

# Decoding Covert Speech from EEG Using a Functional Areas Spatio-Temporal Transformer

Muyun Jiang, Yi Ding, Wei Zhang, Kok Ann Colin Teo, LaiGuan Fong, Shuailei Zhang, Zhiwei Guo, Chenyu Liu, Raghavan Bhuvanakantham, Wei Khang Jeremy Sim, Chuan Huat Vince Foo, Rong Hui Jonathan Chua, Parasuraman Padmanabhan, Victoria Leong, Jia Lu, Balázs Gulyás, Cuntai Guan *Fellow, IEEE*

**Abstract**—Covert speech involves imagining speaking without audible sound or any movements. Decoding covert speech from electroencephalogram (EEG) is challenging due to a limited understanding of neural pronunciation mapping and the low signal-to-noise ratio of the signal. In this study, we developed a large-scale multi-utterance speech EEG dataset from 57 right-handed native English-speaking subjects, each performing covert and overt speech tasks by repeating the same word in five utterances within a ten-second duration. Given the spatio-temporal nature of the neural activation process during speech pronunciation, we developed a Functional Areas Spatio-temporal Transformer (FAST), an effective framework for converting EEG signals into tokens and utilizing transformer architecture for sequence encoding. Our results reveal distinct and interpretable speech neural features by the visualization of FAST-generated activation maps across frontal and temporal brain regions with each word being covertly spoken, providing new insights into the discriminative features of the neural representation of covert speech. This is the first report of such a study, which provides interpretable evidence for speech decoding from EEG. The code for this work has been made public at <https://github.com/Jiang-Muyun/FAST>

**Index Terms**—Brain-Computer Interface (BCI); EEG Signal Analysis; Covert Speech Decoding.

## I. INTRODUCTION

Covert speech is the process of internally articulating speech units, such as words or sentences, without producing audible sounds or movements [1]. This method is particularly

Muyun Jiang, Yi Ding, Shuailei Zhang, Zhiwei Guo, Chenyu Liu and Cuntai Guan are with the College of Computing and Data Science, Nanyang Technological University, Singapore.

Chuan Huat Vince Foo and Rong Hui Jonathan Chua are with DSO National Laboratories, Singapore.

Wei Zhang, Kok Ann Colin Teo, LaiGuan Fong, Raghavan Bhuvanakantham, Wei Khang Jeremy Sim, Parasuraman Padmanabhan, and Balázs Gulyás are with the Cognitive Neuroimaging Centre and Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore.

Kok Ann Colin Teo is also with the Division of Neurosurgery, National University Health System, Singapore, and the IGP-Neuroscience, Interdisciplinary Graduate Programme, Nanyang Technological University, Singapore.

Victoria Leong is with the Division of Psychology, Nanyang Technological University, Singapore, and the Department of Pediatrics, University of Cambridge, United Kingdom.

Jia Lu is with the Yong Loo Lin School of Medicine, National University of Singapore, Singapore.

Balázs Gulyás is also with the Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden.

Cuntai Guan is the Corresponding Author. E-mail: ctguan@ntu.edu.sg

beneficial for individuals with verbal communication impairments due to conditions like stroke, trauma, and amyotrophic lateral sclerosis (ALS), which affect speech production and comprehension [2], [3]. This task also helps provide insights into how the brain forms speech patterns and prepares them for verbal articulation, even when no physical speech occurs. Decoding covert speech has made tremendous progress in using invasive recordings [4], [5]. Decoding from EEG has gained recognition as a method for assisting individuals with severe communication barriers to establish a channel for thought-based interaction [6]. However, the task of decoding covert speech from non-invasive neural recordings such as EEG remains challenging.

Currently, most covert speech BCIs are based on invasive methods, their feasibility has been demonstrated through various studies. This includes utilizing electrocorticography (ECoG) data to synthesize speech using densely connected 3D CNN as shown in [4]. Moreover, [7] demonstrated that the finer details captured by these recordings play a crucial role in understanding and interpreting speech-related neural activity. Additionally, [8] has shown that imagined speech could be decoded from low and cross-frequency features in intracranial electroencephalography (iEEG) signals. Non-invasive technologies have gradually garnered attention for their potential in BCIs. Notably, the use of Magnetoencephalography, as explored in [9], has successfully decoded imagined and spoken phrases. Despite these advancements, studies based on EEG remain scarce due to the challenge of small signal magnitudes, which hinder the ability of decoding methods to learn effective features.

In recent years, with the incorporation of deep learning technologies, particularly CNNs like DeepConvNet [10] and EEG-Net [11], the capability to accurately interpret EEG data has markedly improved in the tasks of motor imagery, emotional cognition, and more [12]–[20]. Building on the success of the transformer model in the fields of computer vision and natural language processing, researchers have also attempted to adopt the transformer structure for the identification and classification of brain activities, leveraging their ability to effectively model long sequences [21]–[24]. However, challenges persist in effectively integrating transformer-based architectures with EEG data. Issues such as high computational demand, the need for large datasets to train effectively, and the sensitivity to the highly variable nature of EEG signals complicate the application.

From the existing literature, numerous studies have focused on EEG classification tasks using pure convolutions or convolutional layers combined with transformer architecture approaches. However, a comprehensive insight into the limiting factors of these approaches reveals a common set of challenges:

- 1) Lack of a high-quality large-scale, multi-utterance, EEG-based covert speech dataset.
- 2) Lack of effective decoding algorithm to tackle challenging covert speech EEG signals.
- 3) Lack of understanding of the covert speech intrinsic information in EEG to enable speech decoding.

Addressing the challenges outlined, we have collected a multi-utterance dataset aimed at exploring the brain's mechanisms during covert speech activities. This dataset is collected from 57 right-handed adult males who engaged in both covert and overt speech tasks. Each participant contributed 1000 utterances, allowing for a detailed analysis of speech processing and brain function.

We developed FAST, a Functional Areas Spatio-temporal Transformer model that integrates a convolutional neural network with a transformer architecture. The model begins with a series of brain functional area convolutional tokenizers designed to extract spatially sensitive features from the data. Tokens from each brain region are first processed through a spatial projection layer to integrate information across different brain regions. These aggregated features then undergo deeper analysis through multiple transformer layers, enhancing the model's ability to decode and interpret complex brain signals. Detailed feature visualization reveals distinct and interpretable speech-neural patterns in the frontal and temporal brain regions for covertly spoken words, providing new insights into the discriminative features underlying the neural representation of covert speech.

Our main contributions can be summarized as follows:

- 1) Proposed FAST, a novel Spatio-temporal Transformer Network inspired by cerebral functional systems for covert speech EEG modeling.
- 2) Provided in-depth feature analysis, revealed discriminative features during covert speech.
- 3) We evaluate the model on a large-scale multi-utterance covert and overt speech dataset from 57 subjects totaling 1000 utterances per participant.

To the best of the authors' knowledge, this is the first study in the literature showing features learned from EEG-based covert speech BCI approaches. This model merges neuroscience-driven feature extraction with a multimodal dataset and a flexible learning strategy.

The structure of this work is outlined as follows: Section II reviews prior research closely related to our study. Section III elaborates on the detailed methodology of our model. Section IV focuses on the procedures for data collection. Section V expresses the training scheme of FAST. Section VI outlines the computational experimental setup and the evaluation of our novel method as well as our insights derived from in-depth feature analysis. Finally, Section VII concludes the study and proposes directions for future research.

## II. RELATED WORK

### A. Neural Foundations and Decoding Advancements of Covert Speech

Recent studies provide neural evidence that covert speech involves an articulatory component [25]. Activation in Broca's area, associated with speech production, and supplementary motor areas underscores its link to articulatory processes, while engagement of auditory imagery networks highlights its reliance on both auditory and motor functions [26], [27]. Research has identified distinct neural signatures for covert speech compared to listening to speech, with overlapping yet specialized activation in key brain regions [28].

Recent advancements in neural decoding of imagined speech have significantly contributed to our understanding of brain processes. For instance, one study [29] employs a diffusion-based model to decode imagined speech from EEG signals, demonstrating the potential of machine learning in enhancing speech recognition accuracy. Another study [30] investigates speech synthesis from brain signals using a generative model, paving the way for neural activity-based communication devices. Additionally, [31] explores transfer learning from overt to imagined speech through a convolutional autoencoder, improving EEG-based imagined speech decoding. A separate framework [32] integrates spoken speech data to refine neural decoding models, providing a more direct interpretation of internal speech representations. Moreover, [33] introduces a deep capsule neural network to decode vowel imagery from EEG, improving the efficiency of imagined speech processing.

Despite these advancements, existing studies often fall short in providing detailed insights into the neurological mechanisms underlying covert speech and lack fine-grained temporal visualizations of neural features. A deeper exploration of neural substrates could bridge these gaps, offering a clearer understanding of covert speech processes and improving BCI design for more effective and intuitive covert speech recognition.

### B. Transformer models for EEG

Building upon the success of transformer models in computer vision and natural language processing, researchers have explored their application in EEG analysis to capture complex spatio-temporal patterns inherent in brain activity. For instance, [23] proposed a transformer-based EEG Conformer that effectively learns spatial and temporal features from EEG data, achieving state-of-the-art performance in emotion recognition tasks. Recently, [34] introduced a multi-anchor space-aware temporal transformer model designed for EEG decoding, demonstrating its efficacy across various datasets. Additionally, [35] developed a transformer architecture that analyzes spatio-temporal dynamics of EEG signals to estimate emotion states.

However, these studies often lack clear visualizations of the tokenization process, making it challenging to interpret and transfer the learned temporal tokens effectively. Furthermore, there is limited investigation into how these temporal tokens interact with spatial features to capture the intricate dynamics

of EEG signals. A more comprehensive exploration of this interplay could provide deeper insights into the spatiotemporal characteristics of brain activity to enhance the interpretability and robustness of neural decoding models.

