Journal of Neural Engineering

PAPER

Decoding imagined speech from EEG signals using hybrid-scale spatial-temporal dilated convolution network

Journal of Neural Engineering

RECEIVED
12 April 2021

REVISED
20 June 2021

ACCEPTED FOR PUBLICATION
13 July 2021

PUBLISHED
11 August 2021

PAPER

Decoding imagined speech from EEG signals using hybrid-scale spatial-temporal dilated convolution network

Fu Li, Weibing Chao, Yang Li*, Boxun Fu, Youshuo Ji, Hao Wu and Guangming Shi

Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an, People's Republic of China

* Author to whom any correspondence should be addressed.

E-mail: liy@xidian.edu.cn

Keywords: brain–computer interface (BCI), EEG-based imagined speech recognition, hybrid-scale, spatial-temporal network

Abstract

Objective. Directly decoding imagined speech from electroencephalogram (EEG) signals has attracted much interest in brain–computer interface applications, because it provides a natural and intuitive communication method for locked-in patients. Several methods have been applied to imagined speech decoding, but how to construct spatial-temporal dependencies and capture long-range contextual cues in EEG signals to better decode imagined speech should be considered. *Approach.* In this study, we propose a novel model called hybrid-scale spatial-temporal dilated convolution network (HS-STDCN) for EEG-based imagined speech recognition. HS-STDCN integrates feature learning from temporal and spatial information into a unified end-to-end model. To characterize the temporal dependencies of the EEG sequences, we adopted a hybrid-scale temporal convolution layer to capture temporal information at multiple levels. A depthwise spatial convolution layer was then designed to construct intrinsic spatial relationships of EEG electrodes, which can produce a spatial-temporal representation of the input EEG data. Based on the spatial-temporal representation, dilated convolution layers were further employed to learn long-range discriminative features for the final classification. *Main results.* To evaluate the proposed method, we compared the HS-STDCN with other existing methods on our collected dataset. The HS-STDCN achieved an averaged classification accuracy of 54.31% for decoding eight imagined words, which is significantly better than other methods at a significance level of 0.05. *Significance.* The proposed HS-STDCN model provided an effective approach to make use of both the temporal and spatial dependencies of the input EEG signals for imagined speech recognition. We also visualized the word semantic differences to analyze the impact of word semantics on imagined speech recognition, investigated the important regions in the decoding process, and explored the use of fewer electrodes to achieve comparable performance.

1. Introduction

Brain–computer interfaces (BCIs) directly convert brain activities into computer control signals to establish connections with the external world [1], which provides an alternative communication method for people suffering from severe neurological diseases [2, 3]. Electroencephalography (EEG) is a non-invasive method that measures electrical activity in the brain. It is one of the most widely used recording modalities for BCIs. Consequently, EEG-based BCI communication systems have become an active topic and have attracted an increasing number of researchers over the past several years. For example, many researchers have focused on various forms of brain signals, such as steady-state visual evoked potential, P300 evoked potential, and motor imagery [4–6]. However, the aforementioned forms of EEG signals always depend on external stimuli or require additional training [7]. Considering that speech is a basic and natural form of human communication, efforts have been focused on attempting to decode speech content from imagined speech EEG signals, which can be utilized as a more intuitive BCI communication system.

The definition of imagined speech, also known as covert speech, inner speech, and verbal thinking, is the internal speech activity without explicitly moving any articulator [8, 9]. Generally, internal speech representations can be produced based on the normal functions of the human brain's cognitive and language activities. Therefore, imagined speech can be used for locked-in patients to convey their intentions directly. From the viewpoint of neuroscience, similar brain activities occur between imagined speech and overt speech because of speech perception and production originating from a variety of brain cortex regions [10, 11]. Wise *et al* found that some activity-related changes occur in Broca's area, Wernicke's area, and the supplementary motor area (SMA, which is speculated to be involved in the motor planning of speech) during the inner verb generation task [12]. High gamma neural power changes were found in the superior temporal gyrus, Wernicke's area, Broca's area, primary motor cortex area, and premotor cortex area during cue presentation and subsequent word articulation [13, 14]. In addition to high-frequency activities, Hermes *et al* [15] also found that theta rhythm oscillations correlate negatively with high-frequency activity in language regions during inner verb generation tasks. Consequently, the speech content can be decoded by modeling the neural representation of the imagery speech from the EEG signals.

Different feature extraction algorithms and classifiers have been used to decode imagined speech from EEG signals in terms of vowels, syllables, phonemes, or words. For example, DaSalla *et al* [16] and Wang *et al* [17] used common spatial pattern (CSP) and support vector machine (SVM) to classify two English vowels and two Chinese characters, respectively. Min *et al* [18] used statistical features, such as mean value, variance, standard deviation, skewness, and extreme learning machine (ELM) to perform binary classification across five different vowels. Nguyen *et al* [19] extracted tangent vectors from a covariance matrix to represent Riemannian manifold features and then used a relevance vector machine (RVM) to classify four different sets of imagined speech tasks, namely, three vowels, two short-long words, two long words, and three short words across different subjects. Sereshkeh *et al* [20] used multilayer perceptron (MLP) based on features extracted using discrete wavelet transform (DWT) to classify the covert repetition of 'yes' and 'no' across multiple sessions. More meaningful words, namely, 'up,' 'down,' 'left,' 'right,' and 'select' [21], or others like 'go', 'back', 'left', 'right', and 'stop' [22] were recognized. In [23], Cooney *et al* used Mel-frequency cepstral coefficient features and SVM to classify seven syllables and four words. Lee *et al* [24] used CSP and three basic classifiers, namely, shrinkage of regularized linear discriminant analysis, random forest, and SVM to classify 12 imagined words with the resting state.

With the success of deep learning methods in many areas [25–27], deep learning methods have been proposed for EEG processing [28, 29] and have improved the classification results. Some studies have also attempted to apply deep learning methods to imagined speech decoding. For instance, Saha *et al* [30] initially computed channel cross-covariance and then used a model framework composed of a convolutional neural network (CNN), a long short-term network, and a deep autoencoder to extract spatial-temporal information. In another work [31], they also employed channel cross-covariance for preprocessing, and a network composed of a hierarchical combination of spatial and temporal CNN cascaded with a deep autoencoder was applied. Cooney *et al* [32] implemented independent component analysis, and deep CNN with transfer learning methodologies were employed to classify five imagined vowels.

Nevertheless, we argue that two issues should be further considered in EEG-based imagined speech recognition tasks. One is how to identify the implicit complex temporal and spatial relationships in EEG signals. For a time-varying EEG signal, it is collected from multiple active electrodes attached to the scalp by arranging a certain spatial layout. Thus, the speech-related information is not only involved in temporal variation, but also in spatial correlations among electrodes. To better decode the imagined speech content, the temporal and spatial dependencies of the EEG signal should be characterized. In particular, in modeling the temporal dependence, neuroscience studies showed that some related brain activities occur in different frequency bands for imagined speech [12–15]. Thus, exploring how to model the temporal frequency oscillation at different levels will contribute to imagined speech recognition. The other issue is how to capture long-range contextual cues. CNNs may fail to capture long-distance temporal dependencies owing to the locality of the convolution layer, while recurrent neural networks suffer from gradient explosion or vanishing issues in processing long-term sequences and cannot be conducted in parallel. However, dilated convolution can avoid these problems by incorporating long-range temporal information in a convolution way, thereby greatly improving the training efficiency. Utilizing this technique is potentially helpful to learn more discriminative features for EEG-based imagined speech recognition.

To address the above issues, in this study, we propose an end-to-end network architecture called hybrid-scale spatial-temporal dilated convolution network HS-STDCN to exploit spatial-temporal discriminative information to decode imagined speech EEG signals. To learn the deep temporal dependencies of the EEG sequences, we adopted a hybrid-scale temporal convolution layer to capture temporal information. Compared with using a single-scale temporal convolution kernel, the hybrid-scale temporal kernel design can learn multiple levels of temporal information. Then, a depthwise spatial convolution layer is used to characterize intrinsic spatial relationships by scanning EEG electrodes along the temporal sequences, which can produce a spatial-temporal representation of the input EEG data. Based on the deep spatial-temporal representation, dilated convolution blocks with incremental receptive fields were employed to learn the discriminative features. Long-range EEG information can be captured by continuously expanding the receptive field. Therefore, the proposed HS-STDCN is capable of modeling temporal and spatial dependencies and learning discriminative information for the final classification.

