Odyssey 2026
23-26 June 2026, Lisbon, Portugal

MEGConformer: Conformer-Based MEG Decoder for Robust Speech and Phoneme Classification

Xabier de Zuazo[1], *Ibon Saratxaga*[1], *Eva Navas*[1]

[1]Communications Engineering, HiTZ Center, University of the Basque Country – UPV/EHU, Spain

xabier.dezuazo@ehu.eus, ibon.saratxaga@ehu.eus, eva.navas@ehu.eus

Abstract

Decoding speech-related information from non-invasive MEG is a key step toward scalable brain-computer interfaces. We present compact Conformer-based decoders on the LibriBrain 2025 PNPL benchmark for two core tasks: Speech Detection and Phoneme Classification. Our approach adapts a compact Conformer to raw 306-channel MEG signals, with a lightweight convolutional projection layer and task-specific heads. For Speech Detection, a MEG-oriented SpecAugment provided a first exploration of MEG-specific augmentation. For Phoneme Classification, we used inverse-square-root class weighting and a dynamic grouping loader to handle 100-sample averaged examples. In addition, a simple instance-level normalization proved critical to mitigate distribution shifts on the holdout split. Using the official Standard track splits and F1-macro for model selection, our best systems achieved 88.9% (Speech) and 73.6% (Phoneme) on the leaderboard, winning the Phoneme Classification Standard track.

Index Terms: speech decoding, MEG, brain-computer interface, neuroscience

1. Introduction

In the emerging fields of deep learning and neuroscience, decoding speech from non-invasive brain signals remains a central problem in neural signal processing and brain-computer interfaces (BCIs) [1, 2, 3]. The recently released LibriBrain dataset [4] and its associated PNPL 2025 Competition [5] provide a unique opportunity to address this challenge by enabling large-scale within-subject modeling of magnetoencephalography (MEG) recordings. LibriBrain contains over 50 hours of continuous MEG data from a single participant listening to the complete Sherlock Holmes corpus, accompanied by precise voice activity and phoneme-level alignments. The competition defines two foundational decoding tasks: Speech Detection, which distinguishes between speech and silence, and Phoneme Classification, which predicts the phoneme being perceived from averaged MEG signals. Both tasks include a Standard and an Extended track, where the Standard track restricts participants to using only the official training data. We participated exclusively in the Standard track, and our system won the Phoneme Classification task in the final competition results.

Our approach explores an initial adaptation of state-of-the-art Automatic Speech Recognition (ASR) architectures to non-invasive neural signals. Specifically, we leverage the Conformer model [6], a convolution-augmented Transformer [7] that combines the global context modeling of self-attention [8] with the local feature extraction strength of convolutional networks [9]. Conformer-based architectures have been and continue to be the state-of-the-art in the ASR field [10, 11, 12, 13, 14]. We hypothesize that their ability to capture both temporal and spectral dependencies may serve as a good starting point for MEG-based speech tasks, which also exhibit structured spatiotemporal patterns over multiple timescales.

In this work, we evaluate a Conformer model adapted to the LibriBrain competition tasks. The models process raw MEG windows to predict either speech activity or phoneme identity, sharing the same backbone while differing in input and output processing. To ensure fair comparison and reproducibility, all experiments follow the official data splits and evaluation metrics provided by the organizers [5].

Our main contributions are fourfold: (i) a unified Conformer architecture jointly optimized for both LibriBrain tasks; (ii) a robust yet straightforward input normalization method that substantially improves holdout generalization; (iii) an effective MEG-specific augmentation; (iv) a dynamic grouping strategy to classify averaged samples in the Phoneme Task.

This study demonstrates that adapting modern ASR architectures for MEG decoding produces competitive, robust results in both voice activity detection and phoneme recognition, emphasizing the growing convergence between speech processing and neural decoding research. For further implementation details, the technical documentation, source code, and checkpoints are available at `https://github.com/neural2speech/libribrain-experiments`.

2. Related work

Decoding speech from non-invasive neural signals has been explored with both Electroencephalography (EEG) and MEG using supervised learning, from linear baselines to deep architectures. EEG studies reconstruct speech features or spectrograms with CNN/Transformer-style models and subject-independent training, showing steady gains with more data [15, 16]. On MEG, prior work demonstrated phrase and word-level decoding and trial-efficiency using wavelet denoising, CNNs, and transfer learning [17, 18, 19], as well as compact end-to-end networks adapted to sensor time series [20], and transformer-based neural encoding that links linguistic context to MEG responses [21]. Other studies have also started to explore phoneme-level decoding in perception and production modalities, comparing traditional and novel model architectures [22, 23]. More recently, LibriBrain introduced an unprecedented amount of within-subject MEG data and standardized tasks for speech detection, phoneme classification, and keyword spotting [4, 5, 24]. Beyond purely supervised setups, contrastive and self-supervised objectives, along with foundation-model guidance, have advanced non-invasive decoding across perception domains, improving retrieval and generation [25, 26, 27]. Recent work on non-invasive brain-to-text systems combines discriminative de-

10.21437/Odyssey.2026-18

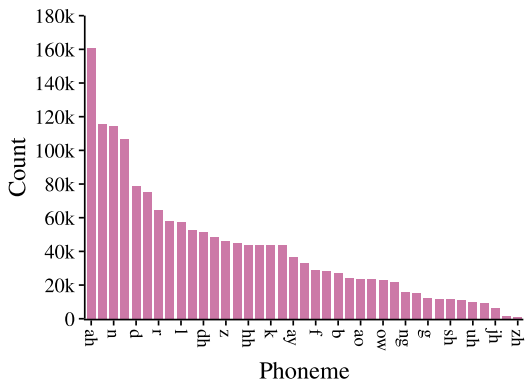

Figure 1: *Total phoneme counts (before 100-sample averaging).*

coders with language-model rescoring and cross-dataset scaling to produce significant score improvements [28].

