Physiological Noise Augmentation Improves Non-Invasive Brain-to-Speech

Benjamin Ballyk [*1] **Teyun Kwon** [*1] **Miran Özdogan** [1] **Oiwi Parker Jones** [1]

Abstract

Non-invasive brain-to-speech decoding aims to restore communication to patients suffering from neurodegenerative disease, without the risks of neurosurgery. Existing MEG and EEG-based methods, while scalable, continue to suffer from high word error rates driven by extremely low signal-to-noise ratios as task-agnostic brain activity dominates recordings. We propose *physiological noise augmentation* (PNA), an ICA-based data augmentation method that explicitly trains decoders to become invariant to artifacts (e.g. ocular and cardiac activity), using NLP-inspired feature remixing to generate biophysically realistic, label-preserving training examples. Combining PNA with multi-trial averaging suppresses residual unmodeled variability that is not phase-locked across repeated trials. We further show that, to first order, PNA approximates anisotropic regularization that penalizes decoder sensitivity along artifact-dominated directions. On MegNIST, a 12-hour imagined-digit MEG dataset, PNA with 10-trial averaging improves decoding accuracy with EEGNet by 8.8% over training on real data alone. Our results suggest that artifact-aware augmentation and trial averaging are complementary tools for improving robustness in non-invasive speech BCIs.

1. Introduction

Neurodegenerative disease, stroke, and cervical spine injuries collectively affect more than 110 million patients globally, and often irreversibly impair one's ability to articulate thoughts into speech (Feigin et al., 2021; Injury et al., 2019; Park et al., 2022). For nearly four decades, brain-to-speech research has sought to restore communication for these patients, yet this goal remains a central challenge at the intersection of neuroscience and biomedical engineering (Farwell & Donchin, 1988; Birbaumer et al., 1999; Wolpaw et al., 2002; Card et al., 2024). Although recent intracranial brain-computer interfaces (BCIs) have achieved increasingly accurate decoding of intended speech from cortical activity (Moses et al., 2021; Willett et al., 2023; Card et al., 2024), these invasive systems carry risks of neurosurgical complications and long-term electrode instability (Leuthardt et al., 2021).

Consequently, there is a growing imperative to develop safer and more accessible non-invasive alternatives based on magnetoencephalography (MEG), electroencephalography (EEG), or functional magnetic resonance imaging (fMRI) (Wolpaw et al., 2002; d'Ascoli et al., 2025). Non-invasive brain recordings, however, suffer from low signal-to-noise ratio (SNR), particularly in the context of imagined speech, which lacks overt articulatory and auditory feedback (Pei et al., 2011; Martin et al., 2018; Panachakel & Ramakrishnan, 2021; Martin et al., 2014; 2016). For several decades, this limitation has been addressed using multi-trial averaging, in which repeated recordings of the same thought are averaged to enhance phase-locked signal relative to independent noise (Sutton et al., 1965). However, this approach is often slow and burdensome, as users must repeat themselves several times, which limits practical applicability.

Concurrently, data collection bottlenecks have motivated augmentation strategies to expose decoding models to a broader range of realistic test-time conditions. While often modestly beneficial, existing augmentations typically operate at the input signal level rather than targeting underlying neurophysiological artifacts, leaving models exposed to structured noise that dominates imagined speech signals (Luo et al., 2021; He et al., 2021).

We address this gap by introducing *physiological noise augmentation* (PNA), a data augmentation framework that isolates artifact-related independent components (e.g., ocular and cardiac activity) and re-injects remixed variants to generate physiologically realistic, label-preserving training examples. By exposing the decoder to a wider range of artifact realizations, PNA encourages invariance to nuisance structure and promotes reliance on task-relevant signals. Combined with multi-trial averaging to reduce residual

[*]Equal contribution [1]PNPL, Department of Engineering Science, University of Oxford. Correspondence to: Benjamin Ballyk <benjamin.ballyk@eng.ox.ac.uk>, Teyun Kwon <teyun.kwon@eng.ox.ac.uk>.

Proceedings of the Workshop on Structured Data for Health at the 43rd International Conference on Machine Learning, Seoul, South Korea. Copyright 2026 by the author(s).

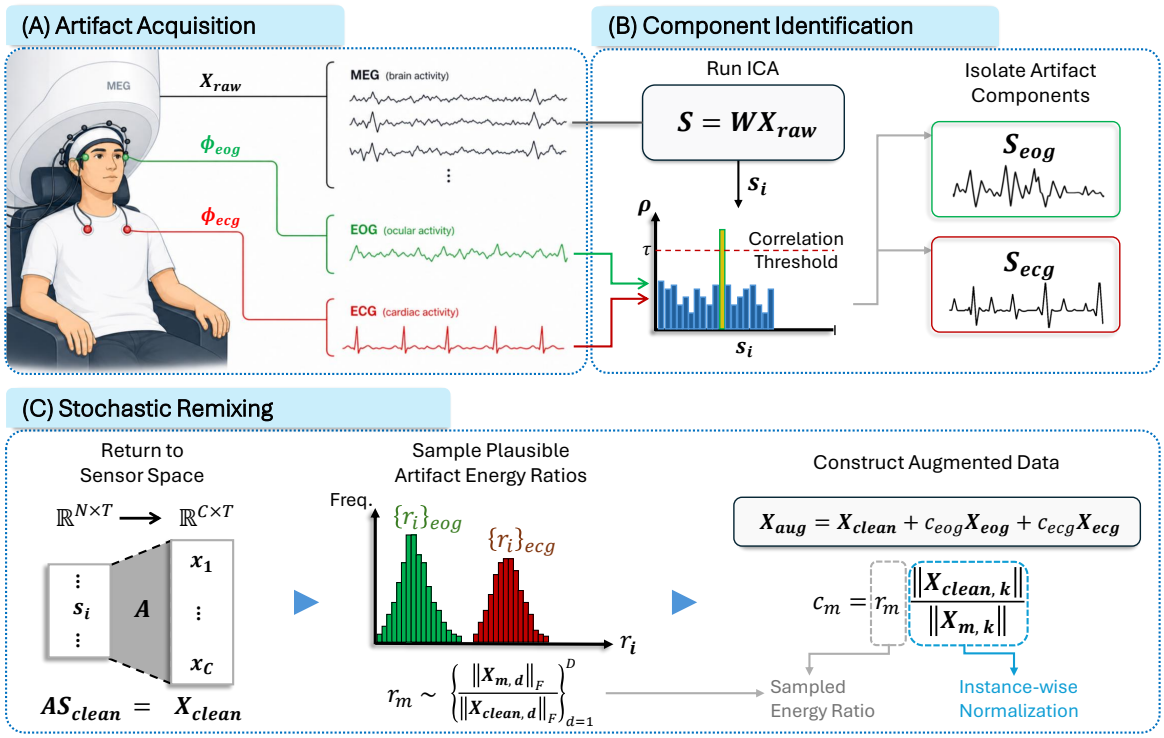

Figure 1. **Methodological breakdown of physiological noise augmentation**.

variability, this approach reduces the number of repetitions required to achieve strong decoding performance.

Contributions. We introduce PNA, an augmentation framework for noninvasive brain-to-speech that promotes invariance to task-agnostic artifacts by remixing artifact-related independent components. PNA generates realistic sensor-space perturbations by scaling artifacts according to the empirical distribution of energy ratios. We provide theoretical support for PNA by showing that, under multi-trial averaging, the method is first-order equivalent to anisotropic weight decay that penalizes weight directions aligned with tracked artifacts. Empirically, PNA substantially improves performance on the MegNIST imagined speech dataset, increasing decoding accuracy by 8.76% over real-data baselines and reaching 70.9% accuracy with EEGNet. A detailed review of prior work is provided in Appendix A.1.

2. Method

We model brain-to-speech decoding as a multiclass classification problem over a fixed vocabulary, $\mathcal{V}$. Given a discrete spatiotemporal recording $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T] \in \mathbb{R}^{C\times T}$, where C is the number of sensors (channels), T is the number of time samples, and $\mathbf{x}_t \in \mathbb{R}^C$ denotes the sensor measurements at time t, our goal is to learn a decoder $f_\theta : \mathbb{R}^{C\times T} \to \mathcal{V}$ that maps $\mathbf{X}$ to a vocabulary item. In practice, $\mathbf{X}$ is an additive mixture of task-relevant neural activity, $\mathbf{X}_{\text{task}}$, and task-agnostic physiological 'artifacts', $\mathbf{X}_{\text{artifact}}$, such as cardiac or ocular activity.

