arXiv:2608.22420v1 [cs.SD] 23 Aug 2026

Cross-Subject Generalization in Decoding Perceived Speech from Non-Invasive Brain Recordings

Aoke Zhang, Bo Wang, Xihong Wu, *Senior Member, IEEE*, Heping Cheng, and Jing Chen

***Abstract*—Decoding perceived speech from non-invasive brain recordings has garnered significant attention in recent years due to its wide range of potential applications. However, existing methods face considerable challenges in cross-subject decoding, primarily due to limited generalizability and the absence of explicit mechanisms for extracting subject-consistent information. These limitations result in high training costs and suboptimal decoding performance. To address these challenges, we propose an innovative Cross-Subject Perceived Speech Decoding (CPSD) framework, which comprises two training stages: source model pre-training and personal specialization. In the source model pre-training stage, contrastive learning is employed to capture shared representations across multiple source subjects. Subsequently, personal specialization initializes the model for the target subject by extracting consistent components from the source model and fine-tuning it using target subject data. Additionally, we introduce the Positional Encoding-based Spatial Attention (PESA) module, which remaps MEG/EEG data into a standardized reference space, thereby enhancing cross-subject consistency and facilitating model training. We evaluate the proposed CPSD framework on three perceived speech neural datasets encompassing different modalities and languages. The results demonstrate that our framework outperforms baseline methods by more than 6.8%, 15.4%, and 15.8% in Top-10 accuracy on the Armeni 2022, PKUEEG 2025, and Broderick 2018 datasets, respectively. Further analyses confirm the effectiveness, efficiency, and robustness of the proposed approach.**

***Index Terms*—brain-computer interface, EEG/MEG, perceived speech decoding, cross-subject decoding, subject consistency.**

I. Introduction

NEURAL speech decoding, which directly transforms neural signals into speech, has made significant progress in recent years [1]–[5]. Among the various paradigms of neural speech decoding, decoding perceived speech has been extensively studied due to its importance in understanding the mechanisms of speech processing [5]–[8], the large volume of available data, and the strong alignment between neural signals and speech stimuli. Furthermore, because of the overlap between the neural mechanisms underlying speech perception and speech production [9]–[12], related research holds potential to advance the development of speech neuroprostheses [2], [3], [13], which can restore communication for patients who have lost the ability to speak. Invasive brain-computer interfaces (BCIs) provide high-quality recordings of speech perception processes. However, they require surgical implantation, which limits accessibility for typical users [14]. Additionally, due to immune rejection reactions, maintaining signal quality over the long term is challenging [15]–[17]. Moreover, the coverage of invasive BCI-implanted electrode arrays is limited. For example, in speech perception research, electrode arrays can be implanted near the superior temporal gyrus and middle temporal gyrus [8]. In contrast, magnetoencephalography (MEG) and electroencephalography (EEG) are widely used non-invasive BCI techniques that offer greater safety, high temporal resolution, and whole-brain recording capabilities [18]. Consequently, a considerable amount of research has focused on this area [4], [19], [20]. The typical task of decoding perceived speech from non-invasive brain recordings involves retrieving the specific speech stimulus from test samples to which a subject is listening [4], [19], [21].

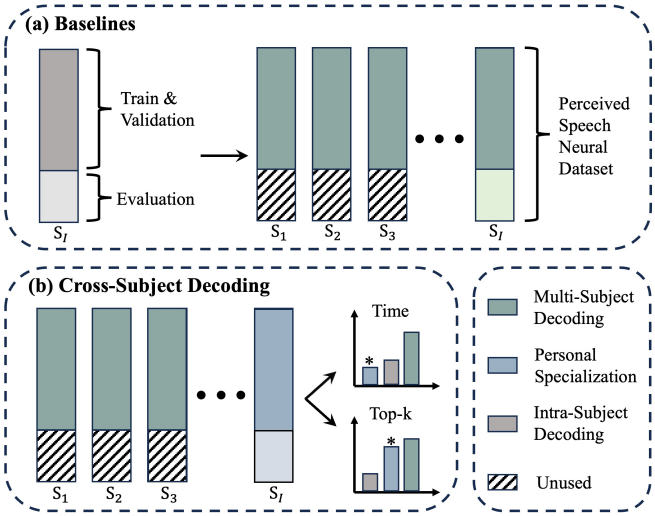

Fig. 1. Different settings for perceived speech decoding include: (a) intra-subject decoding, which involves decoding exclusively within the target subject; multi-subject decoding, where the model is trained on data from all subjects simultaneously and then evaluated on the target subject; and (b) cross-subject decoding, which entails pre-training the model on source subjects followed by generalization and evaluation on the target subject.

This work was supported by the STI 2030-Major Projects (Grant No.2021ZD0201500). (First author: Aoke Zhang.) (Corresponding author: Jing Chen.)

Aoke Zhang is with the Speech and Hearing Research Center, School of Intelligence Science and Technology and the Center for BioMed-X Research, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China.

Bo Wang and Xihong Wu are with the Speech and Hearing Research Center, School of Intelligence Science and Technology, Peking University, Beijing, China, and also with the National Key Laboratory of General Artificial Intelligence, Beijing, China.

Heping Cheng is with the Center for BioMed-X Research, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China.

Jing Chen is with the Speech and Hearing Research Center, School of Intelligence Science and Technology, Peking University, Beijing, China, and also with the Center for BioMed-X Research, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China, and also with the National Key Laboratory of General Artificial Intelligence, Beijing, China. (e-mail: janechenjing@pku.edu.cn)

Decoding speech from MEG/EEG signals remains challenging due to a low signal-to-noise ratio (SNR), limited spatial resolution [22]–[24], and other inherent limitations. These factors hinder the accurate capture of neural activity in speech-related brain regions, thereby reducing the quality of information available for effective decoding. Recent advances in deep learning have shown promise in modeling complex neural signals [21], [25], [26]. However, most existing models focus on intra-subject decoding and face several limitations, as illustrated on the left side of Fig. 1(a). Substantial cross-subject variability in neural responses during speech processing [27]–[29] makes it difficult to train a decoding model using neural signals across subjects. Although some models introduce subject layers [4], [30] to account for this variability, these approaches require simultaneous access to data from all subjects, as shown on the right side of Fig. 1(a). This necessitates retraining the model when new subjects are added, resulting in high computational costs. Moreover, the latent representations of subject layers exhibit low consistency across subjects, complicating training with a unified encoder and limiting the model's overall performance, as shown in Fig. 3(a).

When identical stimuli are presented to different subjects, the shared components of their responses are likely to reflect task-related information [31]–[33], which is the most informative for cross-subject decoding. Models that more effectively extract these subject-consistent representations are therefore expected to generalize better to unseen subjects. However, existing methods typically lack explicit constraints to enforce cross-subject consistency, limiting both decoding performance and generalizability.

Motivated by these observations, we propose an innovative cross-subject perceived speech decoding (CPSD) framework. The motivation behind this framework is further exploring subject-consistent information. We introduce a novel method to remap data from different subjects into a standardized reference space, thereby enhancing subject consistency. This approach not only improves the extraction of task-related information but also enhances generalizability. To further adapt the model to individual subjects, we incorporate a personal specialization stage, which initializes the subject layer using the consistent components learned during pre-training and subsequently aligns the target subject's neural data with the corresponding speech representations.