3. Dataset

We evaluate our models on the LibriBrain dataset and PNPL 2025 competition setup [5]. LibriBrain provides over 50 hours of within-subject, non-invasive MEG recorded from a single participant during naturalistic speech listening (Sherlock Holmes audiobooks), with time-aligned speech activity and phoneme annotations. Recordings were acquired with a 306-channel Elekta/MEGIN system (102 magnetometers, 204 planar gradiometers), originally sampled at 1000 Hz and released with an official preprocessing pipeline (including bad-channel interpolation, head-position correction, signal-space separation, notch, and band-pass filtering), with model inputs downsampled to 250 Hz following the competition protocol.

We follow the official Standard-track splits (train, validation, test, holdout) and evaluation metric (F1-macro) throughout, noting that the holdout split is not publicly available. For the Speech Detection task, labels are derived from voice-activity annotations over continuous recordings, resulting in a binary speech vs. silence classification problem. Across the full dataset, speech occupies ∼76.7% of labeled time (40:19:38 speech vs. 12:16:37 silence). For the Phoneme Classification task, windows are aligned to 39 ARPAbet phoneme labels [29]. Figure 1 reports phoneme occurrence statistics for the phoneme windows, illustrating the long-tailed class distribution. Besides, in the competition setting for the phoneme task, holdout examples are provided as averages over 100 samples to improve SNR, and we mirror this setting in training via dynamic grouping (see Subsection 4.3).

4. Methods

4.1. Model

We use a single Conformer encoder adapted to MEG for both tracks (Speech Detection and Phoneme Classification). The network takes raw sensor time series (306 channels) preprocessed with the official pipeline, where windowed signal segments are downsampled to 250 Hz. For Speech, we use 2.5 s windows (625 samples), and for the Phoneme Task, 0.5 s windows (125 samples). A lightweight 1D convolutional projection adapts the 306 MEG channels to the Conformer input size (144), followed by a dropout of p=0.1, a stack of Conformer blocks, and a linear classifier.

Each Conformer block follows the standard macaron layout: feed-forward, multi-head self-attention, depthwise temporal convolution, and a second feed-forward, with residual connections. We keep the design compact with a hidden size of 144. For Speech, based on Conformer Small, we use 16 layers, 4 heads, a hidden layer dimension of FFNs of 576, and the layer's depthwise convolution layer with a kernel size of 31. For Phonemes, we created a custom-sized Conformer with 7 layers, 12 heads, FFN dimension of 2048, and a kernel size of 127, to better adapt the model to the smaller dataset.

4.2. Speech detection task specifics

The Speech model is a binary classifier with a single-logit head, binary cross-entropy loss [30], and loss-level label smoothing of 0.1 [31]. We found that simple, speech-style augmentation was sufficient for this task. Therefore, we develop and apply a lightweight MEG-specific variant of SpecAugment [32], which we introduce as **MEGAugment**. It includes two operations: (i) *time masking*, which zeroes two random temporal spans (max width T=180 samples at 250 Hz) per window; and (ii) *bandstop masking*, which randomly suppresses narrow frequency bands (Theta, Alpha, Beta, Gamma, HGA [33]) using fourth-order infinite impulse response (IIR) notches with probability p=0.4. Additionally, input windows slide with a 60-sample stride during training to increase diversity via overlapping.

The final layer outputs a probability of speech, which is smoothed during inference by removing speech sequences shorter than 60 samples (240 ms). The block diagram in Figure 2 summarizes the pipeline.

4.3. Phoneme classification task specifics

The Phoneme model uses a linear classification head that outputs 39 logits and is trained with cross-entropy [34]. At inference, we select the class via `argmax`. To mitigate imbalance in the Phoneme Task, we use class weights w_c with the inverse square root of the number of samples (ISNS) [35] rule following Equation 1, where n_c is the count of class c in the training set and C is the total number of classes.

$$w_c \propto \frac{1}{\sqrt{n_c}}, \qquad \frac{1}{C}\sum_c w_c = 1, \tag{1}$$

The architecture (detailed in Figure 3) includes an instance-level input normalization layer directly after the raw input [36, 37]. For each averaged window $x \in \mathbb{R}^{C\times T}$, it normalizes each channel independently using statistics computed over the time axis of that same window (i.e., no dataset or batch-level running mean/variance are accumulated), as defined in Equation 2.

$$y = \frac{x - \mathbb{E}[x]}{\sqrt{\mathrm{Var}[x] + \epsilon}} \cdot \gamma + \beta. \tag{2}$$

Here, $x \in \mathbb{R}^{C\times T}$ is one averaged MEG window (channels × time), $\mathbb{E}[x]$ and $\mathrm{Var}[x]$ denote the mean and (biased) variance computed per window and per channel over the time axis (T), and ϵ is a small constant for numerical stability. We use `InstanceNorm1d` without running statistics, and without affine parameters, so $\gamma = 1$ and $\beta = 0$.