Noninvasive decoding is often confounded by artifacts that co-vary with the stimulus. We therefore seek a decoder that is structurally invariant to the artifact subspace, satisfying

$$f_\theta(v \mid \mathbf{X}_{\text{task}}, \mathbf{X}_{\text{artifact}}) = f_\theta(v \mid \mathbf{X}_{\text{task}}), \quad \forall v \in \mathcal{V}. \quad (1)$$

2.1. Physiological Noise Augmentation

PNA seeks to enforce Equation (1) in three steps (Figure 1): (i) recording nuisance reference channels during data acquisition, (ii) identifying and removing artifact-related ICA components, and (iii) re-injecting scaled artifact projections into the cleaned data to generate physiologically plausible synthetic samples.

Step 1: Measuring artifacts. During recording experiments spanning T_{total} time steps, we collect reference waveforms for task-agnostic artifacts.[1] In MegNIST data used for our experiments, electrooculography (EOG) and electrocardiography (ECG) are recorded to capture ocular and cardiac activity, respectively (Figure 1A). Let $\phi_p \in \mathbb{R}^{T_{\text{total}}}$ denote the reference time series for artifact $p \in \mathcal{P}$.

[1]Recording experiments are generally divided into several sessions, with independent component analysis repeated for each session to account for changes in sensor placement.

Step 2: Removing artifact-correlated sources. PNA assumes that sensor recordings may be separated into statistically independent source components using *independent component analysis* (ICA) (Bell & Sejnowski, 1995), which we overview in Appendix A.2. We fit FastICA (Hyvärinen, 1999) on recordings using MNE-Python (Gramfort et al., 2013), yielding source estimates, $\widehat{\mathbf{S}}$.

Artifact-related components are identified by their correlation with reference channels. For each tracked artifact type $p \in \mathcal{P}$, we define

$$\mathcal{S}_p = \{i \in \{1, \ldots, N\} \ : \ |\rho(\hat{\mathbf{s}}_i, \ \boldsymbol{\phi}_p)| \geq \tau_p\},$$

as the set of associated nuisance component indeces. Here, $\hat{\mathbf{s}}_i \in \mathbb{R}^T$ is the time-series of the i-th ICA component (the i-th row of $\widehat{\mathbf{S}}$), $\boldsymbol{\phi}_p$ denotes the reference time-series for artifact p, and $\rho(\cdot, \cdot)$ is Pearson correlation computed over T_{total} time samples. τ_p are selected correlation thresholds.

Similarly, we define the set of clean indeces, $\mathcal{C}$, as the set of ICA component indeces remaining after removing all artifact-correlated indeces,

$$\mathcal{C} = \{1, \ldots, N\} \setminus \bigcup_{p \in \mathcal{P}} \mathcal{S}_p. \tag{2}$$

The cleaned recording, $\mathbf{X}_{\text{clean}}$, is recovered by linearly projecting the retained clean ICA components back into sensor space. Figure 7 in Appendix D shows a topographical and waveform view of extracted ICA components using EOG (ocular) and ECG (cardiac) reference signals, alongside ICA component correlations.

Step 3: Augmenting with physiological noise. We generate augmented trials, $\mathbf{X}_{\text{aug}}$, using the structure

$$\mathbf{X}_{\text{aug}} = \mathbf{X}_{\text{clean}} + \sum_{p \in \mathcal{P}} c_p \mathbf{X}_p. \tag{3}$$

Here, $\mathbf{X}_p$ represents the sensor-space projection of components indexed in $\mathcal{S}_p^{(q)}$, and c_p is an empirically sampled scaling coefficient. For each augmented trial, we opt to sample c_p to match realistic artifact conditions and to simulate a broad range of signal-to-artifact ratios. We describe this process in Appendix A.3, and we provide pseudocode for the full augmentation procedure in Algorithm 1 (B.1).

Thus far, we have only considered the effect of *tracked* artifacts on decoding. Optionally, after applying PNA, we further apply *multi-trial averaging* to suppress residual untracked artifacts that are not phase-locked to internal speech onset. Multi-trial averaging has long been used to improve SNR in noninvasive brain-to-text (Farwell & Donchin, 1988); we provide a brief overview in Appendix A.4.

2.2. Theoretical Motivation

Proposition 1 (PNA as Regularization)**.** *Let $f(\mathbf{x}; \theta)$ be a differentiable decoder. Let $\bar{\mathbf{x}}$ be the clean K-trial average and let*

$$\delta = \frac{\alpha}{K} \sum_{i=1}^{K} \tilde{\mathbf{n}}_i$$

be the injected, temporally shuffled artifact noise with covariance $\Sigma_\delta = \frac{\alpha^2}{K} \Sigma_n$. For sufficiently small α, minimizing the MSE loss with augmented data is equivalent to minimizing the clean loss plus a penalty on the Frobenius norm of the Jacobian weighted by the artifact covariance:

$$\mathcal{L}_{\text{aug}}(\theta) \ \approx \ \mathcal{L}_{\text{clean}}(\theta) \ + \ \text{Tr}\big(\mathbf{J}_{\bar{\mathbf{x}}} \, \Sigma_\delta \, \mathbf{J}_{\bar{\mathbf{x}}}^\top\big),$$

where $\mathbf{J}_{\bar{\mathbf{x}}} = \frac{\partial f(\bar{\mathbf{x}}; \theta)}{\partial \bar{\mathbf{x}}}$ is the Jacobian of the model outputs with respect to the inputs.

Proof. See Appendix C.

Proposition 1 is related to the equivalence between Gaussian noise and Tikhanov regularization (Bishop, 1995). Under an MSE objective, PNA is first-order equivalent to minimizing the clean (artifact-free) loss plus a covariance-weighted penalty on sensitivity to nuisance directions. It achieves this by expanding label-preserving nuisance realizations: instead of only naturally occurring signal–artifact pairings, the decoder sees many plausible combinations with the same target. In effect, augmentation steers the decoder toward invariance by replacing the dataset-induced nuisance covariance with a broader, deliberately constructed covariance in the tracked artifact subspace.

2.3. Preprocessing & Implementation Details

In our pipeline, we perform data augmentation before preprocessing so that augmentation preserves the original relative scales and variances of physiological signals. We follow the preprocessing strategy of Défossez et al., 2023, applying a robust scaler fitted only to the training data to avoid leakage. Full details are presented in Appendix E.

3. Experiments

We evaluate PNA on the task of classifying imagined digits from MEG data, and compare to a set of common brain-to-speech augmentations.

Dataset. We conduct experiments on MegNIST, a single-subject MEG dataset comprising 12 hours (12,000 trials) of class-balanced imagined digits (0–9). In addition to MEG recordings, MegNIST records EOG and ECG reference signals. Test-time averaging is applied only across consecutive repetitions of the same class, with no trial reuse. Preprocessing and data splits are detailed in Appendix E.

Models. We train two models for each augmentation setting: a simple multilayer perceptron (MLP) and EEGNet (Lawhern et al., 2018). Hyperparameters and training procedures are outlined in Appendix E.

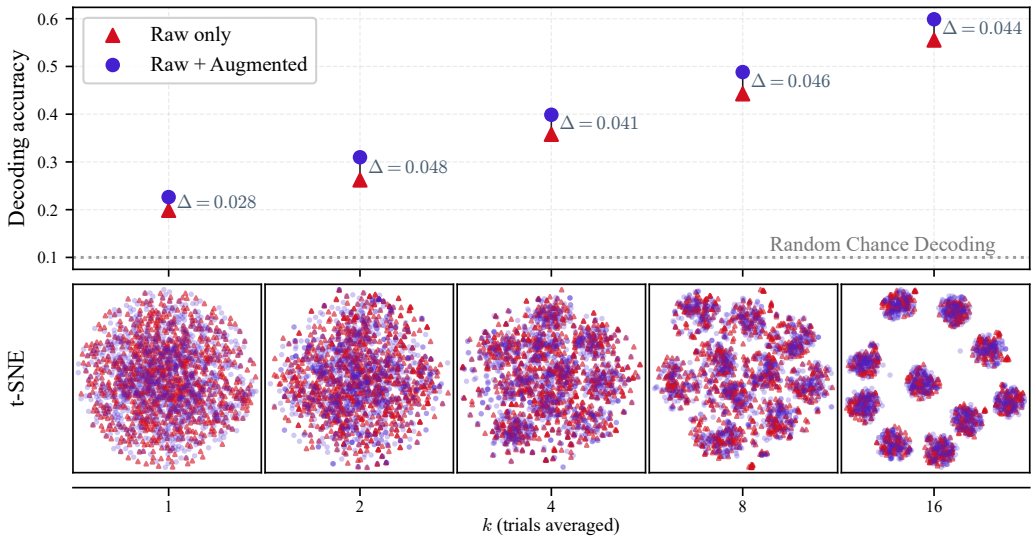

Figure 2. **MLP decoding accuracy uplift via PNA augmentation and associated t-SNE embeddings at various levels of averaging**. The top plot uses 10k training trials for the raw-only runs (red) and an additional 10k augmented trials for raw + augmented runs (blue). The bottom plots show t-SNE embeddings of raw (red) and augmented (blue) data at each level of averaging.