To address these issues, the key contributions of our work are as follows:

- A **C**ross-Subject **P**erceived **S**peech **D**ecoding (**CPSD**) framework is proposed. To the best of our knowledge, this is the first cross-subject decoding framework specifically designed for perceived speech decoding. The CPSD framework consists of two training stages: source model pre-training and personal specialization.
- In this framework, a **P**ositional **E**ncoding-based **S**patial **A**ttention (**PESA**) module and a subject layer initialization method are introduced to explicitly extract subject-consistent information, thereby enabling effective cross-subject decoding.
- The proposed framework was evaluated on three MEG/EEG datasets encompassing different modalities and languages. The results demonstrate consistent performance improvements over baseline methods across subjects. Ablation studies further confirm the effectiveness of the PESA module and the personal specialization stage.
- We conducted extensive analyses, including neural representation consistency evaluation, training time comparison, zero-shot decoding, hyperparameter search, comparison with multi-subject decoding, and PESA analysis. Collectively, these analyses demonstrate the effectiveness, efficiency, and robustness of the proposed framework.

II. Related Work

A. Perceived Speech Decoding

Existing studies include category classification from EEG [34], achieving a decoding accuracy of 61% on this binary classification task. However, this task is constrained by a strict experimental paradigm that differs significantly from the natural language used in daily life. In the context of continuous perceived speech decoding, fMRI has been employed to decode perceived speech for text generation [5], [35], [36]. Nevertheless, due to its low temporal resolution, fMRI struggles to fully capture dynamic information and tends to produce generated texts with a high word error rate. VLAAI [26] reconstructs the speech envelope from EEG signals; however, this method focuses solely on low-level speech features, limiting its decoding performance. Additionally, previous studies have not fully leveraged the relationships between data from different subjects within the dataset. Brainmagic [4] utilizes wav2vec representations as decoding targets and employs a multi-subject training strategy, resulting in significant advancements in perceived speech decoding. However, this method requires simultaneous use of data from all subjects, which hinders its ability to generalize efficiently to new subjects. Furthermore, although Brainmagic employs subject layers to account for subject variability, the latent representations from these layers exhibit low consistency, increasing the difficulty of training with a unified encoder and restricting the model's performance, as shown in Fig. 3(a).

B. Cross-Subject Decoding

Cross-subject decoding has been extensively studied and holds significant practical importance across various applications. One effective approach to addressing this challenge is acquiring transferable representations through self-supervised learning. TF-C [37] leverages the relationship between time and frequency to facilitate transfer between different datasets. Additionally, masked series modeling [38], [39] offers an alternative self-supervised learning approach by employing simple yet effective heuristic methods. Although it is widely acknowledged that self-supervised methods can learn generalizable representations to enhance performance on downstream tasks, these methods require the design of effective data augmentation techniques. However, this remains challenging for MEG/EEG speech decoding due to complex physical interpretations and low SNR. Domain adaptation methods, which

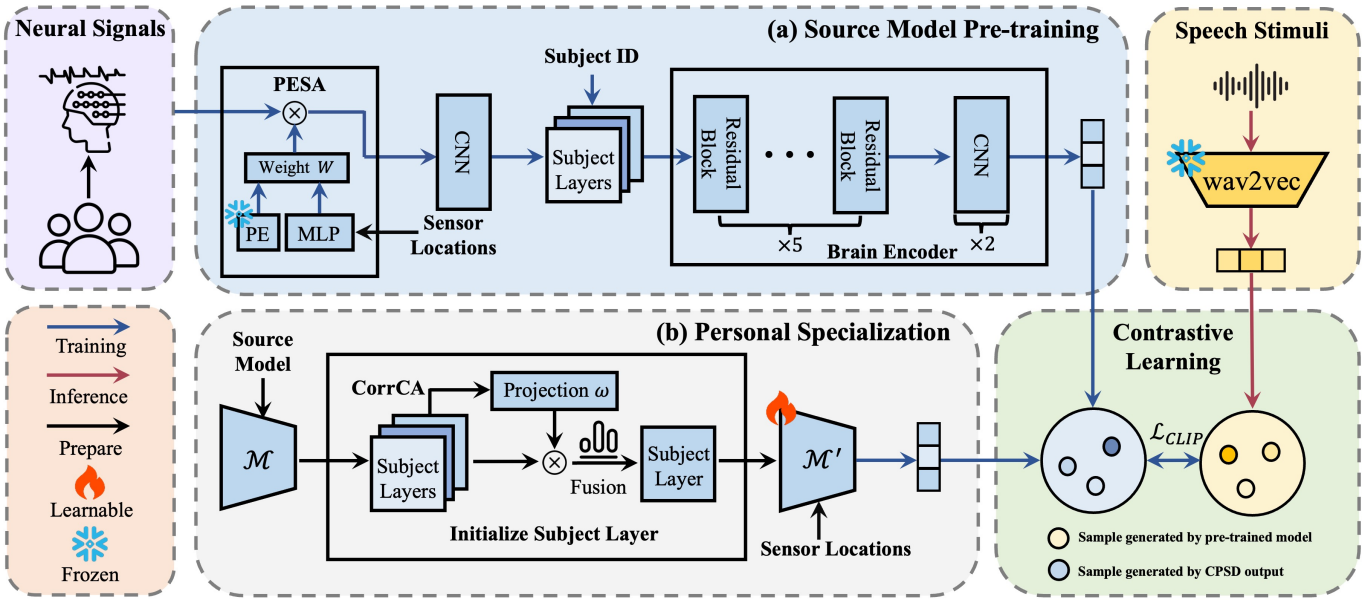

Fig. 2. Illustration of the CPSD framework. (a) During the source model pre-training stage, the model output is aligned with the wav2vec representation. (b) In the personal specialization stage, the subject layer for the target subject is first initialized, and then the representation from the MEG/EEG data of the target subject is realigned with the corresponding wav2vec representation.

learn through adversarial optimization between the encoder and a discriminator, aim to generate subject-invariant representations that can confuse the discriminator [40]–[42], thereby facilitating transfer across subjects. However, acquiring source data can sometimes be more challenging than obtaining pre-trained models, creating a need to fine-tune the model solely on the target data. CL-CS [43] and CL-SSTER [44] utilize contrastive learning to maximize the similarity among different subjects exposed to the same stimulus, thereby acquiring MEG/EEG representations with higher subject consistency. Although this method is applicable to perceived speech, it lacks a design tailored to the corresponding decoding tasks. While MEG/EEG cross-subject decoding has been extensively studied, most related work has focused on emotion recognition, image decoding, and sleep staging. To the best of our knowledge, no framework has been specifically designed for cross-subject perceived speech decoding, a critical area for advancing research in speech processing. Additionally, current mainstream speech decoding models [4], [26] often rely on specific architectures and therefore cannot directly adopt the aforementioned methods.

III. Methodology

A. Problem Formulation

Assume that $\{\mathcal{X}_{\mathcal{S}}^{i}\}_{i=1}^{N}$ and $\{\mathcal{X}_{\mathcal{S}'}^{i}\}_{i=1}^{N}$ represent the source and target data, respectively, where i indexes the MEG/EEG segments, and $\mathcal{S} \in \{1, 2, ..., I-1\}$, $\mathcal{S}' = I$ denote the source and target subject indices. We iteratively select each subject in the dataset as the target subject. Both sets of neural data were evoked by the same stimuli, characterized by features $\{\mathcal{Y}^{j}\}_{j=1}^{N}$. The goal is to efficiently fine-tune the model $\mathcal{M}$, which has been pre-trained on the source data, to achieve high-performance match-mismatch classification on the target data. The task process is illustrated in Fig. 1(b).

B. Model Overview

The model architecture of the CPSD framework is illustrated in Fig. 2. The input data are first mapped to a standardized reference space using the PESA module, followed by a convolutional layer with a kernel size of 1. To fully leverage the source data, subject layers [4] are employed, and the latent representations are processed by the brain encoder to generate the final outputs.

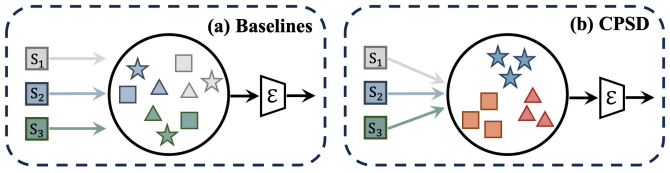

Fig. 3. Motivation for the PESA module. Different shapes within the large circle represent the MEG/EEG features of subjects under various stimuli.