This normalization is applied after the 100-sample averaging (as part of the model forward pass) and is used identically in training and evaluation. We find this choice essential for stable holdout performance: unlike validation and test, the holdout

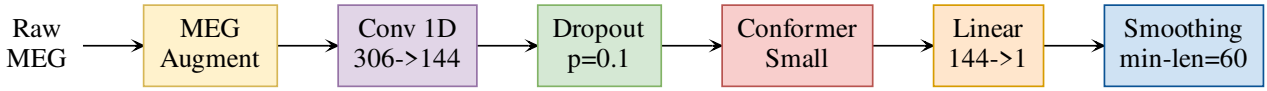

Figure 2: *Conformer-based model architecture used for Speech Task.*

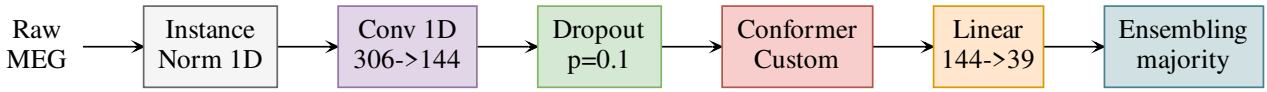

Figure 3: *Conformer-based model architecture used for Phoneme Task.*

split exhibits different statistical characteristics, indicating distributional shift (see Subsection 7.4 for details). Instance normalization fundamentally removes amplitude and scale drift between windows and sessions. Consequently, it closes most of the holdout gap, even if it is sometimes slightly suboptimal on the other splits. Because the competition holdout recordings are pre-averaged over 100 samples to improve SNR, we adapt our training setup accordingly. Particularly, we use a 100-sample dynamic grouping loader that reshuffles groups each epoch, allowing the model to see many independent averages of the same class while preserving temporal locality.

Lastly, we ensemble the five best model seeds to smooth predictions and select the final phonemes using majority voting.

5. Experimental setup

We follow the official LibriBrain data partitions for all experiments, using the provided train, validation, test, and holdout splits. In our experience, re-partitioning or rebalancing the data did not produce consistent improvements, so we retain the original configuration throughout.

Training uses the AdamW optimizer [38] with a learning rate of $1{\cdot}10^{-4}$, weight decay of $5{\cdot}10^{-2}$, batch size of 256, and early stopping based on validation F1-macro with a patience of 10. We monitor F1-macro on the validation set to select the final checkpoint. These hyperparameters are shared across both tasks.

Evaluation follows the competition protocol, using F1-macro as the primary metric on test and holdout splits.

All model seeds are trained on individual NVIDIA A100 and H100 GPUs, and each model configuration is trained with ten random seeds for better comparison.

6. Results

Our Conformer-based models achieved competitive results in both tasks, as shown in Table 1. In Speech Detection, our best model reached an F1-macro of 88.9%, surpassing the official baseline (68.0%). In Phoneme Classification, our model reached 73.6% F1-macro on holdout, which corresponds to the top result in the Standard track and the competition win[1]. Notably, although we participated only in the Standard track, our phoneme results exceed those reported by systems submitted to the Extended track, which were allowed to use additional data sources. These results confirm that Conformer architectures, initially developed for ASR, transfer productively to MEG-based decoding and perform competitively with traditional convolutional and transformer models.

[1] https://neural-processing-lab.github.io/2025-libribrain-competition/prizes/

To contextualize these leaderboard results, we highlight two inference-time effects and the role of instance-level normalization. First, in the Speech task, the post-processing smoothing step produces a small but consistent improvement in F1-macro (from 88.12% to 88.37% in the test set), reaching a final holdout score of 88.90% in Table 1. Second, in the Phoneme task, our single-model performance prior to ensembling is substantially lower (55.09±1.77% across five seeds), and the final majority-vote ensemble is critical to achieving the reported 73.60% on holdout, boosting scores by 33.6%. Ultimately, instance-level normalization was essential for holdout generalization (over +200%), outperforming batch (+17.8%) and layer normalization (+88.2%).

To better understand which components drive performance, and to characterize key dataset effects that impact generalization, we next report additional analyses: ablations with statistical testing, data-size scaling study, frequency band contribution analysis, a distribution-shift assessment across splits in the phoneme task, and experiments on phonetic-feature decoding.

7. Additional analyses

7.1. Ablation study and statistical significance

For the ablation study, we start from our best-performing model and remove each improvement individually while keeping all other settings fixed. Each ablated variant is therefore identical to the full model except for one modification, allowing us to isolate the contribution of that component.

We report per-variant means and standard deviation over ten seeds and assess paired differences with the Wilcoxon signed-rank test [39] over the F1-macro scores in the test set. This non-parametric test does not assume normality and evaluates whether the median of paired score differences differs from zero, indicating a genuine effect rather than random variation [40]. We use a p-value threshold of $p \leq 0.01$ to denote statistical significance.

7.1.1. Speech task ablation

As shown in Table 2, window extension from 0.5 to 2.5 s gives the largest gain (+10.8% relative), and Conformer outperforms SEANet baseline [41] by +9.0%. Both effects are significant: tmax = 0.5 s vs. ours ($W = 0.0$, $p = 0.002$), and SEANet vs. Conformer ($W = 0.0$, $p = 0.002$). Reducing the training stride to 60 also significantly helps (+2.8%; $W = 0.0$, $p = 0.002$). Removing augmentation has a small effect (+0.01%) and is not significant ($W = 13.0$, $p = 0.160$). Although MEGAugment improvements are not significant in the final configuration, it did provide a significant gain of +1.8% in earlier model versions, motivating its inclusion ($W = 1.0$, $p = 0.004$).