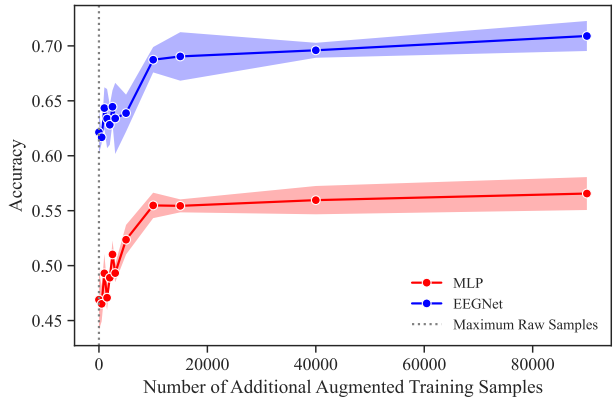

Figure 3. **Accuracy of MLP and EEGNet vs. number of augmented training samples.** The dotted line marks training on 10k raw trials (no augmentation). Shaded regions show 95% SE; all runs use 5 seeds and 10-trial averaging.

3.1. Decoding Performance and the Effect of Averaging

Table 4 shows an ablation of PNA, using. The framework yields modest improvements in the single-trial regime (e.g., EEGNet: $30.9\% \pm 1.2\%$ to $32.1\% \pm 1.3\%$), however its impact is substantially amplified when combined with multi-trial averaging. With 10-trial averaging, accuracies increase from $46.9\% \pm 3.2\%$ to $56.6\% \pm 1.7\%$ for MLP and from $62.1\% \pm 2.3\%$ to $70.9\% \pm 1.6\%$ for EEGNet when augmented samples are included, corresponding to an order-of-magnitude larger gain than in the single-trial setting. This result suggests that augmentation is most effective once averaging has mitigated extreme low-SNR conditions, allowing models to better exploit invariances introduced by PNA. Full results across models and training configurations are provided in Appendix F.1. We use a 1:1 raw-to-augmented data ratio for simplicity and as supported by Figure 3.

3.2. Comparison to Baseline Augmentations

Tables 5 and 6 in Appendix F.2 compares the effects of several augmentation baselines with and without 10-trial averaging, and combines PNA with each of the baseline augmentations, further improving performance over augmenting only on either.

4. Conclusions and Future Work

We present Physiological Noise Augmentation (PNA), an ICA-based framework that enforces invariance to physiological artifacts in non-invasive brain-to-speech decoding. By remixing artifact components to generate diverse, label-preserving samples, PNA improves decoding accuracy by 8.76% over real-data training alone for MEG-based imagined digit classification with EEGNet. We further show that PNA and multi-trial averaging are complementary: PNA reduces sensitivity to tracked nuisances, while averaging suppresses residual task-agnostic variability, together approximating anisotropic regularization. Future work will extend PNA to multi-subject recordings, and will include additional reference signals (e.g. EMG) to accelerate the development of viable non-invasive speech neuroprostheses.

References

Armeni, K., Güçlü, U., van Gerven, M., and Schoffelen, J.-M. A 10-hour within-participant magnetoencephalography narrative dataset to test models of language comprehension. *Scientific Data*, 9(1):278, Jun 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01382-7.

Bell, A. J. and Sejnowski, T. J. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995. doi: 10.1162/neco.1995.7.6.1129.

Birbaumer, N., Ghanayim, N., Hinterberger, T., Iversen, I., Kotchoubey, B., Kübler, A., Perelmouter, J., Taub, E., and Flor, H. A spelling device for the paralysed. *Nature*, 398(6725):297–298, Mar 1999. ISSN 1476-4687. doi: 10.1038/18581. URL https://doi.org/10.1038/18581.

Bishop, C. M. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116, 1995.

Card, N. S., Wairagkar, M., Iacobacci, C., Hou, X., Singer-Clark, T., Willett, F. R., Kunz, E. M., Fan, C., Vahdati Nia, M., Deo, D. R., Srinivasan, A., Choi, E. Y., Glasser, M. F., Hochberg, L. R., Henderson, J. M., Shahlaie, K., Stavisky, S. D., and Brandman, D. M. An accurate and rapidly calibrating speech neuroprosthesis. *New England Journal of Medicine*, 391(7):609–618, 2024. doi: 10.1056/NEJMoa2314132.

Comon, P. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

d'Ascoli, S., Bel, C., Rapin, J., Banville, H., Benchetrit, Y., Pallier, C., and King, J.-R. Towards decoding individual words from non-invasive brain recordings. *Nature Communications*, 16(1):10521, Nov 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-65499-0.

Dawson, G. D. A summation technique for the detection of small evoked potentials. *Electroencephalography & clinical neurophysiology*, 1954.

de Zuazo, X., Saratxaga, I., and Navas, E. Meg-conformer: Conformer-based meg decoder for robust speech and phoneme classification. *arXiv e-prints*, pp. arXiv:2512.01443, December 2025. doi: 10.48550/arXiv.2512.01443.

Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., and King, J.-R. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10):1097–1107, Oct 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00714-5.

Dehgan, A., Abdelhedi, H., Hadid, V., Rish, I., and Jerbi, K. Artificial neural networks for magnetoencephalography: a review of an emerging field. *Journal of Neural Engineering*, 22(3):031001, jun 2025. doi: 10.1088/1741-2552/addd4a.

Farwell, L. A. and Donchin, E. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical Neurophysiology*, 70(6):510–523, 1988.

Feigin, V. L., Stark, B. A., Johnson, C. O., Roth, G. A., Bisignano, C., Abady, G. G., Abbasifard, M., Abbasi-Kangevari, M., Abd-Allah, F., Abedi, V., et al. Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the global burden of disease study 2019. *The Lancet Neurology*, 20(10):795–820, 2021.

Gideoni, Y., Timms, R. C., and Parker Jones, O. Non-invasive neural decoding in source reconstructed brain space. *arXiv e-prints*, pp. arXiv:2410.19838, October 2024. doi: 10.48550/arXiv.2410.19838.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., et al. Meg and eeg data analysis with mne-python. *Frontiers in Neuroinformatics*, 7:267, 2013.

Gwilliams, L., Flick, G., Marantz, A., Pylkkänen, L., Poeppel, D., and King, J.-R. Introducing meg-masc a high-quality magneto-encephalography dataset for evaluating natural speech processing. *Scientific Data*, 10(1):862, Dec 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02752-5.

Han, D. D., Gwon, Y., Lee, A. L., Lee, T., Lee, S. J., Choi, J., Lee, S., Bang, J., Lee, S., Park, D. K., et al. Diver-1: Deep integration of vast electrophysiological recordings at scale. *arXiv preprint arXiv:2512.19097*, 2025.

He, C., Liu, J., Zhu, Y., and Du, W. Data augmentation for deep neural networks model in eeg classification task: a review. *Frontiers in Human Neuroscience*, 15:765525, 2021.

Huang, R., Cho, S., Gohil, C., Parker Jones, O., and Woolrich, M. Meg-gpt: A transformer-based foundation model for magnetoencephalography data. *arXiv e-prints*, pp. arXiv:2510.18080, October 2025. doi: 10.48550/arXiv.2510.18080.

Hyvärinen, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999. doi: 10.1109/72.761722.

Hyvärinen, A. and Oja, E. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5): 411–430, 2000. doi: 10.1016/S0893-6080(00)00026-5.

Injury, G. et al. Global, regional, and national burden of traumatic brain injury and spinal cord injury, 1990-2016: a systematic analysis for the global burden of disease study 2016. *Lancet Neurol*, 18(1):56–87, 2019.

Jayalath, D., Landau, G., and Parker Jones, O. Unlocking non-invasive brain-to-text. *arXiv e-prints*, pp. arXiv:2505.13446, May 2025. doi: 10.48550/arXiv.2505.13446.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of Neural Engineering*, 15 (5):056013, July 2018. doi: 10.1088/1741-2552/aace8c.