1) PESA: To extract subject-consistent information more effectively and facilitate model training, we remap the MEG/EEG segments into a standardized reference space, as illustrated in Fig. 3(b). We select D points in a high-dimensional space to serve as the reference. The positions of these points must satisfy the following properties: (1) each position has a unique embedding vector; (2) the design enables the model to attend to relative offsets rather than absolute positions; and (3) the embedding vectors effectively represent the positions. Positional encoding [45] successfully meets these three properties and is commonly used to supplement positional information in transformer inputs. In our framework, we propose a novel application of positional encoding that utilizes embedding vectors to establish the reference space. The formula for positional encoding is shown in Eq. (1), where r denotes the vectors, and n and i represent the indices of the

vector and its dimension, respectively. We uniformly select D embedding vectors, each with dimension T, with an interval of δ between adjacent vectors. In our experiments, D is set to 270, consistent with [4], and δ is 15.

$$r_{ni} = \begin{cases} \sin(\frac{n}{L^{i/D}}), & \text{if i is even,} \\ \cos(\frac{n}{L^{(i-1)/D}}), & \text{if i is odd.} \end{cases} \tag{1}$$

Next, we need to obtain the weighting coefficients for the different channels. Sensor locations $X_{location} \in \mathbb{R}^{C\times 2}$ obtained from the MNE-Python function find_layout [46] are used to represent the positional information of specific MEG/EEG systems. We then define a multilayer perceptron (MLP) to map these locations into a latent space with dimensions $C \times T$, as shown in Eq. (2).

$$Y = \text{MLP}(X_{location}), Y \in \mathbb{R}^{C\times T} \tag{2}$$

The MLP consists of three blocks, each comprising a linear layer, layer normalization [47], and a GELU activation function [48]. The input and output dimensions of the linear layers are set to T, except for the first layer, which must match the dimension of the input. Then, the similarity between the MLP output and the positional encoding is calculated to measure the contribution of the input data to different positions in the reference space. Subsequently, the softmax function is applied to obtain the final input weights, as shown in Eq. (3), where $P = (r_{n\delta,i})_{n=1,\ldots,D}^{i=1,\ldots,T}$ denotes the stack of positional embedding vectors.

$$W = \text{Softmax}(YP^T), W \in \mathbb{R}^{C\times D} \tag{3}$$

Lastly, the latent representation H is computed from the input $X \in \mathbb{R}^{C\times T}$ using the formula in Eq. (4).

$$H = W^T X, H \in \mathbb{R}^{D\times T} \tag{4}$$

2) Brain Encoder: Due to the superior decoding performance of Brainmagic [4], we employ the same architecture for our encoder. The output of the subject layers is first processed through five residual blocks, each containing three convolutional layers. The first and second convolutional layers have input and output channels of 320, except for the first layer of the brain encoder. The dilation parameters are set to be $2^{2l\%5}$ and $2^{(2l+1)\%5}$, where l is the residual block index, to increase the receptive field of these two convolutional layers. Both convolutional layers are followed by a GELU activation function and batch normalization [49]. The last convolutional layer in each block has an output channel size of 640, and a GLU activation function [50] is used to reduce the dimension back to 320. The kernel size and stride parameters are set to 3 and 1, respectively, for all convolutions. Padding is applied to maintain the temporal dimension T. Finally, two convolutional layers are used, each followed by a GELU activation function. The first convolutional layer increases the input channel size from 320 to 640 with a kernel size of 1, and the subsequent 1×1 convolutional layer adjusts the dimension to match that of the wav2vec representations.

C. Training Pipeline

The CPSD framework involves a two-stage training process: first, pre-training a source model to acquire prior knowledge by fully leveraging data from various subjects; second, personal specialization on the target data to adapt the model to the specific subject, as shown in Algorithm 1:

Algorithm 1 Cross-Subject Perceived Speech Decoding.

INITIALIZATION:
1: Use a leave-one-subject-out approach, treating the remaining subjects in the dataset as source subjects.
2: Get sensor locations using mne.find_layout().
3: Initialize the parameters of CPSD model $\mathcal{M}$.

SOURCE MODEL PRE-TRAINING:
Input: Initialized model $\mathcal{M}$, MEG/EEG segments, sensor locations L, subject ID sid, wav2vec representations W and number of epochs num_1.
Output: Pre-trained source model $\mathcal{M}_{smp}$.
1: **for** $i = 1$ to num_1 **do**
2: $H \leftarrow \mathcal{M}$(MEG/EEG, L, sid)
3: $Loss \leftarrow \mathcal{L}_{CLIP}(H, W)$
4: Optimize the parameters of $\mathcal{M}$.
5: **end for**
6: **return** $\mathcal{M}_{smp}$

PERSONAL SPECIALIZATION:
Input: Source model $\mathcal{M}_{smp}$, MEG/EEG segments, sensor locations L, wav2vec representations W and number of epochs num_2.
Output: Personal specialized model $\mathcal{M}_{ps}$ on the target subject.
1: Initialize subject layer for target subject by Eq. (5) and (6), then get model $\mathcal{M}'_{smp}$.
2: **for** $i = 1$ to num_2 **do**
3: $\mathbf{H} \leftarrow \mathcal{M}'_{smp}$(MEG/EEG, L)
4: $Loss \leftarrow \mathcal{L}_{CLIP}(H, W)$
5: Optimize the parameters of $\mathcal{M}'_{smp}$.
6: **end for**
7: **return** $\mathcal{M}_{ps}$

EVALUATION:
Input: MEG/EEG segments, sensor locations L and wav2vec representations W.
Output: Top-k and rank accuracy.
1: $\mathbf{H} \leftarrow \mathcal{M}_{ps}$(MEG/EEG,$L$)
2: **return** Topk(H, W), Rank(H, W)

1) Source Model Pre-Training: The model $\mathcal{M}$ in the CPSD framework is initially pre-trained on the source data, which includes $I - 1$ subjects. Model $\mathcal{M}$ requires three distinct inputs: the source data, sensor locations, and subject IDs. The source data and sensor locations are first fed into the PESA module, followed by a 1×1 convolutional layer. Subsequently, subject layers process the latent representations along with the subject IDs. Data from different subjects at the same index are randomly selected to facilitate multi-subject training. Finally, a

TABLE I
A BRIEF DESCRIPTION OF THE NEURAL DATASETS RELATED TO SPEECH PERCEPTION. THE TEST SAMPLES INDICATE THE SIZE OF THE RETRIEVAL SET FOR EACH SAMPLE.

Dataset	Language	Modality	Sensors	Sampling rate	Subjects	Folds for CV	Duration	Test samples
Armeni 2022	English	MEG	275	1200	3	3	30.0 h	1024
PKUEEG 2025	Chinese	EEG	64	1000	25	25	68.1 h	128
Broderick 2018	English	EEG	128	512	19	19	19.2 h	128

brain encoder produces the final output, which is aligned with the wav2vec representation extracted from the corresponding stimuli.