Table 1: *F1-macro scores in the holdout for both tasks (Standard track). Our model and scores are marked in bold.*

Model	Speech Detection ↑	Phoneme Classification ↑
Naive Baseline	*45.30%*	*0.47%*
Baseline	*68.04%*	*6.10%*
10th place	88.89%	60.96%
MEGConformer	**88.90%**	**73.60%**
Best (Standard track)	92.18%	73.60%

Table 2: *Speech task variant test scores over ten seeds before smoothing. Bold marks the best scores.*

Variant	F1_{macro} (%)	F1_{pos} (%)	$\text{Accuracy}_{\text{bal}}$ (%)	$\text{AUROC}_{\text{macro}}$ (%)	Jaccard (%)
Our model	**87.59 ± 0.70**	**94.72 ± 0.23**	85.72 ± 1.19	**96.15 ± 0.32**	**78.64 ± 1.02**
tmax=0.5	78.13 ± 1.32	90.62 ± 0.20	75.58 ± 1.74	90.59 ± 0.31	65.87 ± 1.47
SEANet	79.67 ± 0.50	90.97 ± 0.16	77.44 ± 0.59	90.58 ± 0.34	67.70 ± 0.61
No stride	85.18 ± 0.55	93.06 ± 0.14	83.59 ± 1.12	94.35 ± 0.20	75.01 ± 0.75
No augment	87.58 ± 0.55	94.20 ± 0.14	**85.82 ± 1.17**	96.08 ± 0.21	78.53 ± 0.79

SEANet adjustments. To accommodate the longer input window tmax = 2.5 s (625 samples at 250 Hz), we modify only the temporal downsampler, the third `conv1d` (k=50, s=25 in the original), by setting its stride to s=160. This keeps the intermediate length at L_{out}=4, so the subsequent k=4 stem still collapses to 1 (original: $L_{\text{out}} = \lfloor (125-50)/25 \rfloor + 1 = 4$; adjusted: $\lfloor (625-50)/160 \rfloor + 1 = 4$). We also align the classifier and loss with the binary task by replacing the final 1×1 `conv1d` from 512 to 2 with a single-logit 512 to 1 head and using BCE-with-logits plus label smoothing (0.1). Beyond architecture, we match the training and evaluation protocol used for our Conformer: sliding-window training with a stride of 60, MEGAugment, AdamW, early stopping on validation F1-macro, best-checkpoint selection, and validation and test scoring with a stride of 1.

7.1.2. Phoneme task ablation

As shown in Table 3, dynamic grouping (vs. fixed groups) gives the largest improvement (+13.3% relative; $W = 1.0$, $p = 0.004$). Inverse-square-root class weighting (our default) outperforms non-weighted loss by +7.6% (not significant with $W = 10.0$, $p = 0.084$). The custom Conformer provides a small lift over Conformer Small (+0.5%; $W = 16.0$, $p = 0.844$), consistent with similar capacity but with a better fit to the data characteristics.

7.2. Data-size ablation

To assess how the performance scales with within-subject MEG data, we conducted data-size ablations for both tasks using progressively larger subsets of the LibriBrain training set. Figures 4 and 5 show the resulting F1-macro scores as a function of the total hours of MEG data used for training (for phonemes, without sample averaging), with shaded bands indicating one standard deviation across random seeds.

For Speech Detection, performance rapidly improved with additional data and then began to saturate, suggesting that the task may already approach its ceiling with a single subject's data. In contrast, Phoneme Classification, even though with some diminishing returns signs [42], shows an upward trend with no apparent plateau, implying that larger-scale within-subject recordings could still give substantial gains. These

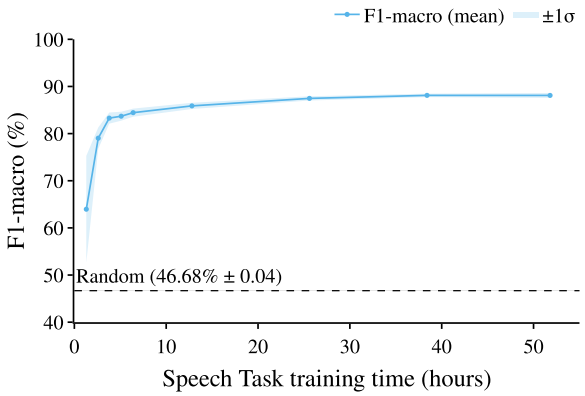

Figure 4: *Data size ablation for Speech Detection Task.*

results align with prior observations [4] that decoding performance continues to scale with the amount of high-quality MEG data available for training.

7.3. Frequency band contribution to the tasks

To probe which parts of the MEG spectrum are most informative for each task, we trained the same decoders while restricting the input to standard neurophysiological bands (Delta, Theta, Alpha, Beta, Gamma, and High-Gamma Activity; HGA) [33, 43]. Figure 6 summarizes the resulting F1-macro scores (mean±std over 10 seeds), together with chance and the unfiltered baseline (dotted line). We assess significance with one-sample Wilcoxon tests against chance.

