Decoding Covert Speech From EEG by Functional Areas Spatio-Temporal Transformer

Muyun Jiang, Wei Zhang, Yi Ding, *Member, IEEE*, Kok Ann Colin Teo, LaiGuan Fong, Shuailei Zhang, *Member, IEEE*, 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, and 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.**

***Index Terms*—Brain-computer interface (BCI), eeg signal analysis, covert speech decoding.**

Received 2 August 2024; revised 24 December 2024, 2 April 2025, 23 July 2025, and 2 September 2025; accepted 19 November 2025. Date of publication 12 January 2026; date of current version 5 June 2026. The work of Victoria Leong was supported by the Ministry of Education, Singapore, Social Science & Humanities Research Fellowship under Grant MOE2020-SSHR-008. This work was supported in part by DSO National Laboratories under Grant DSOCL21193, Singapore and in part by RIE2025 Human Potential Programme Prenatal/Early Childhood under Grant H22P0M0002, *(Corresponding author: Cuntai Guan.)*

Muyun Jiang, Wei Zhang, Yi Ding, Shuailei Zhang, Zhiwei Guo, and Chenyu Liu are with the College of Computing and Data Science, Nanyang Technological University (NTU), Singapore 639798.

Kok Ann Colin Teo is with the Cognitive Neuroimaging Centre and LKC School of Medicine, Nanyang Technological University, Singapore 639798, and also with the Division of Neurosurgery, National University Health System, Singapore 639798.

LaiGuan Fong, Raghavan Bhuvanakantham, and Parasuraman Padmanabhan are with the Cognitive Neuroimaging Centre and LKC School of Medicine, Nanyang Technological University, Singapore 639798.

Wei Khang Jeremy Sim is with the Cognitive Neuroimaging Centre and LKC School of Medicine, Nanyang Technological University, Singapore 639798, and also with IGP-Neuroscience, Interdisciplinary Graduate Programme, Nanyang Technological University, Singapore 639798.

Chuan Huat Vince Foo and Rong Hui Jonathan Chua are with DSO National Laboratories, Singapore 118230.

Victoria Leong is with the Division of Psychology, Nanyang Technological University, Singapore 639798, and also with the Department of Pediatrics, University of Cambridge, CB2 1TN Cambridge, U.K.

Jia Lu is with DSO National Laboratories, Singapore 118230, and with the Yong Loo Lin School of Medicine, National University of Singapore, Singapore 639798, and also with the Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore 639798.

Balázs Gulyás is with the Cognitive Neuroimaging Centre and LKC School of Medicine, Nanyang Technological University, Singapore 639798, and also with the Hungarian Research Network, HUN-REN, 4001 Debrecen, Hungary.

Cuntai Guan is with the College of Computing and Data Science, Nanyang Technological University (NTU), Singapore 639798, and also with the Center for AI in Medicine, LKC School of Medicine, Singapore 308232 (e-mail: ctguan@ntu.edu.sg).

The code for this work has been made public at https://github.com/Jiang-Muyun/FAST.

Data is available on-line at https://github.com/Jiang-Muyun/FAST/blob/main/supplementary.pdf.

Digital Object Identifier 10.1109/JBHI.2026.3653025

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 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]. [7] presented the Chinese Imagined Speech Corpus (Chisco), which contains over 20,000 imagined-speech sentences recorded with high-density EEG. Each subject contributed more than 900 minutes of data, making it the largest per-individual neural language decoding dataset to date. However, the task of decoding covert speech from non-invasive 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, [8] demonstrated that the finer details captured by these recordings play a crucial role in understanding

and interpreting speech-related neural activity. Additionally, [9] has shown that imagined speech could be decoded from low and cross-frequency features in intracranial electroencephalography (iEEG) signals. [10] showed that a transformer model trained on overt-speech ECoG can decode covert-speech sentences with similar accuracy, indicating that the transfer from overt to covert can compensate for the scarcity of covert-speech data. Non-invasive technologies have gradually garnered attention for their potential in BCIs. Notably, the use of Magnetoencephalography, as explored in [11], 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 [12] and EEGNet [13], the capability to accurately interpret EEG data has markedly improved in the tasks of motor imagery, emotional cognition, and more [14], [15], [16], [17], [18], [19], [20], [21]. Graph-based methods are proven effective when learn spatial dependency among brain areas. LGG-Net [22] explicitly models EEG spatial topology by constructing fixed region-level graph structures and applying graph convolutions to capture inter-regional dependencies. EEG-Deformer [23] learns spatial dependencies implicitly through deformable attention, allowing the model to dynamically select informative electrode neighborhoods without relying on predefined graph structures. 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 [24], [25], [26], [27]. However, despite these advancements, current transformer-based methods often lack interpretability. This limitation highlights the need for models that not only capture complex spatiotemporal patterns but also offer explainable and generalizable representations of covert speech tasks.

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 an effective decoding algorithm to tackle challenging covert speech EEG signals.
3) Lack of understanding of the covert speech intrinsic information in the 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. 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) Conducted in-depth feature analysis of temporal activation and spatial distribution, revealing discriminative event-related features in frontal and temporal regions during covert speech.
3) Collected a large-scale multi-utterance dataset comprising 57 subjects, each contributing 1,000 covert and overt speech recordings, and demonstrated that FAST achieves significant accuracy improvements over baseline methods.