2) Personal Specialization: To fine-tune the pre-trained model $\mathcal{M}$ for the target subject, we first initialize a subject layer with parameters tailored to that subject. We continue to leverage subject consistency to address this challenge; the consistent components of the subject layers are expected to capture the shared information across subjects in our decoding task. Therefore, we employ CorrCA [32], [51] to initialize the target subject layer based on the subject layers of the source subjects. The computational steps of the CorrCA algorithm are detailed in Eq. (5):

$$\hat{w} = \arg\max_{w} \frac{w^T X_1 X_2^T w}{\|X_1^T w\| \|X_2^T w\|} \tag{5}$$

where X_1, X_2, and w denote two input data matrices and the weight matrix, respectively, all belonging to $\mathbb{R}^{D\times D}$. The goal of CorrCA is to optimize the weight matrix to maximize the correlation between subjects. Next, a new model $\mathcal{M}^{'}$ is initialized by replacing the subject layers with the average of the consistent components across subjects, as shown in Eq. (6):

$$X_I = \frac{1}{I-1} \sum_{i=1}^{I-1} X_i^T \hat{w} \tag{6}$$

where I is the index of the target subject, and $\{1, ..., I-1\}$ are the indices of the source subjects. Finally, we realign the model outputs from the target subject with the corresponding wav2vec representations to specifically adapt the model for that subject.

D. Loss Function

In our experiments, we employed the CLIP loss [53] during both the source model pre-training stage and the personal specialization stage. This loss function is widely used in various applications [54]–[56] and has demonstrated superior performance in alignment tasks. It aims to align neural representations with corresponding speech features by maximizing the similarity between positive pairs while minimizing the similarity between negative pairs. The loss function is calculated as shown in Eq. (7):

$$\mathcal{L}_{CLIP} = -\frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{\exp\left(\text{sim}(z_i, y_i)\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}(z_i, y_j)\right)} \right) \tag{7}$$

where $\mathcal{L}_{CLIP}$ represents the loss function, N represents the batch size, z_i represents the model outputs, y_i represents the wav2vec representations extracted from the speech segments, and sim(x, y) computes the similarity between x and y.

IV. EXPERIMENTS

A. Datasets

Our framework was evaluated on three neural datasets related to speech perception: Armeni 2022 [57], PKUEEG 2025, and Broderick 2018 [58], as summarized in Table I. **Armeni 2022**: This public dataset includes data from three subjects who listened to English stories for a total of 30 hours. MEG data were recorded using a 275-channel axial gradiometer CTF system with a sampling rate of 1200 Hz. **PKUEEG 2025**: This dataset involves 25 subjects listening to the Chinese story *Romance of the Three Kingdoms*. Sixty-four-channel EEG data were recorded using the NeuroScan system at 1000 Hz. **Broderick 2018**: This public dataset contains EEG data from 19 English-speaking subjects listening to extracts from *The Old Man and the Sea*. A BioSemi ActiveTwo system with 128 channels was used, and the sampling rate was 512 Hz.

B. Data Preprocessing

For the MEG/EEG data, a notch filter was initially applied to eliminate line noise. The data were then resampled to 100 Hz to extract relevant information from the low-frequency band for our decoding task [59]. Independent Component Analysis (ICA) was performed on each trial to remove eye-blink artifacts [60]. The resampled data were segmented into 3-second intervals with a 1.5-second overlap between consecutive segments. Subsequently, the segments were normalized using RobustScaler, and values below -20 or above 20 were clamped to mitigate the influence of outliers [4]. For our match-mismatch classification task, the wav2vec representation was selected as the feature representing the speech stimulus [4], extracted from the final output of wav2vec2-large-xlsr-53 [61].

C. Baseline Methods

Baseline methods were selected from two main categories: intra-subject decoding and cross-subject decoding. Since cross-subject perceived speech decoding has not been addressed in previous work, models and frameworks from related fields were also considered to provide a comprehensive comparison with our framework. For intra-subject decoding,

TABLE II
PERCEIVED SPEECH DECODING RESULTS ACROSS VARIOUS DATASETS. WE AVERAGED THE DECODING RESULTS ACROSS DIFFERENT SUBJECTS AND EVALUATED THE ROBUSTNESS OF PERFORMANCE IMPROVEMENTS USING STATISTICAL TESTS.

	Armeni 2022		PKUEEG 2025		Broderick 2018	
Methods	Top-10(%)	Rankacc(%)	Top-10(%)	Rankacc(%)	Top-10(%)	Rankacc(%)
Random	1.3 ± 0.4	50.6 ± 0.5	7.3 ± 1.2	49.4 ± 2.2	7.6 ± 1.6	50.8 ± 0.3
ATM-S [30]	4.2 ± 1.4	68.8 ± 4.8	9.2 ± 3.2	52.7 ± 2.1	9.0 ± 2.8	51.9 ± 3.3
iTransformer [52]	15.0 ± 12.3	76.6 ± 16.4	11.5 ± 3.5	56.3 ± 3.1	9.3 ± 2.9	52.1 ± 3.3
VLAAI [26]	14.7 ± 13.2	75.0 ± 19.4	14.6 ± 12.4	55.9 ± 11.3	10.5 ± 4.9	55.5 ± 6.5
Brainmagic [4]	17.4 ± 13.8	83.1 ± 7.7	20.8 ± 8.1	65.7 ± 7.0	17.7 ± 7.1	63.0 ± 5.5
BIOT [38]	23.8 ± 16.5	86.8 ± 6.8	22.9 ± 9.7	67.0 ± 8.0	20.0 ± 8.6	63.9 ± 7.3
DAPE [40]	51.5 ± 13.6	95.4 ± 1.9	25.9 ± 12.7	69.1 ± 8.6	22.5 ± 8.5	66.5 ± 6.4
CL-CS [43]	54.5 ± 9.7	96.1 ± 1.5	27.6 ± 10.0	70.8 ± 7.5	24.1 ± 9.6	68.9 ± 6.5
CPSD (Ours)	**61.3** ± 12.4	**96.9** ± 1.5	**43.0** ± 13.6	**80.0** ± 6.7	**39.9** ± 15.3	**77.7** ± 8.0

ATM-S [30], iTransformer [52], VLAAI [26], and Brainmagic [4] were chosen, with the best-performing model selected as the encoder. To compare different methods in cross-subject decoding, BIOT [38], DAPE [40], and CL-CS [43] were selected as baseline methods. For all baselines, training stage hyperparameters were set according to the values recommended in their original papers to ensure optimal performance. Under cross-subject conditions, BIOT, DAPE, and CL-CS use the same training stages as CPSD.

D. Implementation Details

Data from both source and target subjects are partitioned using the same method. To avoid data leakage [62], [63], the dataset is divided into 70%, 10%, and 20% splits based on the order of trials. The stimuli for trials in the training, validation, and test sets do not overlap. The mini-batch size is set to 128 during both the source model pre-training stage and the personal specialization stage. The Adam optimizer [64] is used to minimize the loss, with a learning rate of 3×10^{-4} in both stages. To prevent overfitting on the training set, early stopping is applied if the model's performance does not improve for 10 consecutive epochs on the validation set. The model with the best validation performance is then selected for personal specialization or evaluation. The chance level for these experiments is provided in the main results Table II. We first randomly generate embeddings with the same shape as the wav2vec vectors and evaluate the results on the test set. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

E. Evaluation Metrics

In this study, various evaluation metrics were employed to comprehensively assess the model's performance. Top-10 accuracy was used as the evaluation metric for the match-mismatch classification tasks, measuring whether the target segment appeared among the model's ten most likely predictions, as shown in Eq. (8):

$$Top\text{-}10 = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(l_i \in \hat{l}_{i,1}, ..., \hat{l}_{i,10}) \tag{8}$$

where N represents the number of test samples, $\mathbb{I}$ denotes the indicator function, l_i is the true label of the i-th sample, and $\{\hat{l}_{i,1}, ..., \hat{l}_{i,10}\}$ are the labels sorted in descending order according to their predicted probabilities. Additionally, rank accuracy was used to evaluate the predictive performance on target samples in the test set. The formula for calculating rank accuracy is shown in Eq. (9):

$$Rankacc = 1 - \frac{< rank > -1}{< segments > -1} \tag{9}$$

where $< rank >$ represents the position of the corresponding target sample among the test samples, and $< segments >$ denotes the total number of test samples. These two evaluation metrics are commonly used in speech decoding tasks [4], [65], facilitating comparison with prior work.

