EEG-based Speech Decoding Based on Multi-mode Joint Modeling

Peiran Li[1], *Fei Chen*[2], *Xixin Wu*[1]

[1]The Chinese University of Hong Kong, Hong Kong SAR, China
[2]Southern University of Science and Technology, Shenzhen, China
1155174020@link.cuhk.edu.hk, fchen@sustech.edu.cn, wuxx@se.cuhk.edu.hk

Abstract

Electroencephalography (EEG)-based speech decoding enables the development of non-invasive speech brain-computer interfaces (BCIs) for restoring communication of individuals with speech impairments. Previous work achieves much better performance in decoding spoken and intended speech from EEG signals, with imagined speech decoding lagging far behind. This paper proposes a novel framework to train a unified multi-mode decoding model for EEG signals of imagined, intended and spoken speech modes using a dynamic masking mechanism. Our multi-mode model achieves significantly better four-vowel decoding accuracies than baselines (34.95% vs. 29.18% for imagined speech). Training a single-mode model with a subset of EEG channels selected according to a multi-mode model as inputs provides superior performance than training a single-mode model from all channels. The accuracy improvements and channel selection capability demonstrate the effectiveness of the proposed joint modeling framework.

Index Terms: Brain-computer interface, imagined speech, electroencephalogram, joint modeling

1. Introduction

Speech brain-computer interfaces (BCIs) have emerged as a transformative technology with the potential to restore communication for individuals with speech impairments. [1] [2] [3] [4] Among the various modalities available for brain signal acquisition, electroencephalography (EEG) has gained much attention as a non-invasive and cost-effective method for recording neural activity [5]. Compared to invasive techniques such as electrocorticography (ECoG), EEG is safer to setup, with higher time and spatial resolutions, making it suitable for broader applications [6]. However, the non-invasive nature of EEG introduces challenges, including a lower signal-to-noise ratio (SNR), which complicates the accurate decoding of neural signals [7].

Previous research has predominantly focused on three distinct speech modes, i.e., imagined, intended, and spoken speech [8] [9] [10]. Imagined speech refers to the mental process of formulating and rehearsing words internally, without any accompanying physical movement or sound production. Intended speech involves articulatory movements of the mouth without vocalization, while spoken speech is characterized by the coordinated production of audible words and sounds through articulatory organs such as the tongue, lips, and vocal cords. Each mode provides unique insights into speech-related neural activities, offering opportunities to explore the underlying mechanisms of speech production.

Most existing research has focused on decoding EEG signals from a single speech mode, e.g., imagined speech [11], intended speech [12], and spoken speech [13]. While significant progress has been made in each of these modes, decoding imagined speech remains particularly challenging due to its reliance on internal cognitive processes, which produce weaker and noisier EEG signals compared to the other modes. Various network architectures have been explored to enhance the imagined speech decoding performance. Convolutional neural networks (CNNs) have been extensively employed to capture spatial information in EEG signals [14, 15], while recurrent neural networks (RNNs) specialize in processing temporal dependencies [16]. Other architectures, such as innovative techniques like standardization-refinement domain adaptation (SRDA) [17] and deep belief networks (DBNs) [18], have further improved the development in imagined speech decoding performance. Despite these advancements, previous work mainly focuses on single-mode decoding, and there is a lack of exploration of leveraging relationships between different modes. Recent efforts have been devoted to distinguishing different speech modes in EEG signals [19]. [20] learns a shared representation for three speech modes by reconstruction, however, a single-mode model is still applied to the learnt representation.

In this study, we propose a novel joint-modeling framework that learns a unified model for EEG signals from multiple speech modes, aiming to borrow knowledge from other modes to imagined speech, e.g., which channels are more related to speech decoding. To further encourage shared feature learning and decoding, we also introduce a dynamic masking mechanism that replaces EEG signals at certain time steps and channels with Gaussian noise. Based on the widely used EEGNet architecture, we build an effective decoding model, MSST-EEGNet, by integrating spatial attention (SA), temporal self-attention (TS), and multi-scale convolution (MSC) to capture critical spatial and temporal features at multiple time scales in the EEG signals. Experimental results show that the framework effectively exploits both mode-specific information and shared features, achieving significantly better performance in decoding four Mandarin vowels, especially for the imagined speech mode [21]. We also found that training a single-mode model with the EEG channels selected according to multi-mode model weights as inputs provides superior performance than training a single-mode model using all channels. This demonstrates that joint-modeling successfully learns the important features for speech decoding using knowledge from multiple speech modes.