In Speech Detection task, decoding is consistently above chance in low-frequency bands: Delta (75.86% ±1.94, $W = 55.0$, $p < 0.001$), Theta (59.58% ±2.98, $W = 55.0$, $p < 0.001$), Alpha (76.26% ±1.13, $W = 55.0$, $p < 0.001$), and Beta (72.41% ±1.30, $W = 55.0$, $p < 0.001$). In contrast, Gamma (50.88% ±2.45, $W = 38.0$, $p = 0.161$) and HGA (47.29% ±0.70, $W = 0.0$, $p = 1.000$) are at or below chance. Relative to the full-band baseline (87.59% ±0.70), the best single-band results (Alpha and Delta) retain ∼87% of

Table 3: *Phoneme task test scores over ten seeds before ensembling. Bold marks the best scores.*

Variant	F1_{macro} (%)	$\text{Accuracy}_{\text{bal}}$ (%)	$\text{AUROC}_{\text{macro}}$ (%)
Our model	**44.29 ± 2.84**	47.75 ± 2.70	**96.67 ± 0.31**
Fixed groups	38.39 ± 2.83	40.43 ± 2.77	93.53 ± 1.66
No weights	40.92 ± 3.18	43.88 ± 3.53	95.87 ± 1.02
Conformer Small	44.09 ± 4.56	**48.49 ± 4.19**	96.33 ± 0.60

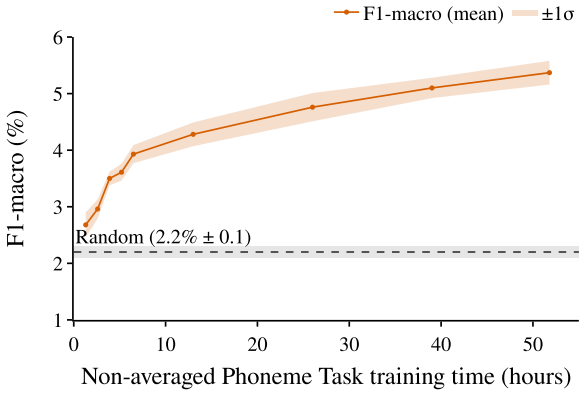

Figure 5: *Data size ablation for Phoneme Classification Task, trained without sample averaging*

.

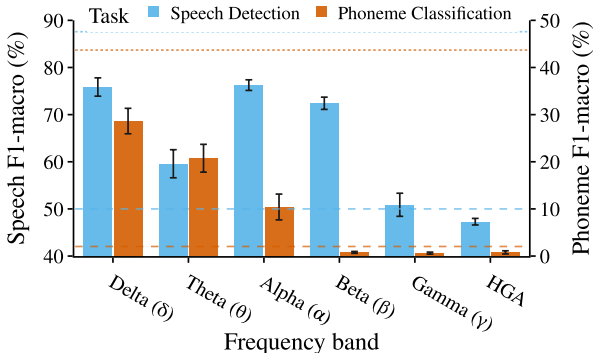

Figure 6: *Decoding mean scores by frequency band for Speech and Phoneme Tasks (left/right y-axes), with black bars showing their standard deviation. The dashed lines represent the chance F1-macro level, and the dotted lines indicate the baseline (ceiling) F1-macro obtained without frequency filtering.*

baseline performance, suggesting that, given the spatiotemporal sensitivity of MEG, speech/silence information accessible to non-invasive recordings is largely conveyed by slow cortical dynamics up to the Beta range, with higher frequencies being either less informative or less reliably measurable.

Similarly, in Phoneme Classification task, decoding is also significantly above chance in Delta (28.64% ±2.70, $W = 55.0$, $p < 0.001$), Theta (20.75% ±2.95, $W = 55.0$, $p < 0.001$), and Alpha (10.41% ±2.74, $W = 55.0$, $p < 0.001$), but falls below chance in Beta, Gamma and HGA, all being not significant ($W = 0.0$, $p = 1.000$). All band-limited models remain substantially below the unfiltered baseline (43.69% ±4.76): Delta retains ∼66% of baseline, Theta ∼48%, and higher bands contribute little in isolation. Significance tests confirm that each band-restricted setting is worse than the baseline in both tasks ($W = 0.0$, $p = 1.000$), indicating that speech decoding benefits from combining information across frequency ranges rather than relying on a single band. However, this effect should be interpreted with caution, as the averaging applied in the phoneme task might have suppressed higher-frequency components.

Overall, both tasks emphasize low-frequency components, with negligible standalone contribution from Gamma and HGA in this dataset. This pattern is consistent with MEG evidence that low-frequency (Delta/Theta) activity supports speech decoding [44], whereas higher-frequency information appears more accessible in invasive recordings [45], and also aligns with earlier frequency-based MEG decoding analyses [23].

7.4. Distribution shift across splits in the phoneme task

To explain the holdout generalization gap observed in the phoneme task, we quantify scale mismatches between partitions by computing the per-window root-mean-square (RMS) energy across channels and time as in Equation 3. The x is the MEG window (C=306 channels, T=125 samples). For validation and test, windows were read with the standard competition preprocessing and standardization enabled. The holdout was read as stored (already standardized on disk). We then formed outline-only histograms using the same bin edges and density normalization across splits.

$$\text{RMS}(x) = \sqrt{mean(x^2)} \quad \text{for } x \in \mathbb{R}^{C\times T}, \tag{3}$$

Figure 7 shows that the holdout distribution is bimodal (two peaks near 0.08 and 0.62), with markedly larger dispersion (mean 0.64, standard deviation 0.22), whereas validation and test are unimodal and tighter (validation 0.54 ± 0.09, test 0.51 ± 0.09). This shift motivated the instance-level normalization used in the Phoneme model to remove amplitude and scale differences window-by-window, thereby improving robustness on the holdout.