F. Experimental Results

1) Main Results: The results of perceived speech decoding are presented in Table II. We first compare the performance of different encoders, including ATM-S, iTransformer,

TABLE III
ABLATION STUDY OF THE CPSD FRAMEWORK ACROSS DIFFERENT DATASETS.
(PS STANDS FOR PERSONAL SPECIALIZATION.)

	Armeni 2022		PKUEEG 2025		Broderick 2018	
	Top-10(%)	Rankacc(%)	Top-10(%)	Rankacc(%)	Top-10(%)	Rankacc(%)
Base	17.4 ± 13.8	83.1 ± 7.7	20.8 ± 8.1	65.7 ± 7.0	17.7 ± 7.1	63.0 ± 5.5
Base+PESA	25.7 ± 25.5	84.7 ± 9.8	27.7 ± 8.5	70.7 ± 5.7	18.1 ± 8.3	62.6 ± 6.9
Base+PS	50.3 ± 12.8	95.1 ± 2.2	29.2 ± 10.9	72.2 ± 6.3	35.3 ± 11.8	75.2 ± 7.0
Base+PESA+PS	**61.3** ± 12.4	**96.9** ± 1.5	**43.0** ± 13.6	**80.0** ± 6.7	**39.9** ± 15.3	**77.7** ± 8.0

VLAAI, and Brainmagic. Brainmagic demonstrates superior performance compared to the other model architectures when decoding data exclusively from the target subject. These results validate our choice of the Brain Encoder. To compare different cross-subject decoding methods, we consider BIOT, DAPE, CL-CS, and CPSD. Our framework surpasses these baseline methods across all three datasets, achieving Top-10 accuracies of 61.3%, 43.0%, and 39.9%, respectively. A pairwise t-test was also conducted, indicating that the improvements are statistically significant for each dataset ($p < 0.001$), demonstrating the reliability of our results.

2) Ablation Study: The results of the ablation study are presented in Table III. To demonstrate the effectiveness of the PESA module and the personal specialization stage, four settings are compared, defined as follows:

- Base: The model was trained and evaluated exclusively on the target data (intra-subject decoding) without utilizing the PESA module.
- Base+PESA: The CPSD framework was trained and evaluated exclusively on the target data.
- Base+PS: CPSD framework was implemented without the PESA module but included two training stages: source model pre-training and personal specialization.
- Base+PESA+PS: We utilized the complete version of the CPSD framework.

The results presented in the first two rows indicate that implementing the PESA module enhances the model's performance in intra-subject decoding for the target subject, thereby validating the initial assumption that PESA optimizes the training process. The two-stage training approach demonstrates significant improvements across three datasets, showing that our method effectively leverages prior knowledge from source subjects' data to facilitate generalization to the target subject. Ultimately, training with the full CPSD framework yields the highest decoding accuracies, confirming that the PESA module improves the model's generalizability. Additionally, pairwise t-tests reveal that all these improvements are statistically significant ($p < 0.001$).

3) Neural Representation Consistency: The subject-consistent information extraction capabilities of pre-trained source models were evaluated using the inter-subject correlation (ISC) metric, calculated via the CorrCA algorithm. The results are presented in Table IV. Two experimental settings were examined: the latent representation from the subject layer and the final output. Compared to Brainmagic, the CPSD framework demonstrated improvements in nearly all settings, with results that were statistically significant ($p < 0.001$).

Although Brainmagic exhibited higher consistency in the subject layer output of the PKUEEG 2025 dataset, the p-value for this result exceeded 0.05. Additionally, we provide the correlation between normalized Top-10 accuracy and normalized ISC on the PKUEEG 2025 dataset, as shown in Fig. 4.

The results indicate that the consistency of neural representations following pre-training is positively correlated with decoding performance, thereby supporting our hypothesis that incorporating subject-consistent information enhances decoding accuracy. This conclusion is further supported by the Broderick 2018 dataset, which demonstrates a correlation coefficient of 0.77.

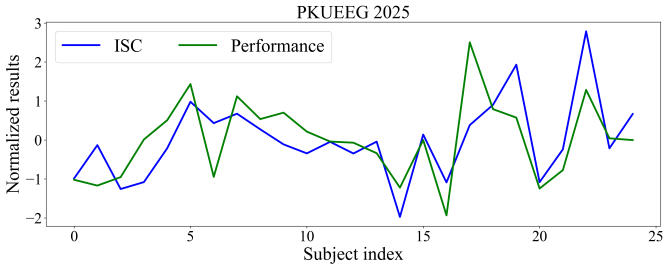

Fig. 4. The correlation between model performance and neural representation consistency ($corr = 0.66$, $p < 0.001$).

4) Training Time Comparison: To demonstrate the generalization efficiency of CPSD for the target subject, we analyze its training costs under different settings, including multi-subject training, intra-subject training, and the personal specialization stage. The results are presented in Fig. 5. We use the number of training steps as the evaluation metric due to interference from other concurrently running processes. Our framework exhibits lower training costs compared to the other settings.

TABLE IV
EVALUATION OF NEURAL REPRESENTATION CONSISTENCY.

Methods	Armeni 2022		PKUEEG 2025		Broderick 2018	
	Latent	Final	Latent	Final	Latent	Final
Brainmagic	0.011 ± 0.001	0.230 ± 0.012	**0.003** ± 0.001	0.093 ± 0.022	0.001 ± 0.001	0.157 ± 0.009
CPSD(Ours)	**0.016** ± 0.002	**0.274** ± 0.012	0.002 ± 0.001	**0.129** ± 0.010	**0.002** ± 0.001	**0.227** ± 0.010

This reduction in computational demand is more pronounced in datasets with a larger number of subjects. Specifically, the difference in training steps for the Armeni 2022 dataset is relatively minor, likely due to the small number of subjects in this dataset. In contrast, the specialization stage requires only 7.5% and 15.6% of the training steps needed for multi-subject training on the PKUEEG 2025 and Broderick 2018 datasets, respectively.

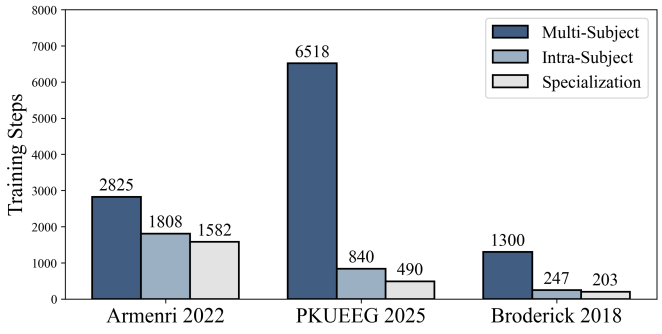

Fig. 5. Training times under different settings.

5) Zero-Shot Decoding: To further evaluate the generalizability of our CPSD framework, we present perceived speech decoding results under a zero-shot setting. In this scenario, we use only the source model pre-training stage and directly evaluate the framework's performance on the test set without any additional specialization. This approach is particularly useful when training data from the target subject are limited [41] and presents a challenging task due to substantial cross-subject variability, which has been underexplored in perceived speech decoding. As shown in Fig. 6, even without personal specialization, our model achieves decoding results significantly above chance level in 44.7% (21/47) of subjects. The chance level of decoding is presented in Table II. The proportion of subjects whose decoding accuracy exceeds the chance level is 33.3% (1/3), 48.0% (12/25), and 42.1% (8/19), respectively. These results highlight the importance of the number of source subjects and the significance of the personal specialization stage for the model. This demonstrates that our framework effectively captures consistent representations from source subjects that generalize to the target subject, underscoring the robustness of our methods.