Leuthardt, E. C., Schalk, G., Roland, J., Rouse, A., and Moran, D. Evolution of brain–computer interfaces: Going beyond classic motor physiology. *Frontiers in Neuroscience*, 15:599549, 2021. doi: 10.3389/fnins.2021.599549.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7. arXiv:1711.05101.

Luo, J., Wang, Y., Xu, R., Liu, G., Wang, X., and Gong, Y. Channel drop out: A simple way to prevent cnn from overfitting in motor imagery based bci. In *International Conference of Pioneering Computer Scientists, Engineers and Educators*, pp. 443–452. Springer, 2021.

Martin, S., Brunner, P., Holdgraf, C., Heinze, H.-J., Crone, N. E., Rieger, J., Schalk, G., Knight, R. T., and Pasley, B. N. Decoding spectrotemporal features of overt and covert speech from the human cortex. *Frontiers in Neuroengineering*, 7:14, 2014. doi: 10.3389/fneng.2014.00014.

Martin, S., Brunner, P., Iturrate, I., Millán, J. d. R., Schalk, G., Knight, R. T., and Pasley, B. N. Word pair classification during imagined speech using direct brain recordings. *Scientific Reports*, 6:25803, 2016. doi: 10.1038/srep25803.

Martin, S., Iturrate, I., Millán, J. d. R., Knight, R. T., and Pasley, B. N. Decoding inner speech using electrocorticography: Progress and challenges toward a speech prosthesis. *Frontiers in Neuroscience*, Volume 12 - 2018, 2018. ISSN 1662-453X. doi: 10.3389/fnins.2018.00422. URL https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2018.00422.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Moses, D. A., Metzger, S. L., Liu, J. R., Anumanchipalli, G. K., Makin, J. G., Sun, P. F., Chartier, J., Dougherty, M. E., Liu, P. M., Abrams, G. M., Tu-Chan, A., Ganguly, K., and Chang, E. F. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227, 2021. doi: 10.1056/NEJMoa2027540.

Özdogan, M., Landau, G., Elvers, G., Jayalath, D., Somaiya, P., Mantegna, F., Woolrich, M., and Parker Jones, O. Libribrain: Over 50 hours of within-subject meg to improve speech decoding methods at scale. *arXiv e-prints*, pp. arXiv:2506.02098, June 2025. doi: 10.48550/arXiv.2506.02098.

Panachakel, J. T. and Ramakrishnan, A. G. Decoding covert speech from eeg-a comprehensive review. *Frontiers in Neuroscience*, Volume 15 - 2021, 2021. ISSN 1662-453X. doi: 10.3389/fnins.2021.642251. URL https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2021.642251.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv e-prints*, pp. arXiv:1904.08779, April 2019. doi: 10.48550/arXiv.1904.08779.

Park, J., Kim, J.-E., and Song, T.-J. The global burden of motor neuron disease: an analysis of the 2019 global burden of disease study. *Frontiers in Neurology*, 13: 864339, 2022.

Pei, X., Barbour, D. L., Leuthardt, E. C., and Schalk, G. Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. *Journal of Neural Engineering*, 8(4):046028, 2011. doi: 10.1088/1741-2560/8/4/046028.

Rommel, C., Paillard, J., Moreau, T., and Gramfort, A. Data augmentation for learning predictive models on eeg: a systematic comparison. *Journal of Neural Engineering*, 19(6):066020, November 2022. doi: 10.1088/1741-2552/aca220.

Song, J., Zhai, Q., Wang, C., and Liu, J. Eeggan-net: enhancing eeg signal classification through data augmentation. *Frontiers in Human Neuroscience*, 18, 2024. ISSN 1662-5161. doi: 10.3389/fnhum.2024.1430086.

Sutton, S., Braren, M., Zubin, J., and John, E. R. Evoked-potential correlates of stimulus uncertainty. *Science*, 150

(3700):1187–1188, 1965. doi: 10.1126/science.150.3700.1187. URL https://www.science.org/doi/abs/10.1126/science.150.3700.1187.

Torma, S. and Szegletes, L. Generative modeling and augmentation of eeg signals using improved diffusion probabilistic models. *Journal of Neural Engineering*, 22(1): 016001, jan 2025. doi: 10.1088/1741-2552/ada0e4.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Wang, Z., Li, S., Chen, X., and Wu, D. Time–frequency transform based eeg data augmentation for brain–computer interfaces. *Knowledge-Based Systems*, 311:113074, 2025. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2025.113074.

Willett, F. R., Kunz, E. M., Fan, C., Avansino, D. T., Wilson, G. H., Choi, E. Y., Kamdar, F., Glasser, M. F., Hochberg, L. R., Druckmann, S., Shenoy, K. V., and Henderson, J. M. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023. doi: 10.1038/s41586-023-06377-x.

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M. Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):767–791, June 2002. doi: 10.1016/S1388-2457(02)00057-3.

Yano, H., Takashima, R., Takiguchi, T., and Nakagawa, S. Representation learning based on variational autoencoders for imagined speech classification. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, pp. 1546–1550, August 2024. doi: 10.23919/EUSIPCO63174.2024.10715355.

A. Additional Background

A.1. Related Work

We organize related work around three threads that define the technical setting of this paper: non-invasive neural decoding, MEG speech datasets, and data augmentation for neural time series.

Non-Invasive Neural Decoding. While invasive BCIs have achieved near-errorless attempted speech decoding performance in the past half-decade (Moses et al., 2021; Willett et al., 2023; Card et al., 2024), noninvasive neural data remains challenged by low signal-to-noise ratios and relatively scarce training data. Défossez et al. (2023) address the data availability issue by adopting a contrastive learning objective that aligns non-invasive brain responses with self-supervised speech representations, identifying the matching 3-s speech segment from MEG with up to 41% top-1 accuracy across participants. d'Ascoli et al. (2025) scale training across participants, devices, languages, and tasks to learn more transferable word-level representations, achieving up to 37% balanced top-10 accuracy over a 250-word retrieval set. Jayalath et al. (2025) incorporate linguistic context through language-model rescoring and out-of-vocabulary predictive infilling, achieving a word error rate of 88% and a character error rate of 68%. Notably, these works focus on decoding perceived speech. In contrast, we study single-subject MEG classification of imagined digits, a setting closer to decoding self-generated internal speech in minimally responsive patients.

Datasets. Recent evidence suggests that data scaling remains the primary bottleneck for high-fidelity noninvasive brain-to-speech systems (Han et al., 2025). Several high-quality public datasets have emerged recently that contain MEG recordings of perceived English speech, including MEG-MASC (Gwilliams et al., 2023) (27 subjects, $\sim$54 total hours), the Deep-MEG Sherlock dataset (Armeni et al., 2022) (3 subjects, $\sim$30 total hours), and LibriBrain (Özdogan et al., 2025) (1 subject, $>$50 total hours). These datasets record neural responses while participants listen to external speech, typically audiobooks, rather than during internal (imagined) speech. As a result, the data is not directly aligned with the setting of brain-to-speech systems that aim to decode self-generated, covert language. Currently, MegNist is the only publicly available MEG internal speech dataset, containing 12 hours of single-subject data of imagined digits (0-9).

Data Augmentation. Data augmentation is a common strategy for improving neural decoders in low-data regimes, but its effectiveness depends strongly on the task and transformation family (Rommel et al., 2022). Existing neural-signal augmentations typically fall into two categories: deterministic signal transformations and learned generative modeling (Dehgan et al., 2025). Signal processing approaches have produced modest single-trial decoding gains through time-frequency transforms (Wang et al., 2025), SpecAugment-style masking (de Zuazo et al., 2025; Park et al., 2019), and spatial dropout or channel masking (Gideoni et al., 2024). Generative approaches instead aim to expand the training distribution by synthesizing neural data, using models such as conditional variational autoencoders (Yano et al., 2024), generative adversarial networks (Song et al., 2024), diffusion models (Torma & Szegletes, 2025), and foundation models for MEG reconstruction and synthesis (Huang et al., 2025). Our proposed method differs from these strategies by infusing additional signal information from artifact electrodes to guide physiologically plausible augmentations.

