Team Null-1 Phoneme Classification System for the NeurIPS 2025 PNPL Competition (Task 2)

Anonymous Author(s)
Affiliation
Address
email

Abstract

Decoding phonemes from non-invasive EEG or MEG brain signals during naturalistic speech listening remains highly challenging due to low signal-to-noise ratio (SNR), complex temporal dynamics, limited data availability, and pronounced phoneme class imbalance. To address these issues, we propose a phoneme-level signal smoothing framework that improves MEG signal quality while explicitly mitigating class imbalance in phoneme classification. Our method integrates fixed and adaptive smoothing strategies through a weighted combination, enabling better preservation of phoneme-related brain signal patterns and stronger generalization to rare phoneme categories. In the NeurIPS 2025 PNPL Competition, our approach achieves state-of-the-art performance in the Phoneme Classification Standard phase, attaining a Macro-F1 score of 73.82% and ranking first on the leaderboard.

1 Introduction

Decoding linguistic information from non-invasive brain signals offers valuable insights into the neural mechanisms of speech processing and enables promising applications in Brain–Computer Interfaces (BCIs)[3, 8]. However, phoneme classification from MEG signals remains challenging due to the inherent noise in MEG recordings, the brief, temporally localized nature of phoneme representations within continuous speech, and the imbalanced class distributions in natural speech.

Signal averaging is a common denoising strategy, as increasing the number of averaged samples typically improves classification accuracy. Yet, it also introduces critical issues: **(i)** averaging may distort the feature space by over-compressing certain phoneme classes while leaving others dispersed and harder to distinguish; and **(ii)** the spatiotemporal patterns and noise characteristics of MEG signals vary substantially across subjects and sessions, causing domain shifts between training and testing data and increasing the risk of overfitting.

We have therefore developed a system that combines fixed and adaptive smoothing strategies. By averaging phoneme samples according to their frequency, we obtained a more balanced dataset. To further enhance performance, we optimized both the data normalization procedure and the model architecture through extensive experiments. Finally, by ensembling models trained under different configurations, our system achieved state-of-the-art performance on the NeurIPS 2025 PNPL Competition Task 2[4], reaching a Macro-F1 of 73.82% and ranking first among 30 submissions.

2 System Description

As shown in Fig. 1(a), the best-performing system employs an ensemble of models. Prior to model input, the MEG signals are normalized and segmented into 0.5s-windows starting from each