6) Hyperparameter Search: We also conducted hyperparameter searches to evaluate the reliability of our results. Our framework introduces two hyperparameters: α, which denotes

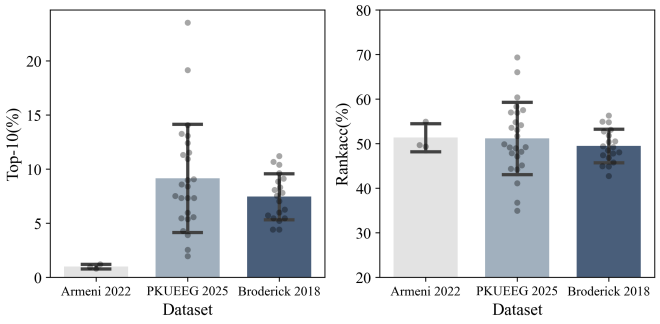

Fig. 6. Zero-shot decoding performance across various perceived speech datasets.

the number of MLP layers in PESA, and δ, which represents the interval between adjacent position embeddings. Due to the high computational cost, we evaluated our framework only on the Broderick 2018 dataset. Fig. 7 shows that the CPSD framework is not sensitive to hyperparameter selection, and we chose the hyperparameters that yielded the highest Top-10 accuracy. The experimental results in Fig. 7 demonstrate that our framework maintains high performance under varying conditions, thereby confirming the robustness of our findings.

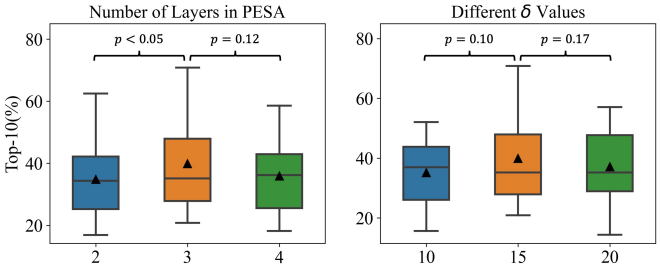

Fig. 7. Decoding results using various hyperparameters on the Broderick 2018 dataset. The triangles in the box plot indicate the mean values.

7) Comparison with Multi-Subject Settings: To highlight the advancements of the PESA module, we evaluated the CPSD framework in a multi-subject setting, which has demonstrated effectiveness in decoding perceived speech [4]. We selected Brainmagic and CPSD as our comparison targets and trained models on all subjects in the dataset simultaneously. By fully leveraging contrastive learning, both models showed improved decoding performance in multi-subject settings, as shown in Table V. However, the performance degradation of

TABLE V
COMPARISON WITH MULTI-SUBJECT SETTINGS.
(MD STANDS FOR MULTI-SUBJECT DECODING.)

	Armeni 2022		PKUEEG 2025		Broderick 2018	
Methods	Top-10(%)	Rankacc(%)	Top-10(%)	Rankacc(%)	Top-10(%)	Rankacc(%)
Brainmagic	17.4 ± 13.8	83.1 ± 7.7	20.8 ± 8.1	65.7 ± 7.0	17.7 ± 7.1	63.0 ± 5.5
Brainmagic-MD	52.7 ± 15.5	95.6 ± 2.1	39.1 ± 10.3	78.8 ± 5.3	39.2 ± 12.4	77.4 ± 7.1
CPSD	61.3 ± 12.4	96.9 ± 1.5	43.0 ± 13.6	80.0 ± 6.7	39.9 ± 15.3	77.7 ± 8.0
CPSD-MD	**63.0** ± 11.8	**97.1** ± 1.3	**45.5** ± 11.5	**81.5** ± 5.5	**46.0** ± 14.5	**81.0** ± 6.6

the CPSD framework was significantly smaller than that of Brainmagic. Compared to the decoding results obtained from multi-subject training, our cross-subject results showed only a performance decrease of 1.7%, 2.5%, and 6.1%, respectively, highlighting the advancement of our source model pre-training stage. When both CPSD and Brainmagic were applied to multi-subject decoding, the Top-10 accuracy improved by 10.3%, 6.4%, and 6.8% across the three datasets, further demonstrating the advantages of our model architecture design.

8) PESA Analysis: To verify that the model effectively utilizes the spatial distribution information of the sensors, we randomly shuffled the sensor locations, then retrained and evaluated our framework using the same data. Due to computational constraints, we conducted experiments only on the Armeni 2022 and Broderick 2018 datasets. The results shown in Fig. 8 indicate that after shuffling, the model's Top-10 accuracy decreased by more than 11% on both datasets. These results are statistically significant, demonstrating the model's effective use of sensor location information.

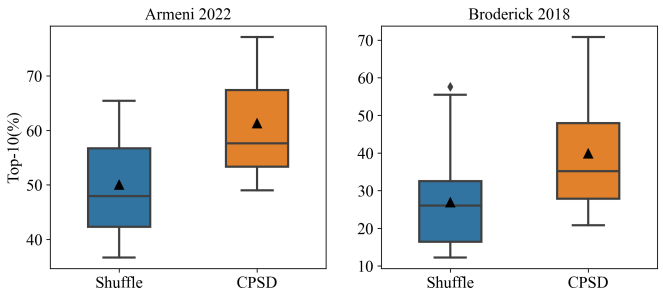

Fig. 8. Verification of the validity of sensor locations as input. The performance improvement of CPSD compared to the Shuffle setting in both figures is statistically significant ($p < 0.01$).

We further interpret the output from the PESA module by visualizing the learned channel weights on MEG/EEG scalp topographies. Let $\{W_i\}_{i \in S}$ represent the PESA weights from models of different subjects. We first average these weights across subjects:

$$W = \frac{1}{|S|} \sum_{i=1}^{|S|} W_i, W \in \mathbb{R}^{C \times D} \tag{10}$$

where $|S|$, C, and D denote the number of subjects, input channels, and hidden dimensions, respectively. To highlight the contribution of different channels to the experimental results, we obtained the topography T from Eq. (11):

$$T = \sqrt{W \odot WE}, T \in \mathbb{R}^{C} \tag{11}$$

where $\odot$ denotes the Hadamard product and $E = \frac{1}{D}(1, 1, \ldots, 1)^T \in \mathbb{R}^D$. The results for the three datasets are shown in Fig. 9. For the perceived speech decoding task, the model assigns higher weights to sensors over the bilateral temporal regions, consistent with their role in auditory processing. These findings demonstrate that the PESA module effectively captures brain regions relevant to specific decoding tasks.

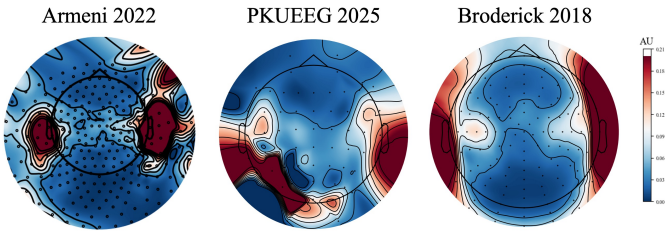

Fig. 9. Visualization of channel weights from the PESA module across various datasets.

V. CONCLUSION

In this work, we propose a cross-subject perceived speech decoding method CPSD that employs a two-stage training approach: source model pre-training and personal specialization. In the source model pre-training stage, neural representations with high subject consistency are extracted through contrastive learning. During the personal specialization stage, consistent components are derived from the subject layer of the source model and adapted to individual differences using target subject data. To enhance the consistency of neural

representations, this study introduces a positional encoding-based spatial attention module PESA. This module remaps MEG/EEG data to a standardized reference space to optimize the model's generalizability and decoding performance. The performance of CPSD was evaluated on three perceived speech neural datasets encompassing different modalities and languages. Experimental results demonstrate that this method outperforms existing baseline models with statistically significant improvements. The effectiveness, robustness, and efficiency of the proposed method were validated through ablation study, hyperparameter search, comparisons with multi-subject setting, zero-shot decoding, neural representation consistency evaluation, sensor location validity analysis, and training time comparison. Finally, visualization of the PESA module's channel weight matrix revealed higher weights near the bilateral temporal lobes, further confirming that the proposed method effectively focuses on critical brain regions associated with speech perception.