### III. METHODOLOGY

In this section, we introduce the model structure of FAST. As depicted in Figure 1, FAST comprises two primary components: the Spatial-temporal Tokenizer (ST) and the Transformer Encoder (TE) Block. ST is responsible for processing information from each brain function area, while the TE blocks utilize a transformer architecture to operate on the tokens generated by ST. The network undergoes a two-step pre-training process followed by a fine-tuning process. The source code of this work has been opened at <https://github.com/Jiang-Muyun/FAST>

#### A. Spatial-temporal Tokenizer

The ST block is composed of multiple independent brain functional area encoders, each responsible for processing information only corresponding to specific brain functions. Given an EEG trial input with  $N$  channels, denoted as  $X^N$ , our preliminary step involves segmenting the EEG trial into shorter segments. This segmentation is achieved by applying a fixed-size window and slide across the trial, ultimately resulting in  $S$  segments. Each segment, particularly the  $i^{th}$  segment, is denoted as  $X_i^N$ , where  $i$  ranges from 1 to  $S$  indicating the segment index. Following this segmentation, the next step is to partition the EEG signals into  $M$  distinct brain areas.

The brain's functions are complex networks with hierarchical organization across neurons, local circuits, and functional areas [17], [36], [37]. EEG electrodes are positioned on the scalp following the 10-10 [38] system, which organizes channels based on their placement over distinct regions of the brain. In this study, we segment the brain into different functional regions according to the spatial locations of EEG channels, similar to LGGNet [15], as illustrated in Figure 2.

This division allows us to examine the specific roles of various brain regions, particularly focusing on the functional responsibilities of the frontal [37] and temporal regions [39]. Specifically, the frontal region is associated with executive functions, decision-making, and attentional control, while the temporal region plays a critical role in auditory processing, memory, and language comprehension. By isolating these areas through the 10-10 electrode system, the network will be able to learn and capture region-specific spatial features more effectively [15].

Mathematically, the division can be represented as a partitioning process

$$P : X_i^N \rightarrow \{x_i^{m_1}, x_i^{m_2}, \dots, x_i^{m_j}\} \quad (1)$$

where  $P$  signifies the partitioning function that takes a subset of channels from each segment  $X_i^N$  to a set of  $M$  brain areas for that segment, and  $x_i^{m_j}$  denotes the subset of channels related to the  $j^{th}$  brain area in the  $i^{th}$  segment.

Each partitioned brain area  $x_i^{m_j}$ , undergoes processing through a specific functional area encoder  $f_j(x)$ , a sequence

of convolutional layers, normalization layers, activation functions, and max pooling operations. A spatial-temporal module is placed at the beginning of the series of convolutions to capture the spatial-temporal information in each functional area. Denoted as  $\text{Conv}_T$  and  $\text{Conv}_S$ . This sequential processing is represented as follows:

$$x_{i,\text{temporal}}^{m_j} = \text{Conv}_T(x_i^{m_j}) \quad (2)$$

$$x_{i,\text{spatial}}^{m_j} = \text{Conv}_S(x_{i,\text{temporal}}^{m_j}) \quad (3)$$

After extracting spatial and temporal features, the data is further processed through Batch Normalization (BN), Gaussian Error Linear Units (GELU) activation functions, and max pooling operations. This is expressed as:

$$x_{i(1)}^{m_j} = \text{MaxPool}(\text{GELU}(\text{BN}(x_{i,\text{temporal}}^{m_j}))) \quad (4)$$

where  $\Phi(x)$  is the cumulative distribution function of the standard Gaussian distribution.

After the first spatial-temporal encoder, a series of temporal encoders are applied to  $x_i^{m_j}$  to extract the short-time information within each functional area. The temporal encoder comprises convolutional layers, normalization layers, activation functions, and pooling operations. The process is outlined as follows:  $l \in [2, L_t]$  and  $L_t$  is the total layer number of the temporal convolution layers:

$$x_{i(l)}^{m_j} = \text{MaxPool}(\text{GELU}(\text{BN}(\text{ConvT}_l(x_{i(l-1)}^{m_j})))) \quad (5)$$

To achieve a representation of the spatial signals that is invariant to the length of the EEG segments, global max pooling is applied across the time dimension, transforming the signal into fixed-dimensional vectors, each representing the most significant signal peak within the duration for its respective brain area:

$$F_i^{m_j} = \text{GlobalMaxPool}_{\text{time}}(x_{i(L_t)}^{m_j}) \quad (6)$$

$F_i^{m_j}$  represent the feature vector of the  $j^{th}$  brain functional area of the  $i^{th}$  segment.

#### B. Transformer Encoder

The TE blocks utilize a transformer architecture that operates on the tokens generated by ST, by first going through spatial projection layers and then transformer blocks. The spatial projection layers operate on the spatial dimension, while transformer blocks operate on the temporal dimension.

The spatial projection involves processing through a series of  $L_s$  layers of multi-head attention mechanisms, each focusing on refining the representation of individual brain functional areas. Formally, for the  $j^{th}$  brain functional area in the  $i^{th}$  segment, the process can be expressed as follows

$$H_{i,(0)}^{m_j} = F_i^{m_j}, \quad (7)$$

The initial feature vector  $H_{i,(0)}^{m_j}$  serves as the input for the first layer. In subsequent layers, the output from the previous layer undergoes a transformation via the multi-head attention

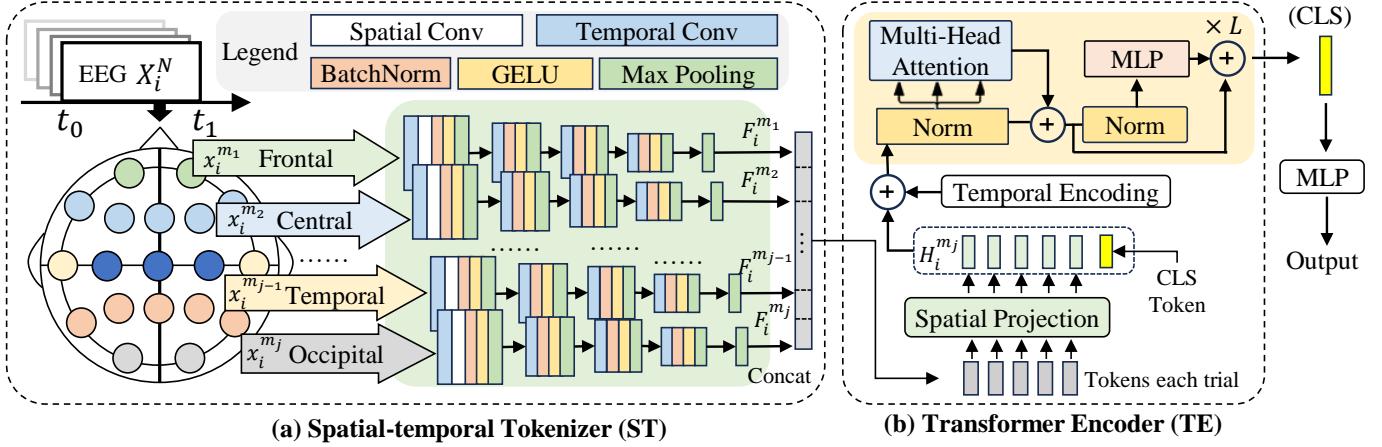


Fig. 1. Overview of proposed FAST. (a) The Spatial-temporal Tokenizer (ST) block illustrates the initial processing of EEG data through spatial and temporal convolutional layers. (b) The Transformer Encoder (TE) block shows the transformer architecture used for tokens generated by ST, which outputs a learned CLS token for classification results.

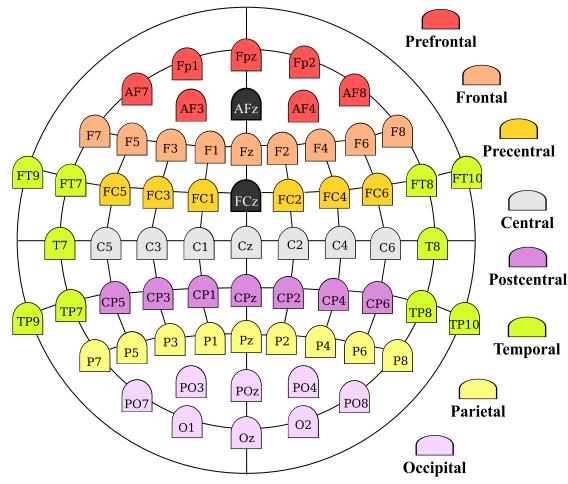


Fig. 2. The divided brain functional regions are based on the spatial locations of EEG channels, where FCz serves as the reference electrode, and AFz serves as the ground electrode.

function  $\text{Multi-Head}(\cdot)$ . The transformation at each layer  $l$  is defined as:

$$H_{i,(l)}^{m_j} = \text{Multi-Head}(H_{i,(l-1)}^{m_j}, H_{i,(l-1)}^{m_j}, H_{i,(l-1)}^{m_j}) \quad (8)$$

$$\delta = \text{Multi-Head}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (9)$$

$$\text{head}_h = \text{Attention}(QW_h^Q, KW_h^K, VW_h^V) \quad (10)$$

where  $Q$ ,  $K$ , and  $V$  represent the query, key, and value matrices,  $W_h^Q$ ,  $W_h^K$ , and  $W_h^V$  are the weight matrices specific to each head for the query, key, and value. The outputs from individual heads are concatenated and then transformed by multiplying with the output weight matrix  $W^O$ .

After each multi-head attention layer, a position-wise feed-forward network (FFN) is added to form a transformer layer. The FFN is formulated as:

$$\text{FFN}(\delta) = W_2 \text{GELU}(W_1 \delta + b_1) + b_2 \quad (11)$$

$$\delta' = \text{LN}(\delta + \text{FFN}(\delta)) \quad (12)$$

where  $W_1$ ,  $W_2$  are the weight matrices and  $b_1$ ,  $b_2$  are the biases for the FFN. Layer normalization (LN) and residual connections are applied at the end of FFN, where  $\delta$ , the output of the Multi-Head Attention, serves as the input to the FFN, and  $\delta'$  represents the output of the FFN.

Then,  $L_s$  transformer layers in the spatial projection block are applied iteratively to process the data from each brain region, forming an output of  $H_{i,(L_s)}^{m_j}$ . By concatenating them for each region  $j \in \{1, \dots, J\}$  into a single feature vector, we obtain  $G_i$ , a refined representation of all the brain functional areas within the  $i^{th}$  segment. This process can be expressed as:

$$G_i = \bigoplus_{j=1}^J H_{i,(L_s)}^{m_j} \quad (13)$$

where  $\bigoplus$  denotes the concatenation operation. This results in a comprehensive feature vector  $G_i$  that encapsulates the functional dynamics of all the brain regions for the  $i^{th}$  segment.