7.5. Phonetic-feature decoding as a linguistically grounded probe

As an additional analysis to probe linguistically grounded representations for MEG decoding, we trained separate binary classifiers targeting core phonetic features related to manner of articulation and voicing without sample averaging. Each model was adapted from the Phoneme Classification configuration described in Subsection 4.3, using the same backbone and training hyperparameters, but modified for binary output. Notably, we replaced the 39-class head with a single-logit output, used a binary cross-entropy loss with label smoothing and automatic

Table 4: *Decoding performance for binary phonetic features in the non-averaged test set (in %).*

Feature	Positives (%)	$F1_{macro}$ (%)	$F1_{pos}$ (%)	$Accuracy_{bal}$ (%)	$AUROC_{macro}$ (%)
Speech	*76.76*	*88.10 ± 0.52*	*94.42 ± 0.17*	*86.38 ± 0.97*	*96.38 ± 0.18*
Phoneme	–	*5.37 ± 0.21*	*4.00 ± 1.08*	*5.86 ± 0.22*	*64.35 ± 1.05*
Voicing	77.38	57.77 ± 0.56	83.23 ± 1.30	57.38 ± 0.69	63.88 ± 0.45
Plosive	18.82	55.58 ± 0.53	25.54 ± 1.86	55.31 ± 0.57	62.09 ± 1.22
Fricative	18.55	53.78 ± 0.78	23.85 ± 2.07	53.70 ± 0.71	57.71 ± 1.34
Affricate	0.97	49.69 ± 0.03	0.00 ± 0.00	49.98 ± 0.06	60.29 ± 1.92
Nasal	11.26	51.74 ± 0.54	12.98 ± 1.06	51.67 ± 0.50	55.97 ± 1.47
Liquid	7.51	50.80 ± 1.02	8.16 ± 3.03	51.02 ± 0.57	55.81 ± 1.20
Glide	3.62	52.16 ± 0.81	6.70 ± 1.65	51.83 ± 0.70	60.93 ± 2.25

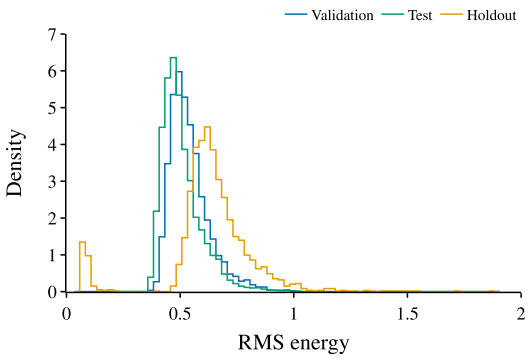

Figure 7: *RMS energy distribution per split in Phoneme Task.*

positive-class weighting following Equation 1 to address class imbalance (based on the Speech model in Subsection 4.2).

This approach follows the ideas proposed by the competition authors in their blog on language-inspired phoneme classification [46], where phonetic features are suggested to handle dataset imbalance and improve infrequent phoneme decoding.

Based on [47], we defined six phonetic features based on manner of articulation: Plosive, Fricative, Affricate, Nasal, Liquid, Glide, and Voicing that distinguishes voiced from voiceless phonemes, including vowels and consonants. Table 4 reports preliminary decoding results using F1-macro, positive-class F1, balanced accuracy, and AUROC (mean ± std across ten seeds).

Following Subsection 7.1, statistical significance is evaluated with a one-sample Wilcoxon signed-rank test against chance (0.5 for F1-macro).

Among these features, Voicing showed the strongest decoding signal, approaching 58% F1-macro (significant with $W = 55.0$, $p = 0.001$). Despite its similarity in label balance to the Speech Task, scores are not on par, implying that decoding some fine-grained speech features may not be straightforward. Manner-based features such as Plosive and Fricative were moderately decodable (both significant with $W = 55.0$, $p = 0.001$), while Affricate, a composite and under-represented category (0.97%), remained at chance (not significant with $W = 0.0$, $p = 1.0$). Attempts to address the affricate issue through ensembling of Plosive and Fricative models, transfer learning from these features, and larger effective batch sizes (via gradient accumulation) did not improve the convergence. This supports the view that decomposing phonemes into binary linguistic features alone is insufficient to overcome the data scarcity, as some linguistic features may also be rare enough not to be easily decodable, like affricates here.

Overall, these experiments suggest that feature-based representations are promising for improving interpretability and possibly scaling, but further research is needed to address underrepresented classes and explore multitask formulations that jointly learn shared articulatory subspaces.

8. Discussion and conclusion

The Conformer architecture proves suitable to decode the spatiotemporal nature of MEG signals, combining convolutional blocks that capture local temporal structure with self-attention for long-range dependencies. Using very similar models across both tasks, we achieved top-10 leaderboard performance, with minor input preprocessing differences between Speech and Phoneme decoding. In particular, our approach won the Phoneme Classification track in the LibriBrain Competition 2025. Instance-level normalization is indispensable to mitigate distributional shifts between training and holdout splits in the Phoneme Task, substantially improving generalization. The additional analyses in Section 7 suggest that this robustness is closely tied to distribution shift across splits. Moreover, our frequency-band analysis shows that both speech and phoneme decoding draw most of their discriminative power from low-frequency components (Delta–Beta), while Gamma and HGA contribute little in isolation. This suggests that, in the LibriBrain setting, effective MEG decoding benefits from integrating information across slow cortical dynamics, and high-frequency activity contribution is insignificant. Furthermore, a simple SpecAugment-based [32] augmentation was effective for speech detection, but showed limited benefit for the more complex, imbalanced phoneme classification task. On the other hand, as LibriBrain is single-subject, cross-subject generalization remains an open challenge for future work.