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 expressed 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 of Covert Speech

Recent studies provide converging neural evidence that covert speech engages both articulatory and auditory systems [28]. Activation in Broca's area and supplementary motor areas indicates that even silent, internal speech relies on motor planning and articulatory representations. Simultaneously, the involvement of auditory imagery networks suggests that individuals "hear" their internal speech through neural pathways similar to those used in external auditory processing [29], [30]. Additionally, research has shown that covert speech and listening to speech, while engaging overlapping brain regions, activate distinct neural signatures [31]. This distinction highlights the different processing demands of internal speech versus external. Beyond the classical language network, covert speech also recruits the executive

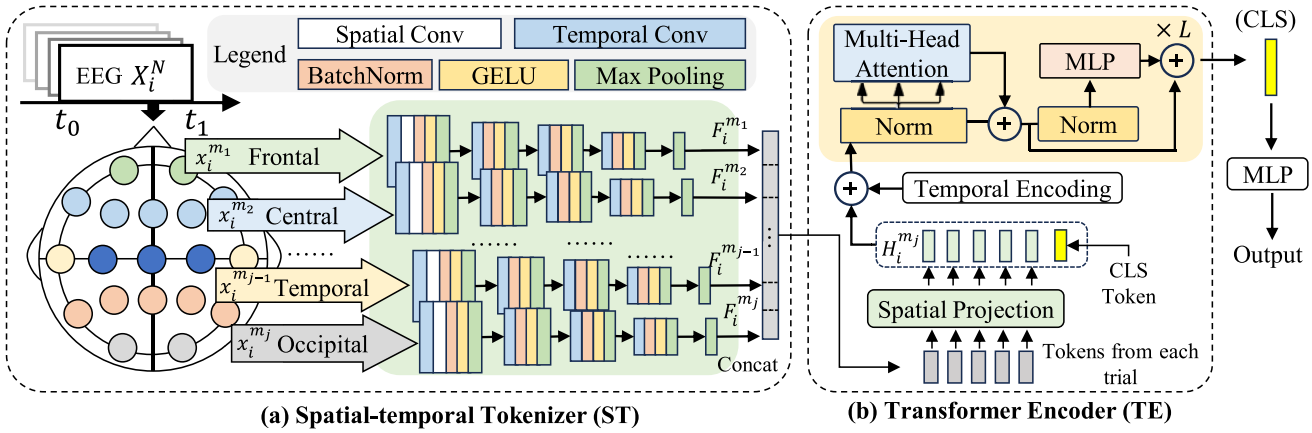

Fig. 1. Overview of proposed FAST. (a) The Spatial-temporal Tokenizer (ST) block illustrates the initial processing of EEG data through spatial and temporal convolutional layers. (b) The Transformer Encoder (TE) block shows the transformer architecture used for tokens generated by ST, which outputs a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

control network, including regions such as the dorsolateral prefrontal cortex (DLPFC) and anterior cingulate cortex (ACC), reflecting its reliance on attention control and working memory processes [32], [33].

Together, these findings suggest that covert speech is a complex, distributed process occurring across distinct spatiotemporal scales. These studies highlight the need for spatiotemporal feature extraction techniques that can capture both the sequential (temporal) and distributed (spatial) neural patterns underlying covert speech, a key research gap that the proposed model aims to address.

B. Decoding Advancements of Covert Speech

Recent advancements in neural decoding of imagined speech have significantly contributed to our understanding of brain processes. For instance, one study [34] 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 [35] investigates speech synthesis from brain signals using a generative model, paving the way for neural activity-based communication devices. Additionally, [36] explores transfer learning from overt to imagined speech through a convolutional autoencoder, improving EEG-based imagined speech decoding. A separate framework [37] integrates spoken speech data to refine neural decoding models, providing a more direct interpretation of internal speech representations. Moreover, [38] introduces a deep capsule neural network to decode vowel imagery from EEG, improving the efficiency of imagined speech processing. Lastly, [39] introduces a new 120-hour EEG silent speech dataset and proposes a Large Brain Language Model (LBLM) with a novel Future Spectro-Temporal Prediction (FSTP) pretraining paradigm, achieving 39.6% in word-level classification accuracy in silent speech decoding.

Despite progress, many studies lack detailed insights into the neural mechanisms of covert speech and provide limited temporal visualizations of neural features. Clearer exploration of neural substrates could improve understanding and BCI design. Additionally, the tokenization process is often not well visualized, making learned temporal tokens harder to interpret and transfer. How these tokens interact with spatial features also remains underexplored. Investigating this interplay could offer deeper insights into spatiotemporal brain dynamics and enhance model interpretability.

III. Methodology

In this section, we introduce the model structure of FAST. As depicted in Fig. 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 sliding 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 [18], [40], [41]. EEG electrodes are positioned on the scalp following the 10-10 [42] 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

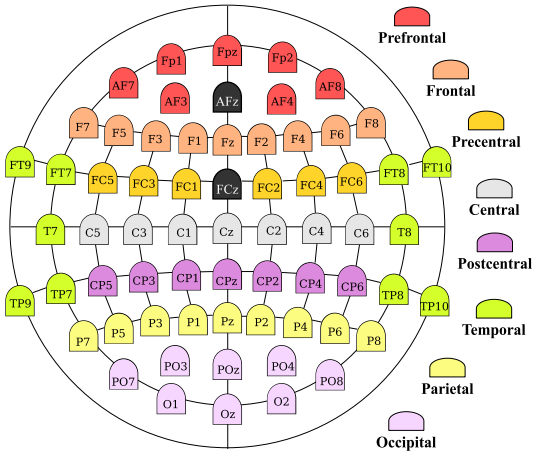

Fig. 2. The divided brain functional regions are based on the spatial locations of EEG channels, where FCz serves as the reference electrode, and AFz serves as the ground electrode.

according to the spatial locations of EEG channels, similar to LGGNet [22], as illustrated in Fig. 2.