Following spatial projection, a transformer block comprising  $L$  layers, which leverages self-attention and FFNs as previously described, is employed on the sequence of extracted feature vectors  $G_i$  to decode temporal information across trials. To better capture temporal dynamics within the transformer framework, a temporal encoding, denoted by  $T_i$ , is introduced to each input feature vector. This encoding is initially randomized and subsequently learned during training. Additionally, a classification token represented as  $F_{cls}$  is introduced and similarly learned during training. The input tokens for this transformer are presented as follows:

$$G = \{G_1 + T_1, G_2 + T_2, \dots, G_i + T_i, F_{cls} + T_{i+1}\}, \quad (14)$$

The transformer operates iteratively as depicted:

$$G^L = \text{Transformer}_l(G^{l-1}) \quad (15)$$

for  $l \in \{1, \dots, L\}$ . Each transformer block applies self-attention across the entire sequence of feature vectors, including the classification token, facilitating interactions among various segments of the input. The final classification token,  $G_{cls}^L$ ,

is derived after multiple layers of attention and subsequently processed through a classification head for predictions:

$$\hat{y} = \text{MLP}(G_{\text{cls}}^L) \quad (16)$$

where  $\hat{y}$  represents the predicted probabilities, with a shape of (B,5) corresponding to the 5 classes in our classification task. The algorithm detailed above, which we will refer to as Algorithm 1, can be briefly summarized as follows:

---

#### Algorithm 1 Data Processing and Feature Extraction

---

**Require:** EEG input  $X^N$  with  $N$  channels

Segment  $X^N$  into  $S$  segments  $X_i^N$ ,  $i = 1$  to  $S$

Partition  $X_i^N$  into  $M$  areas:  $P : X_i^N \rightarrow \{x_i^{m_1}, \dots, x_i^{m_M}\}$

**for** each area  $m_j$  in segment  $i$  **do**

    Encode  $x_i^{m_j}$  using  $f_j(x)$

    Apply spatial-temporal processing to  $x_i^{m_j}$

    Normalize, activate, and pool:  $x_{i(1)}^{m_j}$

    Process for invariant representation:  $x_{i(L_t)}^{m_j}$

    Global max pooling:  $F_i^{m_j}$

**end for**

Apply spatial projection to get  $H_{i,(L_s)}^{m_j}$

Concatenate to get  $G_i$  for each segment

Prepare sequence  $G$  with  $G_i$  and  $G_{\text{cls}}$

Encode sequence with transformer

Predict  $\hat{y} = \text{MLP}(G_{\text{cls}}^L)$

Return  $\hat{y}$

---

## IV. DATASET

### A. Protocol Settings

We collected a multi-utterance dataset during overt and covert speech activities. The dataset comprises recordings from 57 right-handed native English-speaking adult males, each performing covert and overt speech tasks by repeating the same word in five utterances within a ten-second duration, with a total of 1000 utterances per individual. The average age of the subjects was  $24.2 \pm 3.6$  years. The information of the collected dataset is shown in Table I. The data collection adhered to the guidelines of the Declaration of Helsinki and took place at the Cognitive Neuroimaging Centre (CoNiC) at Nanyang Technological University (NTU) in Singapore. Written informed consent was obtained from all participants. The study received ethical approval from the Institutional Review Board (IRB) of NTU under the approval number IRB-2022-040.

As shown in Fig 3, the experimental protocol comprises a total of 10 blocks, including alternating overt speech blocks and covert speech blocks, totaling 5 blocks for each condition. Every block lasts about 5.5 minutes, followed by a short rest of 45 seconds. There are a total of 20 trials in each block, participants were instructed to speak overtly or covertly, 5 repetitions in each 10-second trial. Between trials, the words were presented in a pseudo-random order to reduce predictability. During the overt speech blocks, participants were directed to speak each word loudly and clearly, keeping the natural tone of their daily speech. During the covert speech blocks,

TABLE I  
SUMMARY OF DATA COLLECTION SETTINGS

Feature	Details
Subjects	57
Modality	EEG
Experiment Type	Overt and covert and word-speaking
Total Blocks	10 (5 blocks for each condition)
Block Duration	5.5 minutes
Rest Between Blocks	45 seconds
Trials per Block	20
Words per Trial	5 utterances
Selected Words	"Go there", "Distract Target", "Follow me", "Explore Here", "Terminate"

participants were instructed to imagine the pronunciation of the word without any physical movement or production of any audible sound. We chose to collect a multi-utterance dataset to ensure more stable and robust neural responses. The single utterance is prone to variability due to factors like attention lapses or insufficient context, which can adversely affect the decoding and induction of cognitive signals.

We used 5 words for speaking in this experiment based on the following criteria: 1) Adequate length featuring a polysyllabic structure. 2) Distinct articulation between terms. 3) To enhance the system's applicability for BCI, the chosen words should relate to robot control functions. The chosen words are "Go there", "Distract Target", "Follow me", "Explore Here", and "Terminate".

### B. Data Collection and Preprocessing

EEG data were captured using a Brain Product BrainCap MR device equipped with 64 ring-type electrodes, recording at a sampling rate of 5,000 Hz. Electrode positioning follows the standard 10-10 system, with FCz as the reference electrode and AFz as the grounding electrode. Impedance checks were conducted before each experiment to ensure that the impedance values of all electrodes remained below 5 kΩ.

The continuously captured EEG data are subjected to band-pass filtering from 1 Hz to 40 Hz using an FIR filter to eliminate slow drifts and high-frequency noise. Additionally, a 50 Hz notch filter was applied to exclude line noise, followed by downsampling to 200 Hz [15], [40]. Baseline correction was performed using the 1-second period before the cue. ICA decomposition was then applied to the EEG data to identify and remove components associated with eye movement (EoG) artifacts linked to the Fp1 and Fp2 electrodes. Additionally, muscle-related (EMG) artifacts were automatically detected and removed using MNE. Subsequently, the continuous EEG data were segmented into epochs aligned with the markers indicating word onset.

## V. EXPERIMENT SETTINGS

### A. Training Scheme

The training scheme for FAST and baseline models involves a subject-independent pre-training phase followed by a subject-dependent fine-tuning phase. Leave-one-subject-out (LOSO) approach is used in the pre-training phase, where the model is trained on data from all subjects except one, ensuring

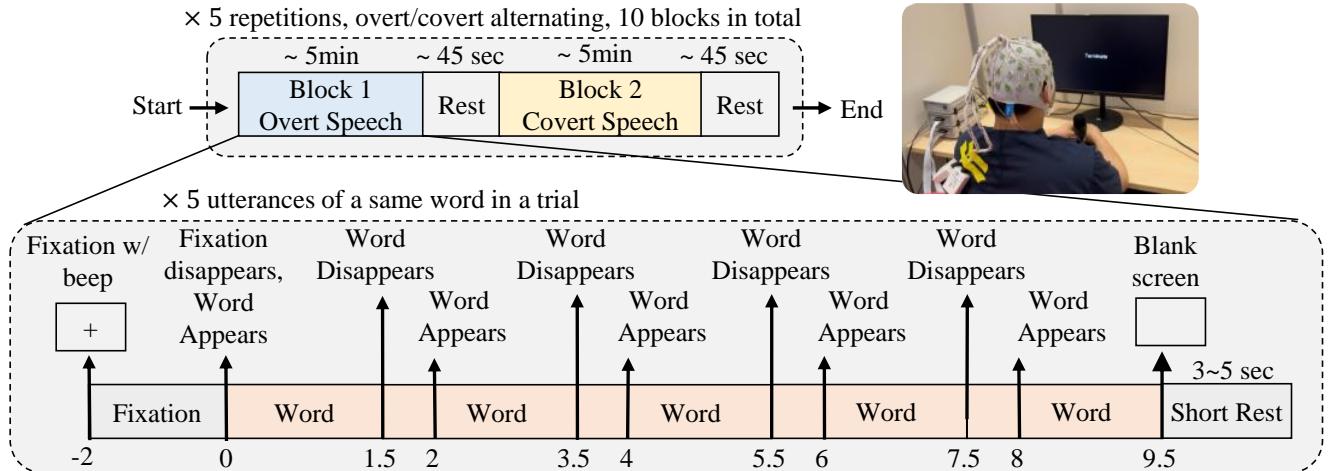


Fig. 3. Protocol of the experiment. (Top): The experiment is structured into blocks. Each subject will complete 10 blocks, alternating between overt and covert EEG experiments. Each block consists of 20 trials, presenting five words in a pseudo-random order. (Bottom): In each trial, the same words were displayed on the screen at predetermined intervals ( $T = 0, 2, 4, 6, 8$  seconds) and vanished at  $T + 1.5$  seconds. Subjects are instructed to overtly pronounce or covertly imagine pronouncing the words five times following the blink of the words. A brief resting period, with a random duration ranging from 3 to 5 seconds, is provided after each trial.

independence from the excluded subject. During fine-tuning, a leave-one-block-out (LOBO) [41] cross-validation method is applied by dividing each subject's data into five blocks, each containing 20 trials. In each fine-tuning iteration, four blocks serve as the training set, while the remaining block is held out for testing. This cycle is repeated until each block has been used as the test set, providing a thorough fine-tuning and evaluation across the full dataset.

Each model was trained for a fixed 200 epochs, after which training was stopped, and the accuracy of the test set was evaluated. We made sure the test dataset was only used once in the final assessment. Model performance is assessed using multi-class accuracy, F1-score, Cohen-Kappa, and AUC, present in Table II. The equations for calculating these metrics are provided in the supplementary material.

## VI. RESULTS

In this section, we detail the performance of FAST for the covert speech task. We present detailed results on the 5-utterances cases in Table II. We evaluate FAST against a list of commonly used models, including ST-Transformer [44], BIOT [42], EEGViT [43], EEGNet [11], DCN [10], EEG-Conformer [23], and TSception [46], as well as more recent architectures such as EEG-Deformer [45]. All baseline models are implemented with the same pre-training and fine-tuning strategy as the proposed method. Box plot of the accuracy visualization is shown in Figure 4. The chance level was determined using a binomial distribution model [47] with a random guessing probability of  $p = 0.2$ , and the standard deviation was calculated as  $\sigma = \sqrt{p(1-p)/n}$ . Using a normal approximation, the 95% confidence interval for random accuracy was [0.1896, 0.2104].