As well as the proposed method for EEG-based imagined speech recognition, we also investigated word semantics based on the HS-STDCN model. In previous studies, the attributes of words could also affect the decoding performance. For example, Nguyen *et al* [19] analyzed the impact of words' sound, meaning, and complexity on classification performance. In their conclusions, words with higher complexity in terms of length can be discriminated more easily. Lee *et al* [24] investigated the impacts of the number of syllables and word properties (concrete or abstractness), but their experiments showed that the number of syllables or letters did not affect the performance. A previous study [33] argued that imagined speech retains deep-level features, such as lexical and semantic information, but does not represent surface-level information, such as phonological detail. Inspired by this and considering semantics is the core attribute of word expressions, in this paper, we visualized the semantic differences of different imagery words and performed different category classifications according to the visualization results to investigate the impact of word semantics on imagined speech decoding. In addition, we investigated which brain regions are more correlated to imagined speech activity and explore the use of fewer electrodes to maintain comparable performance.

We conducted experiments on our collected imagined speech dataset. Our HS-STDCN achieved an averaged classification accuracy of 54.31% for decoding eight imagined words. This result is significantly better than other methods at a significance level of 0.05, which proves the superiority of HS-STDCN in imagined speech decoding.

The remainder of this paper is organized as follows. Section 2 presents the experimental paradigm, signal acquisition and preprocessing, and the proposed HS-STDCN model. In section 3, classification experiments are conducted to evaluate the proposed model, and the results are discussed. Finally, section 4 concludes the paper.

2. Method

2.1. Subjects

Nine healthy native Chinese subjects were recruited (all males; one left-handed; mean age: 22.89 ± 1.20; range: 22–26 years). All subjects reported normal vision or normal vision after correction, and no history of neuropathy or language disorder. None of them had prior experience with imagined speech tasks. Each subject was informed of the experimental content and agreed to participate in the experiment.

2.2. Experimental design and paradigm

The experimental paradigm design may affect the reliability of the decoding results. A previous study [34] has revealed that oscillations in the normal human brain at time scales reflect a memory of the dynamics of the system for more than a few seconds, which shows temporal correlations both in short and long terms in EEG signals. And this temporal correlation artifacts can result in illusory high performance. For example, Wester [35] collected EEG trials for the same word continuously over a period and reported a high recognition rate. Porbadnigk *et al* [36] later proved that the achieved high classification accuracy in [35] is due to the recognition of temporally correlated artifacts instead of words. A similar phenomenon was found in human neural activities in image tasks [37, 38]. Concretely, if the words are presented in blocks, that is, the collected EEG trials in one block come from one category, the classification accuracy may be illusory high due to the recognition of temporal correlated artifacts in block recordings. To avoid confounding block-level effects with experimental results, experiments should be designed across multiple blocks. And in each block, EEG signals of different categories should be collected randomly.

Previous studies [19, 22, 24] used text or audio as the prompts of imagined speech task. When using auditory cue for imagined speech production, the auditory cortex can be activated [24]. The speech production by picture encompasses at least three broad stages, i.e. picture identification, word activation, and verbal response [39, 40]. Compared with text stimuli, pictures are more effective as primes and are more primeable as targets [39]. Moreover, considering that visual presentation can better induce learning, recall, and retrieval of information [41, 42], we used pictures as prompt for speech content and selected ten pictures for each category to avoid the effect of single picture information.

To mitigate the effect of potential temporal correlation artifacts and make the experimental results more reliable, we collected imagined speech EEG trials in a random order across different blocks. Specifically, eight two-character Chinese words (apple, banana, grape, watermelon, baboon, lion, rhinoceros, and zebra) were utilized in our experiments. Before the formal experiment began, each subject was given

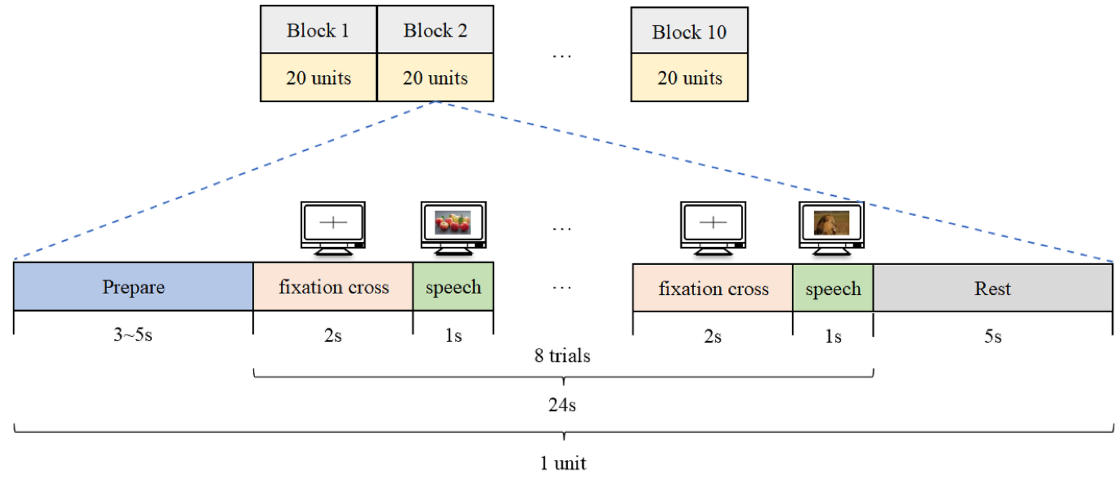


Figure 1. The experiment consisted of ten blocks, each consisting of 20 units. One unit includes 3–5 s preparation, eight imagined speech trials with a total time of 24 s, and 5 s relaxation. Each trial consisted of 2 s of fixation and 1 s of task execution.

some instructions to ensure that they had a clear understanding of the experiment details. The subjects were asked to judge a picture first and then pronounced the corresponding Chinese words in their minds without any overt vocalization or muscle movement. And each subject would practice 3–5 units of the imaginary speech task to learn how to perform imagined speech. In the formal experiment, each subject was asked to perform ten experimental blocks. There was a short break between two consecutive blocks, and the duration was determined by the subject. Each block consists of 20 units. In each unit, each of the eight words was executed once in a random order. In this way, the collected EEG trials for each category come from separated time periods and different blocks; thus, the impact of temporal correlation artifacts on classification results is minimized.

Figure 1 shows the detailed experimental procedure for a unit within one block. At the beginning of each unit, there was a 3–5 s preparation to allow the subject to adjust for a comfortable sitting posture. Subsequently, a fixation cross was displayed in the center of the screen for 2 s to inform the subject focus on the screen, and a picture was then randomly selected from the ten pictures corresponding to each category; the picture was displayed for 1 s for imagined speech task execution. During the 1 s picture display time, subjects judged the picture first and then pronounced the corresponding Chinese words in their minds without any overt vocalization or muscle movement. These 1 s EEG data were used for imagined speech decoding. After each of the eight words was executed once, the subject was allowed for a 5 s break.

2.3. Data acquisition and processing

Subjects were seated 0.5 m in front of the screen in a closed quiet chamber to isolate the external noises and electromagnetic interference. The experimental paradigm was designed using Psychopy [43]. A 64-channel Biosemi ActiveTwo^{TM} System with AgCl electrodes placed following the 10–20 international system was used to record imagined speech EEG signals. The sampling rate was set to 1024 Hz. We collected 200 trials for each category in ten blocks. Thus, 1600 trials were conducted for each subject.

The raw signal was preprocessed using the MNE-Python package [44]. Specifically, a 1–90 Hz FIR bandpass filter with zero phase was applied to the raw EEG signal to retain the information of significant frequency bands that may be associated with imagined speech activities and remove possible artifacts related to muscle movement. A FIR notch filter at 50 Hz was then applied to remove the line noise. Subsequently, the EEG data were resampled to 256 Hz, and z-score normalization was performed to normalize the EEG data. The normalization can be formed as

$$\hat{x} = \frac{x - \mu}{\sigma}, \tag{1}$$

where x is the EEG data vector from one channel, and μ and σ are the mean and standard deviation, respectively. Except for the preprocessing operations mentioned above, there are no extra operations such as artefact or trial removal. The collected 1600 trials (200 trials per category) for each subject are all used in our experiments.