This division allows us to examine the specific roles of various brain regions, particularly focusing on the functional responsibilities of the frontal [41] and temporal regions [43]. 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 [22].

Mathematically, the division can be represented as a partitioning process

$$P : X_i^N \rightarrow \{x_i^{m_1}, x_i^{m_2}, \ldots, x_i^{m_j}\} \tag{1}$$

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 Conv_T and Conv_S. This sequential processing is represented as follows:

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

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

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:

$$x_{i(1)}^{m_j} = \text{MaxPool}\left(\text{GELU}\left(\text{BN}\left(x_{i,\text{spatial}}^{m_j}\right)\right)\right) \tag{4}$$

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:

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

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:

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

$F_i^{m_j}$ 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

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

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 function Multi-Head$(\cdot)$. After each multi-head attention layer, a position-wise feed-forward network (FFN) is added to form a transformer layer. 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 $j \in \{1, \ldots, J\}$ 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:

$$G_i = \bigoplus_{j=1}^{J} H_{i(L_{sp})}^{m_j} \tag{8}$$

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

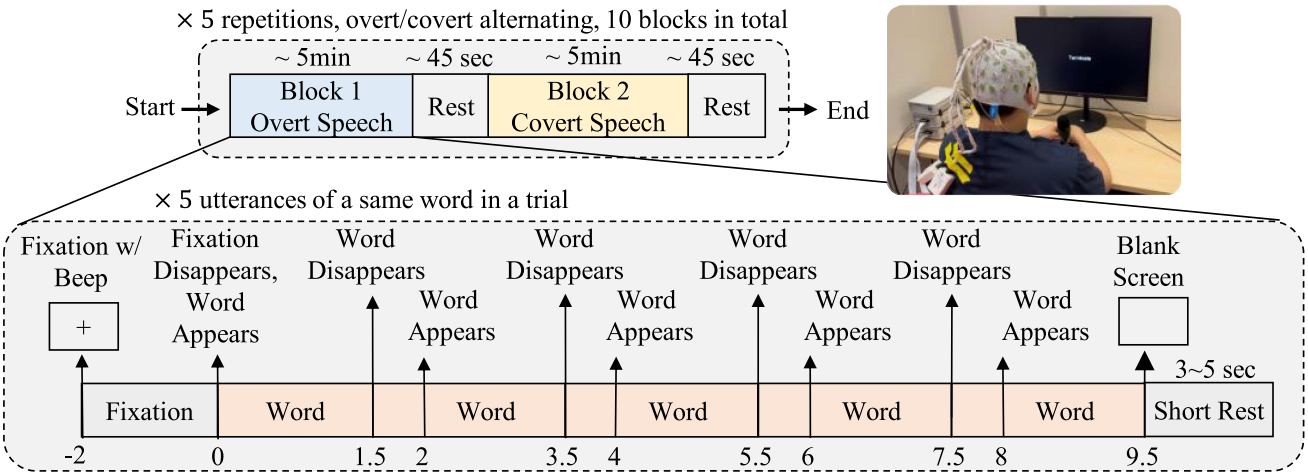

Fig. 3. Protocol of the experiment. (Top) The experiment is structured into blocks. Each subject will complete 10 blocks, alternating between overt and covert EEG experiments. Each block consists of 20 trials, presenting five words in a pseudo-random order. (Bottom) In each trial, the same words were 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.

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"

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:

$$G = \{G_1 + T_1, G_2 + T_2, \ldots, G_i + T_i, F_{\text{cls}} + T_{i+1}\}, \quad (9)$$

The transformer operates iteratively as depicted:

$$G^L = \text{Transformer}_l(G^{l-1}) \quad (10)$$

for $l \in \{1, \ldots, L\}$. 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^L_{cls}, is derived after multiple layers of attention and subsequently processed through a classification head for predictions:

$$\hat{y} = \text{MLP}(G^L_{\text{cls}}) \quad (11)$$

where $\hat{y}$ represents the predicted probabilities, with a shape of (B,5) corresponding to the 5 classes in our classification task.

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 ± 3.6 years. Information about 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, 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

TABLE II
PERFORMANCE COMPARISONS IN THE PRE-TRAIN AND FINE-TUNE PHASE FOR COVERT SPEECH IN 5 UTTERANCES OF THE SAME WORD. THE 5-CLASS CHANCE LEVEL IS 20%, WITH A 95% BINOMIAL CONFIDENCE INTERVAL OF [18.96%, 21.04%].NOTE *** DENOTES A p-VALUE LESS THAN 0.001, ** FOR LESS THAN 0.01, AND * FOR LESS THAN 0.05.