### A. Results and Findings

Table II displays the average classification accuracies, F1-score, Cohen-Kappa, AUC, and the respective standard devi-

ations for FAST and the comparative baseline models for 5-utterances covert speech classification. Wilcoxon signed-rank test is performed between FAST and each of the baseline methods, with statistical significance levels indicated with asterisks (\*), and the top-performing results highlighted in bold. The results of the test reveal that FAST significantly outperforms some baseline models during pre-training and outperforms most models except TSception during fine-tuning.

Fine-tuning improves the performance of all models across the evaluated metrics. For instance, FAST's accuracy increases from 26.9% in the pre-training phase to 34.7% after fine-tuning. Similarly, baseline models such as EEG-Deformer, DCN, EEGNet, and Conformer also show notable improvements. FAST achieves the highest accuracy of 34.7%, outperforming all baseline models. TSception, EEG-Deformer, the closest competitor, achieves 31.2%, showing a clear advantage for FAST. FAST attains the highest F1-Score of 0.340, indicating a better balance between precision and recall than other models, including TSception (0.287). FAST records the highest Cohen-Kappa value (0.184), indicating stronger agreement compared to TSception (0.140) and EEG-Deformer (0.126).

Figure 4 (b) illustrates the fine-tuning accuracy distributions for each model. FAST demonstrates a higher median accuracy and reduced variability, indicating its stability and efficiency. While TSception and EEG-Deformer remain competitive. The reason for the lower effectiveness of baseline models is primarily due to their design limitations. These models were not originally intended to handle the complexities of speech-related data. Additionally, CNN-only models struggle with handling long input sequences, which limits their performance in our task.

### B. Visualization and Analysis

In this section, we present the visualization and analysis of the ST features extracted from covert speech EEG data.

TABLE II

PERFORMANCE COMPARISONS IN THE PRE-TRAIN AND FINE-TUNE PHASE FOR COVERT SPEECH IN 5 UTTERANCES OF THE SAME WORD

Stage	Model	Accuracy(%)	F1-score	Cohen-Kappa	AUC
Pretrain	BIOT [42]	21.8 ± 4.6 ***	0.187 ± 0.048 *	0.023 ± 0.058 ***	0.508 ± 0.017 ***
	EEGViT [43]	23.2 ± 4.9 **	0.221 ± 0.048	0.039 ± 0.062 **	0.535 ± 0.045 ***
	DCN [10]	25.0 ± 5.7	0.240 ± 0.057	0.063 ± 0.071	0.561 ± 0.063 ***
	EEGNet [11]	25.1 ± 5.8	0.238 ± 0.060	0.064 ± 0.072	0.570 ± 0.052 ***
	ST-Transformer [44]	24.6 ± 5.1 *	0.228 ± 0.060	0.057 ± 0.064 *	0.557 ± 0.048 ***
	EEG-Conformer [23]	26.1 ± 6.3	0.189 ± 0.087	0.077 ± 0.079	0.608 ± 0.083
	EEG-Deformer [45]	26.2 ± 6.4	<b>0.241 ± 0.070</b>	0.077 ± 0.080	0.591 ± 0.061 *
	TSeption [46]	25.6 ± 6.2	0.187 ± 0.074 *	0.070 ± 0.078	0.618 ± 0.075
	FAST	<b>26.9 ± 6.9</b>	0.221 ± 0.091	<b>0.087 ± 0.086</b>	<b>0.627 ± 0.083</b>
Finetune	BIOT [42]	24.0 ± 6.5 ***	0.234 ± 0.063 ***	0.050 ± 0.081 ***	0.530 ± 0.041 ***
	EEGViT [43]	24.6 ± 5.6 ***	0.244 ± 0.056 ***	0.057 ± 0.070 ***	0.558 ± 0.059 ***
	DCN [10]	26.5 ± 6.5 ***	0.264 ± 0.066 ***	0.081 ± 0.082 ***	0.577 ± 0.072 ***
	EEGNet [11]	27.2 ± 6.5 ***	0.271 ± 0.064 ***	0.089 ± 0.081 ***	0.589 ± 0.064 ***
	ST-Transformer [44]	27.2 ± 6.4 ***	0.269 ± 0.063 ***	0.090 ± 0.080 ***	0.584 ± 0.059 ***
	EEG-Conformer [23]	28.6 ± 7.0 ***	0.221 ± 0.047 ***	0.107 ± 0.087 ***	0.604 ± 0.082 ***
	EEG-Deformer [45]	30.1 ± 7.9 *	0.290 ± 0.082 **	0.126 ± 0.099 *	0.615 ± 0.076 **
	TSeption [46]	31.2 ± 8.3	0.287 ± 0.076 **	0.140 ± 0.103	0.623 ± 0.090 *
	FAST	<b>34.7 ± 10.7</b>	<b>0.340 ± 0.108</b>	<b>0.184 ± 0.134</b>	<b>0.662 ± 0.097</b>

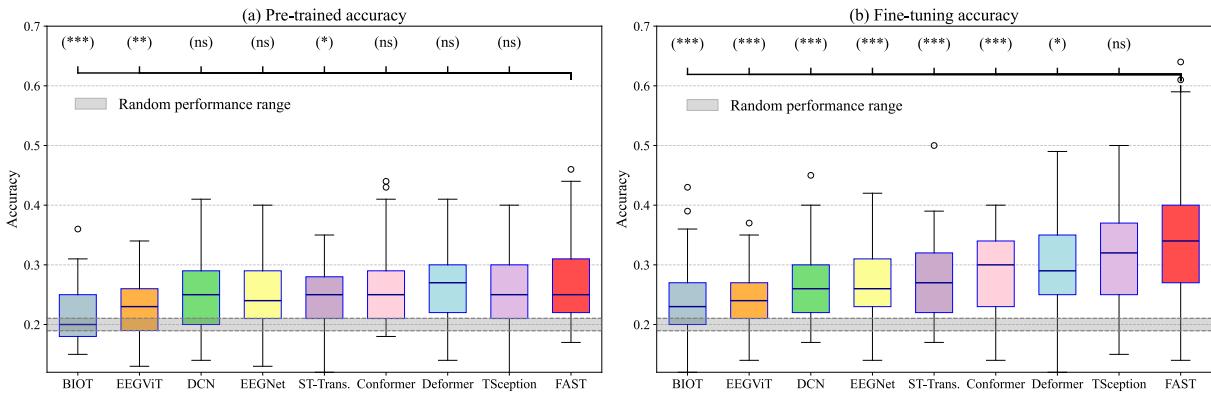
Note \*\*\* denoting a  $p$ -value less than 0.001, \*\* for less than 0.01, and \* for less than 0.05

Fig. 4. Box plot illustrating the accuracy of covert speech recognition for all subjects ordered by the median accuracy: (a) Accuracy from pre-trained models; (b) Accuracy after fine-tuning. Significance levels are compared between FAST and each of the baselines, indicated with asterisks (\*), where (ns) denotes  $p$ -values  $>0.05$ . The random performance range is indicated in the gray bar.

The aim is to gain a deeper understanding of the temporal and spatial activation patterns associated with covert speech processing across different brain regions.

In Fig. 5(a), visualizations of the covert speech features averaged across all subjects are shown. This averaging helps smooth out individual variability and noise, resulting in a clearer and more interpretable representation of the data. The features were derived from the left-out subject following each round of leave-one-subject-out training. The visualization is conducted by densely sliding windows across each EEG trial with a step size of 0.02 seconds. These segments are fed into pre-trained ST for feature extraction for each leave-out subject after each round of the LOSO training. Since the features output by ST have already passed through an activation layer, their values inherently represent the strength of the activation. We then averaged these features across all subjects and applied z-score normalization along the trial dimension for the plot shown in Figure 5(a). The normalized activation per class shown in Figure 5(b) is computed by averaging the features on a per-class basis, with a single line representing the activation for each word and brain region. This process captures

the fine-grained temporal and spatial activation during each covert speech trial. To more clearly visualize the difference in activation between words, a one-vs-all strategy is adopted, with the results illustrated in Figure 5(c).

To interpret model predictions, the Integrated Gradients [48] method is employed, which assigns an importance score to each input feature by approximating the integral of the model's output gradients shown in Figure 6. A detailed algorithm for generating this figure is described in the supplementary material.

The activation map shown in Figure 5(a) highlights task-related responses occurring at both the onset and offset of word cues. Elevated rhythmic responses are observed in frontal and temporal regions, crucially implicated in speech processing [49], as well as in the occipital lobe, predominantly associated with visual processing [50]. A stronger response is observed during the first cue onset. Figures 5(b) and (c) illustrate that activation within the frontal and temporal lobes remains stable prior to  $t = 0$  sec and becomes discriminative during the cue onset period (marked by the gray bar). Once the cue disappears at  $t = 10$  sec, activation levels revert to baseline. Conversely,