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

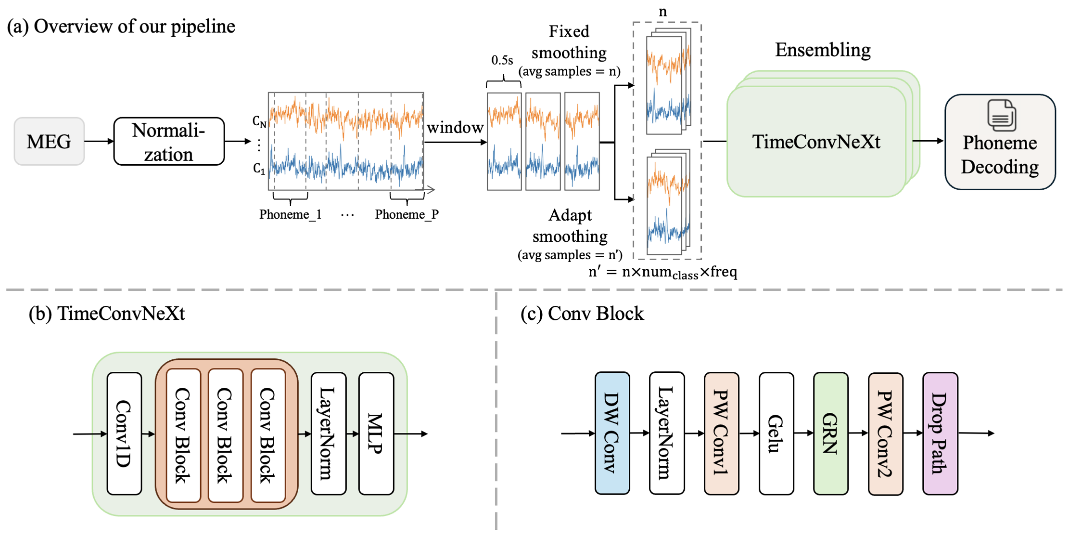

Figure 1: (a) Overview of our proposed system: the normalized MEG signals are segmented into short windows following each phoneme onset, then grouped and averaged before model training. (b) Architecture of our TimeConvNeXt model[11]. (c) Structure of a single Conv Block within the model.

phoneme onset. Afterward, the data processed with different smoothing strategies are concatenated and fed into the model.

2.1 Normalization

Given the temporal drift and amplitude fluctuations of MEG signals across sessions, we applied channel-wise z-score normalization. The normalization used the global mean computed from the entire training set and the standard deviation estimated exclusively from the phoneme samples. The channel-wise standard deviation σ_c was computed as $\sigma_c = \frac{1}{N}\sum_{i=1}^{N} \text{std}\left(\mathbf{x}_i^{(c)}\right)$, here $\mathbf{x}_i^{(c)}$ represents the time series of the i-th sample in the c-th channel, and N is the number of samples. This approach provides a robust measure of phoneme-related variability.

2.2 Smoothing Strategy

As shown in Fig. 1(a), we employed a combination of fixed and adaptive smoothing strategies to construct a more balanced dataset.

Fixed smoothing strategy. Following the official PNPL competition tutorial, we grouped samples by their phoneme labels and averaged a fixed number of 100 samples per class. Overlaps between adjacent groups were allowed, with the degree of overlap controlled by a stride parameter to enhance the diversity of the averaged signals.

Adaptive smoothing strategy. To further mitigate class imbalance, we adopted an adaptive smoothing strategy. For each phoneme class, we computed its relative frequency and determined the number of grouped samples as $100 \times num_{class} \times freq_i$, where $freq_i$ denotes the relative frequency of the i-th phoneme. This approach ensured sufficient representation of low-frequency phonemes during training. Similar to the fixed strategy, overlapping was applied to enhance variability.

2.3 Model Architecture

Fig. 1(b) illustrates the principal architecture of models. The smoothed MEG signals are first down-sampled using a 1D convolutional layer, followed by three ConvNeXt V2–inspired blocks that extract linguistic representations from the MEG features[11]. A LayerNorm layer is then applied[1], and the resulting phoneme-level features are passed through an MLP module to produce the output logits. The detailed structure of the ConvNeXt V2–inspired blocks is shown in Fig. 1(c)[11].

2.4 Loss function

For model optimization, we employed the standard Cross-Entropy (CE) loss. Given K phoneme classes, the CE loss for a single sample is defined as $\mathcal{L}_{CE} = -\sum_{k=1}^{K} y_k \log(p_k)$, where y_k is the one-hot target label (1 if k is the true class and 0 otherwise), and p_k is the model's predicted probability for class k (the output of the softmax layer).

2.5 Ensemble Strategy

We employed a diversity-weighted ensemble strategy that adaptively combines multiple models trained under different configurations. Each model's contribution was determined by its prediction diversity relative to the others.

We first computed the pairwise Spearman correlations between model predictions to form a similarity matrix $S \in \mathbb{R}^{M \times M}$[7], where $S_{ij} \in [0, 1]$ denotes the correlation between model i and model j. The diversity score of model i was defined as $d_i = (1 - \bar{s}_i)^2$, where $\bar{s}_i = \frac{1}{M-1} \sum_{j \neq i} S_{ij}$. Normalized ensemble weights were obtained via a temperature-controlled softmax as $w_i = \frac{d_i^{1/T}}{\sum_{j=1}^{M} d_j^{1/T}}$, where T controls the sharpness of the weight distribution. The final ensemble prediction was computed as $\hat{y} = \sum_{i=1}^{M} w_i y_i$. This strategy emphasizes models that provide complementary information while down-weighting redundant ones, yielding more stable and accurate phoneme classification across sessions.