2. Methodology

2.1. Multi-mode Joint-modeling Framework

The proposed joint-modeling framework, as shown in Figure 1, is designed to simultaneously model EEG signals from three distinct speech modes of imagined, intended, and spoken

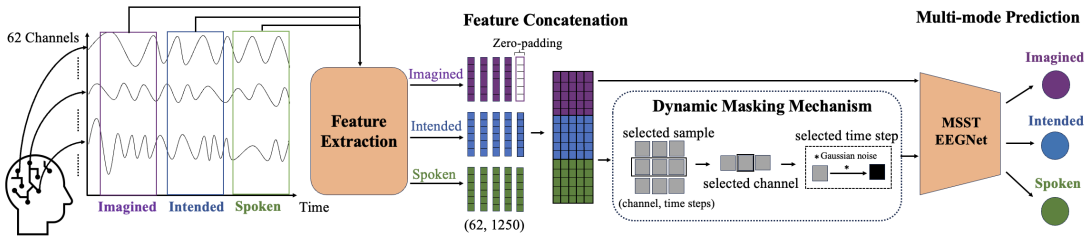

Figure 1: *Joint-modeling framework based on MSST-EEGNet with a dynamic masking mechanism.*

speech, with an aim to learn shared features, optimize the decoding model with more data, and avoid mode-specific noise contamination, e.g., the undesired signal noise introduced by actual articulatory movements in the spoken speech mode. The joint-modeling framework consists of the following modules:

- **EEG Signal Feature Extraction and Concatentation**: Channel-wise EEG features are extracted and preprocessed independently. The extracted channel-wise EEG features of three modes at each time step are concatenated to obtain multi-mode feature inputs. Vowels may be different for different modes. In our dataset, both of intended and spoken speech modes were recorded for 2.5 seconds at a sampling rate of 500Hz, resulting in 1250 time steps. In contrast, the imagined speech mode was recorded for 2 seconds. Hence, zero values are padded to the end of imagined speech features to obtain 1250 time steps for concatenation.
- **Speech Decoding with Enhanced EEGNets**: The multi-mode features are fed to an multi-scale spatial and temporal attention-entanced EEGNet (MSST-EEGNet) structure to obtain mode-wise decoded speech outputs, as described in Sec. 2.2.
- **Training with Dynamic Masking Mechanism**: To force the MSST-EEGNet to learn shared and mode-specific features, we propose a novel dynamic masking mechanism that randomly replaces input features at certain time steps and channels with Gaussian noise, as illustrated in Sec. 2.3.

2.2. MSST-EEGNet Architecture

The proposed speech decoding model is improved from the widely used EEGNet structure [22], with novel enhancements on multi-scale spatial and temporal information capturing. As shown in Figure 2, a spatial attention (SA) block and a temporal self-attention (TS) block are introduced to better capture spatial and temporal dependencies in EEG signals, following [12]. The SA outputs are fed to the TS block, and the TS outputs are fed to the enhanced EEGNet structure. Compared to the original EEGNet structure, three multi-scale convolutional layers are incorporated to capture features across multiple temporal scales, based on the consideration that EEG signals often contain patterns at varying temporal resolutions, corresponding to short-, medium- and long-term neural activities with different signal frequencies. Three parallel softmax layers are applied to the outputs of EEGNet to obtain decoded vowels for the three speech modes.