VI. Future Work

In future research, we will aim to improve the model's zero-shot decoding performance and explore additional applications. Potential solutions for extending the model to a broader range of applications include the following two points:

Studies have confirmed an overlap between the neural mechanisms underlying speech perception and speech production. For example, Hickok et al. demonstrated that Broca's area is involved both in motor planning for speech production and in recognizing and parsing speech signals during perception [9]. Additionally, there is a correlation between different speech production tasks: models trained on overt speech outperform those trained directly on imagined speech, indicating that the neural representations elicited by these two tasks overlap [66], [67]. Therefore, developing a transfer learning framework that leverages different experimental settings may represent a promising technical approach for achieving non-invasive speech neuroprostheses.

Lastly, although brain foundation models have been extensively studied [68]–[71], their application to perceived speech decoding still faces significant limitations. These models rely heavily on tokenization techniques to achieve cross-device training [38]; however, the effectiveness of tokenizing perceived speech neural data remains insufficiently validated. Consequently, developing effective methods for cross-device decoding has become a critical challenge in implementing relevant foundation models. The PESA module proposed in this article decouples the input data dimensions from the model parameters, enabling control over the transformation of input data through channel position information. This approach allows a unified-dimensional neural representation to be fed into the encoder for further processing. Therefore, it is feasible to construct a brain foundation model based on the PESA module and evaluate the effectiveness of this method in speech decoding tasks.

References

[1] J. Lévy, M. Zhang, S. Pinet, J. Rapin, H. Banville, S. d'Ascoli, and J.-R. King, "Noninvasive decoding of typed sentences from human brain activity," *Nature Neuroscience*, pp. 1–7, 2026.

[2] M. Wairagkar, N. S. Card, T. Singer-Clark, X. Hou, C. Iacobacci, L. M. Miller, L. R. Hochberg, D. M. Brandman, and S. D. Stavisky, "An instantaneous voice-synthesis neuroprosthesis," *Nature*, pp. 1–8, 2025.

[3] N. S. Card, M. Wairagkar, C. Iacobacci, X. Hou, T. Singer-Clark, F. R. Willett, E. M. Kunz, C. Fan, M. Vahdati Nia, D. R. Deo *et al.*, "An accurate and rapidly calibrating speech neuroprosthesis," *New England Journal of Medicine*, vol. 391, no. 7, pp. 609–618, 2024.

[4] 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.

[5] J. Tang, A. LeBel, S. Jain, and A. G. Huth, "Semantic reconstruction of continuous language from non-invasive brain recordings," *Nature Neuroscience*, vol. 26, no. 5, pp. 858–866, 2023.

[6] A. Bhattacharjee, Z. Zada, H. Wang, B. Aubrey, W. Doyle, P. Dugan, D. Friedman, O. Devinsky, A. Flinker, P. J. Ramadge *et al.*, "Aligning brains into a shared space improves their alignment with large language models," *Nature Computational Science*, vol. 6, no. 2, pp. 169–178, 2026.

[7] J. Zou, D. Poeppel, and N. Ding, "Constituent-constrained word prediction during language comprehension," *Nature Neuroscience*, vol. 29, pp. 1498–1509, 2026.

[8] H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani, "Towards reconstructing intelligible speech from the human auditory cortex," *Scientific reports*, vol. 9, no. 1, p. 874, 2019.

[9] G. Hickok and D. Poeppel, "The cortical organization of speech processing," *Nature reviews neuroscience*, vol. 8, no. 5, pp. 393–402, 2007.

[10] J. I. Skipper, H. C. Nusbaum, and S. L. Small, "Listening to talking faces: motor cortical activation during speech perception," *Neuroimage*, vol. 25, no. 1, pp. 76–89, 2005.

[11] D. J. Kraemer, C. N. Macrae, A. E. Green, and W. M. Kelley, "Sound of silence activates auditory cortex," *Nature*, vol. 434, no. 7030, pp. 158–158, 2005.

[12] M. E. Wheeler, S. E. Petersen, and R. L. Buckner, "Memory's echo: vivid remembering reactivates sensory-specific cortex," *Proceedings of the National Academy of Sciences*, vol. 97, no. 20, pp. 11 125–11 129, 2000.

[13] A. B. Silva, K. T. Littlejohn, J. R. Liu, D. A. Moses, and E. F. Chang, "The speech neuroprosthesis," *Nature Reviews Neuroscience*, vol. 25, no. 7, pp. 473–492, 2024.

[14] E. C. Leuthardt, D. W. Moran, and T. R. Mullen, "Defining surgical terminology and risk for brain computer interface technologies," *Frontiers in Neuroscience*, vol. 15, p. 599549, 2021.

[15] L. Chen, H. Zhong, L. Wang, L. Xu, W. Fan, Y. Zhao, H. Zhang, Y. Shen, K. Wu, X. Fu *et al.*, "Long-term stable subdural recordings enabled by fibrosis-resistant hydrogel-integrated μecog arrays," *Advanced Science*, vol. 12, no. 47, p. e15453, 2025.

[16] D. Yang, G. Tian, J. Chen, Y. Liu, E. Fatima, J. Qiu, N. A. N. N. Malek, and D. Qi, "Neural electrodes for brain-computer interface system: From rigid to soft," *BMEMat*, vol. 3, no. 3, p. e12130, 2025.

[17] V. S. Polikov, P. A. Tresco, and W. M. Reichert, "Response of brain tissue to chronically implanted neural electrodes," *Journal of Neuroscience Methods*, vol. 148, no. 1, pp. 1–18, 2005.

[18] J. Peksa and D. Mamchur, "State-of-the-art on brain-computer interface technology," *Sensors*, vol. 23, no. 13, p. 6001, 2023.

[19] A. Zhang, B. Wang, X. Wu, and J. Chen, "A novel multimodal method for decoding speech perception from brain activities," *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025.

[20] B. Wang, X. Xu, L. Zhang, B. Xiao, X. Wu, and J. Chen, "Semantic reconstruction of continuous language from meg signals," *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2190–2194, 2024.

[21] B. Wang, X. Xu, B. Xiao, L. Zheng, X. Wu, H. Cheng, and J. Chen, "Hierarchical decoding of perceived speech from non-invasive brain recordings," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 34, pp. 2349–2360, 2026.

[22] S. Sanei and J. A. Chambers, "Eeg signal processing," *John Wiley & Sons*, 2013.

[23] F.-H. Lin, J. W. Belliveau, A. M. Dale, and M. S. Hämäläinen, "Distributed current estimates using cortical orientation constraints," *Human brain mapping*, vol. 27, no. 1, pp. 1–13, 2006.

[24] J. Sarvas, "Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem," *Physics in Medicine & Biology*, vol. 32, no. 1, pp. 11–22, 1987.

[25] 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," *2024 IEEE International Conference on*

Acoustics, Speech, and Signal Processing Workshops (ICASSPW), pp. 113–114, 2024.

[26] 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, 2023.

[27] E. Myers, M. Phillips, and E. Skoe, "Individual differences in the perception of phonetic category structure predict speech-in-noise performance," *The Journal of the Acoustical Society of America*, vol. 156, no. 3, pp. 1707–1719, 2024.

[28] N. Giovannone and R. M. Theodore, "Individual differences in lexical contributions to speech perception," *Journal of Speech, Language, and Hearing Research*, vol. 64, no. 3, pp. 707–724, 2021.

[29] A. G. Huth, W. A. De Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant, "Natural speech reveals the semantic maps that tile human cerebral cortex," *Nature*, vol. 532, no. 7600, pp. 453–458, 2016.