2.4. Proposed method

Figure 2 shows the framework of the proposed HS-STDCN model to specify the proposed method clearly. The goal of the proposed framework was to capture long range temporal and spatial information within multichannel EEG signals. Thus, in HS-STDCN, we adopted four steps to extract discriminative information from imagined speech EEG signals. The first step was to characterize the temporal dependencies at multiple levels for each electrode

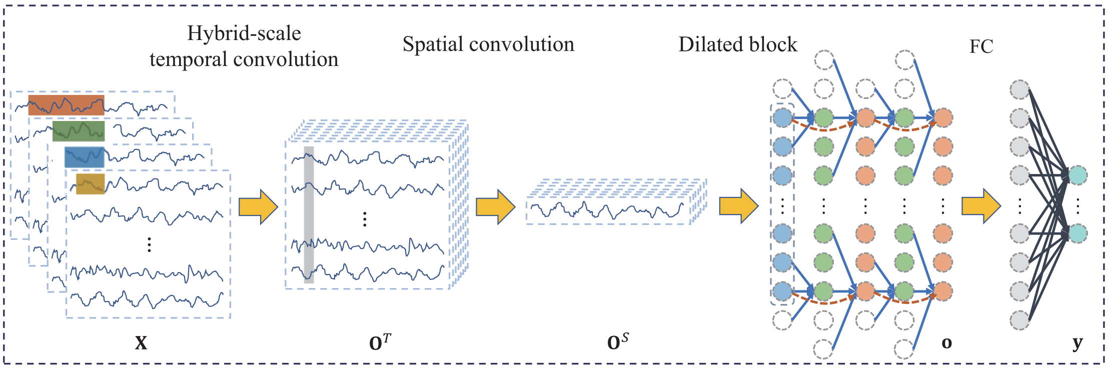


Figure 2. Overall visualization of the HS-STDCN framework. The network first used hybrid-scale temporal convolutions and a spatial convolution to learn spatial-temporal representations. Subsequently, dilated convolutions with incremental receptive fields were used to further exploit high-level discrepancy information. Finally, a fully connected layer was used for classification.

EEG data. The intrinsic spatial relationships between multiple electrodes were then constructed to obtain the spatial-temporal representations. Subsequently, the long-range cues were captured by dilated convolution blocks. Finally, a fully connected layer was used for the classification. The overall process is described as follows.

2.4.1. Character temporal dependencies

Note that EEG signals can be treated as multichannel time series; thus, we first applied temporal convolution on each channel of imagined speech EEG signals to model the temporal variations. In particular, related information was found between brain activities and imagined speech in different frequency bands [12–15]. We used hybrid-scale rather than single-scale temporal filters on the input EEG data to learn the temporal frequency information at different levels. Specifically, for an imagined speech EEG sample $\mathbf{X} = [X_{1,1}, X_{1,2}, \ldots, X_{1,T}; \ldots; X_{C,1}, X_{C,2}, \ldots, X_{C,T}] \in \mathbb{R}^{C\times T}$ where C is the number of electrodes, and T is the sequence length, the temporal convolution operation can be formulated as follows:

$$\mathcal{F}^{t_k}(X_{c,t}) = (\mathbf{X} * f^{t_k})(X_{c,t}) = \sum_{i=1}^{l_k} f^{t_k}_{1,i} \cdot X_{c,t+i-(l_k-1)/2}, \tag{2}$$

where $\mathcal{F}^{t_k}$ denotes the temporal convolution operation, $f^{t_k} \in \mathbb{R}^{1\times l_k}$ is the temporal convolution filter with kernel length l_k, $k \in [1,2,\ldots,K]$, $c \in [1,2,\ldots,C]$, $t \in [1,2,\ldots,T]$, and the subscript $t + i - (l_k - 1)/2$ denotes the convolution position of the temporal convolution on a single-channel EEG sequence. L et $\widehat{\mathbf{X}}^{t_k} = \left[\hat{X}^{t_k}_{1,1}, \hat{X}^{t_k}_{1,2}, \ldots, \hat{X}^{t_k}_{1,T}; \ldots; \hat{X}^{t_k}_{C,1}, \hat{X}^{t_k}_{C,2}, \ldots, \hat{X}^{t_k}_{C,T}\right]$ denotes the output of the temporal convolution operation $\mathcal{F}^{t_k}$. By fusing the different levels of temporal frequency information, we obtained the deep temporal representation of each electrode, which kept the spatial location structural information:

$$\mathbf{O}^T = \sum_{k=1}^{K} \widehat{\mathbf{X}}^{t_k}, \tag{3}$$

where $\mathbf{O}^T \in \mathbb{R}^{C\times T}$ denotes the final output feature map after fusing different levels of temporal information.

2.4.2. Construct spatial relationship

After obtaining the multiple-level temporal information of each electrode, to construct the functional spatial relationship between EEG electrodes, we introduced a spatial convolution to capture this potential structure relationship to distinguish different imagined speech contents. Specifically, on the temporal feature matrix $\mathbf{O}^T$, a depthwise convolution with a spatial convolution filter $f^{t_s} \in \mathbb{R}^{C\times 1}$ was adopted to learn the intrinsic spatial information. The operation $\mathcal{F}^s$ can be expressed as follows:

$$\mathcal{F}^s\left(O^T_{:,t}\right) = \left(\mathbf{O}^T * f^s\right)\left(O^T_{:,t}\right) = \sum_{i=1}^{C} f^s_{i,1} \cdot O^T_{i,t}. \tag{4}$$

Herein, we obtained the spatial-temporal representation $\mathbf{O}^S \in \mathbb{R}^T$ for the input data $\mathbf{X}$.

2.4.3. Capture long-range cues

After obtaining the spatial-temporal representations, the multichannel EEG sequence was squeezed into a one-dimensional sequence. To capture the long-range cues, we used one-dimensional dilated convolution blocks that gradually expand the receptive fields by increasing the dilation rate to extract high-level features on the spatial-temporal representation $\mathbf{O}^S$. The dilated convolution operation $\mathcal{F}^d$ is defined as follows:

$$\mathcal{F}^d\left(O^S_t\right) = \left(\mathbf{O}^S *_r f^d\right)\left(O^S_t\right) = \sum_{i=1}^{l} f^d_i \cdot O^S_{t+r\cdot(i-(l-1)/2)}, \tag{5}$$

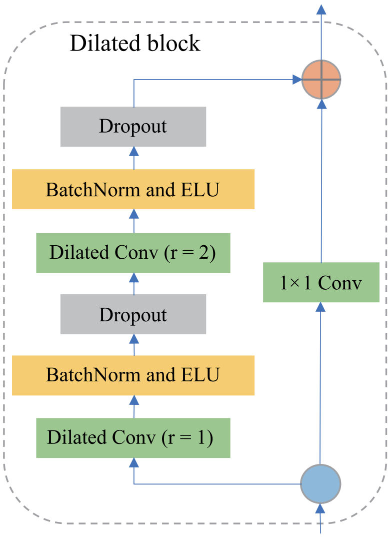


Figure 3. Detailed description for the dilated block.