2.2.1. Spatial Attention Block

Speech production involves the activation of multiple cortical regions in the brain. The spatial patterns of neural activity at different EEG channels are critical for distinguishing different

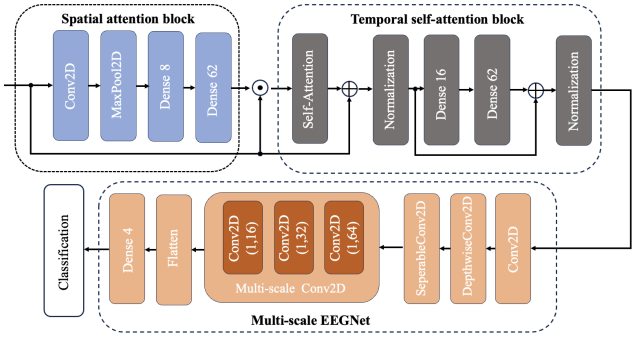

Figure 2: *Architecture of MSST-EEGNet.*

speech content. The SA block addresses this challenge by assigning adaptive weights to EEG channels, enabling the model to focus on the most informative channels for speech decoding. Inspired by the STAnet framework [23], the SA block consists of a convolutional layer, a max-pooling layer, and two fully connected layers with 8 and 62 neurons, respectively. To mitigate overfitting problem, a dropout rate of 0.5 is applied before each fully connected layer. The exponential linear unit (ELU) is used as the activation function throughout the block, ensuring stable and efficient training.

2.2.2. Temporal Self-Attention Block

Capturing temporal dynamics of EEG signals is important for speech decoding. Inspired by the transformer architecture [24] and the TA-EEGNet [12], the TS block incorporates a self-attention mechanism that enables the model to focus on key information from long-distance steps. The block consists of a self-attention layer, followed by dropout with a probability of 0.5, layer normalization, and two fully connected layers with 16 and 62 neurons, respectively. Another dropout layer with a probability of 0.5 and a final layer normalization step are applied. The fully connected layer employs the ReLU activation function. Residual connections are also integrated to mitigate gradient vanishing issues.

2.2.3. Multi-Scale Convolution Block

To capture diverse temporal scales, the model incorporates multi-scale convolutional layers [25] [26] with different kernel sizes of1×16, 1×32 and 1×64 for short-, medium-, and long-term patterns of neural activities, respectively. The outputs of these convolutional layers are concatenated, followed by batch normalization, ELU activation, average pooling, and dropout with a probability of 0.5.

2.3. Dynamic Masking Mechanism

To facilitate the learning of meaningful shared features across different modes, inspired by [27][28], a dynamic masking mechanism is employed during joint training process, by replacing a specified proportion of sample points at certain time steps and certain channels with Gaussian noise. The dynamic masking mechanism is expected to disrupt mode-specific information for trivial decoding without modeling the brain signals (e.g., decoding based on noise from device movements), and encourage the model to extract mode-invariant features. The masking mechanism involves the following steps:

- **Sample Selection**: A portion of EEG samples are randomly selected to be masked.
- **Channel Selection**: For each selected EEG sample, a subset of channels is randomly selected for masking.
- **Time Segment Masking**: For each selected channel, a specified number of time segments are randomly selected for masking. Each segment corresponds to a continuous portion of the time steps, simulating the partial loss of temporal information. To ensure a smooth transition and avoid abnormal changes at the segment boundaries, a smoothed masking function $\mathbf{m}(t)$ is applied to each selected segment:

$$\mathbf{m}(t) = \begin{cases} \frac{t - t_{\text{start}}}{l_{\text{smooth}}}, & t \in [t_{\text{start}}, t_{\text{start}} + l_{\text{smooth}}], \\ 0, & t \in [t_{\text{start}} + l_{\text{smooth}}, t_{\text{end}} - l_{\text{smooth}}], \\ \frac{t - (t_{\text{end}} - l_{\text{smooth}})}{l_{\text{smooth}}}, & t \in [t_{\text{end}} - l_{\text{smooth}}, t_{\text{end}}], \end{cases}$$

  Where t_{start} and t_{end} denote the starting and ending time points of the selected time segment, respectively, and l_{smooth} is the length of transition region.
- **Gaussian Noise Adding**: For each masked segment, a Gaussian noise $\mathbf{n}(t) \sim \mathcal{N}(\mu, \sigma \cdot \eta)$ is added to artificially pollute the input signals. The noise mean μ and standard deviation σ are estimated from the original signal in the masked region. η is a noise scaling factor. To preserve continuity and avoid abrupt changes, the generated noise is further smoothed using a Savitzky-Golay filter [29]. The final augmented signal $\widetilde{\mathbf{X}}(t)$ for the masked region is computed as:

$$\widetilde{\mathbf{X}}(t) = \mathbf{X}(t) \cdot \mathbf{m}(t) + \mathbf{n}(t) \cdot (1 - \mathbf{m}(t)).$$

  where the masking function $\mathbf{m}(t)$ determines the proportion of the original signal $\mathbf{X}(t)$ and the injected noise $\mathbf{n}(t)$ at each time step t.