This strategy emphasizes models that provide complementary information while down-weighting redundant ones, yielding more stable and accurate phoneme classification across sessions.

3 Experimental Setup

3.1 Dataset

The NeurIPS 2025 PNPL Competition used the LibriBrain dataset[12], which is divided into four subsets: training, validation, test, and holdout. The holdout set is used for final evaluation, and its reference labels are not publicly available. Each speech segment is annotated with 39 ARPAbet phoneme classes[10]. No additional datasets were used in our experiments.

3.2 Implementations

The hyperparameters are listed in Table 4 in Appendix A. We trained all the models for 30 epochs using the AdamW optimizer with an initial learning rate of 1e-4. We also applied a cosine scheduler with a weight decay of 0.01. During training and validation, we set a fixed random seed and applied both fixed and adaptive smoothing, with an overlap stride of 50. To further augment the training data, we re-applied smoothing every two epochs, allowing the model to see a larger variety of smoothed MEG signals despite the limited dataset size. The model checkpoint of the epoch with the lowest loss on the validation set is used for evaluation. The implementations are based on DeepSpeed [5].

4 Results

Table 1: Dataset entropy comparison with and without adaptive smoothing.

	w/ *adaptive smoothing*	w/o *adaptive smoothing*
Entropy	**5.165**	4.833

We evaluated our framework using four architectures: TCN[2], CLDNN[6], Transformer[9], and our proposed TimeConvNeXt model. As shown in Table 2, our model consistently achieved the highest phoneme classification performance across most of the metrics[11].

To further evaluate the contribution of our adaptive smoothing strategy, we first analyzed the dataset entropy with and without adaptive smoothing. As shown in Table 1, adaptive smoothing yields a more balanced data distribution. We then compared models trained with and without this component on the test set. As presented in Table 2, removing adaptive averaging results in a notable drop in Macro-F1 and balanced accuracy, indicating that this strategy effectively mitigates phoneme imbalance and enhances data diversity. Although the model with adaptive averaging achieves higher classification accuracy, it shows a slight decrease in AUROC, reflecting the inherent difference between threshold-dependent metrics (e.g., F1) and ranking-based metrics (e.g., AUROC).

Finally, by combining multiple models through our diversity-weighted ensemble, we obtained the best overall performance, showing improved robustness and generalization across sessions. The official leaderboard Macro-F1 scores are reported in Table 3.

Table 2: Performance comparison across different models on the test set. Our method shows superior balanced accuracy and F1. "Single(ours)" refers to our ConvNeXt model shown in Fig. 1, and "Ensemble(ours)" refers to our ensemble model with ConvNeXt, CLDNN, and TCN structures etc.

Method	**Mi-F1**	**Ma-F1**	**BACC**	**Mi-AUROC** $\uparrow$	**Ma-AUROC** $\uparrow$
			Ours		
TCN [2]	50.45 ± 1.12	44.46 ± 1.02	46.43 ± 1.00	$\underline{94.25 \pm 0.49}$	$\underline{93.84 \pm 0.47}$
CLDNN [6]	48.83 ± 1.85	43.49 ± 1.93	46.90 ± 2.29	93.48 ± 0.25	93.27 ± 0.20
Transformer [9]	49.66 ± 2.09	46.00 ± 2.25	48.13 ± 1.78	91.80 ± 1.08	91.43 ± 1.11
Single(ours)	$\underline{54.96 \pm 3.35}$	$\underline{50.70 \pm 3.71}$	$\underline{52.98 \pm 3.36}$	93.33 ± 0.33	93.46 ± 0.27
Ensemble(ours)	**61.31**	**54.90**	**59.23**	**96.91**	**96.88**
			w/o *Adaptive Smoothing*		
TCN [2]	$\underline{52.25 \pm 2.43}$	$\underline{38.03 \pm 1.77}$	$\underline{40.19 \pm 2.29}$	96.73 ± 0.40	$\underline{96.20 \pm 0.41}$
CLDNN [6]	51.61 ± 1.33	37.33 ± 1.55	40.11 ± 1.82	$\mathbf{97.03 \pm 0.35}$	$\mathbf{96.77 \pm 0.44}$
Transformer [9]	45.79 ± 4.48	30.15 ± 5.10	34.08 ± 5.71	94.16 ± 0.51	91.91 ± 0.61
Single(ours)	$\mathbf{52.79 \pm 2.68}$	$\mathbf{39.35 \pm 1.54}$	$\mathbf{41.79 \pm 1.91}$	$\underline{96.78 \pm 0.55}$	96.08 ± 1.26