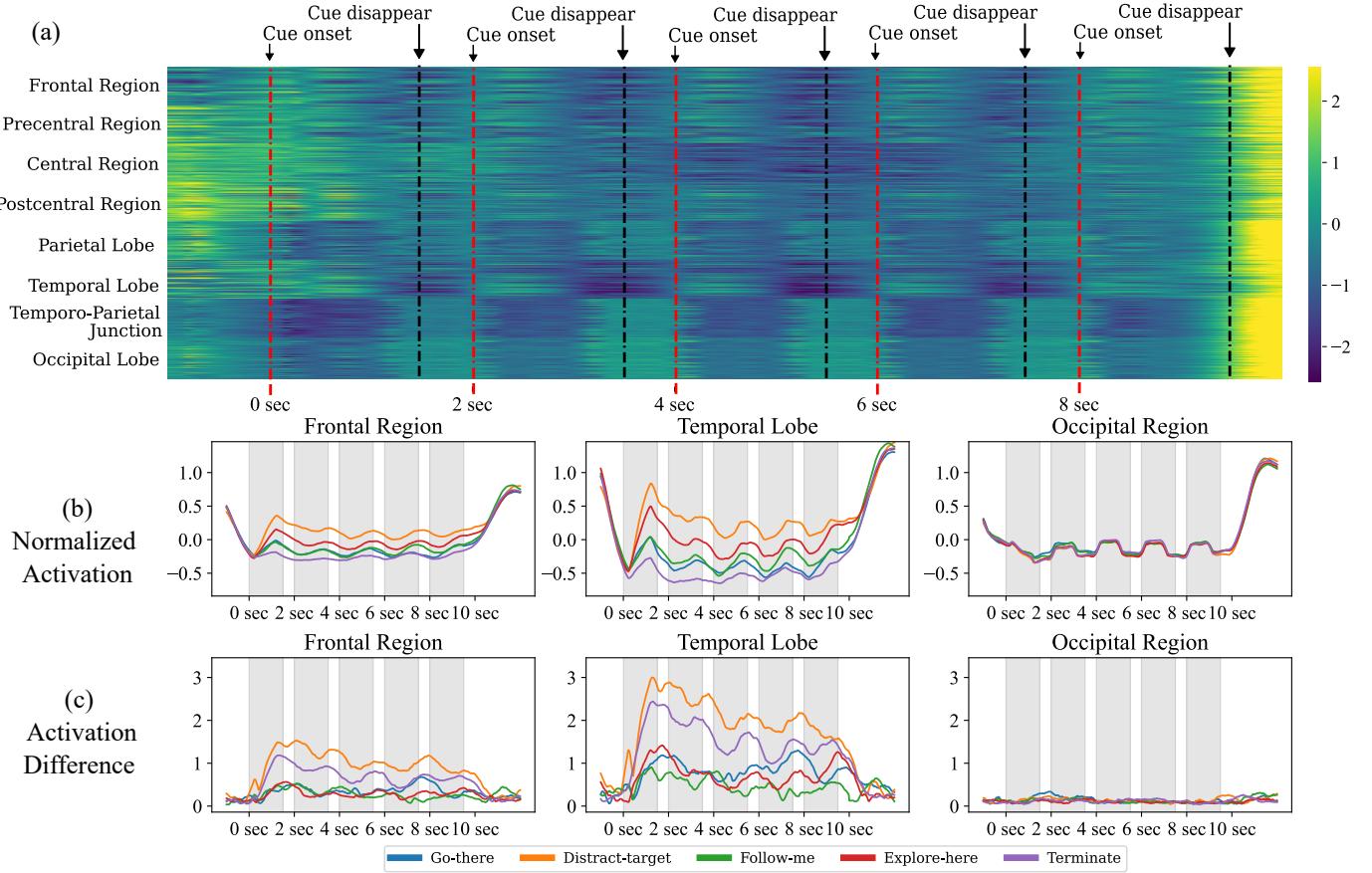


Fig. 5. Feature visualization of the ST on covert speech EEG data averaged across all subjects. The features were extracted from the leave-out subject after each round of leave-one-subject-out training. (a) A heatmap illustrates the features generated by the ST layers along with the corresponding time-locked features. (b) Normalized activation maps highlight the relative activation scores across the frontal, temporal, and occipital lobes. (c) Difference activation maps show the one-versus-all contrast for each region.

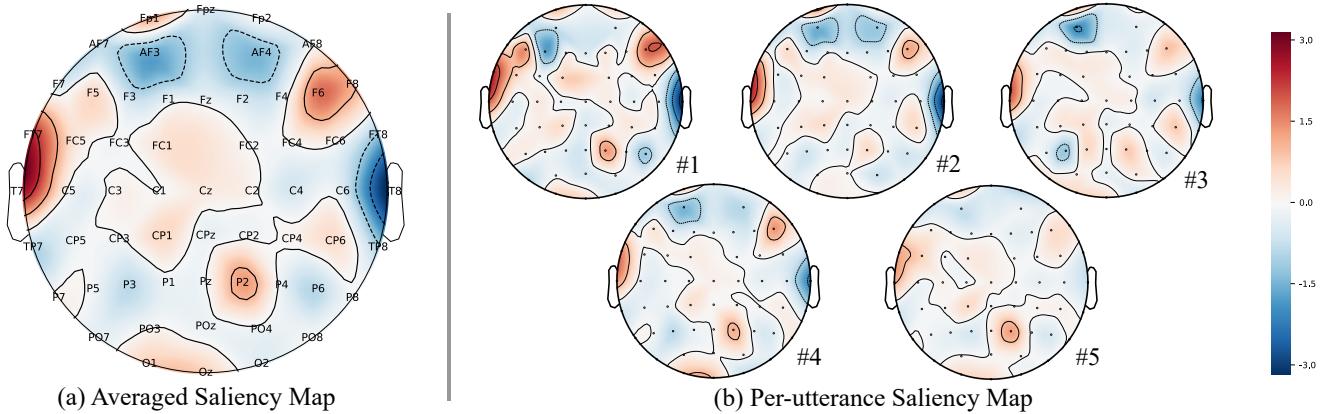


Fig. 6. EEG saliency maps during covert speech (a): Activation is prominently observed in the left hemisphere electrodes (T7, FT7, FC5, C5, F5), corresponding primarily to Broca's area, suggesting involvement in speech motor planning and articulatory rehearsal. Additional activation is visible in electrodes F6 and F8 on the right hemisphere homologue of Broca's area, possibly reflecting prosodic processing or cognitive control mechanisms. No activation was detected in Wernicke's or occipital areas. (b) Activation intensity gradually decreases across successive utterances.

occipital lobe activation aligns with cue onset but does not show discriminative responses between the words. This finding indicates that visual information does not contribute to decoding performance. This activation pattern supports the critical role of frontal and temporal lobes [51] in covert speech processing [49], whereas occipital activation reflects visual responses to the cue rather than linguistic content.

Figure 6 illustrates the EEG saliency maps derived from the Integrated Gradients [48] method for the covert speech classification task. The averaged saliency map Figure 6(a) demonstrates pronounced neural activation predominantly localized within the left hemisphere, particularly at electrodes (T7, FT7, FC5, C5, F5), with additional notable activations at AF7 and AF3. These electrode sites correspond primarily to

Broca's area and adjacent premotor regions [52], suggesting robust involvement in speech motor planning, articulatory rehearsal [49], and phonological processing [53] during covert speech tasks. In addition to the left hemisphere activations, activations in the right frontal electrodes F6 and F8, which may correspond anatomically to the right hemisphere homologue of Broca's area [54]. This right hemisphere activation likely indicates neural processes associated with prosodic modulation [55]–[58], emotional aspects [59] of speech, or cognitive control [60] mechanisms during covert speech.

The per-utterance saliency map depicted in Figure 6(b) indicates a gradual reduction in neural activation intensity across successive utterances. This diminishing pattern suggests reduced cognitive and articulatory effort with repeated covert rehearsal of speech content.

### C. Ablation Study

To investigate whether repetitions of a word enhance decoding accuracy, the utterances split paradigm is shown in Fig 7. The input to the model was controlled—specifically, 2 to 10 seconds, corresponding to 1 to 5 repetitions of the word. The effect of the number of utterances of the same word on model performance is present in Table III. The results demonstrate a consistent improvement in decoding accuracy with an increasing number of utterances during both pre-training and fine-tuning stages. The improvement could be attributed to several factors. First, multiple utterance attempts reduce the noise and variability inherent in single-utterance data recording [61], [62], providing a more stable signal. Second, they offer richer contextual information, enabling the model to capture and learn stable cognitive patterns. Third, the aggregation of data across multiple utterances enhances the signal-to-noise ratio, leading to improved decoding accuracy.

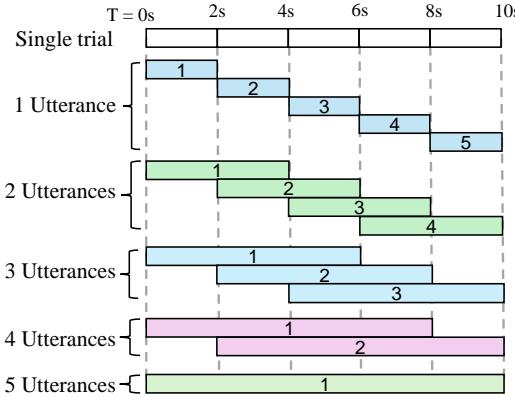


Fig. 7. Illustration of experiments using various utterance splits, where each 10-second trial consists of five repetitions of the same word.

To investigate the individual contributions of model components and training strategies to the overall performance by selectively disabling specific elements. Two different ablation experiments are conducted. To investigate the contributions of the TE, we performed an ablation study comparing classification accuracy using only the ST and connecting the classification MLP directly after ST. The results, presented in Figure 8 (a), reveal that incorporating TE consistently

TABLE III  
PERFORMANCE COMPARISONS OF FAST FOR COVERT SPEECH  
ACROSS 1 TO 5 UTTERANCES OF THE SAME WORD

N utterances	Pre-trained accuracy	Fine-tuned accuracy
1 Utterance	23.9 ± 3.9	28.4 ± 6.3
2 Utterances	24.9 ± 4.4	30.5 ± 7.0
3 Utterances	25.5 ± 4.8	32.6 ± 7.9
4 Utterances	26.2 ± 5.7	33.3 ± 8.3
5 Utterances	26.9 ± 6.3	34.7 ± 10.7

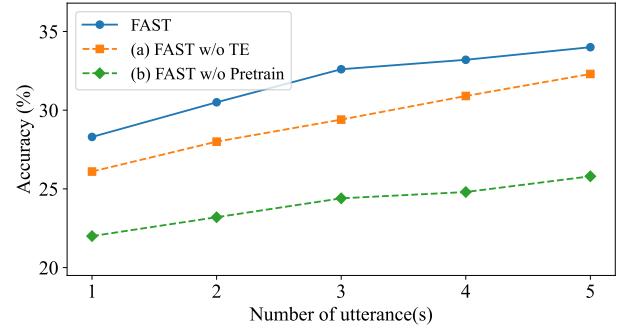


Fig. 8. The average covert speech accuracies of FAST under different ablation cases. (b) An ablation study excludes TE blocks, where the classification head is directly connected after ST. (c) Ablation study without fine-tuning, where weights are randomly initialized instead of loading from LOSO pretraining.

improves classification accuracy across the fine-tuning phase with respect to different utterances.