where $f^d \in \mathbb{R}^l$ is the dilated convolution filter with kernel length l, r is the dilation rate, and the subscript $t + r \cdot (i - (l-1)/2)$ denotes the convolution position of the dilated convolution on the $\mathbf{O}^S$. To avoid losing the continuity of information, the dilation rate was first set to 1. A detailed description of the dilated block is presented in figure 3. The dilated block consisted of two dilated convolutions with dilation rates of 1 and 2, and a skip connection that adds the input to the output of the dilated convolutions was used to ease network training, which can be represented as follows:

$$\mathbf{o}^d = x + \mathcal{F}^d(x), \qquad (6)$$

where x and $\mathbf{o}^d$ denote the input and output of the dilated block, respectively. Two dilated blocks were employed in HS-STDCN. The final feature vector, denoted by $\mathbf{o} \in \mathbb{R}^L$, of the EEG sample $\mathbf{X}$ containing deep-level spatial-temporal information is expressed as follows:

$$\mathbf{o} = [o_1, o_2, \ldots, o_L] \qquad (7)$$

2.4.4. Prediction

Finally, we added a supervision term to the network by applying the softmax function to the output feature vector o to predict the class label. It can be shown as follows:

$$\mathbf{y} = \mathbf{oW} + \mathbf{b} = \{y_1, y_2, \ldots, y_N\} \in \mathbb{R}^{1\times N}, \qquad (8)$$

where $\mathbf{W} \in \mathbb{R}^{L\times N}$ and $\mathbf{b} \in \mathbb{R}^{1\times N}$ are the transform matrices, and N is the number of imagery word classes. Finally, the output vector $\mathbf{y}$ is fed into the softmax layer for imagery word classification, which can be expressed as follows:

$$P(n \mid \mathbf{X}) = \exp(y_n) / \sum_{i=1}^{N} \exp(y_i), \qquad (9)$$

where $P(n \mid \mathbf{X})$ denotes the predicted probability that the input sample $\mathbf{X}$ belongs to the nth class. As a result, label $\tilde{l}$ of sample $\mathbf{X}$ is predicted as follows:

$$\tilde{l} = \mathrm{argmax}_{\mathrm{n}} P(n \mid \mathbf{X}). \qquad (10)$$

3. Results

3.1. Experiment settings

In the experiment, the ten-block EEG data were split into the training set and testing set at a ratio of 8:2. The first eight-block EEG data were used for training, and the last two blocks were used for testing. In this way, the training samples and test samples come from different blocks, which further mitigates the effect of temporal correlation on classification and makes the experimental results more credible.

All 64-electrode EEG data with a 256 Hz resampling rate were used to decode the imagery words; thus, the size $C \times T$ of the input sample $\mathbf{X}$ is 64×256. Four different scale temporal convolution filters (1×63, 1×31, 1×19, 1×7) were employed to extract different levels of temporal information, and the number of each scale temporal filter F_1 was set to 8. The number and kernel size of the dilated convolution filter in the dilated block were set to 16 and 3, respectively. Specifically, we implemented HS-STDCN using Keras, trained, and tested on an NVIDIA TITAN XP GPU with CUDA 10 and cuDNN v7. The categorical cross-entropy loss function and the SGD optimizer were used to optimize the model parameters. We ran 200 epochs on the training set. Empirically, the learning rate of the first 100 epochs was set to 0.01, and the last 100 epochs were set to 0.0001 to finely optimize the network. The other settings of the HS-STDCN model are listed in table 1.

3.2. Experimental results and discussion

3.2.1. Results on imagined speech decoding task

To evaluate the proposed HS-STDCN model, we conducted experiments on the collected imagined speech EEG signals. In the experiment, the mean accuracy (ACC) and standard deviation (STD) were used as the evaluation criteria for all subjects in the dataset. In addition, to validate the classification superiority of HS-STDCN, we conducted the same experiments using hand-crafted features and other deep learning methods that have achieved significant performance in previous studies, including CSP and SVM [17], DWT and MLP [20], statistic features and ELM [18], tangent vectors and RVM [19], and EEGNet [28]. All classification results are summarized in table 2.

Table 2 clearly shows that the proposed HS-STDCN model outperforms all the comparison methods, which verifies the superiority of HS-STDCN in imagined speech EEG signal decoding. In particular, the proposed method improved the state-of-the-art traditional method Tangent + RVM

Table 1. HS-STDCN architecture.

Layer	Filters	Kernel	Output	Options
Input			$(C, T, 1)$	
Conv2D	F_1	$(1, 7)$	(C, T, F_1)	mode = same
	F_1	$(1, 19)$	(C, T, F_1)	mode = same
	F_1	$(1, 31)$	(C, T, F_1)	mode = same
	F_1	$(1, 63)$	(C, T, F_1)	mode = same
BatchNorm			(C, T, F_1)	
Add			(C, T, F_1)	
DepthwiseConv2D	$F_1 \times 2$	$(C, 1)$	$(1, T, F_1)$	mode = valid, depth = 2
BatchNorm			$(1, T, F_1)$	
ELU			$(1, T, F_1)$	
AvgPool2D		$(1, 16)$	$(1, T//16, F_1)$	
Dropout			$(1, T//16, F_1)$	p = 0.2
Reshape			$(T//16, F_1)$	
Dilated block	F_2	3	$(T//16, F_2)$	mode = same, p = 0.3
Dilated block	F_2	3	$(T//16, F_2)$	mode = same, p = 0.3
Flatten			$(T//16 \times F_2)$	
Dense and softmax			N	max_norm = 0.25

Table 2. Acc ± std of the classification performance (%). The Bold value indicates the proposed method HS-STDCN achieved highest performance.

Method	Acc ± Std
CSP + SVM [17]	16.43 ± 1.38
DWT + MLP [20]	24.20 ± 4.12
Statistic features + ELM [18]	23.68 ± 4.28
Tangent + RVM [19]	39.45 ± 6.14
EEGNet [28]	49.93 ± 4.95
STDCN	52.64 ± 5.02
HS-STDCN	**54.31** ± 5.22

Table 3. Paired t-test between HS-STDCN and other methods at a significance level of 0.05.

Method	p-Value
HS-STDCN vs CSP + SVM [17]	<0.001
HS-STDCN vs DWT + MLP [20]	<0.001
HS-STDCN vs Statistic features + ELM [18]	<0.001
HS-STDCN vs Tangent + RVM [19]	<0.001
HS-STDCN vs EEGNet [28]	<0.001
HS-STDCN vs STDCN	0.009

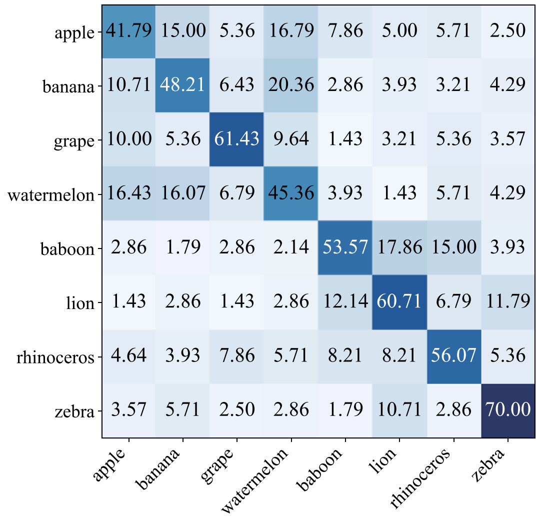
by 14.86%. Compared with the deep learning model EEGNet, the improvement was 4.38%. This result shows the superiority of the proposed network architecture for EEG-based imagined speech recognition. In addition, to further verify the importance of the multiple-level temporal information obtained by HS-STDCN in the performance improvement, we conducted additional experiments on the collected data sets by ablating the hybrid-scale temporal kernel design of HS-STDCN, which is denoted as STDCN in table 2, and obtained 52.64% accuracy. This result is also clearly superior to all the five comparison methods. More importantly, we observed that the performance can be further improved with hybrid-scale temporal convolution, which showed the superiority of employing different levels of temporal information in imagined speech decoding tasks.

To verify that the hybrid-scale temporal kernel design is statistically significantly better than the single temporal kernel, we performed a paired t-test statistical analysis between HS-STDCN and STDCN at a significance level of 0.05. Meanwhile, we also performed paired t-test statistical analysis between HS-STDCN and other methods to evaluate the superiority of our method. As shown in table 3, HS-STDCN was significantly better than the single temporal scale method STDCN and other methods in decoding imagined speech.

To better elucidate the confusion of HS-STDCN in recognizing different imagery words, the confusion matrices of the above experiments are depicted in figure 4. From this, we obtained two observations:

(a) Animal words were easier to recognize than fruit words. In the eight imagery words, the animal word ‘zebra’ achieved the highest recognition rate and was 8% higher than the fruit word ‘grape’, which showed the second highest recognition rate. The recognition rates of the remaining three animal words were also better than those of the remaining three fruit words. This shows that word attributes have an important impact on the decoding performance.