A.2. Independent Component Analysis

Independent component analysis (ICA) is modeled on the assumption that a matrix of spatiotemporal sensor recordings, $\mathbf{X}$, arises from instantaneous linear mixing of N independent sources, $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_T] \in \mathbb{R}^{N \times T}$ according to

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t, \qquad \text{for } t = 1, \ldots, T,$$

or equivalently,

$$\mathbf{X} = \mathbf{A}\mathbf{S}, \tag{4}$$

where $\mathbf{A} \in \mathbb{R}^{C \times N}$ is an unknown mixing matrix. *Independent component analysis* (ICA) seeks to recover latent independent sources (rows of $\mathbf{S}$), by estimating an unmixing matrix $\mathbf{W} \in \mathbb{R}^{N \times C}$ which yields component estimates,

$$\widehat{\mathbf{S}} = \mathbf{W}\mathbf{X}, \tag{5}$$

such that the rows of $\widehat{\mathbf{S}}$ are as statistically independent as possible (Comon, 1994). A canonical objective is to pick the optimal unmixing matrix, $\mathbf{W}^*$, to minimize the mutual information of the component random variables:

$$\mathbf{W}^\star \in \underset{\mathbf{W} \in \mathbb{R}^{N \times C}}{\arg\min}\, I(\hat{\mathbf{s}}_{1,:}, \ldots, \hat{\mathbf{s}}_{N,:}), \tag{6}$$

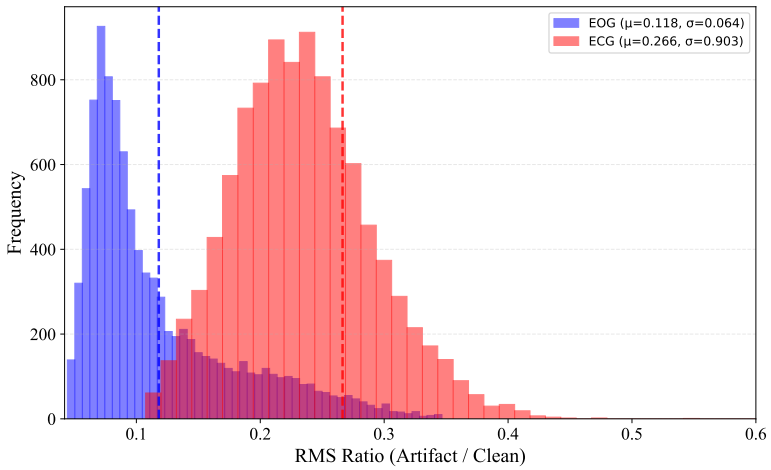

Figure 4. **Empirical distributions of Amplitude ratios of EOG-to-clean signal (blue) and ECG-to-clean signal (red)**. Zeros have been removed for clarity. We observe that cardiac components tend to correspond to a larger magnitude or neural activity relative to EOG components

where $\hat{\mathbf{s}}_{i,:} \in \mathbb{R}^{1\times T}$ is the i-th row of $\widehat{\mathbf{S}}$.

Notably, the estimation of $\mathbf{W}$ was originally formulated as training a single-layer neural network for unsupervised learning, as in the Infomax ICA algorithm (Bell & Sejnowski, 1995). However, current implementations typically use the FastICA algorithm, which improves computational efficiency by using a fixed-point iteration to maximize a non-Gaussianity contrast as proxy for statistical independence (Hyvärinen, 1999; Hyvärinen & Oja, 2000).[2]

A.3. Stochastic Amplitude Sampling

We propose to sample the artifact scaling coefficients by explicitly matching the empirical distribution of artifact-to-signal energy ratios observed in the data, reflecting realistic physiological conditions.

Crucially, as magnetometers and gradiometers differ in measurement units and scale (approx. 10^{-15} T vs. 10^{-13} T/m), a global Frobenius norm would be dominated by gradiometer readings. We therefore compute energy ratios separately for each sensor type. For brevity, however, we show the formulation in the single-sensor case.

Concretely, let $\|\cdot\|_F$ denote the Frobenius norm over sensors (of the same type) and time, and let $\mathbf{X}_{p,d}$ represent the reconstructed data for tracked artifact p from training trial $d \in \{1, ..., D\}$. We compute the empirical amplitude ratios

$$r_{p,d} = \frac{\|\mathbf{X}_{p,d}\|_F}{\|\mathbf{X}_{\text{clean},d}\|_F + \varepsilon},$$

where $\varepsilon > 0$ is a small constant for numerical stability. These ratios characterize the distribution of observed artifact strengths relative to the clean neural signal.

During training, we sample each r_p from the empirical distributions $\{r_{p,d}\}_{d=1}^{D}$. Figure 4 shows the empirical distributions of EOG and ECG energy ratios, on the same axes. We note that ECG signal tends to carry a higher relative energy, roughly twice that of EOG.

During the forward pass, given clean trial $\mathbf{X}_{\text{clean}}$, and randomly selected artifact snippets $\mathbf{X}_p$ for all $p \in \mathcal{P}$, we set

$$c_p = r_p \frac{\|\mathbf{X}_{\text{clean}}\|_F}{\|\mathbf{X}_p\|_F + \varepsilon}, \tag{7}$$

This procedure ensures that augmented samples preserve realistic artifact structure while spanning the observed range of artifact-to-signal ratios. Sampling from the full distribution of physiological noise strengths allows the training objective to

[2]Most ICA algorithms also approximate the mixing matrix, $\mathbf{A}$, by taking an inverse or pseudoinverse of $\mathbf{W}$.

more closely match the expected risk under realistic test-time corruption. As a result, the learned decoder becomes robust to variable ocular and cardiac contamination.

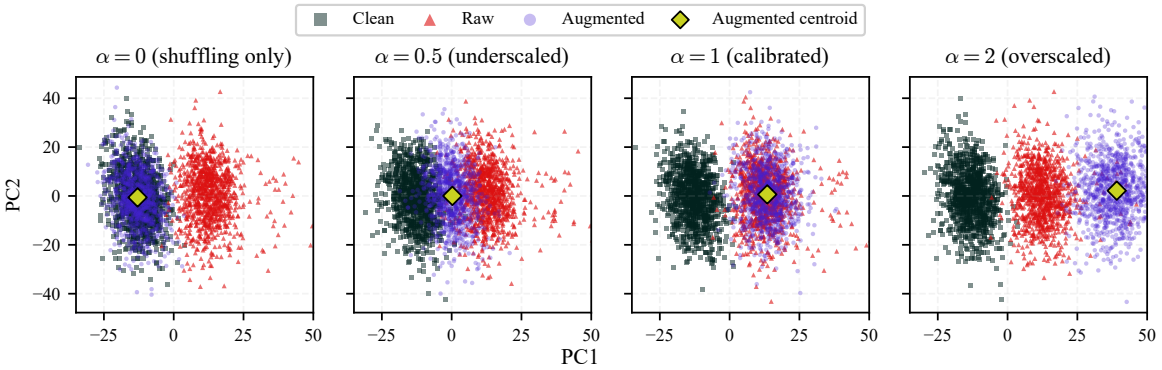

Figure 5. **PCA Projections of 15-trial Averaged MEG Data for Imagined Digits at Various Augmentation Scalings**. Raw and clean data form distinct clusters, while augmented data effectively interpolates between them for scaling parameter $\alpha \in [0, 1]$. At $\alpha = 1$, the projection matches the calibration strategy defined in (7). For $\alpha > 1$, embeddings shift away from the support of the biological region. The misalignment between augmented and clean data at $\alpha = 0$ is attributed to the random indexing in the augmentation step (3) and the visualization of a sample subset.

A.4. Multi-Trial Averaging and Resampling

We define the K-trial average for a set of trials $\{\mathbf{X}^{(k)}\}_{k=1}^{K}$ as the spatiotemporal mean:

$$\overline{\mathbf{X}} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{X}^{(k)}. \tag{8}$$

In brain-to-speech decoding, K-trial averaging at inference requires the patient to imagine the target stimulus K times. This approach assumes that each trial consists of a phase-locked neural signal plus independent stochastic noise; consequently, averaging preserves the coherent signal component while the noise variance is attenuated by a factor of K. For K condition-matched trials with covariance $\mathbf{\Sigma}$, the sample average $\overline{\mathbf{X}}$ satisfies $\mathrm{Cov}(\overline{\mathbf{X}}) = \mathbf{\Sigma}/K$ under the assumption of independence, yielding a $\sqrt{K}$ scaling in Signal-to-Noise Ratio (SNR) (Dawson, 1954).

During training, we consider two averaging strategies:

1. **Averaging without resampling:** Trials are partitioned into disjoint groups of size K and averaged. While simple, this reduces the effective training set size by a factor of K, limiting input variability.

2. **Averaging with resampling:** Subsets of size K are randomly drawn (without replacement within each subset) from the full training set. For a dataset of size N, there are $\binom{N}{K}$ possible subsets, a combinatorial explosion (when $\min(K, N-K) >> 0$) that allows for massive data augmentation while preserving the original dataset's coverage.