To evaluate the impact of our pre-training and fine-tuning strategy, we conducted an ablation study without pre-training, where model parameters are randomly initialized rather than loaded from the pre-trained checkpoints. As shown in Figure 8 (b), the results demonstrate that the model without pre-training exhibits lower accuracy across different utterances. This suggests that pre-training plays a crucial role in enabling the model to learn generalizable representations of covert speech.

## VII. CONCLUSION AND FUTURE WORKS

In this paper, we proposed FAST an innovative covert speech architecture that incorporates ST blocks and TE blocks, demonstrating superior performance to existing baseline models. To overcome the scarcity of large-scale datasets in covert speech tasks for EEG-based BCI, we have collected a large-scale multi-utterance dataset with 57 subjects, each performing 1000 utterances per word. Analysis of extracted features from the model revealed particularly high responsiveness in the frontal temporal region of the brain, providing new insights for the understanding of neural dynamics during covert speech. We also validate the effectiveness of our algorithm on the publicly available BCI competition dataset.

While our study presents novel patterns related to covert speech, it has certain limitations. Notably, we included only male participants to minimize EEG variability, limiting our findings' generalizability to females. Neural responses during covert speech may differ across genders, and future research should incorporate a more diverse participant pool to enhance robustness. Additionally, expanding the dataset to different

languages and demographics would further validate our approach to make a widely applicable solution for EEG-based covert speech recognition.

FAST's interpretability may offer practical benefits for other BCI applications in the future. For example, in neurofeedback, its activation patterns can help identify cognitive state-specific signatures to design personalized training protocols. In speech rehabilitation, these insights can decode neural correlates of speech, guiding the development of tailored therapeutic interventions. Overall, FAST's detailed mapping of brain activity may provide a robust foundation for developing more targeted and effective BCI applications in the future.

### ACKNOWLEDGMENT

This study is supported by the DSO National Laboratories (DSOCL21193), Singapore.

- [1] S. L. Metzger, J. R. Liu, D. A. Moses, M. E. Dougherty, M. P. Seaton, K. T. Littlejohn, J. Chartier, G. K. Anumanchipalli, A. Tu-Chan, K. Ganguly *et al.*, "Highly generalizable spelling using a silent-speech bci in a person with severe anarthria," in *Brain-Computer Interface Research: A State-of-the-Art Summary 11*. Springer, 2024, pp. 21–28.
- [2] 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.
- [3] A. B. Silva, K. T. Littlejohn, J. R. Liu, D. A. Moses, and E. F. Chang, "The speech neuroprosthesis," *Nature Reviews Neuroscience*, pp. 1–20, 2024.
- [4] M. Angrick, C. Herff, E. Mugler, M. C. Tate, M. W. Slutzky, D. J. Krusinski, and T. Schultz, "Speech synthesis from ecog using densely connected 3d convolutional neural networks," *Journal of neural engineering*, vol. 16, no. 3, p. 036019, 2019.
- [5] J. Lu, Y. Li, Z. Zhao, Y. Liu, Y. Zhu, Y. Mao, J. Wu, and E. F. Chang, "Neural control of lexical tone production in human laryngeal motor cortex," *Nature Communications*, vol. 14, no. 1, p. 6917, 2023.
- [6] X. Gu, Z. Cao, A. Jolfaei, P. Xu, D. Wu, T.-P. Jung, and C.-T. Lin, "Eeg-based brain-computer interfaces (bcis): A survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 18, no. 5, pp. 1645–1666, 2021.
- [7] S. Duraivel, S. Rahimpour, C.-H. Chiang, M. Trampis, C. Wang, K. Barth, S. C. Harward, S. P. Lad, A. H. Friedman, D. G. Southwell *et al.*, "High-resolution neural recordings improve the accuracy of speech decoding," *Nature communications*, vol. 14, no. 1, p. 6938, 2023.
- [8] 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.
- [9] D. Dash, P. Ferrari, and J. Wang, "Decoding imagined and spoken phrases from non-invasive neural (meg) signals," *Frontiers in neuroscience*, vol. 14, p. 290, 2020.
- [10] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.
- [11] 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, 2018.
- [12] K. P. Thomas, C. Guan, L. C. Tong, and V. A. Prasad, "An adaptive filter bank for motor imagery based brain computer interface," in *2008 30th Annual international conference of the IEEE engineering in medicine and biology society*. IEEE, 2008, pp. 1104–1107.
- [13] R. Mahamune and S. H. Laskar, "Classification of the four-class motor imagery signals using continuous wavelet transform filter bank-based two-dimensional images," *International Journal of Imaging Systems and Technology*, vol. 31, no. 4, pp. 2237–2248, 2021.
- [14] M. A. Rahman, A. Anjum, M. M. H. Milu, F. Khanam, M. S. Uddin, and M. N. Mollah, "Emotion recognition from eeg-based relative power spectral topography using convolutional neural network," *Array*, vol. 11, p. 100072, 2021.
- [15] Y. Ding, N. Robinson, C. Tong, Q. Zeng, and C. Guan, "Lggnnet: Learning from local-global-graph representations for brain–computer interface," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [16] C. Ieracitano, N. Mammone, A. Hussain, and F. C. Morabito, "A novel multi-modal machine learning based approach for automatic classification of eeg recordings in dementia," *Neural Networks*, vol. 123, pp. 176–190, 2020.
- [17] J. D. Power, A. L. Cohen, S. M. Nelson, G. S. Wig, K. A. Barnes, J. A. Church, A. C. Vogel, T. O. Laumann, F. M. Miezin, B. L. Schlaggar *et al.*, "Functional network organization of the human brain," *Neuron*, vol. 72, no. 4, pp. 665–678, 2011.
- [18] T. Song, W. Zheng, P. Song, and Z. Cui, "Eeg emotion recognition using dynamical graph convolutional neural networks," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 532–541, 2018.
- [19] P. Zhong, D. Wang, and C. Miao, "Eeg-based emotion recognition using regularized graph neural networks," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1290–1301, 2020.
- [20] D. Klepl, F. He, M. Wu, D. J. Blackburn, and P. Sarigiannis, "Eeg-based graph neural network classification of alzheimer's disease: An empirical evaluation of functional connectivity methods," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 2651–2660, 2022.
- [21] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *International conference on machine learning*. PMLR, 2018, pp. 4055–4064.
- [22] J. Xie, J. Zhang, J. Sun, Z. Ma, L. Qin, G. Li, H. Zhou, and Y. Zhan, "A transformer-based approach combining deep learning network and spatial-temporal information for raw eeg classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 2126–2136, 2022.
- [23] Y. Song, Q. Zheng, B. Liu, and X. Gao, "Eeg conformer: Convolutional transformer for eeg decoding and visualization," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 710–719, 2022.
- [24] S. Bagchi and D. R. Bathula, "Eeg-convtransformer for single-trial eeg-based visual stimulus classification," *Pattern Recognition*, vol. 129, p. 108757, 2022.
- [25] L. Lu, M. Han, G. Zou, L. Zheng, and J.-H. Gao, "Common and distinct neural representations of imagined and perceived speech," *Cerebral Cortex*, vol. 33, no. 10, pp. 6486–6493, 2023.
- [26] C. Fernyhough and A. M. Borghi, "Inner speech as language process and cognitive tool," *Trends in cognitive sciences*, 2023.
- [27] X. Tian, J. M. Zarate, and D. Poeppel, "Mental imagery of speech implicates two mechanisms of perceptual reactivation," *Cortex*, vol. 77, pp. 1–12, 2016.
- [28] L. Nalborczyk, M. Longcamp, M. Bonnard, V. Serveau, L. Spieser, and F.-X. Alario, "Distinct neural mechanisms support inner speaking and inner hearing," *Cortex*, vol. 169, pp. 161–173, 2023.
- [29] S. Kim, Y. E. Lee, S. H. Lee, and S. W. Lee, "Diff-e: Diffusion-based learning for decoding imagined speech eeg," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2023, 2023, pp. 1159–1163.
- [30] Y.-E. Lee, S.-H. Kim, S.-H. Lee, J.-S. Lee, S. Kim, and S.-W. Lee, "Speech synthesis from brain signals based on generative model," in *2023 11th International Winter Conference on Brain-Computer Interface (BCI)*. IEEE, 2023, pp. 1–4.
- [31] Z. Guo, M. Jiang, C. Liu, M. Wu, J. Lu, B. Gulyás, and C. Guan, "Enhancing eeg-based covert speech decoding through knowledge transfer," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [32] S.-H. Lee, Y.-E. Lee, S. Kim, B.-K. Ko, and S.-W. Lee, "Towards neural decoding of imagined speech based on spoken speech," in *2023 11th International Winter Conference on Brain-Computer Interface (BCI)*. IEEE, 2023, pp. 1–4.
- [33] J. A. Ramirez-Quintana, J. M. Macias-Macias, G. Ramirez-Alonso, M. I. Chacon-Murguia, and L. F. Corral-Martinez, "A novel deep capsule neural network for vowel imagery patterns from eeg signals," *Biomedical Signal Processing and Control*, vol. 81, p. 104500, 2023.
- [34] Y. Ding, S. Zhang, C. Tang, and C. Guan, "Masa-tcn: Multi-anchor space-aware temporal convolutional neural networks for continuous and discrete eeg emotion recognition," *IEEE Journal of Biomedical and Health Informatics*, 2024.