(b) Imagery words with smaller semantic differences were more difficult to distinguish. The misclassified fruit words were mainly recognized as other fruit words, whereas animal words also tended to be incorrectly classified as other animal words. For example, ‘apple’ is easier to recognize as ‘banana’ and ‘watermelon’, and ‘baboon’ is more easily misclassified as ‘lion’ and ‘rhinoceros’. A study [45] revealed the effects of semantic relatedness on verbal working memory performance and that semantic relatedness has a negative impact on word memory [46]. Considering that the four fruit words are close in semantics and so do the four animal words, we can infer that



	apple	banana	grape	watermelon	baboon	lion	rhinoceros	zebra
apple	41.79	15.00	5.36	16.79	7.86	5.00	5.71	2.50
banana	10.71	48.21	6.43	20.36	2.86	3.93	3.21	4.29
grape	10.00	5.36	61.43	9.64	1.43	3.21	5.36	3.57
watermelon	16.43	16.07	6.79	45.36	3.93	1.43	5.71	4.29
baboon	2.86	1.79	2.86	2.14	53.57	17.86	15.00	3.93
lion	1.43	2.86	1.43	2.86	12.14	60.71	6.79	11.79
rhinoceros	4.64	3.93	7.86	5.71	8.21	8.21	56.07	5.36
zebra	3.57	5.71	2.50	2.86	1.79	10.71	2.86	70.00

Figure 4. Confusion matrix based on the classification results of HS-STDCN. *x*-axis indicates the predicted class, and *y*-axis displays the ground-truth class.

semantics is an important factor in discriminating imagined speech content.

3.2.2. Semantics impact analysis

In a previous literature [47], specific information processing is induced by the semantic meaning of the word. In [48], Skrandies *et al* revealed that the electrical brain activation between different semantic word classes showed significant differences. To further understand the impact of word semantics on imagined speech decoding performance, we explored the semantic relations of different types of words in this section. The relations were constructed by performing multidimensional scaling (MDS) operation [49] on the output vectors **o** of HS-STDCN in (7) on all testing samples. By MDS operation, the output multidimensional vectors **o** were mapped into two-dimensional scatterplots to visualize the distances between different words. Then we calculated the centroids for each type of word based on the results of MDS operation. The results are presented in figure 5. From this, we obtained three observations:

(a) The semantic gap was clear between the fruit words and animal words for all subjects, which is consistent with the normal semantic perception of humans. This observation indicates that the HS-STDCN can learn rich semantic information from imagined speech EEG signals.

(b) The semantic difference between the four animal words was larger than that of the four fruit words. Considering that the four selected animals are not common in daily life and the connections between them are not closer than the four fruits, this also reflects the latent semantic perception in imagined speech activities.

(c) The subjects showed semantic perception differences in imagery words. For both the fruit and animal words, the semantic differences between them differed for each subject. For example, the distance between the four fruit words of S9 was closer than that of other subjects, and the distance between the four animal words of S3 was more compact than that of other subjects, which may be due to the individual differences in semantic processing [50].

Based on the observation above, we conducted additional experiments on the data referring to the word semantic differences. Specifically, we classified the four fruit words, four animal words, two fruit words and two animal words. The classification results are presented in table 4. We can see that the proposed method HS-STDCN also achieved the highest performance in both experiments, which further verifies the superiority of HS-STDCN in decoding imagined speech. Another point can be seen that the classification performance of two fruit words and two animal words is higher than that of four fruit words and four animal words. Due to the large semantic differences between fruit words and animal words, it is much easier to discriminate between fruit words and animal words. However, the four fruit words are similar in semantics and so do the four animal words, thus these two types of four classes classification tasks are much harder than two fruit words and two animal words. In addition, the classification performance of the four animal words was better than that of the four fruit words, which may be because the four selected fruit words were more common to the subjects and induced more similar EEG signals.

3.2.3. Activity of EEG electrodes

The electrode activity maps are depicted in figure 6 by visualizing the temporal features extracted by the hybrid-scale temporal convolution layers of HS-STDCN on all testing samples to explore the contribution of different brain regions for decoding imagined speech. The contribution was evaluated by calculating the L2 norm of each row of the temporal features $\mathbf{O}^T$ in equation (3) and mapping these values into the corresponding electrodes. As shown in figure 6, the parietal lobe, left posterior temporal lobe, and left inferior frontal lobe make more contributions, which have been proven to be related to semantic processing, word reading, and comprehension [51–53]. This result shows that our HS-STDCN focuses on the activities in the brain language function areas, which verifies the superiority of the HS-STDCN from the viewpoint of neuroscience.

Moreover, to further investigate where the obtained discriminative information in HS-STDCN comes from in terms of different imagery words,

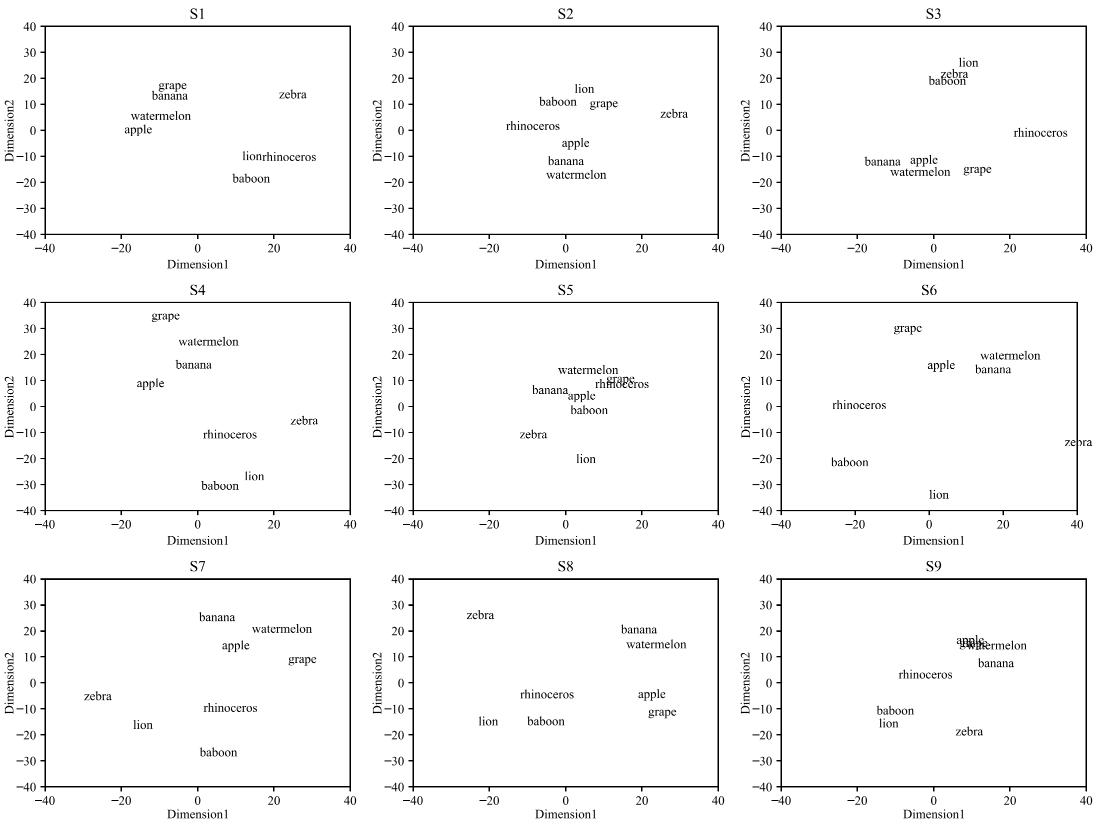


Figure 5. Semantic difference visualization between different imagery words for all the subjects.

Table 4. Acc ± std of the classification performance (%) for different four words. The Bold values indicate the proposed method HS-STDCN achieved highest performance.

Method/Word choice	Fruit words	Animal words	Two fruit and two animal words
CSP + SVM [17]	27.51 ± 2.78	29.38 ± 1.68	29.10 ± 5.24
DWT + MLP [20]	31.95 ± 3.88	37.29 ± 5.27	38.48 ± 5.43
Statistic features + ELM [18]	33.48 ± 3.40	37.16 ± 4.06	40.42 ± 4.82
Tangent + RVM [19]	42.99 ± 6.49	50.21 ± 4.50	58.96 ± 6.27
EEGNet [28]	53.20 ± 5.23	63.54 ± 5.67	71.81 ± 3.73
STDCN	55.21 ± 5.22	65.49 ± 6.11	74.10 ± 4.98
HS-STDCN	**56.32** ± 5.49	**67.71** ± 6.08	**76.18** ± 4.45

the electrode activity maps corresponding to each imagery word are depicted in figure 7. Based on this figure, we had two observations:

(a) For the four fruit words, the activated brain regions were roughly the same. Figures 7(a)–(d) shows that the parietal lobe made a greater contribution than the other areas. For ‘watermelon’, the left posterior temporal lobe had the same contribution as the parietal lobe.