Stage	Model	Accuracy(%)	F1-score	Cohen-Kappa	AUC
Pretrain	BIOT [46]	21.8 ± 4.6 ***	0.187 ± 0.048 *	0.023 ± 0.058 ***	0.508 ± 0.017 ***
	EEGViT [47]	23.2 ± 4.9 **	0.221 ± 0.048	0.039 ± 0.062 **	0.535 ± 0.045 ***
	DCN [12]	25.0 ± 5.7	0.240 ± 0.057	0.063 ± 0.071	0.561 ± 0.063 ***
	EEGNet [13]	25.1 ± 5.8	0.238 ± 0.060	0.064 ± 0.072	0.570 ± 0.052 ***
	ST-Transformer [48]	24.6 ± 5.1 *	0.228 ± 0.060	0.057 ± 0.064 *	0.557 ± 0.048 ***
	EEG-Conformer [26]	26.1 ± 6.3	0.189 ± 0.087	0.077 ± 0.079	0.608 ± 0.083
	EEG-Deformer [23]	26.2 ± 6.4	$\mathbf{0.241 \pm 0.070}$	0.077 ± 0.080	0.591 ± 0.061 *
	TSception [49]	25.6 ± 6.2	0.187 ± 0.074 *	0.070 ± 0.078	0.618 ± 0.075
	FAST	$\mathbf{26.9 \pm 6.9}$	0.221 ± 0.091	$\mathbf{0.087 \pm 0.086}$	$\mathbf{0.627 \pm 0.083}$
Finetune	BIOT [46]	24.0 ± 6.5 ***	0.234 ± 0.063 ***	0.050 ± 0.081 ***	0.530 ± 0.041 ***
	EEGViT [47]	24.6 ± 5.6 ***	0.244 ± 0.056 ***	0.057 ± 0.070 ***	0.558 ± 0.059 ***
	DCN [12]	26.5 ± 6.5 ***	0.264 ± 0.066 ***	0.081 ± 0.082 ***	0.577 ± 0.072 ***
	EEGNet [13]	27.2 ± 6.5 ***	0.271 ± 0.064 ***	0.089 ± 0.081 ***	0.589 ± 0.064 ***
	ST-Transformer [48]	27.2 ± 6.4 ***	0.269 ± 0.063 ***	0.090 ± 0.080 ***	0.584 ± 0.059 ***
	EEG-Conformer [26]	28.6 ± 7.0 ***	0.221 ± 0.047 ***	0.107 ± 0.087 ***	0.604 ± 0.082 ***
	EEG-Deformer [23]	30.1 ± 7.9 *	0.290 ± 0.082 **	0.126 ± 0.099 *	0.615 ± 0.076 **
	TSception [49]	31.2 ± 8.3	0.287 ± 0.076 **	0.140 ± 0.103	0.623 ± 0.090 *
	FAST	$\mathbf{34.7 \pm 10.7}$	$\mathbf{0.340 \pm 0.108}$	$\mathbf{0.184 \pm 0.134}$	$\mathbf{0.662 \pm 0.097}$

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 [22], [44]. 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. The 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 independence from the excluded subject. During fine-tuning, a leave-one-block-out (LOBO) [45] 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, as presented 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-utterance cases in Table II. We evaluate FAST against a list of commonly used models, including ST-Transformer [48], BIOT [46], EEGViT [47], EEGNet [13], DCN [12], EEG-Conformer [26], and TSception [49], as well as more recent architectures such as EEG-Deformer [23]. 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 Fig. 4. The chance level was determined using a binomial distribution model [50] 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 deviations for FAST and the comparative baseline models for 5-utterances covert speech classification. Wilcoxon signed-rank

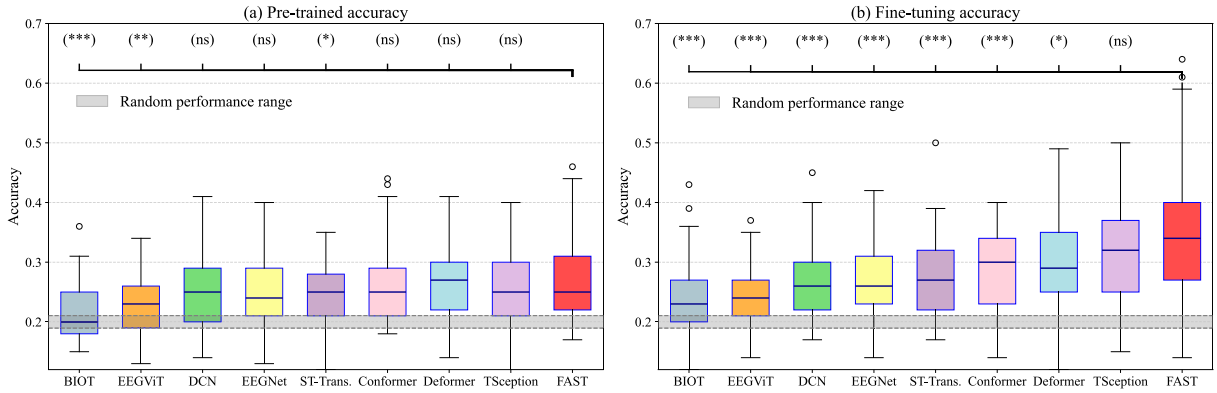

Fig. 4. Box plot illustrating the accuracy of covert speech recognition for all subjects ordered by the median accuracy: (a) Accuracy from pre-trained models; (b) Accuracy after fine-tuning. Significance levels are compared between FAST and each of the baselines, indicated with asterisks (*), where (ns)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

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, corresponding to a 7.8% improvement. Similarly, baseline models such as EEG-Deformer, DCN, EEGNet, and Conformer also show notable improvements. FAST achieves the highest accuracy of 34.7%, outperforming the closest baseline (TSception, 31.2%) by 3.5%. 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).