- [35] Y. Ding, C. Tong, S. Zhang, M. Jiang, Y. Li, K. L. J. Liang, and C. Guan, “Emt: A novel transformer for generalized cross-subject eeg emotion recognition,” *arXiv preprint arXiv:2406.18345*, 2024.
- [36] H. Kober, L. F. Barrett, J. Joseph, E. Bliss-Moreau, K. Lindquist, and T. D. Wager, “Functional grouping and cortical–subcortical interactions in emotion: a meta-analysis of neuroimaging studies,” *Neuroimage*, vol. 42, no. 2, pp. 998–1031, 2008.
- [37] J. J. Allen, P. M. Keune, M. Schönenberg, and R. Nusslock, “Frontal eeg alpha asymmetry and emotion: From neural underpinnings and methodological considerations to psychopathology and social cognition,” p. e13028, 2018.
- [38] L. Koessler, L. Maillard, A. Benhadid, J. P. Vignal, J. Felblinger, H. Vespiagnani, and M. Braun, “Automated cortical projection of eeg sensors: anatomical correlation via the international 10–10 system,” *Neuroimage*, vol. 46, no. 1, pp. 64–72, 2009.
- [39] M. Oliveri, G. Koch, and C. Caltagirone, “Spatial–temporal interactions in the human brain,” *Experimental Brain Research*, vol. 195, pp. 489–497, 2009.
- [40] D.-Y. Lee, M. Lee, and S.-W. Lee, “Decoding imagined speech based on deep metric learning for intuitive bci communication,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1363–1374, 2021.
- [41] K. Zhang, N. Robinson, S.-W. Lee, and C. Guan, “Adaptive transfer learning for eeg motor imagery classification with deep convolutional neural network,” *Neural Networks*, vol. 136, pp. 1–10, 2021.
- [42] 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.
- [43] R. Yang and E. Modesitt, “Vit2eeg: leveraging hybrid pretrained vision transformers for eeg data,” *arXiv preprint arXiv:2308.00454*, 2023.
- [44] Y. Song, X. Jia, L. Yang, and L. Xie, “Transformer-based spatial-temporal feature learning for eeg decoding,” *arXiv preprint arXiv:2106.11170*, 2021.
- [45] Y. Ding, Y. Li, H. Sun, R. Liu, C. Tong, C. Liu, X. Zhou, and C. Guan, “Eeg-deformer: A dense convolutional transformer for brain-computer interfaces,” *IEEE Journal of Biomedical and Health Informatics*, 2024.
- [46] Y. Ding, N. Robinson, S. Zhang, Q. Zeng, and C. Guan, “Tsception: Capturing temporal dynamics and spatial asymmetry from eeg for emotion recognition,” *IEEE Transactions on Affective Computing*, 2022.
- [47] L. D. Brown, T. T. Cai, and A. DasGupta, “Interval estimation for a binomial proportion,” *Statistical science*, vol. 16, no. 2, pp. 101–133, 2001.
- [48] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.
- [49] W. Zhang, M. Jiang, K. A. C. Teo, R. Bhuvanakantham, L. Fong, W. K. J. Sim, Z. Guo, C. H. V. Foo, R. H. J. Chua, P. Padmanabhan, V. Leong, J. Lu, B. Gulyás, and C. Guan, “Revealing the spatiotemporal brain dynamics of covert speech compared with overt speech: A simultaneous eeg-fmri study,” *NeuroImage*, vol. 293, p. 120629, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1053811924001241>
- [50] P. C. Vijn, B. W. van Dijk, and H. Spekreijse, “Topography of occipital eeg-reduction upon visual stimulation,” *Brain Topography*, vol. 5, pp. 177–181, 1992.
- [51] S. C. Blank, S. K. Scott, K. Murphy, E. Warburton, and R. J. Wise, “Speech production: Wernicke, broca and beyond,” *Brain*, vol. 125, no. 8, pp. 1829–1838, 2002.
- [52] G. Nasios, E. Dardiotis, L. Messinis *et al.*, “From broca and wernicke to the neuromodulation era: insights of brain language networks for neurorehabilitation,” *Behavioural neurology*, vol. 2019, 2019.
- [53] G.-J. Rutten, “Broca-wernicke theories: A historical perspective,” *Handbook of Clinical Neurology*, vol. 185, pp. 25–34, 2022.
- [54] C. Code, “Can the right hemisphere speak?” *Brain and Language*, vol. 57, no. 1, pp. 38–59, 1997.
- [55] S. Luthra, “The role of the right hemisphere in processing phonetic variability between talkers,” *Neurobiology of Language*, vol. 2, no. 1, pp. 138–151, 2021.
- [56] S. Boklina and A. Batalov, “Improvement of speech function in patients with aphasia: the right hemisphere, an enemy or a friend?” *Human Physiology*, vol. 44, pp. 161–169, 2018.
- [57] A. N. LaCroix, N. Blumenstein, M. Tully, L. C. Baxter, and C. Rogalsky, “Effects of prosody on the cognitive and neural resources supporting sentence comprehension: A behavioral and lesion-symptom mapping study,” *Brain and language*, vol. 203, p. 104756, 2020.
- [58] J. I. Skipper, S. Goldin-Meadow, H. C. Nusbaum, and S. L. Small, “Speech-associated gestures, broca’s area, and the human mirror system,” *Brain and language*, vol. 101, no. 3, pp. 260–277, 2007.
- [59] K. M. Hartikainen, “Emotion-attention interaction in the right hemisphere,” *Brain sciences*, vol. 11, no. 8, p. 1006, 2021.
- [60] B. C. Preisig and M. Meyer, “Predictive coding and dimension-selective attention enhance the lateralization of spoken language processing,” *Neuroscience & Biobehavioral Reviews*, p. 106111, 2025.
- [61] K. Grill-Spector, R. Henson, and A. Martin, “Repetition and the brain: neural models of stimulus-specific effects,” *Trends in cognitive sciences*, vol. 10, no. 1, pp. 14–23, 2006.
- [62] P. Gagnepain, G. Chételat, B. Landau, J. Dayan, F. Eustache, and K. Lebreton, “Spoken word memory traces within the human auditory cortex revealed by repetition priming and functional magnetic resonance imaging,” *Journal of Neuroscience*, vol. 28, no. 20, pp. 5281–5289, 2008.
- [63] “2020 international bci competition,” <https://osf.io/pq7vb/>, 2020, accessed: Apr. 2020.
- [64] N. Nieto, V. Peterson, H. L. Rufiner, J. E. Kamienkowski, and R. Spies, “Thinking out loud, an open-access eeg-based bci dataset for inner speech recognition,” *Scientific Data*, vol. 9, no. 1, p. 52, 2022.
- [65] G. A. P. Coretto, I. E. Gareis, and H. L. Rufiner, “Open access database of eeg signals recorded during imagined speech,” in *12th International Symposium on Medical Information Processing and Analysis*, vol. 10160. SPIE, 2017, p. 1016002.

## SUPPLEMENTARY MATERIALS

### A. Comparison with Public Datasets

As shown in Table IV, the distinctiveness of our dataset lies in the following key aspects when compared to existing freely available datasets:

**Subject Count:** Our dataset includes data from 57 subjects, which is notably larger than the referenced datasets ([63]–[65]). A larger subject pool improves the generalizability of the findings and allows for more robust statistical analyses, particularly in exploring inter-subject variability.

**Repetition Count Per Subject:** Each participant in our dataset completed 1,000 utterances, providing a higher volume of data per subject compared to existing datasets ([63]–[65]). This high repetition count enables better training and validation of machine learning models, as it provides a denser dataset for both individual-specific and generalized decoding tasks.

**Multi-utterances:** We chose to collect a multi-utterance dataset over a single-utterance dataset to ensure more stable and robust neural responses.

### B. Modeling Details

In the implementation of FAST on our self-collected dataset, we used  $L_t = 4$ ,  $M = 8$  and  $F = 32$ , for the  $F_i$  feature vector, its dimension will be  $8 \times 32 = 256$ , and  $L = 4$  for the transformer layers. The AdamW optimizer was employed with an initial learning rate of 0.001. The learning rate schedule follows a cosine decay pattern with a warm-up phase. During the first 10 epochs, the learning rate gradually increases from 10% to 100%, enabling a smooth warm-up period. Following this, the learning rate decays from 100% to 10% over the remaining training epochs. To accelerate the pre-training phase, FP16 automatic mixed precision training was used, conserving memory and enhancing GPU throughput. For the fine-tuning phase, FP32 precision was utilized without mixed precision. We use two A100 GPUs, the experiment can be completed in approximately 5 hours.

### C. Metrics for Evaluation

$$\text{Accuracy: } \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision (Macro): } \frac{1}{C} \sum_{i=1}^C \frac{TP_i}{TP_i + FP_i}$$

$$\text{Recall (Macro): } \frac{1}{C} \sum_{i=1}^C \frac{TP_i}{TP_i + FN_i}$$

$$\text{F1-score (Macro): } \frac{1}{C} \sum_{i=1}^C \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

$$\text{Cohen's Kappa: } \kappa = \frac{p_o - p_e}{1 - p_e}$$

$$p_o = \frac{1}{N} \sum_{i=1}^N \delta(y_i, \hat{y}_i), \quad p_e = \sum_{k=1}^C \left( \frac{n_k}{N} \cdot \frac{m_k}{N} \right)$$

Where  $p_o$  is the observed agreement,  $p_e$  is the expected agreement,  $n_k$  and  $m_k$  are the actual and predicted counts for class  $k$ , respectively.

$$\text{AUC (One-vs-Rest): } \text{AUC} = \frac{1}{C} \sum_{i=1}^C \text{AUC}_i$$

Where  $\text{AUC}_i$  is the area under the ROC curve for class  $i$ , computed using one-vs-rest strategy.

### Explanation of Symbols:

- $C$ : Number of classes in the classification task.
- $\delta(y_i, \hat{y}_i)$ : Indicator function that is 1 if  $y_i = \hat{y}_i$  (correct prediction) and 0 otherwise.
- $N$ : Total number of samples in the dataset.
- $n_k$ : Actual number of samples belonging to class  $k$ .
- $m_k$ : Predicted number of samples assigned to class  $k$ .
- $p_o$ : Observed agreement - the proportion of times the predicted label matches the true label.
- $p_e$ : Expected agreement - the agreement expected by chance, based on class distributions.
- $\text{AUC}_i$ : The area under the ROC curve for class  $i$ , computed using the one-vs-rest strategy.