(b) For the four animal words, the activated areas showed some differences. Figures 7(e) and (f) shows that for ‘baboon’ and ‘zebra’, both the parietal lobe and the left posterior temporal lobe made significant contributions. For ‘lion’, the most activated areas were concentrated in the left posterior temporal lobe. For ‘rhinoceros’, the parietal lobe and left inferior frontal lobe contributed the most.

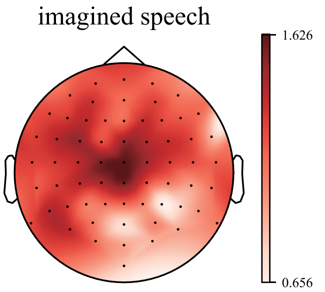


Figure 6. EEG electrode activity maps by mapping the temporal features extracted by the hybrid-scale temporal convolution layers of HS-STDCN to the corresponding electrodes. Darker red denotes more significant contribution.

To show the activity divergences more clearly between the fruit word and animal word, we depicted the electrode activity maps of the whole four fruit

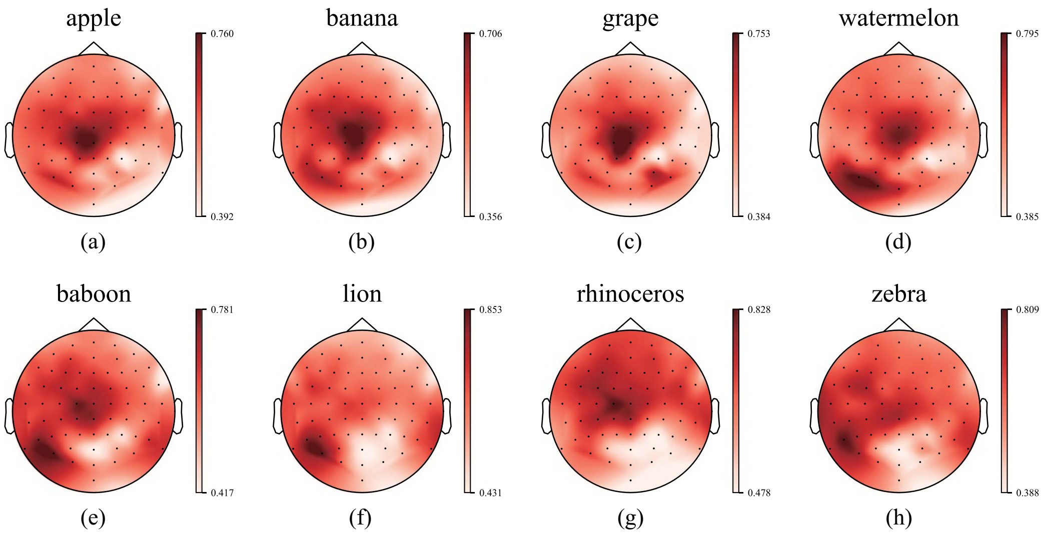


Figure 7. EEG electrode activity maps of different imagery words based on HS-STDCN.

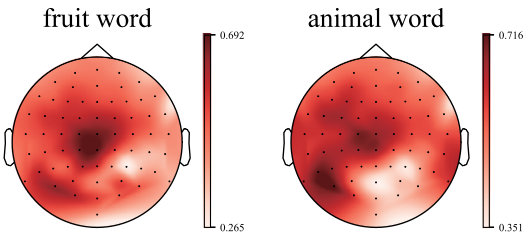


Figure 8. EEG electrode activity maps of fruit word and animal word based on HS-STDCN.

words and the whole four animal words in figure 8. It can be seen that for the fruit word, the parietal lobe makes more contribution than other areas, while for the animal word, the most obvious characteristics are concentrated on the left posterior temporal lobe. Besides, in the right hemisphere, the activities for animal words are mainly distributed in the right temporal lobe, while the activities for fruit words are mainly near the right occipital lobe.

3.2.4. *Performance based on less electrodes*

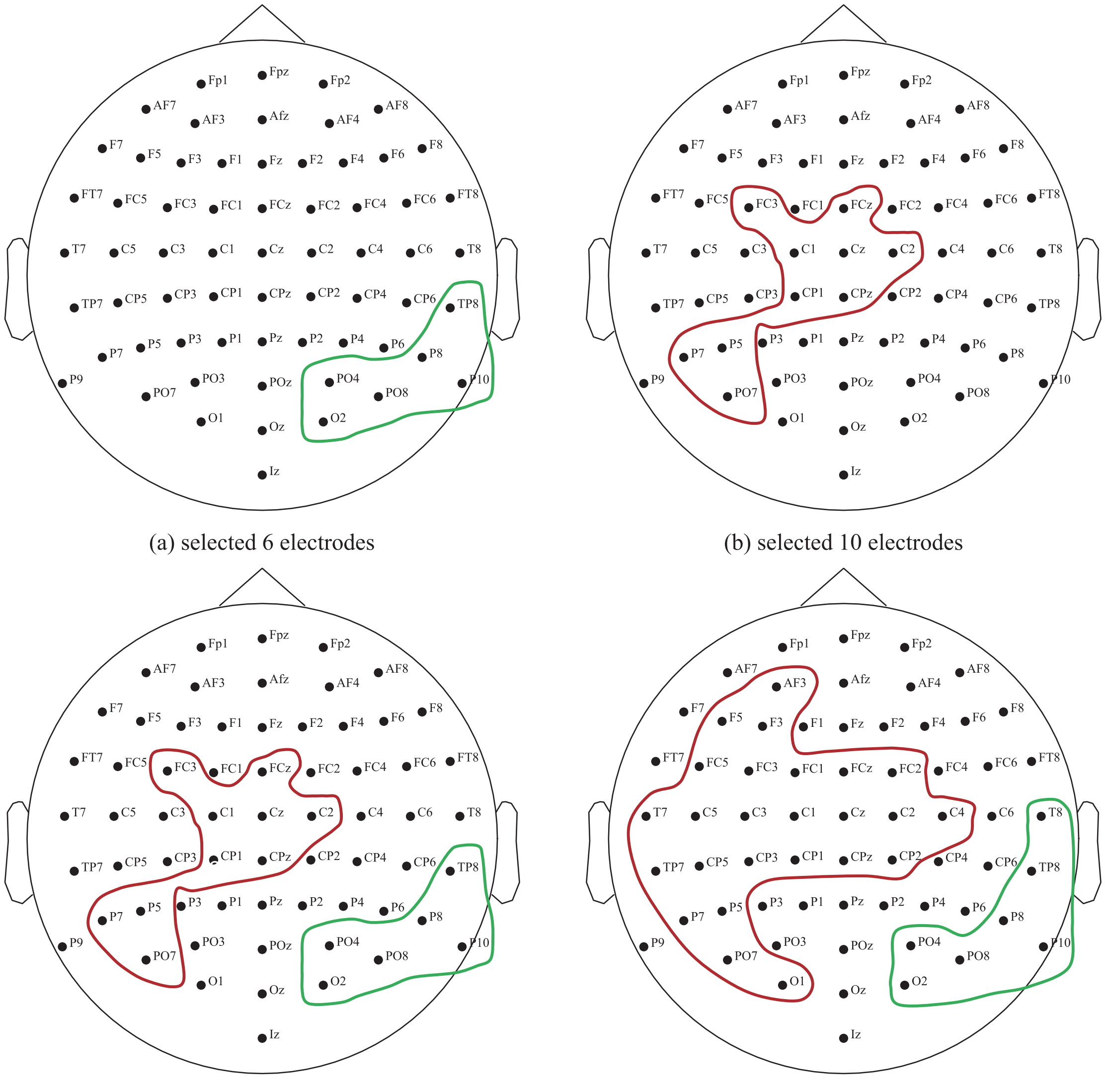
For the brain–computer interface system, using fewer electrodes can greatly improve the feasibility and convenience of application. In this section, we investigated how the decoding performance of the proposed HS-STDCN model varies with fewer electrodes. Based on the results shown in figures 6–8, we selected different numbers of electrodes from more contributed areas to imagined speech recognition. These selected electrodes were referred to the location of the parietal lobe, left posterior temporal lobe, and left inferior frontal lobe. In addition, we selected several electrodes from the right temporal lobe and right occipital lobe because of the differences between the fruit words and animal words. Figure 9 depicts the selected electrodes, and the classification results are shown in table 5. We have two observations from table 5:

(a) The use of fewer electrodes achieved comparable results to those obtained using all 64 electrodes. The performance based on 32 electrodes decreased by only 4% compared with all 64 electrodes. Similarly, using 16 electrodes also achieved a significant decoding performance, which was 2% lower than using 32 electrodes. The results verify that the language function areas are activated and contribute more to the imagined speech decoding task. Moreover, at the expense of a small performance, the number of electrodes can be greatly reduced by selecting the most important electrodes.