In this work, we employ averaging with resampling. To ensure a fair comparison with single-trial models, we fix the number of generated samples to match the original training set cardinality N.

Crucially, averaging implicitly infuses task awareness into output data, as phase-locked portions of the signal remain relatively unaffected, while other components destructively interfere.[3] Figure 6 shows t-SNE embeddings (Van der Maaten & Hinton, 2008) of MegNIST data before and after averaging with resampling, within label classes. We observe that without averaging, t-SNE is unable to locate local clusters separating imagined digit classes, however, clear clusters appear neatly along class boundaries after 15-trial averaging. We repeat this experiment in Figure 8 of Appendix D using UMAP (McInnes et al., 2018) to establish a more faithful representation of inter-class global structure.

[3]MegNIST and similar datasets prompt participants to imagine each label at a screen-cued moment, enabling precise temporal alignment across trials.

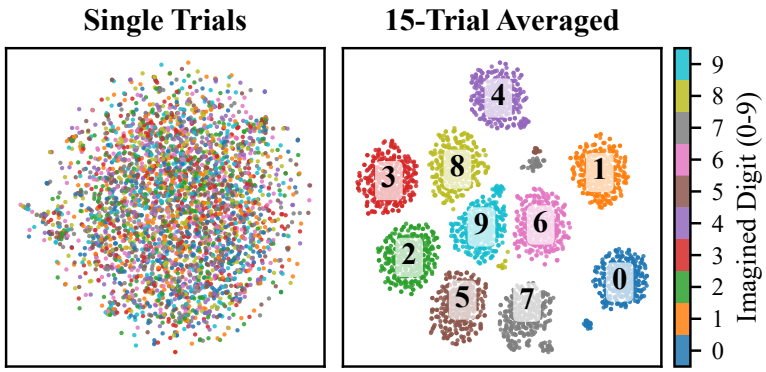

Figure 6. t-SNE embeddings of single-trial (left) and 15-trial averaged (right) intra-patient MEG recordings of imagined digits (0–9). Averaging is performed with resampling to preserve dataset cardinality. t-SNE parameters: perplexity = 30.

B. Pseudo-code

B.1. PNA Pseudo-code

Algorithm 1 Physiological Noise Augmentation

Require: Session data $\{\mathbf{X}^{(q)}\}$, artifact reference signals $\{\phi_p^{(q)}\}_{p\in\mathcal{P}}$, thresholds $\{\tau_p\}$, stability constant ε

// 1. Extract Artifacts & Clean Subspace
for session $q = 1 \dots Q$ **do**
  $\mathbf{W}, \mathbf{A} \leftarrow \text{FastICA}(\mathbf{X}^{(q)})$ ▷ Obtain mixing matrix
  $\widehat{\mathbf{S}}^{(q)} \leftarrow \mathbf{W}\mathbf{X}^{(q)}$ ▷ Identify ICA components
  $\mathcal{S}_p^{(q)} \leftarrow \{i : |\rho(\hat{\mathbf{s}}_i, \phi_p)| \geq \tau_p\}_{p\in\mathcal{P}}$ ▷ Obtain artifact indices via correlation
  $\mathcal{C}^{(q)} \leftarrow \{1 \dots N\} \setminus \left(\bigcup_{p\in\mathcal{P}} \mathcal{S}_p^{(q)}\right)$ ▷ Obtain remaining clean indeces
  $\mathbf{X}_p^{(q)} \leftarrow \mathbf{A}_{:,\mathcal{S}_p}\widehat{\mathbf{S}}_{\mathcal{S}_p,:} \quad \forall p \in \mathcal{P}$ ▷ Restore artifacts to sensor space
  $\mathbf{X}_{\text{clean}}^{(q)} \leftarrow \mathbf{A}_{:,\mathcal{C}}\widehat{\mathbf{S}}_{\mathcal{C},:}$ ▷ Restore artifact-free data to sensor space
end for

// 2. Empirical Ratio Calibration
for trial $d = 1 \dots D$, artifact $p \in \mathcal{P}$ **do**
  $r_{k,d} \leftarrow \|\mathbf{X}_{k,d}\|_F / (\|\mathbf{X}_{\text{clean},d}\|_F + \varepsilon)$
end for

// 3. Augmentation Forward Pass
function AUGMENT($\mathbf{X}_{\text{clean}}$)
  Sample trial index $j \in \{1, \dots, D\}$
  $c_p \leftarrow r_{p,j} \cdot \|\mathbf{X}_{\text{clean}}\|_F / (\|\mathbf{X}_{p,j}\|_F + \varepsilon) \quad \forall p \in \mathcal{P}$
  return $\mathbf{X}_{\text{clean}} + \sum_{p\in\mathcal{P}} c_p \mathbf{X}_{p,j}$
end function

C. Proof of Proposition 1

Proof. **Step 1: First-order Taylor expansion.** We approximate the model output on augmented data $\mathbf{x}_{\text{aug}} = \bar{\mathbf{x}} + \delta$ via a Taylor expansion around the clean average $\bar{\mathbf{x}}$:

$$f(\bar{\mathbf{x}} + \delta;\, \theta) \;\approx\; f(\bar{\mathbf{x}};\, \theta) + \mathbf{J}_{\bar{\mathbf{x}}}\, \delta,$$

where higher-order terms $O(\|\delta\|^2)$ are neglected under the assumption that α is small.

Step 2: Expansion of the augmented loss. The augmented loss is the expectation over the joint distribution of signal, label, and noise:

$$\mathcal{L}_{\text{aug}} \;=\; \mathbb{E}_{\mathbf{x},\,y,\,\delta}\left[\,\left\|y - f(\bar{\mathbf{x}} + \delta;\,\theta)\right\|^2\right].$$

Substituting the Taylor approximation gives

$$\mathcal{L}_{\text{aug}} \;\approx\; \mathbb{E}\Big[\,\left\|y - f(\bar{\mathbf{x}};\,\theta) - \mathbf{J}_{\bar{\mathbf{x}}}\,\delta\right\|^2\Big].$$

Step 3: Distributing the expectation. Define the clean residual $\mathbf{e} = y - f(\bar{\mathbf{x}};\,\theta)$. Expanding the squared norm yields

$$\mathcal{L}_{\text{aug}} \;\approx\; \mathbb{E}\big[\mathbf{e}^\top\mathbf{e}\big] \;-\; 2\,\mathbb{E}\big[\mathbf{e}^\top\mathbf{J}_{\bar{\mathbf{x}}}\,\delta\big] \;+\; \mathbb{E}\big[\delta^\top\mathbf{J}_{\bar{\mathbf{x}}}^\top\mathbf{J}_{\bar{\mathbf{x}}}\,\delta\big].$$

Step 4: The cross-term vanishes. Because $\tilde{\mathbf{n}}$ is temporally shuffled, δ is independent of the clean signal $\bar{\mathbf{x}}$ and the label y, and hence independent of the residual $\mathbf{e}$ and the Jacobian $\mathbf{J}_{\bar{\mathbf{x}}}$. By independence,

$$\mathbb{E}\big[\mathbf{e}^\top\mathbf{J}_{\bar{\mathbf{x}}}\,\delta\big] \;=\; \mathbb{E}\big[\mathbf{e}^\top\mathbf{J}_{\bar{\mathbf{x}}}\big]\cdot\mathbb{E}[\delta].$$

Since the noise components are zero-mean, $\mathbb{E}[\delta] = \mathbf{0}$, and the entire cross-term vanishes.

