**University of Macau**

**Institute of Collaborative Innovation**


澳門大學

UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

# EEG-based Imagined Speech Recognition Using Attention-based Multi-scale Fusion Convolutional Neural Network

***by***

**Qihao Xu, MC364202**

Graduation Project Report submitted in partial fulfillment
of the requirements of the Degree of
Master of Science in Cognitive Neuroscience

Project Supervisor

Prof. Feng Wan

23 04 2025

DECLARATION

I sincerely declare that:

1. I am the sole authors of this report,
2. All the information contained in this report is certain and correct to the best of my knowledge,
3. I declare that the thesis here submitted is original except for the source materials explicitly acknowledged and that this thesis or parts of this thesis have not been previously submitted for the same degree or for a different degree, and
4. I also acknowledge that I am aware of the Rules on Handling Student Academic Dishonesty and the Regulations of the Student Discipline of the University of Macau.

Signature : 

Name : Qihao Xu

Student No. : MC364202

Date : 23 April 2025

# Abstract

Imagined Speech, as a subfield of Brain-Computer Interface (BCI), involves the utilization of devices such as EEG, ECoG, fMRI, fNIRs to capture signals from the human brain. Specifically, Imagined Speech focuses on the extraction of imagined spoken words within the human brain. Deep Learning has found extensive applications in the field of Brain-Computer Interface, with Convolutional Neural Networks (CNNs) being the prevailing methodologies. However, traditional single-scale CNNs are currently unable to meet the high-accuracy requirements. This is attributed to their incapability of extracting sufficient information from complex EEG signals. As a result, multi - scale CNNs, as novel approaches, have increasingly attracted the attention of researchers. In this project, an innovative multi-scale CNN method, namely the Attention-based Multi-scale Fusion Convolutional Neural Network (AMFCNN) developed from Attention-based Dual-scale Fusion Convolutional Neural Network (ADFCNN), is proposed for the Imagined Speech classification task. This method extracts spectral and spatial features from EEG signals across three distinct temporal scales. Branch-I extracts global and detailed spatial information at a large temporal scale. Branch-II extracts abundant and detailed spatial information at a medium temporal scale. Branch-III extracts abundant and detailed spatial information at a small temporal scale. Subsequently, the extracted features are fused together, and with the aid of the self-attention mechanism, the data is classified into different categories. Within-subject and cross-subject experiments were carried out on the ASU imagined speech dataset. The proposed AMFCNN method demonstrated improvements of up to 10.44% and 5.93% in the within-subject and cross-subject cases respectively, when compared

with the baseline models. In the within-subject case, AMFCNN achieved average accuracies of 55.44%, 64.72%, 76.18%, 42.40%, and 38.33% for the average, long words average, short long words average, short words average, and vowels average respectively. In the cross-subject case, AMFCNN achieved average accuracies of 48.67%, 54.71%, 54.52%, 44.80%, and 40.33% for the average, long words average, short long words average, short words average, and vowels average respectively. Also, T-SNE visualization and ablation experiments were conducted, which further validated the effectiveness and superiority of AMFCNN over ADFCNN and other baseline methods in the task of imagined speech.

# Table of Contents

# List of Tables and Figures

# CHAPTER 1. Introduction

## 1.1 Brain Computer Interface

Brain-Computer Interface (BCI) represents an innovative technological approach that constructs a non-muscular pathway between the human brain and external devices [11]. There have been many studies exploring the applications of BCI in different fields, such as healthcare, entertainment, gaming, communication, control, and research. BCI has become one of the frontiers of future research interest.

In contemporary implementations, BCIs typically capture signals acquired with devices such as Electrocorticography (ECoG), Magnetoencephalography (MEG), Electroencephalography (EEG), Functional Magnetic Resonance Imaging (fMRI), and Functional Near-Infrared Spectroscopy (fNIRS). Then the data collected by these devices are then will be processed with machine learning or deep learning methods and translated into commands for controlling computers, prosthetic arms, and other devices.

