 DOI: 10.3217/978-3-99161-050-2-065

MindVoice: A Multimodal Framework for EEG-Based Speech Decoding

Anarghya Das[1]*, **Matthew Li**[2], **Wenyao Xu**[1]
[1] *University at Buffalo, Buffalo, USA;* [2] *Purdue University, West Lafayette, USA;*
E-mail: anarghya@buffalo.edu

Introduction: Speech impairments pose challenges for automatic speech recognition (ASR) systems, which struggle with slurred pronunciation, unpredictable pauses, and variability in speech patterns [1]. We propose a contrastive learning framework combining speech and EEG, designed to decode speech from neural signals using minimal EEG channels, enabling scalable solutions for speech impairments.

Material, Methods and Results: We developed a robust data collection protocol to record EEG and speech data from 13 participants, which consisted of 20 English words carefully selected to maximize phonemic diversity and three paragraphs commonly used in speaking tests. The stimuli were designed to ensure a broad representation of English phonemes, enabling the model to learn the phonetic patterns of the language. Stimuli presentation was fully automated using PsychoPy [2] and time-synchronized using the Lab Streaming Layer (LSL). A multimodal model was trained to encode EEG and speech data into a shared latent space, employing cross-attention to enrich EEG embeddings with temporally aligned speech audio and a contrastive loss to align these embeddings effectively. Trained on 1,366 samples (1,093 training, 273 validation), the model achieved an impressive top-1 validation accuracy of 98.94%, highlighting its capacity to capture complex EEG-speech relationships. A channel ablation study confirmed robustness, showing minimal accuracy decline to 97.90% even when using a single EEG channel (F4) selected from the cortical region involved in speech production. This minimal performance drop underscores the model's suitability and potential for deployment in practical, low-channel wearable devices. The overall workflow is illustrated in Figure 1.

Conclusion: This study demonstrates a generalizable contrastive learning framework for decoding speech from EEG signals, extending our previous work [3]. Our proposed model provides robust shared embeddings between EEG and speech, facilitating downstream decoding tasks such as zero-shot classification and EEG-to-text applications. Its effectiveness with fewer EEG channels further highlights its potential for scalable deployment in low-channel wearable devices, extending the reach of brain-computer interface (BCI) applications beyond assistive communication.

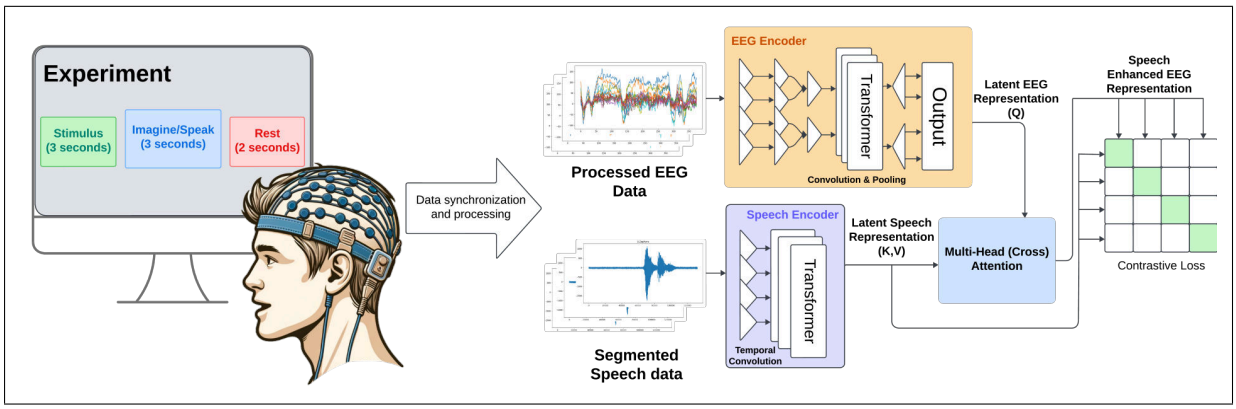

Figure 1: *Overview of experimental setup and pipeline for EEG and audio data collection, synchronization, encoding, and multimodal alignment for speech decoding.*

References:

[1] S. Alharbi, M. Alrazgan, A. Alrashed, T. Alnomasi, R. Almojel, R. Alharbi, S. Alharbi, S. Alturki, F. Alshehri, and M. Almojil. utomatic Speech Recognition: Systematic Literature Review. In *Proceedings of the 14th Annual International Conference of the IEEE/EMBS*, vol. 9, pp. 131 858–131 876, 2021.

[2] Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. PsychoPy2: experiments in behavior made easy.*Behavior Research Methods*, 2019.

[3] Das, P. Soni, M.-C. Huang, F. Lin, and W. Xu. ultimodal speech recognition using EEG and audio signals: A novel approach for enhancing ASR systems. *Smart Health*, vol. 32, p. 100477, 2024.

Published by Verlag der Technischen Universität Graz