### D. Visualization Details

To interpret model predictions, the Integrated Gradients (IG) method [48] is employed. This method assigns an importance score to each input feature by computing the path integral of the gradients of the model’s output with respect to the input, along a straight line path from a pre-defined baseline to the actual input. The IG method satisfies two key axioms: Sensitivity, which ensures that features contributing differently to the output are assigned different importance scores, and Implementation Invariance, which guarantees that two functionally equivalent models yield the same importance scores. The IG computation is formalized as:

$$IG_i(x) = (x_i - x'_i) \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$

where  $x$  is the input,  $x'$  is the baseline,  $F$  is the model’s output function, and  $i$  denotes the  $i$ -th feature. This method provides an intuitive understanding of the contribution of each input feature to the model’s prediction, as illustrated in Figure 6.

### E. Results on Overt Speech Decoding

The decoding accuracy for overt speech is higher compared to covert speech, as seen in Table V. This difference can be attributed to two key factors. First, overt speech likely produces stronger and more consistent neural activity due to the physical act of speaking, which enhances the reliability of the decoding process. Second, the muscle movements involved in overt speech may generate electromyographic signals, providing additional information that aids in classification.

### F. Ablation on Brain Areas Partitioning Configurations

In addition to the M=8 in the manuscript, here we add the following configurations: (a) M=5 by merging the electrodes from prefrontal into frontal, precentral, and postcentral into

TABLE IV  
COMPARISON OF DATASETS BASED ON SUBJECT COUNT, REPETITIONS, AND UTTERANCES PER TRIAL

Dataset	Classes	Subjects	Repetitions per Subject	Utterances per Trial
BCI Competition (Imagined Speech Classification) [63]	5	15	350	1
Thinking Out Loud Dataset [64]	5	10	223 (average)	1
Croetto DB [65]	6	15	240 (average)	1
Ours	5	57	1,000	5

TABLE V  
PERFORMANCE COMPARISONS COVERT/OVERT SPEECH DECODING FOR 5 UTTERANCES

Stage	Condition	Accuracy	F1-score	Cohen-Kappa	AUC
Pretrain	Covert	$26.9 \pm 6.3$	$0.221 \pm 0.091$	$0.087 \pm 0.086$	$0.627 \pm 0.083$
	Overt	$30.1 \pm 8.8$	$0.234 \pm 0.113$	$0.126 \pm 0.111$	$0.681 \pm 0.096$
Finetune	Covert	$34.7 \pm 10.7$	$0.340 \pm 0.108$	$0.184 \pm 0.134$	$0.662 \pm 0.097$
	Overt	$51.3 \pm 12.9$	$0.523 \pm 0.104$	$0.496 \pm 0.089$	$0.721 \pm 0.058$

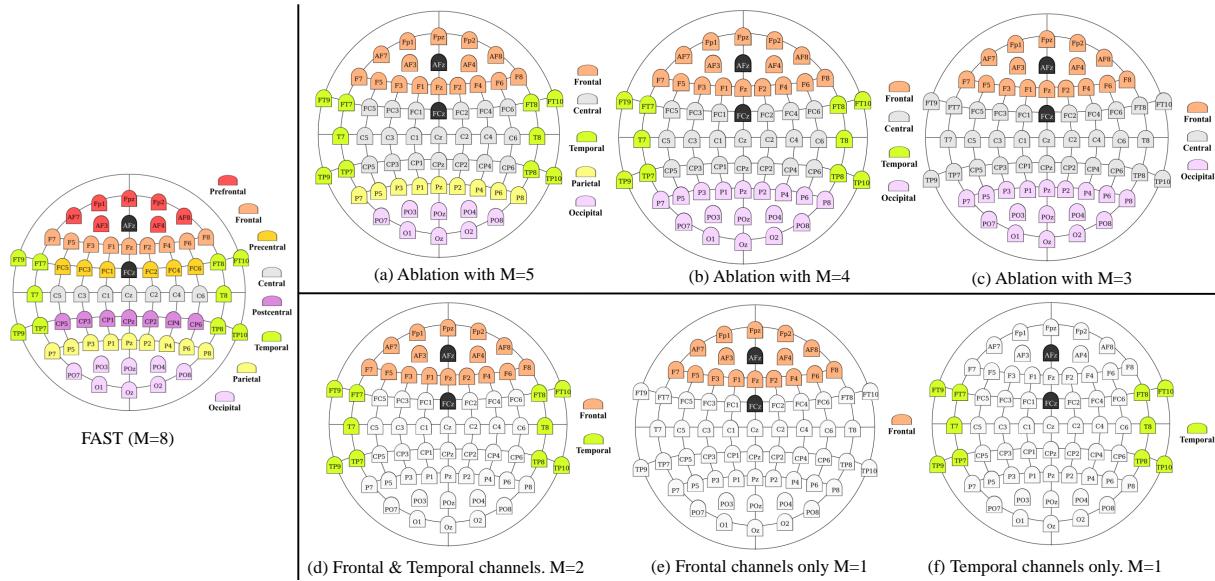


Fig. 9. illustration of different M and configurations. (a) M=5 by merging the electrodes from the prefrontal into frontal, precentral, and postcentral central; (b) M=4 by additionally merging parietal into occipital; (c) M=3 by only keeping frontal, central, and the occipital. The performance is shown in the Table VI (a)(b)(c). The performance metrics generally decrease as the number of regions M decreases from FAST to conditions (a)(b) and (c).

central; (b) M=4 by additionally merging parietal into occipital; (c) M=3 by only keeping the frontal, central and the occipital. The performance is shown in the Table VI (a)(b)(c). The performance metrics generally decrease as the number of regions M decreases from FAST to conditions (a)(b) and (c).

To examine the impact of using only speech-related electrodes, three additional ablation studies were conducted: (d) using both Frontal and Temporal electrodes, (e) using Frontal electrodes only, and (f) using Temporal electrodes only. As shown in Table VI (d)(e)(f), the baseline FAST model (M=8) outperforms all ablated configurations in both pre-train and finetune stages. Among the ablations, temporal electrodes were shown more impactful during finetuning as shown in (c).

#### G. Additional Public Dataset

In addition to the self-collected dataset, to validate the effectiveness of FAST, we included the publicly available multi-class imagined speech classification dataset from the 2020 International BCI Competition [63], this dataset consists

of EEG recordings from 15 subjects, aged 20-30 years old. The subjects were instructed to imagine the silent pronunciation of five conversational words/phrases: 'hello,' 'help me,' 'stop,' 'thank you,' and 'yes.' given the word as if they were performing real speech, without moving any articulators or making the sound. Each subject participated in 70 trials per class, yielding a total of 350 trials (60 trials per class for training and 10 trials per class for validation). EEG signals were recorded from 64 electrodes following the 10-20 international system. Ground and reference electrodes were placed at Fpz and FCz.

For this dataset, the limited subject count (15 subjects) led to subject-independent pre-training being ineffective for both FAST and baseline models. Therefore, we proceeded directly with subject-dependent training, performing a 5-fold cross-validation on each subject's 350 trials.

#### H. Results on public dataset

Table VII provides the classification accuracies and standard deviations across various models on the BCI competition

TABLE VI  
PERFORMANCE COMPARISONS WITH DIFFERENT M AND CONFIGURATIONS

Stage	Model	Accuracy	F1-score	Cohen-Kappa	AUC
Pretrain	FAST (M=8)	26.9 ± 6.9	0.221 ± 0.091	0.087 ± 0.086	0.627 ± 0.083
	(a) M=5	25.5 ± 5.2	0.163 ± 0.067	0.069 ± 0.065	0.633 ± 0.073
	(b) M=4	25.6 ± 5.5	0.168 ± 0.065	0.071 ± 0.068	0.634 ± 0.076
	(c) M=3	25.9 ± 5.6	0.178 ± 0.074	0.074 ± 0.070	0.637 ± 0.074
	(d) M=2 Frontal + Temporal	26.9 ± 5.5	0.195 ± 0.072	0.087 ± 0.069	0.638 ± 0.072
	(e) M=1 Frontal Only	26.4 ± 5.1	0.180 ± 0.068	0.080 ± 0.064	0.637 ± 0.071
	(f) M=1 Temporal Only	27.4 ± 6.4	0.215 ± 0.079	0.092 ± 0.080	0.642 ± 0.072
	FAST (M=8)	34.7 ± 10.7	0.340 ± 0.108	0.184 ± 0.134	0.662 ± 0.097
	(a) M=5	33.1 ± 7.0	0.291 ± 0.073	0.164 ± 0.087	0.645 ± 0.080
Finetune	(b) M=4	32.7 ± 7.8	0.288 ± 0.082	0.159 ± 0.097	0.643 ± 0.081
	(c) M=3	33.4 ± 7.2	0.293 ± 0.072	0.167 ± 0.090	0.652 ± 0.077
	(d) M=2 Frontal + Temporal	33.7 ± 8.1	0.299 ± 0.083	0.172 ± 0.102	0.652 ± 0.083
	(e) M=1 Frontal Only	33.8 ± 8.4	0.289 ± 0.087	0.172 ± 0.105	0.649 ± 0.078
	(f) M=1 Temporal Only	34.2 ± 8.7	0.316 ± 0.087	0.178 ± 0.109	0.651 ± 0.078

TABLE VII  
PERFORMANCE COMPARISONS ON 2020 INTERNATIONAL BCI  
COMPETITION DATASET, TRACK 3: IMAGINED SPEECH CLASSIFICATION

Method	Accuracy (1 utterance)
LSTM	36.5 ± 4.8 ***
GRU	38.3 ± 4.7 ***
DeepConvNet [10]	33.2 ± 5.8 ***
EEGNet [11]	40.6 ± 6.2 ***
EEG-Conformer [23]	43.3 ± 5.1 ***
Dong-Yeon et al. [40]	48.1 ± 3.6 (NA)
TSception [46]	52.2 ± 8.8 ***
FAST	<b>54.8 ± 9.1</b>

Note \*\*\* denoting a  $p$ -value less than 0.001, \*\* for less than 0.01, and \* for less than 0.05

dataset, the dataset only contains a single utterance per trial. The table also presents the results of the Wilcoxon signed-rank test, with significant p-values marked by asterisks. The individualized accuracy for [40] is unavailable and is indicated as (NA). FAST achieved the highest average accuracy at 54.8% with a standard deviation of 9.1%. Among the models evaluated, EEG-Conformer, [40], and TSception also demonstrated competitive performance, achieving average accuracies of 43.38%, 48.10%, and 52.26% respectively. The FAST model outperformed all other approaches.