Looking ahead, adapting speech-model architectures for MEG decoding could open the way toward end-to-end speech reconstruction using sequential objectives such as CTC or seq2seq heads, and even enable speech synthesis. Another promising direction is the use of linguistically grounded feature-based representations [46], which may help tackle data imbalance and improve interpretability by decomposing phoneme classification into binary articulatory features (see Subsection 7.5 for the challenges found). Finally, our scaling experiments (Subsection 7.2) indicate that while speech decoding performance starts to saturate, phoneme classification continues to improve with additional data, suggesting further gains could be achieved as larger MEG datasets become available.

9. Acknowledgements

The Basque and Spanish Governments fund this research (IKUR-IKA-23/18, and AIA2025-163317-C31 by MICIU/AEI / 10.13039/501100011033/). The authors thank both the DIPC Supercomputing Center and the technological management body of the Basque Government, EJIE, for their technical and human support. This work was conducted as part of the #neural2speech and brAIn2lang teams. Last but not least, special thanks to Ekain Arrieta for inspiring the application of Conformer and ASR-based architectures to MEG decoding.

10. References

[1] J. T. Panachakel and A. G. Ramakrishnan, "Decoding covert speech from EEG-a comprehensive review," *Frontiers in Neuroscience*, vol. Volume 15 - 2021, 2021. [Online]. Available: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2021.642251

[2] A. Dehgan, H. Abdelhedi, V. Hadid, I. Rish, and K. Jerbi, "Artificial neural networks for magnetoencephalography: a review of an emerging field," *Journal of Neural Engineering*, vol. 22, no. 3, p. 031001, jun 2025. [Online]. Available: https://doi.org/10.1088/1741-2552/addd4a

[3] S. A. Murad and N. Rahimi, "Unveiling thoughts: A review of advancements in EEG brain signal decoding into text," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 17, no. 1, pp. 61–76, 2025.

[4] M. Özdogan, G. Landau, G. Elvers, D. Jayalath, P. Somaiya, F. Mantegna, M. Woolrich, and O. P. Jones, "Libribrain: Over 50 hours of within-subject MEG to improve speech decoding methods at scale," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. [Online]. Available: https://openreview.net/forum?id=smlCP7jHN3

[5] G. Landau, M. Özdogan, G. Elvers, F. Mantegna, P. Somaiya, D. Jayalath, L. Kurth, T. Kwon, B. Shillingford, G. Farquhar, M. Jiang, K. Jerbi, H. Abdelhedi, Y. Mantilla Ramos, C. Gulcehre, M. Woolrich, N. Voets, and O. Parker Jones, "The 2025 PNPL competition: Speech detection and phoneme classification in the LibriBrain dataset," *NeurIPS Competition Track*, 2025.

[6] 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 *Interspeech*, 2020, pp. 5036–5040.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://Proc..neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[8] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=BJC_jUqxe

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[10] M. Burchi and V. Vielzeuf, "Efficient conformer: Progressive downsampling and grouped attention for automatic speech recognition," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 8–15.

[11] S. Kim, A. Gholami, A. Shaw, N. Lee, K. Mangalam, J. Malik, M. W. Mahoney, and K. Keutzer, "Squeezeformer: An efficient transformer for automatic speech recognition," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9361–9373, 2022.

[12] Y. Peng, S. Dalmia, I. Lane, and S. Watanabe, "Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding," in *International Conference on Machine Learning*. PMLR, 2022, pp. 17 627–17 643.

[13] K. Kim, F. Wu, Y. Peng, J. Pan, P. Sridhar, K. J. Han, and S. Watanabe, "E-branchformer: Branchformer with enhanced merging for speech recognition," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 84–91.

[14] D. Rekesh, N. R. Koluguri, S. Kriman, S. Majumdar, V. Noroozi, H. Huang, O. Hrinchuk, K. Puvvada, A. Kumar, J. Balam *et al.*, "Fast conformer with linearly scalable attention for efficient speech recognition," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[15] B. Accou, J. Vanthornhout, H. V. hamme, and T. Francart, "Decoding of the speech envelope from EEG using the VLAAI deep neural network," *Scientific Reports*, vol. 13, no. 1, p. 812, Jan 2023. [Online]. Available: https://doi.org/10.1038/s41598-022-27332-2

[16] X. Xu, B. Wang, Y. Yan, H. Zhu, Z. Zhang, X. Wu, and J. Chen, "ConvConcatNet: a deep convolutional neural network to reconstruct mel spectrogram from the EEG," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2024, pp. 113–114.

[17] D. Dash, P. Ferrari, S. Malik, A. Montillo, J. A. Maldjian, and J. Wang, "Determining the optimal number of MEG trials: A machine learning and speech decoding perspective," in *Brain Informatics*, S. Wang, V. Yamamoto, J. Su, Y. Yang, E. Jones, L. Iasemidis, and T. Mitchell, Eds. Cham: Springer International Publishing, 2018, pp. 163–172.