(b) The right hemisphere also played an important role in imagined speech recognition. The performance based on six electrodes selected from the right hemisphere, which showed an activity difference, had a 2% improvement over that of ten electrodes. This result indicates that the right hemisphere is also involved in the processing of language-related activities, which is consistent with the observation of [54].

3.2.5. *Analysis of image reaction and imagined speech production*

In our imagined speech experiments, the subjects were asked to generate the corresponding word in their minds when they saw a picture. The induced brain activities were quite different from just reacting to the exposure of different types of pictures. From the electrode activity maps in figures 6–8, we can see that the parietal lobe, left posterior temporal lobe, and left inferior frontal lobe make more contributions for decoding. These areas have been proven to be related


Fpz Fp1 Fp2
AF7 AF3 Afz AF4 AF8
F7 F5 F3 F1 Fz F2 F4 F6 F8
FT7 FC5 FC3 FC1 FCz FC2 FC4 FC6 FT8
T7 C5 C3 C1 Cz C2 C4 C6 T8
TP7 CP5 CP3 CP1 CPz CP2 CP4 CP6 TP8
P9 P7 P5 P3 P1 Pz P2 P4 P6 P8 P10
PO7 PO3 POz PO4 PO8
O1 Oz O2
Iz

(a) selected 6 electrodes

(b) selected 10 electrodes

(c) selected 16 electrodes

(d) selected 32 electrodes

Figure 9. Position of selected electrodes. Red line encloses the electrodes that make more contribution for imagined speech, and green line encloses the electrodes that show different activity distributions in the right hemisphere between the fruit words and animal words.

Table 5. Acc ± std of the classification performance (%) using different numbers of electrodes. The Bold value indicates using all 64 electrodes achieved highest performance.

Number of electrodes	Acc ± Std
6	41.91 ± 5.50
10	39.90 ± 7.85
16	48.23 ± 6.85
32	50.28 ± 6.86
64	**54.31** ± 5.22

to semantic processing, word reading, and comprehension [51–53]. To some extent, this verified that the classification results are based on imagined speech activities.

To show the differences between the image reaction and imagined speech production, we designed a comparative experiment to collect the EEG signal reacting to different types of images. Concretely, the subjects are informed to only react to the exposure of pictures without pronouncing corresponding words in their mind. Note that except for the mental task difference, the procedure of data acquisition and processing is consistent with imagined speech experiments. We conducted classification experiments on the collected EEG data using our proposed method HS-STDCN. The comparative classification results are shown in table 6. We can see that the classification performance of just reacting to different types of pictures is quite poorer than the imagined speech tasks. We performed paired t-test statistical analysis at a significance level of 0.05 for the experimental results of HS-STDCN between the image reaction group and imagined speech group. The test result (p = 0.009) shows that the results of imagined speech group are significantly better than the results of image reaction group. The comparative classification results and paired t-test statistical analysis validate our

Table 6. Acc of the classification performance (%) for image reaction and imagined speech based on HS-STDCN. The Bold values indicate imagined speech achieved higher performance.

Subject	Image reaction	Imagined speech
S1	36.25	**52.19**
S8	35.63	**57.50**
S9	30.63	**48.75**

classification results are the recognition of imagined speech production process rather than the response to the exposure of different types of pictures.

4. Conclusion

In this study, we proposed an end-to-end network called HS-STDCN to learn temporal and spatial dependencies from imagined speech EEG signals. To model the temporal variations well, we used hybrid-scale temporal convolution layers on the input EEG data to learn temporal information at different levels. A depthwise spatial convolution layer was then used to characterize the intrinsic spatial relationships. Based on the deep spatial-temporal representation, dilated convolution blocks were employed to learn the long-range discriminative features. The experimental results demonstrated that the proposed HS-STDCN method achieved a better recognition performance than the other comparison methods. Based on the HS-STDCN, we analyzed the impact of word semantics on the decoding performance of imagery words, investigated the important brain regions for imagined speech decoding, and explored the use of fewer electrodes to achieve comparable performance. In future work, we will conduct cross-subject experiments to explore a more general and efficient imaginary speech decoding mechanism.

Data availability statement

The data generated and/or analysed during the current study are not publicly available for legal/ethical reasons but are available from the corresponding author on reasonable request.

Acknowledgment

This study was supported in part by the National Key Research and Development Project of China under Grant 2018YFB2202400; the NSFC under Grant 61672404, Grant 61875157, Grant 61751310, Grant 61836008, and Grant 61632019; the National Defense Basic Scientific Research Program of China under Grant JCKY2017204B102; the Science and Technology Plan of Xi'an under Grant 20191122015KYPT011JC013; the Scientific Research Program Funded by the Shaanxi Provincial Education Department under Grant 20JY022; the Fundamental Research Funds of the Central Universities of China under Grant QTZX2107, Grant RW200141 and JB211907; the Industry-University-Academy Cooperation Program of Xidian University-Chongqing IC Innovation research Institute under Grant No. CQIRI-2021CXY-Y14; the Natural Science Basic Research Program of Shaanxi under Grant 2021JQ-193; and the Project funded by China Postdoctoral Science Foundation under Grants 2021M692504 and 2021TQ0259.

ORCID iD

Yang Li https://orcid.org/0000-0002-5093-2151

References

[1] Wolpaw J R, Birbaumer N, McFarland D J, Pfurtscheller G and Vaughan T M 2002 Brain–computer interfaces for communication and control *Clin. Neurophysiol.* **113** 767–91
[2] Brumberg J S, Nieto-Castanon A, Kennedy P R and Guenther F H 2010 Brain–computer interfaces for speech communication *Speech Commun.* **52** 367–79
[3] Chaudhary U, Birbaumer N and Ramos-Murguialday A 2016 Brain–computer interfaces for communication and rehabilitation *Nat. Rev. Neurol.* **12** 513–25
[4] Chen X, Wang Y, Gao S, Jung T P and Gao X 2015 Filter bank canonical correlation analysis for implementing a high-speed SSVEP-based brain–computer interface *J. Neural Eng.* **12** 046008
[5] Li Q, Liu S, Li J and Bai O 2015 Use of a green familiar faces paradigm improves P300-speller brain–computer interface performance *PLoS One* **10** 1–15
[6] Doud A J, Lucas J P, Pisansky M T and He B 2011 Continuous three-dimensional control of a virtual helicopter using a motor imagery based brain–computer interface *PLoS One* **6** 1–10
[7] Nicolas-Alonso L F and Gomez-Gil J 2012 Brain computer interfaces, a review *Sensors* **12** 1211–79
[8] Alderson-Day B and Fernyhough C 2015 Inner speech: development, cognitive functions, phenomenology, and neurobiology *Psychol. Bull.* **141** 931–65
[9] Martin S, Iturrate I, Millán J D R, Knight R T and Pasley B N 2018 Decoding inner speech using electrocorticography: progress and challenges toward a speech prosthesis *Front. Neurosci.* **12** 422
[10] Penfield W and Roberts L 2014 *Speech and Brain Mechanisms* vol 62 (Princeton, NJ: Princeton University Press)
[11] Gracco V L, Tremblay P and Pike B 2005 Imaging speech production using fMRI *Neuroimage* **26** 294–301
[12] Wise R, Chollet F, Hadar U, Friston K, Hoffner E and Frackowiak R 1991 Distribution of cortical neural networks involved in word comprehension and word retrieval *Brain* **114** 1803–17
[13] Pei X, Leuthardt E C, Gaona C M, Brunner P, Wolpaw J R and Schalk G 2011 Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition *Neuroimage* **54** 2960–72
[14] Leuthardt E, Pei X M, Breshears J, Gaona C, Sharma M, Freudenburg Z, Barbour D and Schalk G 2012 Temporal evolution of gamma activity in human cortex during an overt and covert word repetition task *Front. Hum. Neurosci.* **6** 99
[15] Hermes D, Miller K J, Vansteensel M J, Edwards E, Ferrier C H, Bleichner M G, van Rijen P C, Aarnoutse E J and Ramsey N F 2014 Cortical theta wanes for language *Neuroimage* **85** 738–48
[16] DaSalla C S, Kambara H, Sato M and Koike Y 2009 Single-trial classification of vowel speech imagery using common spatial patterns *Neural Netw.* **22** 1334–9