3. Experiments

3.1. Dataset

The dataset used in this study comprises EEG signals collected from 10 healthy adults with a mean age of 22.6 years. All participants were native Mandarin speakers with normal hearing and speech abilities. During the data collection, as shown in Figure 3, each trial contains six stages:

- Rest: Once the trial started, 3 seconds of baseline data were recorded for reference and processing.
- Listen: auditory stimuli were presented through headphones. Each stimulus lasted approximately 700ms, followed by a silent interval. The stimuli were recorded by a native Mandarin-speaking woman at a sampling rate of 16kHz.
- Imagine: Participants were instructed to imagine themselves speaking the stimulus from the previous stage, including mimicking mouth movements and sound production mentally.
- Intend: Participants silently mouthed the stimulus without producing any vocal cord vibrations.
- Speak: Participants verbally articulated the stimulus. Their speech was recorded using a microphone at a sampling rate of 16 kHz.
- Report: After completing the experiment, participants were asked to report on their performance during each stage.

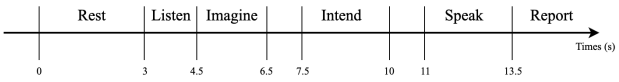

Figure 3: *EEG signal recording procedure for each trial.*

The EEG signals were recorded for four Mandarin vowels, i.e., /a/, /i/, /u/, /ü/, using a Neuroscan system at a sampling rate of 500Hz. Preprocessing was conducted using EEGLAB [30], including re-referencing, bandpass filtering (0.5-70Hz), and conducting independent component analysis (ICA) to remove artifacts. The dataset comprises a total number of 9,034 samples, including 3,043 samples for both imagined and intended speech mode, and 2,948 samples for spoken speech mode. The number of samples for vowels /a/, /i/, /u/, and /ü/ are 2,567, 2,603, 2,462, and 1,042, respectively, respectively. Each EEG sample has a size of (62, 1,250), where 62 represents the number of channels and 1,250 the time steps.

3.2. Experimental Setup

The proposed framework was evaluated using 10-fold cross validation. For each training-test split, 90% of the whole dataset was used as training data and 10% as testing set. 10% of the training set was further held as the validation set. The model was trained on an NVIDIA A100 GPU. During training, a batch size of 64 was used. The Adam optimizer was employed to minimize the cross-entropy loss, which measures the difference between predicted vowels and ground-truth vowels. For the individual mode training, the number of epoches was set to 300, while it was set to 500 in multi-mode training. The learning rate was automatically adjusted by learning rate scheduler `ReduceLROnPlateau`. We use masking ratios of 0.5, 0.25 and 0.15 for sampling of sample, channel, and time segment, respectively.

3.3. Results and Analysis

Table 1 shows the speech decoding accuracies of single- and multi-mode models, i.e., mean and standard deviation of accuracies of 10 testing folds. We can observe that compared to single-mode models, the multi-mode models, either joint modeling two modes or three modes, generally achieve significantly better performance on all three modes. It can also be found that simultaneously modeling all three modes provides the best performance. Our tri-mode model outperforms the baseline of reconstruction-based representation learning [20] with large margins. This demonstrates the effectiveness of the proposed joint-modeling framework.

Results of applying dynamic masking mechanism to joint-training are provided in Table 2. It can be observed that gener-

Table 1: *Single- and multi-mode speech decoding accuracies*

Training Mode			Testing Mode Accuracy (%)		
Img	**Int**	**Spk**	**Img**	**Int**	**Spk**
✓			28.07±0.98	-	-
	✓		-	47.14±1.76	-
		✓	-	-	56.83±1.64
✓	✓		29.53±1.15	47.35±1.43	-
✓		✓	32.36±2.95	-	58.79±1.47
	✓	✓	-	47.65±1.12	57.26±1.45
✓	✓	✓	**32.69**±1.08	**47.93**±1.35	**61.41**±2.27
WMCnet [20]			29.18±1.59	44.92±2.89	57.43±1.77

Note: Img, Int and Spk stand for imagined, intended, and spoken modes, respectively. "-" indicates the model was not tested on this mode.