Step 5: Evaluating the penalty term. We are left with

$$\mathcal{L}_{\text{aug}} \;\approx\; \mathcal{L}_{\text{clean}} \;+\; \mathbb{E}\big[\delta^\top\mathbf{J}_{\bar{\mathbf{x}}}^\top\mathbf{J}_{\bar{\mathbf{x}}}\,\delta\big].$$

Applying the trace identity $\mathbb{E}[\mathbf{z}^\top\mathbf{A}\,\mathbf{z}] = \mathrm{Tr}(\mathbf{A}\,\Sigma_z)$, which holds for any zero-mean random vector $\mathbf{z}$ with covariance Σ_z and deterministic matrix $\mathbf{A}$, with $\mathbf{z} = \delta$ and $\mathbf{A} = \mathbf{J}_{\bar{\mathbf{x}}}^\top\mathbf{J}_{\bar{\mathbf{x}}}$, gives

$$\mathbb{E}\big[\delta^\top\mathbf{J}_{\bar{\mathbf{x}}}^\top\mathbf{J}_{\bar{\mathbf{x}}}\,\delta\big] \;=\; \mathrm{Tr}\big(\mathbf{J}_{\bar{\mathbf{x}}}^\top\mathbf{J}_{\bar{\mathbf{x}}}\,\Sigma_\delta\big).$$

Applying the cyclic property of the trace and substituting $\Sigma_\delta = \frac{\alpha^2}{K}\Sigma_n$ yields

$$\mathrm{Tr}\big(\mathbf{J}_{\bar{\mathbf{x}}}^\top\mathbf{J}_{\bar{\mathbf{x}}}\,\Sigma_\delta\big) \;=\; \mathrm{Tr}\big(\mathbf{J}_{\bar{\mathbf{x}}}\,\Sigma_\delta\,\mathbf{J}_{\bar{\mathbf{x}}}^\top\big),$$

and therefore

$$\mathcal{L}_{\text{aug}}(\theta) \;\approx\; \mathcal{L}_{\text{clean}}(\theta) \;+\; \mathrm{Tr}\big(\mathbf{J}_{\bar{\mathbf{x}}}\,\Sigma_\delta\,\mathbf{J}_{\bar{\mathbf{x}}}^\top\big). \qquad \square$$

D. Additional Data Visualization

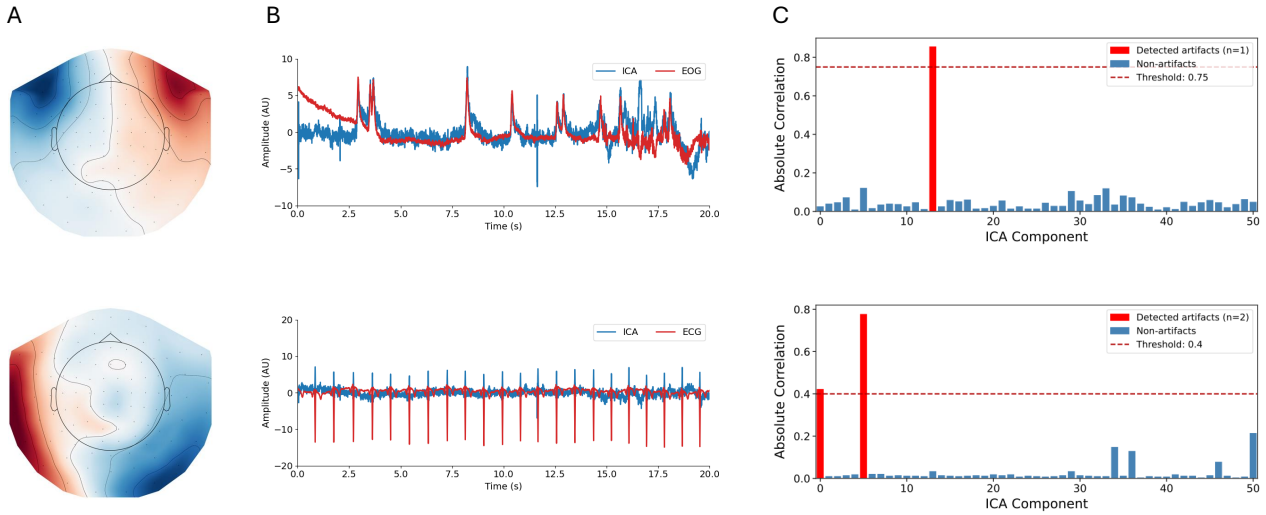

Figure 7. **ICA Artifact Component Selection for EOG and ECG Signals**. The topographies captured in column **A** show the relative strengths of the ICA mapping from source to sensor space for a sample EOG (top) and ECG (bottom) component. The corresponding ICA component waveforms are captured in column **B**, where they are compared to the reference EOG and ECG signals. Finally, column **C** show the absolute Pearson correlations between the first 50 ICA components and the artifact waveform for EOG (top) and ECG (bottom).

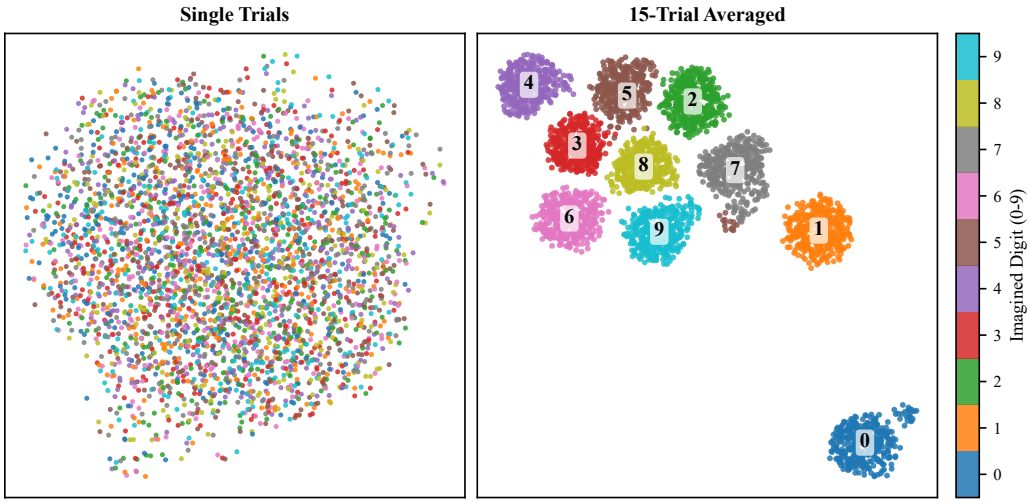

Figure 8. UMAP embeddings of intra-patient MEG recordings for imagined digits (0–9), shown for single trials (left) and 15-trial averages (right). Averaging improves cluster separability, and—unlike t-SNE, UMAP preserves aspects of global structure, revealing a separation between representations of digit 0 and digits 1–9

Unlike t-SNE, UMAP preserves some global structure, and the observed separation of digit 0 from digits 1–9 may reflect a systematic difference in neural representation rather than a purely local clustering artifact. This could arise from semantic differences between zero and non-zero numerals or from task-related strategies. While intgriguing, the effect is preliminary and must be validated across multiple subjects and modalities.

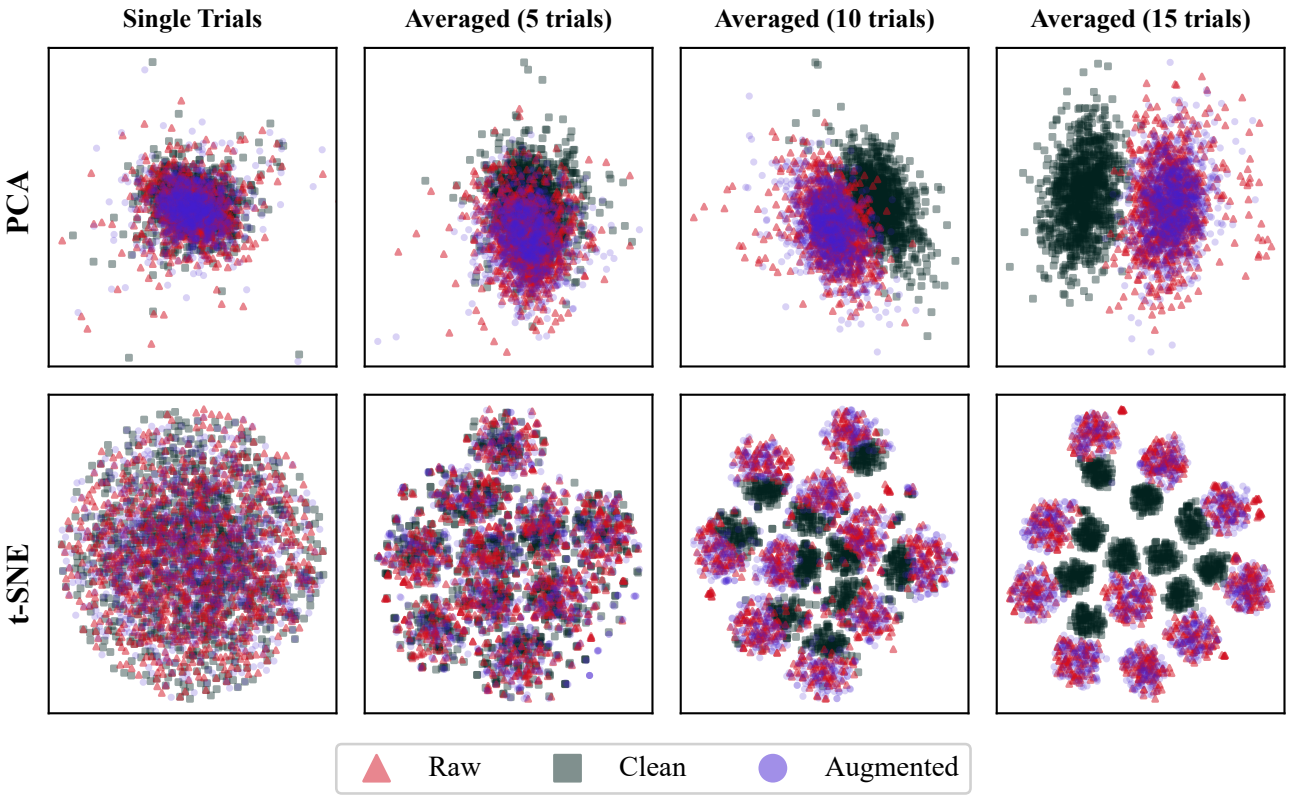

Figure 9. **PCA and t-SNE Embeddings of Raw, Augmented (ocular and cardiac), and Clean Data at Various levels of Averaging**. While raw and augmented data remain closely aligned, the cluster of clean data separates with additional averaging, indicating that PNA correctly restores the corruption to artifact-cleaned data that would be seen at test time. Under t-SNE, label-wise clusters emerge for substantial averaging.