Table 3: Official Macro-F1 results of different models on the leaderboard. "Single(ours)" refers to our ConvNeXt model shown in Fig. 1, and "Ensemble(ours)" refers to our ensemble model with ConvNeXt, CLDNN, and TCN structures etc.

Method	**Macro-F1**
TCN [2]	48.31
CLDNN [6]	53.16
Transformer [9]	46.80
Single(ours)	$\underline{65.42}$
Ensemble(ours)	**73.82**

5 Conclusion

In this work, we introduce a system constructed for participation in the NeurIPS 2025 PNPL Competition[4]. The system showed the best performance among a total of 30 submissions to the competition. Through phoneme segment-based normalization and adaptive smoothing, we improve both the stability and discriminability of neural representations. Moreover, the diversity-weighted ensemble further enhanced robustness across recording sessions. Overall, our findings highlight that carefully designed preprocessing, data averaging strategies, and hybrid modeling can substantially advance neural speech decoding from non-invasive brain recordings, paving the way for more generalizable and interpretable BCI systems.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization. *arXiv preprint arXiv:1607.06450v1*, 2016.

[2] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[3] Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10):1097–1107, 2023.

[4] Oiwi Parker Jones, Gilad Landau, Miran Özdogan, Gereon Elvers, Mantegna Francesco, Pratik Somaiya, Dulhan Jayalath, Brendan Shillingford, Greg Farquhar, Minqi Jiang, Karim Jerbi, Hamza Abdelhedi, Caglar Gulcehre, and Mark Woolrich. The 2025 PNPL Competition: speech detection and phoneme classification in the LibriBrain dataset. In *Proc. NeurIPS*, 2025.

[5] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proc. KDD*, pages 3505–3506, 2020.

[6] Tara N Sainath, Oriol Vinyals, Andrew Senior, and Haşim Sak. Convolutional, long short-term memory, fully connected deep neural networks. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4580–4584. Ieee, 2015.

[7] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[8] Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[10] R. L. Weide. The carnegie mellon pronouncing dictionary. `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`, 1998.

[11] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proc. CVPR*, 2023.

[12] Miran Özdogan, Gilad Landau, Gereon Elvers, Dulhan Jayalath, Pratik Somaiya, Francesco Mantegna, Mark Woolrich, and Oiwi Parker Jones. Layer Normalization. *arXiv preprint arXiv:2506.02098*, 2025.

A Hyperparameters for model

This section presents the model parameters of our ConvNeXt V2-inspired model, as well as the training hyperparameters.

Table 4: Hyperparameters for our model

Hyperparameters		**Values**
In Channels		306
Init Channels		128
Hidden Dimensions		[128, 256, 512]
Ratios		[5, 5, 5]
Groups		[2, 4, 8]
Epochs		30
Weight decay		1e-2
Label smoothing		0.1
Optimizer	Type	AdamW
	Betas	[0.9, 0.99]
	Eps	1e-5
	Type	WarmupCosineLR
Scheduler	Warmup ratio	0.01
	Cos min ratio	0.01