Table 2: *Decoding accuracies of multi-mode models that are trained with different speech modes masked*

Masked Mode			Testing Mode Accuracy (%)		
Img	**Int**	**Spk**	**Img**	**Int**	**Spk**
			32.69±1.08	47.93±1.35	**61.41**±2.27
		✓	**34.95**±2.34	**49.16**±2.59	56.08±1.94
	✓		33.65±2.57	42.54±2.12	59.52±1.08
✓			29.82±2.01	44.11±0.85	59.85±1.62
	✓	✓	33.86±1.04	43.12±1.68	55.34±1.81
✓	✓	✓	32.78±1.09	42.54±2.92	55.11±2.28

Note: Checkmarks indicate which modes were masked. Empty row represents no masking applied.

ally applying masking mechanism to one mode hurts that mode but benefits the other modes, e.g., masking the spoken mode improves imagined speech decoding accuracy from 32.69% to 34.92%, and intended speech from 47.93% to 49.16%.

Tables 3 and 4 present the confusion matrices of the imagined and intended speech decoding models using joint-modeling with spoken mode masked. Table 5 shows the joint-modeling spoken speech decoding model without masking. For imagined and intended speech, decoding of the vowel /i/ consistently achieves higher accuracy compared to other vowels. All the models demonstrate limited performance in classifying the vowel /ü/, particularly in imagined speech, probably due the smaller size of /ü/ samples in the dataset.

Table 3: *Confusion matrix (%) of imagined speech decoding using joint-modeling with spoken mode masked*

	/a/	**/i/**	**/u/**	**/ü/**
/a/	**32.23**±2.65	31.47±2.93	27.36±1.74	8.97±2.35
/i/	22.14±1.69	**45.66**±3.86	24.43±1.89	7.57±3.52
/u/	26.43±4.17	31.26±1.93	**33.87**±2.13	9.42±2.79
/ü/	35.11±2.41	28.33±3.70	20.95±1.63	**16.59**±3.94

To further analyze the effect of joint-modeling, we select a subset of channels according to the channel importance for prediction by a multi-mode model or a single-mode model, and use the selected channels as inputs to train another single-mode model from scratch to compare with the original training of single-mode models with all channels as inputs. We aim to verify that the multi-mode joint-modeling enhances the selection of more informative channels for speech decoding. More

Table 4: *Confusion matrix (%) of intended speech decoding using joint-modeling with spoken mode masked*

	/a/	**/i/**	**/u/**	**/ü/**
/a/	**45.35**±2.47	35.47±2.85	17.28±1.73	2.93±3.82
/i/	12.53±1.64	**61.36**±1.82	22.83±1.88	3.43±2.49
/u/	12.47±3.27	27.43±2.95	**56.63**±2.52	3.59±1.87
/ü/	14.79±2.35	40.39±2.79	12.77±3.58	**31.94**±2.42

Table 5: *Confusion matrix (%) of spoken speech decoding using joint-modeling without masking*

	/a/	**/i/**	**/u/**	**/ü/**
/a/	**62.29**±2.97	22.39±3.33	13.58±4.78	2.61±2.57
/i/	16.31±2.36	**60.61**±3.97	18.17±2.60	5.19±3.46
/u/	15.32±3.48	19.69±2.99	**60.42**±2.65	4.77±2.74
/ü/	13.74±2.77	30.05±2.75	19.92±3.04	**36.37**±2.36

specifically, the importance is calculated by comparing the difference between original predicted accuracy and the perturbed accuracy, which is obtained with a tested channel set to zeros in the inputs. As showed in Figure 4, the newly trained single-mode models using selected subsets of channels according a to multi-mode model (highlighted as blue points) produce better decoding performance for imagined speech, compared to using all 62 channels, and using channel selection according to a single-mode model (yellow points). This supports our speculation that joint-modeling can borrow knowledge from other speech modes to enhance modeling of the target mode.