[18] D. Dash, P. Ferrari, S. Malik, and J. Wang, "Overt speech retrieval from neuromagnetic signals using wavelets and artificial neural networks," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2018, pp. 489–493.

[19] D. Dash, P. Ferrari, D. Heitzman, and J. Wang, "Decoding speech from single trial MEG signals using convolutional neural networks and transfer learning," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 5531–5535.

[20] M. Sarma, C. Bond, S. Nara, and H. Raza, "MEGNet: A MEG-based deep learning model for cognitive and motor imagery classification," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2023, pp. 2571–2578.

[21] A. Klimovich-Gray, G. Di Liberto, L. Amoruso, A. Barrena, E. Agirre, and N. Molinaro, "Increased top-down semantic processing in natural speech linked to better reading in dyslexia," *NeuroImage*, vol. 273, p. 120072, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811923002185

[22] X. de Zuazo, E. Navas, I. Saratxaga, M. Bourguignon, and N. Molinaro, "Phone pair classification during speech production using MEG recordings," in *IberSPEECH 2024*, 2024, pp. 76–80.

[23] X. de Zuazo, E. Navas, I. Saratxaga, M. Bourguignon, and N. Molinaro, "Decoding phone pairs from MEG signals across speech modalities," *Computer Speech & Language*, vol. 99, p. 101939, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230826000033

[24] G. Elvers, G. Landau, and O. P. Jones, "Elementary, my dear watson: Non-invasive neural keyword spotting in the libribrain dataset," in *Data on the Brain & Mind Findings*, 2025. [Online]. Available: https://openreview.net/forum?id=gRJ9dd07QF

[25] 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.

[26] Y. Benchetrit, H. Banville, and J.-R. King, "Brain decoding: toward real-time reconstruction of visual perception," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=3y1K6buO8c

[27] H. J. Banville, Y. Benchetrit, S. d'Ascoli, J. Rapin, and J.-R. King, "Scaling laws for decoding images from brain activity," *CoRR*, vol. abs/2501.15322, January 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2501.15322

[28] D. Jayalath, G. Landau, and O. P. Jones, "Unlocking non-invasive brain-to-text," in *ICML 2025 Generative AI and Biology (GenBio) Workshop*, 2025. [Online]. Available: https://openreview.net/forum?id=lRVOPe1EZy

[29] Carnegie Mellon University, "The CMU pronouncing dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict, 1993, version 0.7b.

[30] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[31] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/f1748d6b0fd9d439f71450117eba2725-Paper.pdf

[32] 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 *Interspeech 2019*, 2019, pp. 2613–2617.

[33] P. K. Mandal, A. Banerjee, M. Tripathi, and A. Sharma, "A comprehensive review of magnetoencephalography (MEG) studies for brain functionality in healthy aging and alzheimer's disease (AD)," *Frontiers in Computational Neuroscience*, vol. 12, p. 60, Aug 2018. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fncom.2018.00060

[34] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.

[35] B. Bakirarar and A. H. Elhan, "Class weighting technique to deal with imbalanced class problem in machine learning: Methodological research," *Türkiye Klinikleri Biyoistatistik*, vol. 15, no. 1, pp. 19–29, 2023.

[36] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2017. [Online]. Available: https://arxiv.org/abs/1607.08022

[37] ——, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4105–4113.

[38] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[39] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945. [Online]. Available: http://www.jstor.org/stable/3001968

[40] G. Santafe, I. Inza, and J. A. Lozano, "Dealing with the evaluation of supervised classification algorithms," *Artificial Intelligence Review*, vol. 44, no. 4, pp. 467–508, 12 2015. [Online]. Available: https://doi.org/10.1007/s10462-015-9433-y

[41] M. Tagliasacchi, Y. Li, K. Misiunas, and D. Roblek, "Seanet: A multi-modal speech enhancement network," in *Interspeech 2020*, 2020, pp. 1126–1130.

[42] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," 2020. [Online]. Available: https://arxiv.org/abs/2001.08361

[43] J. Gruenwald, S. Sieghartsleitner, C. Kapeller, J. Scharinger, K. Kamada, P. Brunner, and C. Guger, "Characterization of high-gamma activity in electrocorticographic signals," *Frontiers in Neuroscience*, vol. 17, 2023. [Online]. Available: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2023.1206120

[44] G. B. Cogan and D. Poeppel, "A mutual information analysis of neural coding of speech by low-frequency MEG phase information," *Journal of neurophysiology*, vol. 106, no. 2, pp. 554–563, 2011.

[45] T. Proix, J. Delgado Saa, A. Christen, S. Martin, B. N. Pasley, R. T. Knight, X. Tian, D. Poeppel, W. K. Doyle, O. Devinsky *et al.*, "Imagined speech can be decoded from low-and cross-frequency intracranial EEG features," *Nature communications*, vol. 13, no. 1, p. 48, 2022.

[46] T. Kwon, S. Cho, G. Elvers, F. Mantegna, and O. Parker Jones, "Language-inspired approaches to phoneme classification," 2025, blog post. [Online]. Available: https://neural-processing-lab.github.io/2025-libribrain-competition/blog/language-inspired-approaches-phoneme-classification

[47] N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang, "Phonetic feature encoding in human superior temporal gyrus," *Science*, vol. 343, no. 6174, pp. 1006–1010, 2014. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1245994