Fig. 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. The aim is to gain a deeper understanding of the temporal and spatial activation patterns associated with covert speech processing across different brain regions. Showing the features learned by the model is crucial for interpreting and validating its behaviour in the context of covert speech decoding. By visualizing these features, we can assess whether the model captures meaningful spatiotemporal patterns that align with established neuroscientific knowledge.

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 a 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 Fig. 5(a). The normalized activation per class shown in Fig. 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 Fig. 5(c).

To interpret model predictions, the Integrated Gradients [51] method is employed, which assigns an importance score to each input feature by approximating the integral of the model's output gradients shown in Fig. 6. A detailed algorithm for generating this figure is described in the supplementary material.

The activation map in Fig. 5(a) reveals task-related responses following both the onset and offset of word cues, prominently localized in the frontal and temporal regions that play key roles in speech processing, similar to [52]. A stronger response is observed at the onset of the first cue, with temporal lobe activation showing a distinct pattern compared to other regions. Fig. 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, occipital lobe activation aligns with

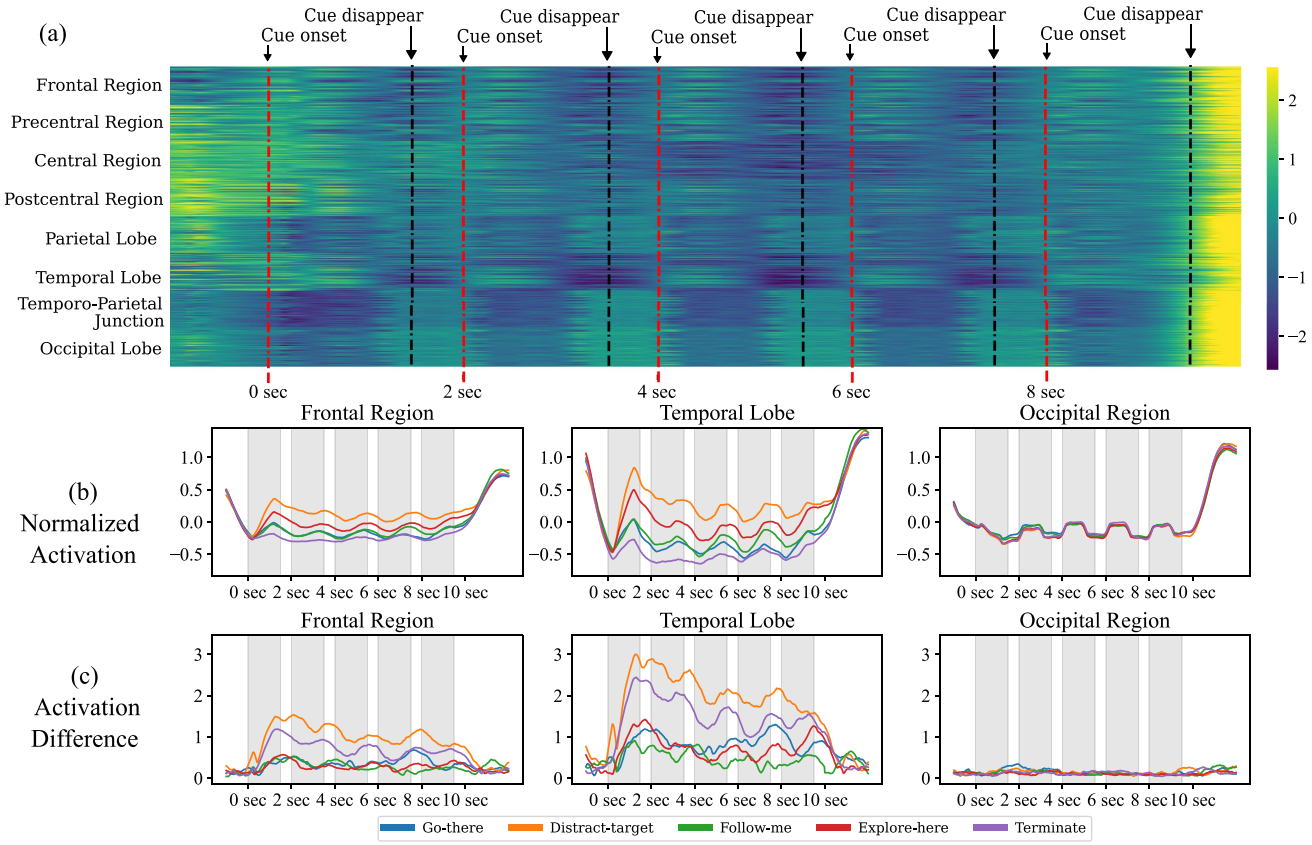

Fig. 5. Feature visualization of the ST on covert speech EEG data averaged across all subjects. The features were extracted from the leave-out subject after each round of leave-one-subject-out training and z-score normalized to highlight relative temporal fluctuations. (a) A heatmap illustrates the features 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.