[17] Wang L, Zhang X, Zhong X and Zhang Y 2013 Analysis and classification of speech imagery EEG for BCI *Biomed. Signal Process. Control* **8** 901–8

[18] Min B, Kim J, Park H J and Lee B 2016 Vowel imagery decoding toward silent speech BCI using extreme learning machine with electroencephalogram *BioMed Res. Int.* **2016** 2618265

[19] Nguyen C H, Karavas G K and Artemiadis P 2017 Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features *J. Neural Eng.* **15** 016002

[20] Sereshkeh A R, Trott R, Bricout A and Chau T 2017 EEG classification of covert speech using regularized neural networks *IEEE-ACM Trans. Audio Speech Lang.* **25** 2292–300

[21] González-Castañeda E F, Torres-García A A, Reyes-García C A and Villaseñor-Pineda L 2017 Sonification and textification: proposing methods for classifying unspoken words from EEG signals *Biomed. Signal Process. Control* **37** 82–91

[22] Qureshi M N I, Min B, Park H J, Cho D, Choi W and Lee B 2017 Multiclass classification of word imagination speech with hybrid connectivity features *IEEE Trans. Biomed. Eng.* **65** 2168–77

[23] Cooney C, Folli R and Coyle D 2018 Mel frequency cepstral coefficients enhance imagined speech decoding accuracy from EEG *2018 29th Irish Signals and System Conf. (ISSC)* (IEEE) pp 1–7

[24] Lee S H, Lee M and Lee S W 2020 Neural decoding of imagined speech and visual imagery as intuitive paradigms for BCI communication *IEEE Trans. Neural Syst. Rehabil. Eng.* **28** 2647–59

[25] Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z 2016 Rethinking the inception architecture for computer vision *2016 Conf. on Computer Vision and Pattern Recognition (CVPR)* pp 2818–26

[26] Chiu C C *et al* 2018 State-of-the-art speech recognition with sequence-to-sequence models *2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE) pp 4774–8

[27] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I 2017 Attention is all you need *Advanced Neural Information Processing Systems* pp 6000–10

[28] Lawhern V J, Solon A J, Waytowich N R, Gordon S M, Hung C P and Lance B J 2018 EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces *J. Neural Eng.* **15** 056013

[29] Schirrmeister R T, Springenberg J T, Fiederer L D J, Glasstetter M, Eggensperger K, Tangermann M, Hutter F, Burgard W and Ball T 2017 Deep learning with convolutional neural networks for EEG decoding and visualization *Hum. Brain Mapp.* **38** 5391–420

[30] Saha P, Fels S and Abdul-Mageed M 2019 Deep learning the EEG manifold for phonological categorization from active thoughts *ICASSP 2019–2019 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE) pp 2762–6

[31] Saha P, Abdul-Mageed M and Fels S 2019 SPEAK YOUR MIND! towards imagined speech recognition with hierarchical deep learning *Proc. Interspeech 2019* pp 141–5

[32] Cooney C, Folli R and Coyle D 2019 Optimizing layers improves CNN generalization and transfer learning for imagined speech decoding from EEG *2019 IEEE Int. Conf. on Systems, Man and Cybernetics (SMC)* (IEEE) pp 1311–16

[33] Oppenheim G M and Dell G S 2008 Inner speech slips exhibit lexical bias, but not the phonemic similarity effect *Cognition* **106** 528–37

[34] Linkenkaer-Hansen K, Nikouline V V, Palva J M and Ilmoniemi R J 2001 Long-range temporal correlations and scaling behavior in human brain oscillations *J. Neurosci.* **21** 1370–7

[35] Wester M 2006 Unspoken speech-speech recognition based on electroencephalography Master's Thesis Universitat Karlsruhe (TH)

[36] Porbadnigk A, Wester M, Calliess J and Schultz T 2009 EEG-based speech recognition: impact of temporal effects *BIOSIGNALS 2009—Proc. Int. Conf. on Bio-Inspired Systems and Signal Processing* (SciTePress) pp 376–81

[37] Spampinato C, Palazzo S, Kavasidis I, Giordano D, Souly N and Shah M 2017 Deep learning human mind for automated visual classification *2017 Conf. on Computer Vision and Pattern Recognition (CVPR)* pp 6809–17

[38] Li R, Johansen J S, Ahmed H, Ilyevsky T V, Wilbur R B, Bharadwaj H M and Siskind J M 2021 The perils and pitfalls of block design for EEG classification experiments *IEEE Trans. Pattern Anal. Mach. Intell.* **43** 316–33

[39] Glaser W R 1992 Picture naming *Cognition* **42** 61–105

[40] Johnson C J, Paivio A and Clark J M 1996 Cognitive components of picture naming *Psychol. Bull.* **120** 113–39

[41] Defeyter M A, Russo R and McPartlin P L 2009 The picture superiority effect in recognition memory: a developmental study using the response signal procedure *Cogn. Dev.* **24** 265–73

[42] Hockley W E 2008 The picture superiority effect in associative recognition *Mem. Cogn.* **36** 1351–9

[43] Peirce J W 2007 PsychoPy—Psychophysics software in Python *J. Neurosci. Methods* **162** 8–13

[44] Gramfort A *et al* 2013 MEG and EEG data analysis with MNE-Python *Front. Neurosci.* **7** 267

[45] Kowialiewski B and Majerus S 2020 The varying nature of semantic effects in working memory *Cognition* **202** 104278

[46] Ishii T 2015 Semantic connection or visual connection: investigating the true source of confusion *Lang. Teach Res.* **19** 712–22

[47] Skrandies W 1998 Evoked potential correlates of semantic meaning—a brain mapping study *Cogn. Brain Res.* **6** 173–83

[48] Skrandies W, Chiu M J and Lin Y 2004 The processing of semantic meaning in Chinese words and evoked brain topography *Brain Topogr.* **16** 255–9

[49] Kaneshiro B, Perreau Guimaraes M, Kim H S, Norcia A M and Suppes P 2015 A representational similarity analysis of the dynamics of object processing using single-trial EEG classification *PLoS One* **10** 1–27

[50] Pexman P M and Yap M J 2018 Individual differences in semantic processing: insights from the calgary semantic decision project *J. Exp. Psychol.: Learn. Mem. Cogn.* **44** 1091–112

[51] Bzdok D, Hartwigsen G, Reid A, Laird A R, Fox P T and Eickhoff S B 2016 Left inferior parietal lobe engagement in social cognition and language *Neurosci. Biobehav. Rev.* **68** 319–34

[52] Buchsbaum B R, Hickok G and Humphries C 2001 Role of left posterior superior temporal gyrus in phonological processing for speech perception and production *Cogn. Sci.* **25** 663–78

[53] D'Ausilio A, Craighero L and Fadiga L 2012 The contribution of the frontal lobe to the perception of speech *J. Neurolinguistics* **25** 328–35

[54] Alexandrou A M, Saarinen T, Mäkelä S, Kujala J and Salmelin R 2017 The right hemisphere is highlighted in connected natural speech production and perception *Neuroimage* **152** 628–38