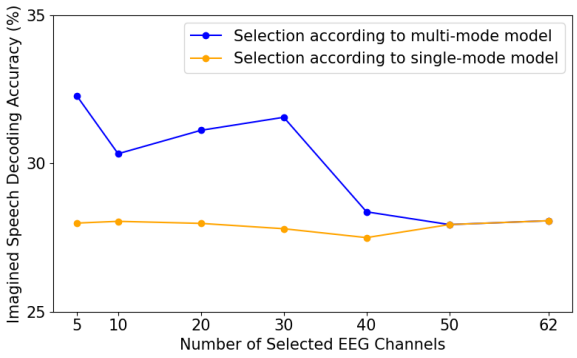

Figure 4: *Accuracy of imagined speech decoding models using a subset of EEG channels as inputs, which are selected according to channel importance for prediction by a multi-mode model with spoken mode masked or a single-mode model.*

4. Conclusions

In this study, we proposed a novel joint-modeling framework for building EEG-based speech decoding models by simultaneously modeling EEG signals of different speech modes, including imagined, intended and spoken speech, based on a dynamic masking mechanism and a multi-scale EEGNet structure with spatial and temporal attention. Compared to single-mode decoding, our multi-mode approach with dynamic masking mechanism improved the decoding accuracy across all speech modes significantly. These improvements demonstrate that leveraging the relationship between different speech modes could be a promising direction for advancing EEG-based speech decoding systems. Future work could further explore this direction by investigating consonant and tone decoding tasks.

5. Acknowledgements

This study was supported in part by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd., a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government.

6. References

[1] Q. Rabbani, G. Milsap, and N. E. Crone, "The potential for a speech Brain-Computer Interface using chronic electrocorticography," *Neurotherapeutics*, vol. 16, no. 1, pp. 144–165, 2019.

[2] C. Herff and T. Schultz, "Automatic speech recognition from neural signals: a focused review," *Frontiers in neuroscience*, vol. 10, p. 429, 2016.

[3] S. L. Metzger, K. T. Littlejohn, A. B. Silva, D. A. Moses, M. P. Seaton, R. Wang, M. E. Dougherty, J. R. Liu, P. Wu, M. A. Berger *et al.*, "A high-performance neuroprosthesis for speech decoding and avatar control," *Nature*, vol. 620, no. 7976, pp. 1037–1046, 2023.

[4] M. Angrick, S. Luo, Q. Rabbani, D. N. Candrea, S. Shah, G. W. Milsap, W. S. Anderson, C. R. Gordon, K. R. Rosenblatt, L. Clawson *et al.*, "Online speech synthesis using a chronically implanted Brain–Computer Interface in an individual with ALS," *Scientific reports*, vol. 14, no. 1, p. 9617, 2024.

[5] R. Abiri, S. Borhani, E. W. Sellers, Y. Jiang, and X. Zhao, "A comprehensive review of EEG-based Brain-Computer Interface paradigms," *Journal of neural engineering*, vol. 16, no. 1, p. 011001, 2019.

[6] B. J. Edelman, S. Zhang, G. Schalk, P. Brunner, G. Müller-Putz, C. Guan, and B. He, "Non-invasive Brain-Computer Interfaces: state of the art and trends," *IEEE Reviews in Biomedical Engineering*, 2024.

[7] D. M. Goldenholz, S. P. Ahlfors, M. S. Hämäläinen, D. Sharon, M. Ishitobi, L. M. Vaina, and S. M. Stufflebeam, "Mapping the signal-to-noise-ratios of cortical sources in magnetoencephalography and electroencephalography," *Human brain mapping*, vol. 30, no. 4, pp. 1077–1086, 2009.

[8] C. Cooney, R. Folli, and D. Coyle, "Neurolinguistics research advancing development of a direct-speech Brain-Computer Interface," *IScience*, vol. 8, pp. 103–125, 2018.

[9] Z. Zhang, X. Ding, Y. Bao, Y. Zhao, X. Liang, B. Qin, and T. Liu, "Chisco: An EEG-based BCI dataset for decoding of imagined speech," *Scientific Data*, vol. 11, no. 1, p. 1265, 2024.

[10] X. Chen, R. Wang, A. Khalilian-Gourtani, L. Yu, P. Dugan, D. Friedman, W. Doyle, O. Devinsky, Y. Wang, and A. Flinker, "A neural speech decoding framework leveraging deep learning and speech synthesis," *Nature Machine Intelligence*, pp. 1–14, 2024.