[30] D. Li, C. Wei, S. Li, J. Zou, and Q. Liu, "Visual decoding and reconstruction via eeg embeddings with guided diffusion," *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pp. 102 822–102 864, 2024.

[31] U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, and R. Malach, "Intersubject synchronization of cortical activity during natural vision," *science*, vol. 303, no. 5664, pp. 1634–1640, 2004.

[32] J. P. Dmochowski, P. Sajda, J. Dias, and L. C. Parra, "Correlated components of ongoing eeg point to emotionally laden attention–a possible marker of engagement?" *Frontiers in human neuroscience*, vol. 6, p. 112, 2012.

[33] S. A. Nastase, V. Gazzola, U. Hasson, and C. Keysers, "Measuring shared responses across subjects using intersubject correlation," pp. 667–685, 2019.

[34] I. Simanova, M. Van Gerven, R. Oostenveld, and P. Hagoort, "Identifying object categories from event-related eeg: toward decoding of conceptual representations," *PloS one*, vol. 5, no. 12, p. e14465, 2010.

[35] W. Lu, D. Nie, P. Xue, Z. Cui, P. Li, D. Zhang, and X. Wen, "Brain-inspired fmri-to-text decoding via incremental and wrap-up language modeling," *Advances in Neural Information Processing Systems*, vol. 38, pp. 149 986–150 009, 2026.

[36] X. Chen, C. Du, C. Liu, Y. Wang, and H. He, "Open-vocabulary auditory neural decoding using fmri-prompted llm," *arXiv preprint arXiv:2405.07840*, 2024.

[37] X. Zhang, Z. Zhao, T. Tsiligkaridis, and M. Zitnik, "Self-supervised contrastive pre-training for time series via time-frequency consistency," *Advances in neural information processing systems*, vol. 35, pp. 3988–4003, 2022.

[38] C. Yang, M. Westover, and J. Sun, "Biot: Biosignal transformer for cross-data learning in the wild," *Advances in Neural Information Processing Systems*, vol. 36, pp. 78 240–78 260, 2023.

[39] J. Dong, H. Wu, H. Zhang, L. Zhang, J. Wang, and M. Long, "Simmtm: A simple pre-training framework for masked time-series modeling," *Advances in Neural Information Processing Systems*, vol. 36, pp. 29 996–30 025, 2023.

[40] D. Bethge, P. Hallgarten, T. Grosse-Puppendahl, M. Kari, R. Mikut, A. Schmidt, and O. Özdenizci, "Domain-invariant representation learning from eeg with private encoders," *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1236–1240, 2022.

[41] Y. Wang, B. Zhang, and Y. Tang, "Dmmr: Cross-subject domain generalization for eeg-based emotion recognition via denoising mixed mutual reconstruction," *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 1, pp. 628–636, 2024.

[42] B.-Q. Ma, H. Li, W.-L. Zheng, and B.-L. Lu, "Reducing the subject variability of eeg signals with adversarial domain generalization," *International Conference on Neural Information Processing*, pp. 30–42, 2019.

[43] M. Hu, D. Xu, K. He, K. Zhao, and H. Zhang, "Cross-subject emotion recognition with contrastive learning based on eeg signal correlations," *Biomedical Signal Processing and Control*, vol. 104, p. 107511, 2025.

[44] X. Shen, L. Tao, X. Chen, S. Song, Q. Liu, and D. Zhang, "Contrastive learning of shared spatiotemporal eeg representations across individuals for naturalistic neuroscience," *NeuroImage*, vol. 301, p. 120890, 2024.

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

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

[47] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[48] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[49] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *International conference on machine learning*, pp. 448–456, 2015.

[50] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," *International conference on machine learning*, pp. 933–941, 2017.

[51] L. C. Parra, S. Haufe, and J. P. Dmochowski, "Correlated components analysis-extracting reliable dimensions in multivariate data," *arXiv preprint arXiv:1801.08881*, 2018.

[52] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, "itransformer: Inverted transformers are effective for time series forecasting," *The Twelfth International Conference on Learning Representations*, 2024.

[53] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," *International conference on machine learning*, pp. 8748–8763, 2021.

[54] Y. Benchetrit, H. Banville, and J.-R. King, "Brain decoding: toward real-time reconstruction of visual perception," *International Conference on Learning Representations*, vol. 2024, pp. 7846–7858, 2024.

[55] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *International conference on machine learning*, pp. 19 730–19 742, 2023.

[56] N. Mu, A. Kirillov, D. Wagner, and S. Xie, "Slip: Self-supervision meets language-image pre-training," *European conference on computer vision*, pp. 529–544, 2022.

[57] K. Armeni, U. Güçlü, M. van Gerven, and J.-M. Schoffelen, "A 10-hour within-participant magnetoencephalography narrative dataset to test models of language comprehension," *Scientific Data*, vol. 9, no. 1, p. 278, 2022.

[58] M. P. Broderick, A. J. Anderson, G. M. Di Liberto, M. J. Crosse, and E. C. Lalor, "Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech," *Current Biology*, vol. 28, no. 5, pp. 803–809, 2018.

[59] M. Hämäläinen, R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa, "Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain," *Reviews of modern Physics*, vol. 65, no. 2, p. 413, 1993.

[60] T.-P. Jung, S. Makeig, A. J. Bell, and T. J. Sejnowski, "Independent component analysis of electroencephalographic and event-related potential data," *Central auditory processing and neural modeling*, pp. 189–197, 1998.

[61] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[62] X. Xu, B. Wang, B. Xiao, Y. Niu, Y. Wang, X. Wu, H. Cheng, and J. Chen, "The impacts of temporal autocorrelations on eeg decoding," *Biomedical Signal Processing and Control*, vol. 113, p. 108783, 2026.

[63] R. Li, J. S. Johansen, H. Ahmed, T. V. Ilyevsky, R. B. Wilbur, H. M. Bharadwaj, and J. M. Siskind, "The perils and pitfalls of block design for eeg classification experiments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 316–333, 2020.

[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[65] F. Pereira, B. Lou, B. Pritchett, S. Ritter, S. J. Gershman, N. Kanwisher, M. Botvinick, and E. Fedorenko, "Toward a universal decoder of linguistic meaning from brain activation," *Nature communications*, vol. 9, no. 1, p. 963, 2018.

[66] S. Komeiji, T. Mitsuhashi, Y. Iimura, H. Suzuki, H. Sugano, K. Shinoda, and T. Tanaka, "Feasibility of decoding covert speech in ecog with a transformer trained on overt speech," *Scientific Reports*, vol. 14, no. 1, p. 11491, 2024.

[67] 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.

[68] Q. Xiao, Z. Cui, C. Zhang, S. Chen, W. Wu, A. Thwaites, A. Woolgar, B. Zhou, and C. Zhang, "Brainomni: A brain foundation model for unified eeg and meg signals," *Advances in Neural Information Processing Systems*, vol. 38, pp. 41 179–41 212, 2026.

[69] D. Liu, Y. Chen, Z. Chen, Z. Cui, Y. Wen, J. An, J. Luo, and D. Wu, "Eeg foundation models: Progresses, benchmarking, and open problems," *arXiv preprint arXiv:2601.17883*, 2026.

[70] J. Wang, S. Zhao, Z. Luo, Y. Zhou, H. Jiang, S. Li, T. Li, and G. Pan, "Cbramod: A criss-cross brain foundation model for eeg decoding," *International conference on learning representations*, vol. 2025, pp. 75 310–75 346, 2025.

[71] W.-B. Jiang, L. Zhao, and B.-L. Lu, "Large brain model for learning generic representations with tremendous eeg data in bci," *International Conference on Learning Representations*, vol. 2024, pp. 16 405–16 426, 2024.