Among the several BCI signal acquisition devices, EEG is more frequently used in BCI because it can be easily set up in any environment where measurements are required [14]. It measures the electrical potential of neurons that fire simultaneously by placing electrodes on the scalp [15]. Over the years, several paradigms of EEG-based BCI have been developed, including steady-state visual evoked potentials (SSVEP), event related potentials (ERP), emotion, and motor imagery (MI) [9].

## 1.2 Imagined Speech Recognition

Despite the substantial advancements achieved by BCI systems employing the aforementioned paradigms in recent years, they are confronted with challenges such as slow processing speed and high user effort requirements [11]. These challenges impede their practical applicability in real-world scenarios. To tackle these issues, a novel BCI paradigm, namely imagined speech, has been proposed. Imagined speech, also known as speech imagery, covert speech, inner speech, or verbal thinking, refers to the activity in which individuals produce first-person motor imagery of speaking without moving any articulators [12].

As a relatively new BCI paradigm, imagined speech captures brain signals from cerebral regions associated with speaking. These signals are then decoded into either voice or text, enabling convenient device control or communication for individuals with speech impairments. Although the transmission of information from a computer to the brain remains currently unfeasible, the transfer of signals from the human brain to a computer has reached a practical stage.

This technology offers paralyzed individuals an additional means to communicate and control devices. The technique can not only be used by paralyzed people, but for the general population, imagined speech based BCI can also enhance work efficiency due to its convenient control and communication capabilities. In the region of scientific research, as the analysis of brain signals continues to evolve, it also holds the potential to deepen our understanding of neuroscience.

Numerous methods have been devised for the EEG-based imagined speech recognition task. In the field of EEG-based BCI, traditional approaches typically

encompass two distinct components: feature extraction and classification [9]. Feature extraction techniques for imagined speech predominantly involve signal processing methods like common spatial pattern (CSP) [30], fast Fourier transform (FFT) [25], power spectral density (PSD) [23], discrete wavelet transform (DWT) [16], wavelet packet decomposition (WPD) [29], difference mode decomposition (DMD) [31]. Despite the traditional signal processing methods, feature extraction methods based on deep learning, statistical features [17], are also popular.

In recent years, the classifiers in imagined speech brain-computer interfaces (BCIs) predominantly rely on Machine Learning or Deep Learning techniques. In the field of Deep Learning, methods such as convolutional neural networks (CNNs), long-short term memory networks (LSTMs) [23] [25] [30], and transformers [4] are commonly used. In Machine Learning, algorithms like random forest (RF) [31], support vector machine (SVM) [6], and k-nearest neighbors (KNN) [6] are frequently employed.

For example, Lee et al. [1] introduced an end-to-end framework Siamese neural network. Datta et al. [2] proposed a method using a multi-channel convolutional neural network (MC-CNN) to identify the grammatical classes (verbs or nouns) of imagined speech. Panachakel et al. [3] applied Mean phase coherence (MPC) with a shallow neural network. Ahn et al. [4] proposed a multiscale convolutional transformer that operates on the spatial, spectral, and temporal features of imagined speech electroencephalogram (EEG) signals. Kamble et al. [5] designed a method combining smoothed pseudo-Wigner-Ville distribution (SPWVD) and CNN to classify imagined words. In another study by Kamble et al. [6], they proposed an adaptive Rational Dilation Wavelet Transform (RADWT) method with particle swarm optimization (PSO) adaptively tuning parameters. This method was then

experimentally combined and evaluated with six different machine learning algorithms: Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF), Bagging, AdaBoost, and Rotation Forest. Kamble et al. [7] used time-frequency representation (TFR) plots of EEG signals with smoothed pseudo-Wigner-Ville distribution (SPWVD) and classified them using a convolutional neural network (CNN). Macias et al. [8] proposed a method named Capsules for Speech Imagery Analysis (CapsK-SI) based on a Capsule Neural Network.

## 1.3 Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is a method extract global and local features from the input data and holds the features in a smaller size output. It uses a convolution operation which use a small kernel screening through the whole data matrix and conduct a weighted sum of the datapoint in the kernel to get a single output in each step of the screening.