[11] D. Lopez-Bernal, D. Balderas, P. Ponce, and A. Molina, "A state-of-the-art review of EEG-based imagined speech decoding," *Frontiers in human neuroscience*, vol. 16, p. 867281, 2022.

[12] X. Wang, Y.-H. Lai, and F. Chen, "Intended Speech Classification with EEG Signals Based on a Temporal Attention Mechanism: A Study of Mandarin Vowels," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2024, pp. 1–4.

[13] X. Wang, M. Li, H. Li, S. H. Pun, and F. Chen, "Cross-Subject Classification of Spoken Mandarin Vowels and Tones with EEG Signals: A Study of End-to-End CNN with Fine-Tuning," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2023, pp. 535–539.

[14] C. Cooney, R. Folli, and D. Coyle, "Optimizing layers improves CNN generalization and transfer learning for imagined speech decoding from EEG," in *2019 IEEE international conference on systems, man and cybernetics (SMC)*. IEEE, 2019, pp. 1311–1316.

[15] C. Cooney, A. Korik, R. Folli, and D. Coyle, "Evaluation of hyperparameter optimization in machine and deep learning methods for decoding imagined speech EEG," *Sensors*, vol. 20, no. 16, p. 4629, 2020.

[16] S. Chengaiyan, A. S. Retnapandian, and K. Anandan, "Identification of vowels in consonant–vowel–consonant words from speech imagery based EEG signals," *Cognitive Neurodynamics*, vol. 14, no. 1, pp. 1–19, 2020.

[17] M. Jiménez-Guarneros and P. Gómez-Gil, "Standardization-refinement domain adaptation method for cross-subject EEG-based classification in imagined speech recognition," *Pattern Recognition Letters*, vol. 141, pp. 54–60, 2021.

[18] T.-J. Lee and K.-B. Sim, "Vowel classification of imagined speech in an Electroencephalogram using the deep belief network," *Journal of Institute of Control, Robotics and Systems*, vol. 21, no. 1, pp. 59–64, 2015.

[19] J.-S. Lee, H.-N. Jo, and S.-H. Lee, "Towards Unified Neural Decoding of Perceived, Spoken and Imagined Speech from EEG Signals," *arXiv preprint arXiv:2411.09243*, 2024.

[20] R. Sharon, M. Sur, and H. Murthy, "Harnessing the Multi-phasal Nature of Speech-EEG for Enhancing Imagined Speech Recognition," *IEEE Open Journal of Signal Processing*, 2025.

[21] T. L. Gottfried and T. L. Suiter, "Effect of linguistic experience on the identification of Mandarin Chinese vowels and tones," *Journal of Phonetics*, vol. 25, no. 2, pp. 207–231, 1997.

[22] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for EEG-based Brain–Computer Interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, p. 056013, Jul. 2018. [Online]. Available: http://dx.doi.org/10.1088/1741-2552/aace8c

[23] E. Su, S. Cai, L. Xie, H. Li, and T. Schultz, "STAnet: A spatiotemporal attention network for decoding auditory spatial attention from EEG," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 7, pp. 2233–2242, 2022.

[24] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[25] J. Wang, J. Zhuang, L. Duan, and W. Cheng, "A multi-scale convolution neural network for featureless fault diagnosis," in *2016 international symposium on flexible automation (isfa)*. IEEE, 2016, pp. 65–70.

[26] T. K. Dutta, D. R. Nayak, and Y.-D. Zhang, "Arm-net: Attention-guided residual multiscale cnn for multiclass brain tumor classification using mr images," *Biomedical Signal Processing and Control*, vol. 87, p. 105421, 2024.

[27] J. Crabbé and M. Van Der Schaar, "Explaining time series predictions with dynamic masks," in *International Conference on Machine Learning*. PMLR, 2021, pp. 2166–2177.

[28] 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," *arXiv preprint arXiv:1904.08779*, 2019.

[29] J. Chen, P. Jönsson, M. Tamura, Z. Gu, B. Matsushita, and L. Eklundh, "A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky–Golay filter," *Remote sensing of Environment*, vol. 91, no. 3-4, pp. 332–344, 2004.

[30] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of neuroscience methods*, vol. 134, no. 1, pp. 9–21, 2004.