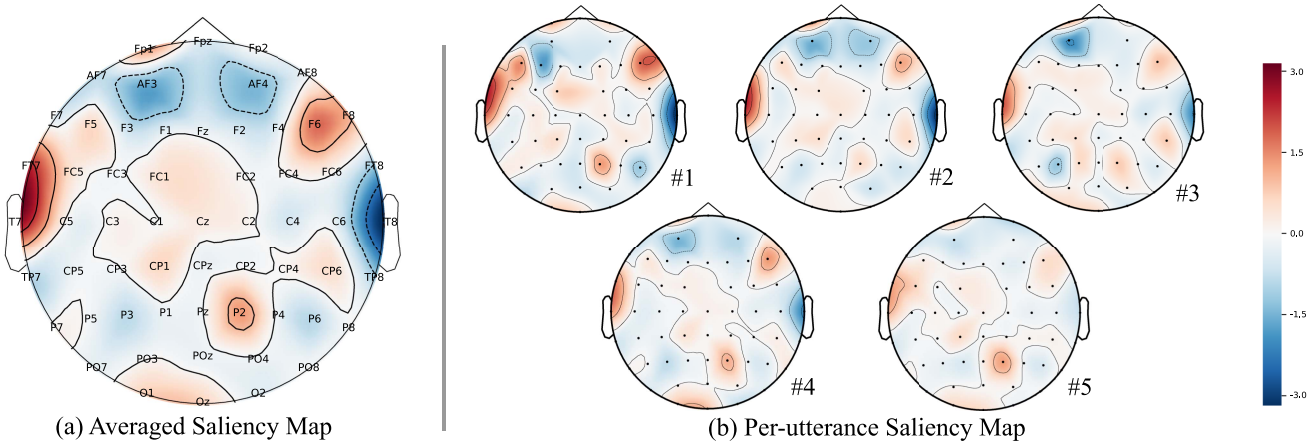

Fig. 6. EEG saliency maps during covert speech, with color scale representing the normalized integrated gradients importance. (a): Activation is prominently observed in the left hemisphere electrodes (T7, FT7, FC5, C5, F5), corresponding primarily to Broca's area, suggesting involvement in speech motor 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.

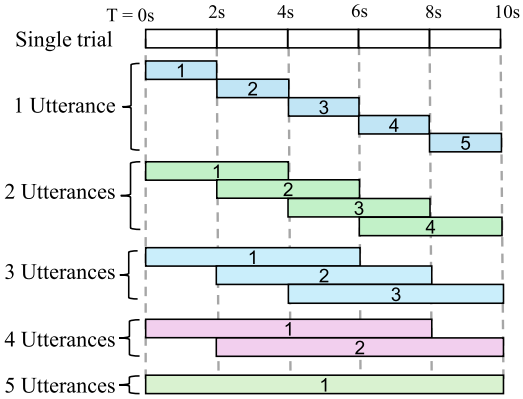

Fig. 7. Illustration of experiments using various utterance splits, where each 10-second trial consists of five repetitions of the same word.

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 [53] in covert speech processing [52], whereas occipital activation reflects visual responses to the cue rather than linguistic content.

Fig. 6 illustrates the EEG saliency maps derived from the Integrated Gradients [51] method for the covert speech classification task. The averaged saliency map Fig. 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 [54], suggesting robust involvement in speech motor planning, articulatory rehearsal [52], and phonological processing [55] 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 [56]. This right hemisphere activation likely indicates neural processes associated with prosodic modulation [57], [58], [59], [60], emotional aspects [61] of speech, or cognitive control [62] mechanisms during covert speech.

The per-utterance saliency map depicted in Fig. 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

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

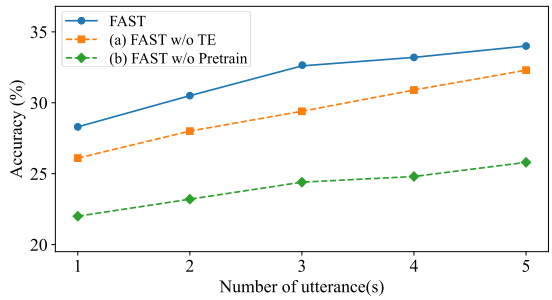

Fig. 8. The average covert speech accuracies of FAST under different ablation cases. (a) An ablation study excludes TE blocks, where the classification head is directly connected after ST. (b) Ablation study without fine-tuning, where weights are randomly initialized instead of loading from LOSO pretraining.

improvement could be attributed to several factors. First, multiple utterance attempts reduce the noise and variability inherent in single-utterance data recording [63], [64], 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.

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 Fig. 8(a), reveal that incorporating TE consistently 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 Fig. 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, our study is focused on isolated command decoding for BCI usage, such as robot control. Extending FAST to larger vocabularies or continuous-speech decoding remains an important direction for future work. Moreover, although FAST achieves clear improvements over all baselines, the absolute accuracy remains modest, reflecting the intrinsic difficulty of covert-speech EEG decoding. Using overt-speech EEG as an auxiliary modality through overt-to-covert transfer learning may further improve covert-speech decoding by offering richer articulatory supervision during pretraining.

The finding from this study that speech-related information is primarily linked to activation in the frontal and temporal brain regions could inform neurofeedback strategies by providing real-time feedback based on activity in these areas, helping participants consciously adjust relevant neural patterns. For example, reinforcing frontal-temporal activity during speech imagery tasks could help strengthen speech-related neural pathways [65]. This type of targeted training may support rehabilitation in conditions such as expressive (Broca) aphasia, where deficits in frontal lobe [66] function contribute to difficulties in speech production. 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.