E. Model, Pre-processing and Baseline Augmentation Parameters

Parameter	Value
Data Splits	
Train : Val : Test	10:1:1
Averaging	
Number of Samples to Average	10
# Training Samples (using resampling)	10,000
Architectural Parameters	
MLP # Hidden Layers	1
MLP Hidden Layer Width	128
EEGNet # Feature Maps (F_1)	16
EEGNet # Spatial Filters per Feature Map (D)	4
EEGNet # Pointwise Filters (F_2)	64
EEGNet Kernel Length	25
Training	
Batch Size	64
Maximum # Training Epochs	200
Early Stopping	After 25 Epochs
MLP Dropout Rate	0.35
MLP Learning Rate	0.0001
MLP Weight Decay	0.0005
EEGNet Learning Rate	0.0005
EEGNet Weight Decay	0.001
Optimizer	AdamW (Loshchilov & Hutter, 2019)

Table 1. Hyperparameters used in our experiments.

Parameter	Value
Input MEG Data	
# Gradiometer Channels	102
# Magnetometer Channels	204
Original Sampling Rate	1000 Hz
Preprocessing	
Downsampled rate	250 Hz
High-Pass Filter	0.1 Hz
Low-Pass Filter	120 Hz
Channel-wise Standardization	IQR = [-1, 1]

Table 2. Preprocessing used on MEG data. We apply channel-wise scaling to account for different magnitudes of gradiometer (T/m) and magnetometer (T) channels.

Parameter	Value
Baseline Augmentation	
White Noise std	0.1
Smooth Time Mask Length	50
Frequency Shift	0.5 Hz
Temporal Shift	1 step
Amplitude Scaling	uniformly drawn from [0.9, 1.1]

Table 3. Baseline augmentation hyperparameters, tuned to optimize single-trial validation set performance.

F. Additional Experimental Results

F.1. Physiological Noise Augmentation

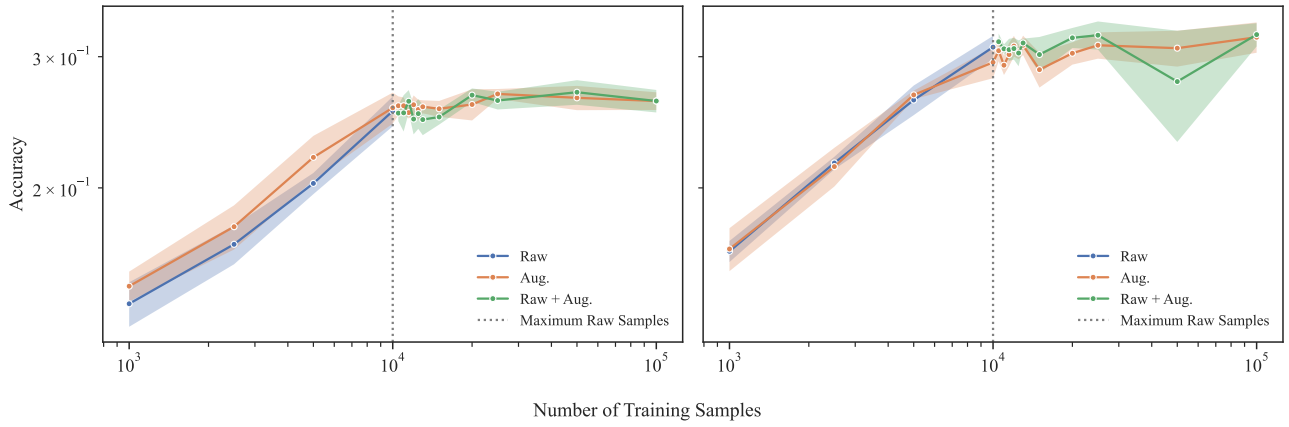

Figure 10. Classification accuracy of an MLP (left) and EEGNet (right) trained on different numbers of single-trial training samples (no averaging). The shaded area shows the standard error with 95% confidence across 5 seeds.

Data Strategy	MLP	EEGNET
Raw Only	0.2536 ± 0.0133	0.3090 ± 0.0124
+ AVERAGING	0.3720 ± 0.0354	0.5200 ± 0.0494
Augmented Only	0.2674 ± 0.0047	0.3186 ± 0.0170
+ AVERAGING	0.5040 ± 0.0301	0.6260 ± 0.0150
Raw + Augmented	0.2688 ± 0.0117	0.3212 ± 0.0132
+ AVERAGING	$\mathbf{0.5200 \pm 0.0447}$	$\mathbf{0.6360 \pm 0.0206}$

Table 4. **Ablation of Data Augmentation Strategies.** *Augmented* refers to the PNA pipeline (Section 2.1). *Raw Only* and *Augmented Only* each consist of 10k training trials, while for *Raw + Augmented*, we report the best result from datasets consisting of 10k *Raw* and 10k of *Augmented* trials (20k total). Averaging uses 10 trials, with resampling during training. For reference, random chance decoding yields an expected accuracy of 0.1. Best results are in **bold**; standard errors are in scriptsize.

For "Raw + Aug.", we concatenate varying amounts of augmented samples to 10,000 raw samples, to create final dataset sizes of 10,500, 11,000, 11,500, 12,000, 12,500, 13,000, 15,000, 20,000, 25,000, 50,000 and 100,000.

F.2. Baseline Augmentations

	MLP		EEGNet	
Augmentation Type	**Single-trial**	**10-trial**	**Single-trial**	**10-trial**
None	0.254 ± 0.013	0.372 ± 0.035	0.309 ± 0.012	0.520 ± 0.049
White Noise	0.229 ± 0.007	0.410 ± 0.016	0.315 ± 0.005	0.582 ± 0.030
Smooth Time Mask	0.235 ± 0.015	0.494 ± 0.022	0.311 ± 0.011	0.570 ± 0.024
Frequency Shift	0.247 ± 0.007	$\mathbf{0.506} \pm 0.023$	$\mathbf{0.329} \pm 0.006$	0.582 ± 0.029
Temporal Shift	0.238 ± 0.015	0.418 ± 0.031	0.306 ± 0.008	0.572 ± 0.034
Amplitude Scaling	0.231 ± 0.008	0.416 ± 0.042	0.313 ± 0.011	0.570 ± 0.040
PNA (Ours)	$\mathbf{0.256} \pm 0.014$	0.462 ± 0.023	0.295 ± 0.016	$\mathbf{0.586} \pm 0.010$

Table 5. **Comparison to Augmentation Baselines**. We train models on 10k augmented trials without raw data, and 10k raw trials without augmentations for *None*. Best results are in **bold**; standard errors are in scriptsize.

Augmentation Type (After PNA)	MLP	EEGNet
White Noise	0.248 ± 0.014	0.330 ± 0.004
Smooth Time Mask	0.252 ± 0.011	0.315 ± 0.009
Frequency Shift	$\mathbf{0.263 \pm 0.003}$	0.323 ± 0.008
Temporal Shift	0.250 ± 0.012	$\mathbf{0.333 \pm 0.015}$
Amplitude Scaling	0.245 ± 0.009	0.327 ± 0.013

Table 6. **Stacked PNA and Baseline Augmentation Results for Single-Trial Training**. We train models on 10,000 augmented trials without raw data. PNA is performed before each baseline augmentation. Best results are in **bold**; standard errors are in scriptsize.