal classification (CTC) [29] parameterizes a conditional distribution over frame-level labels:

$$p(s_{1:T} \mid \mathbf{X}_{1:T}) = \prod_{t=1}^{T} p(s_t \mid \mathbf{X}_{1:t}), \quad s_t \in \mathcal{V} \tag{1}$$

where $s_{1:T}$ denotes an alignment sequence. The CTC loss defines the probability of a target sequence $\mathbf{y}$ as the sum over all valid alignments that collapse to $\mathbf{y}$:

$$p(\mathbf{y} \mid \mathbf{X}_{1:T}) = \sum_{s_{1:T} \in \mathcal{A}(\mathbf{y})} p(s_{1:T} \mid \mathbf{X}_{1:T}) \tag{2}$$

where $\mathcal{A}(\mathbf{y})$ is the set of frame-level paths that reduce to $\mathbf{y}$ under the compression operator, which collapse the consecutive repeat of the token $v \in \mathcal{V}$ or ignore blank token ϵ [29, 30]. The maximum likelihood decoding is then given by:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \; p(\mathbf{y} \mid \mathbf{X}_{1:T}) \tag{3}$$

which empirically often corresponds to selecting the alignment

$$\hat{s}_{1:T} = \arg\max_{s_{1:T}} p(s_{1:T} \mid \mathbf{X}_{1:T}) \tag{4}$$

(the Viterbi alignment), and then applying compression [30, 32].

A.1 Probabilistic Fusing with Language Model

In state-of-art Brain-to-Text BCI, probabilistic fusing with an external language model is commonly performed through a weighted finite-state transducer (WFST) framework [20]. Instead of searching over sequences using only the decoder probabilities, WFST-based decoding composes three transducers:

- Token (V): Encodes the posterior distribution from the neural decoder over tokens $\mathcal{V}$. Transition cost are given by the negative log-probabilities

$$-\log P_{\text{ND}}(s_t \mid x_{1:T}) \tag{5}$$

- Lexicon (L): The lexicon L maps token sequences to words, $\mathcal{V}^* \rightarrow \mathcal{W}$. Valid mappings carry zero cost, and words with multiple pronunciations are represented by multiple parallel paths, each corresponding to a valid token sequence.
- Grammar (G): Encodes the language model as an n-gram finite-state transducer over words. Transition costs correspond to negative log-probabilities

$$-\log P_{\text{LM}}(w_t \mid w_{t-n+1:t-1}) \tag{6}$$

The composed graph $V \circ L \circ G$ constrains decoding to valid word sequences, while integrating both neural decoder (ND) scores and language model (LM) priors:

$$\text{score}(\mathbf{y}) = \log p_{\text{ND}}(\mathbf{y} \mid \mathbf{X}_{1:T}) + \beta \log p\text{LM}(\mathbf{y}) \tag{7}$$

where β is the LM scaling weight. Beam search over this graph yields a ranked list of top-K candidate hypotheses. For real-time decoding top-1 hypothesis fused with 3-gram LM model is often used while the user produce the utterance.

After sentence completion, stronger models of linguistic regularity—such as large language models (LLMs), can be applied post hoc, either for top-K hypothesis re-ranking or for generative error correction of the decoder output [22–26]. The re-ranking score may include an additional LLM term:

$$\text{score}(\mathbf{y}) = \log p_{\text{ND}}(\mathbf{y}|\mathbf{X}_{1:T}) + \beta \log p\text{LM}(\mathbf{y}) + \alpha \log p_{\text{LLM}}(\mathbf{y}) \tag{8}$$

while generative error correction leverages the top-K hypotheses as input to an large language model (LLM) that synthesizes the transcription. Because this correction depends on sentence-level context, it is typically performed only after completion of the utterance.

B Uncertainty and Informativeness Quantification

To evaluate the effectiveness of probabilistic fusing between the neural decoder and the language model, we quantify uncertainty and informativeness of the decoder's predicted distributions. We focus on integration with an n-gram LM (via WFST decoding), as it is the most practical solution to satisfy the BCI user real-time decoding need.

B.1 Sequence-level analysis

To evaluate whether useful information is present in lower-ranked hypotheses. Two oracle-style metrics are employed:

- $\min$ WER: the minimum WER among the top-K hypotheses, representing the upper bound achievable by reranking methods such as large language model (LLM) rescoring [22, 23].
- Word-Recall: The number of words not recovered after composing all predicted words across the top-K hypotheses. This represents the upper bound of corrections achievable if downstream models exploit all alternatives present in the N-best list [24–26].

B.2 Frame-level analysis

To analyze frame-wise distributions relative to the Viterbi alignment $\hat{s}_{1:T}$ (empirically, $\hat{s}_{1:T} \in \mathcal{A}(\hat{\mathbf{y}})$ with CTC-trained models).

- Uncertainty: measured as normalized entropy of the frame-level predicted distribution:

$$H_t^{\text{norm}} = -\frac{1}{\log|\mathcal{V}|} \sum_{c \in \mathcal{V}} p(s_t = c \mid \mathbf{X}_{1:t}) \log p(s_t = c \mid \mathbf{X}_{1:t}) \tag{9}$$

- Informativeness: measured by comparing the predicted probability assigned to the hypothesized token $\hat{s}t$ versus the reference token $s_t^\star$. The reference token at each frame, $s_t^\star$ is derived with forced alignment.

C Oracle Manipulation of Neural Decoder Outputs to Represent Probabilistic Regimes

For class probability distribution of the neural decoder at time t over $\mathcal{V}$ phoneme classes

$$\mathbf{p}t = \left[p_t^{(1)}, p_t^{(2)}, \ldots, p_t^{(\mathcal{V})}\right], \quad \sum_{v=1}^{\mathcal{V}} p_t^{(v)} = 1 \tag{10}$$

The predicted class is given by $\hat{s}_t = \arg\max_v p_t^{(v)}$. To probe how different uncertainty structures affect probabilistic fusing with downstream language models, we consider oracle manipulations of $\mathbf{p}_t$ into three distinct regimes, inspired by prior work on confidence measures and posterior distributions in ASR:

C.1 Over-confident distribution

All probability mass is concentrated on the predicted class:

$$p_t^{(v)} = \begin{cases} 1, & v = \hat{s}_t, \\ 0, & v \neq \hat{s}_t. \end{cases} \tag{11}$$

This regime communicates only the maximum-likelihood decision, analogous to top-1 decoding, with no expression of uncertainty or alternatives.

C.2 Uncertainty-aware distribution

A calibrated distribution communicates confidence in the predicted class but spreads the residual mass uniformly across all non-predicted classes:

$$p_t^{(v)} = \begin{cases} p, & v = \hat{s}_t, \\ \dfrac{1-p}{\mathcal{V}-1}, & v \neq \hat{s}_t, \end{cases} \quad 0 < p < 1 \tag{12}$$

This regime captures overall uncertainty, but the alternatives are uninformative since all classes are treated equally regardless of similarity.

C.3 Alternative-rich distribution

The residual probability mass is distributed non-uniformly, emphasizing plausible confusions (e.g., acoustically or articulatorily similar phonemes) while down-weighting implausible classes:

$$p_t^{(v)} = \begin{cases} p, & v = \hat{s}_t, \\ \alpha_v, & v \neq \hat{s}t, \end{cases} \quad \text{with } \sum v \neq \hat{s}_t \alpha_v = 1 - p \tag{13}$$

This regime provides the richest signal for downstream fusion, as it conveys both calibrated uncertainty and structured alternatives.