References

[1] S. L. Metzger 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*. Berlin, Germany: Springer, 2024, pp. 21–28.

[2] S. L. Metzger 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 Rev. Neurosci.*, vol. 25, pp. 473–492, 2024.

[4] M. Angrick et al., "Speech synthesis from ECOG using densely connected 3 d convolutional neural networks," *J. Neural Eng.*, vol. 16, no. 3, 2019, Art. no. 36019.

[5] J. Lu et al., "Neural control of lexical tone production in human laryngeal motor cortex," *Nature Commun.*, vol. 14, no. 1, 2023, Art. no. 6917.

[6] X. Gu et al., "EEG-based brain-computer interfaces (BCIs): A survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 5, pp. 1645–1666, Sep./Oct. 2021.

[7] Z. Zhang et al., "Chisco: An EEG-based bci dataset for decoding of imagined speech," *Sci. Data*, vol. 11, no. 1, 2024, Art. no. 1265.

[8] S. Duraivel et al., "High-resolution neural recordings improve the accuracy of speech decoding," *Nature Commun.*, vol. 14, no. 1, 2023, Art. no. 6938.

[9] T. Proix et al., "Imagined speech can be decoded from low-and cross-frequency intracranial eeg features," *Nature Commun.*, vol. 13, no. 1, 2022, Art. no. 48.

[10] S. Komeiji et al., "Feasibility of decoding covert speech in ecog with a transformer trained on overt speech," *Sci. Rep.*, vol. 14, no. 1, 2024, Art. no. 11491.

[11] D. Dash, P. Ferrari, and J. Wang, "Decoding imagined and spoken phrases from non-invasive neural (MEG) signals," *Front. Neurosci.*, vol. 14, 2020, Art. no. 290.

[12] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[13] 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," *J. Neural Eng.*, vol. 15, no. 5, 2018, Art. no. 056013.

[14] K. P. Thomas, C. Guan, L. C. Tong, and V. A. Prasad, "An adaptive filter bank for motor imagery based brain computer interface," in *Proc. 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2008, pp. 1104–1107.

[15] R. Mahamune and S. H. Laskar, "Classification of the four-class motor imagery signals using continuous wavelet transform filter bank-based two-dimensional images," *Int. J. Imag. Syst. Technol.*, vol. 31, no. 4, pp. 2237–2248, 2021.

[16] 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, 2021, Art. no. 100072.

[17] 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 Netw.*, vol. 123, pp. 176–190, 2020.

[18] J. D. Power et al., "Functional network organization of the human brain," *Neuron*, vol. 72, no. 4, pp. 665–678, 2011.

[19] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 532–541, Jul.–Sep. 2020.

[20] P. Zhong, D. Wang, and C. Miao, "EEG-based emotion recognition using regularized graph neural networks," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1290–1301, Jul.–Sep. 2022.

[21] D. Klepl, F. He, M. Wu, D. J. Blackburn, and P. Sarrigiannis, "EEG-based graph neural network classification of alzheimer's disease: An empirical evaluation of functional connectivity methods," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 2651–2660, 2022.

[22] Y. Ding, N. Robinson, C. Tong, Q. Zeng, and C. Guan, "LGGNet: Learning from local-global-graph representations for brain–computer interface," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 7, pp. 9773–9786, Jul. 2024.

[23] Y. Ding et al., "EEG-deformer: A dense convolutional transformer for brain-computer interfaces," *IEEE J. Biomed. Health Informat.*, vol. 29, no. 3, pp. 1909–1918, Mar. 2025.

[24] N. Parmar et al., "Image transformer," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4055–4064.

[25] J. Xie et al., "A transformer-based approach combining deep learning network and spatial-temporal information for raw EEG classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 2126–2136, 2022.

[26] Y. Song, Q. Zheng, B. Liu, and X. Gao, "EEG conformer: Convolutional transformer for EEG decoding and visualization," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 710–719, 2022.

[27] S. Bagchi and D. R. Bathula, "EEG-convtransformer for single-trial EEG-based visual stimulus classification," *Pattern Recognit.*, vol. 129, 2022, Art. no. 108757.

[28] L. Lu, M. Han, G. Zou, L. Zheng, and J.-H. Gao, "Common and distinct neural representations of imagined and perceived speech," *Cereb. Cortex*, vol. 33, no. 10, pp. 6486–6493, 2023.

[29] C. Fernyhough and A. M. Borghi, "Inner speech as language process and cognitive tool," *Trends Cogn. Sci.*, vol. 27, pp. 1180–1193, 2023.

[30] 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.

[31] 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.

[32] K. Christoff, A. M. Gordon, J. Smallwood, R. Smith, and J. W. Schooler, "Experience sampling during fMRI reveals default network and executive system contributions to mind wandering," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 21, pp. 8719–8724, 2009.

[33] A. Morin and J. Michaud, "Self-awareness and the left inferior frontal gyrus: Inner speech use during self-related processing," *Brain Res. Bull.*, vol. 74, no. 6, pp. 387–396, 2007.

[34] S. Kim, Y. E. Lee, S. H. Lee, and S. W. Lee, "Diff-e: Diffusion-based learning for decoding imagined speech EEG," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2023, pp. 1159–1163.

[35] 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 *Proc. 11th Int. Winter Conf. Brain-Comput. Interface*, 2023, pp. 1–4.

[36] Z. Guo et al., "Enhancing EEG-based covert speech decoding through knowledge transfer," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2025, pp. 1–5.

[37] 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 *Proc. 11th Int. Winter Conf. Brain-Comput. Interface*, 2023, pp. 1–4.

[38] 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," *Biomed. Signal Process. Control*, vol. 81, 2023, Art. no. 104500.

[39] J. Zhou et al., "Pretraining large brain language model for active bci: Silent speech," in *Proc. 33rd ACM Int. Conf. Multimedia*. 2025.

[40] 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.

[41] 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," *Psychophysiology*, vol. 55, 2018, Art. no. e13028.

[42] L. Koessler et al., "Automated cortical projection of eeg sensors: anatomical correlation via the international 10–10 system," *Neuroimage*, vol. 46, no. 1, pp. 64–72, 2009.

[43] M. Oliveri, G. Koch, and C. Caltagirone, "Spatial–temporal interactions in the human brain," *Exp. Brain Res.*, vol. 195, pp. 489–497, 2009.

[44] D.-Y. Lee, M. Lee, and S.-W. Lee, "Decoding imagined speech based on deep metric learning for intuitive BCI communication," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1363–1374, 2021.

[45] K. Zhang, N. Robinson, S.-W. Lee, and C. Guan, "Adaptive transfer learning for eeg motor imagery classification with deep convolutional neural network," *Neural Netw.*, vol. 136, pp. 1–10, 2021.

[46] C. Yang, M. Westover, and J. Sun, "Biot: Biosignal transformer for cross-data learning in the wild," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 78240–78260.

[47] R. Yang and E. Modesitt, "Vit2EEG: Leveraging hybrid pretrained vision transformers for eeg data," 2023, *arXiv:2308.00454*.

[48] Y. Song, X. Jia, L. Yang, and L. Xie, "Transformer-based spatial-temporal feature learning for eeg decoding," 2021, *arXiv:2106.11170*.

[49] Y. Ding, N. Robinson, S. Zhang, Q. Zeng, and C. Guan, "TSception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2238–2250, Jul.–Sep. 2023.

[50] L. D. Brown, T. T. Cai, and A. DasGupta, "Interval estimation for a binomial proportion," *Stat. Sci.*, vol. 16, no. 2, pp. 101–133, 2001.

[51] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328.

[52] W. Zhang et al., "Revealing the spatiotemporal brain dynamics of covert speech compared with overt speech: A simultaneous EEG-fmri study," *NeuroImage*, vol. 293, 2024, Art. no. 120629. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811924001241

[53] 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.

[54] G. Nasios et al., "From broca and wernicke to the neuromodulation era: Insights of brain language networks for neurorehabilitation," *Behav. Neurol.*, vol. 2019, 2019, Art. no. 9894571.

[55] G.-J. Rutten, "Broca-wernicke theories: A historical perspective," *Handbook Clin. Neurol.*, vol. 185, pp. 25–34, 2022.

[56] C. Code, "Can the right hemisphere speak?" *Brain Lang.*, vol. 57, no. 1, pp. 38–59, 1997.

[57] S. Luthra, "The role of the right hemisphere in processing phonetic variability between talkers," *Neurobiol. Lang.*, vol. 2, no. 1, pp. 138–151, 2021.

[58] S. Buklina and A. Batalov, "Improvement of speech function in patients with aphasia: The right hemisphere, an enemy or a friend?" *Hum. Physiol.*, vol. 44, pp. 161–169, 2018.

[59] 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 Lang.*, vol. 203, 2020, Art. no. 104756.

[60] 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 Lang.*, vol. 101, no. 3, pp. 260–277, 2007.

[61] K. M. Hartikainen, "Emotion-attention interaction in the right hemisphere," *Brain Sci.*, vol. 11, no. 8, 2021, Art. no. 1006.

[62] B. C. Preisig and M. Meyer, "Predictive coding and dimension-selective attention enhance the lateralization of spoken language processing," *Neurosci. Biobehavioral Rev.*, vol. 172, 2025, Art. no. 106111.

[63] K. Grill-Spector, R. Henson, and A. Martin, "Repetition and the brain: Neural models of stimulus-specific effects," *Trends Cogn. Sci.*, vol. 10, no. 1, pp. 14–23, 2006.

[64] P. Gagnepain, G. Chételat, B. Landeau, 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," *J. Neurosci.*, vol. 28, no. 20, pp. 5281–5289, 2008.

[65] A. C. Rampinini et al., "Functional and spatial segregation within the inferior frontal and superior temporal cortices during listening, articulation imagery, and production of vowels," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 17029.

[66] D. R. Fontoura, J. C. Rodrigues, L. Mansur, A. M. Monção, and J. F. Salles, "Neuropsycholinguistic profile of patients poststroke in the left hemisphere with expressive aphasia," *Revista Neuropsicologia, Neuropsiquiatria y Neurociencias*, vol. 13, no. 2, pp. 91–110, 2013.

[67] "2020 international bci competition," 2020, Accessed: Apr. 2020. [Online]. Available: https://osf.io/pq7vb/