In the field of brain-computer interfaces (BCIs), CNN has been widely used to extract spatial and spectral features from electroencephalogram (EEG) signals. For instance, one-dimensional convolution along the temporal dimension can effectively extract spectral information present in EEG signals, while one-dimensional convolution along the channel dimension is capable of extracting spatial features.

A number of studies have applied CNN in BCI research like PSD+CNN [19], Entropy+CNN [20], Spectro-Spatio-Temporal EEG Representation+CNN [32], Length Classifier with Words CNNs [33], FBCSP+CNN [21], FFT+CNN [22], DWT+CNN [39], TCNN with CNN [34], APIT-MEMD+CNN with Voting [35], CNN with PCA+LDA [36], CSP+CNN[24], Time-frequency and Statistical

Characteristics+CNN [38], CWT+CNN [26], STFT+CNN [27], STFT+Multi-Channel CNN [28], Shift-Invariant Sparse Coding+CNN [37] and so on.

Several CNN architectures have been developed and are widely adopted in BCI tasks, even serving as baseline methods in the BCI field. Representative examples include EEGNet [42], DeepConvNet [43], ShallowConvNet [43], and FBCNet [44]. Currently, a large number of studies use CNN for feature extraction in their BCI studies. The prevalence of CNN in BCI applications reflects its effectiveness in handling BCI related tasks.

## 1.4 Multi-scale CNN

However, most of the aforementioned studies rely solely on single-scale Convolutional Neural Networks (CNNs) for imagined speech recognition. Given the complexity of electroencephalogram (EEG) signals, which encompass information across multiple frequencies, time scales, and regions within the human brain, the exploration of multi-scale CNNs has recently been noticed by researchers. For example, a hybrid-scale CNN has been proposed to account for inter subject differences [39]. A deep multi-scale CNN is designed to extract frequency and time features at multiple scales [40], while a 3D 3-scale CNN is used to extract temporal-spatial features directly from raw EEG signals [41]. An Attention-based Dual-scale Fusion CNN (ADFCNN) is developed which combines multi-scale CNN with self-attention mechanism [9]. These multi-scale CNNs have demonstrated their effectiveness in extracting features in different scales, outperforming traditional single-scale CNNs in accuracy. Consequently, multi-scale CNNs have the potential to become the dominant approach in future BCI research.

The Attention-based Dual-scale Fusion CNN (ADFCNN) is a deep learning model that has been demonstrated to be more effective in the Motor Imagery recognition task [9] compared to traditional single-scale CNNs. It is a convolutional neural network that fuses two scales of EEG data based on an attention mechanism. The incorporation of a self-attention mechanism in the multi-scale CNN, when concatenating data of different scales, endows it with high accuracy, setting it apart among multi-scale CNNs.

There is evidence of the similarity between the motor imagery task and the imagined speech task. In both processes, the Primary Motor Cortex and Premotor Cortex are active [13]. Also, both types of brain-computer interfaces (BCIs) extract spectral and temporal features from EEG signals and distinguish between different representations using labels. Given the potential of ADFCNN in this area, it is necessary to validate the effectiveness of ADFCNN in imagined speech tasks. But the ADFCNN method has its imperfections, the scale of feature extraction in its Branch-II is not small enough, which may result in significant information loss. Considering the limitation of ADFCNN, where only two scales of feature extraction cannot capture all the information across all scales, we have developed a method based on ADFCNN, named AMFCNN. AMFCNN adds an additional, smaller scale to ADFCNN, aiming to extract more complete and abundant information from EEG signals for imagined speech recognition.

The con

## 1.5 Objectives

The objectives of the project are to:

- Validate the effectiveness of motor imagery method ADFCNN on imagined speech recognition task to show the potential practicability of transferring methods between motor imagery and imagined speech tasks.

- Propose a new deep learning method AMFCNN fixing a drawback of ADFCNN for imagined speech recognition BCI and verify the effectiveness of AMFCNN by comparing it with other deep learning methods to prove the effectiveness of multi-scale CNNs for imagined speech.

## 1.6 Organization

The organization of the report is as follows: Chapter 2 is Materials and Methods including dataset description, data preprocessing, model architecture and experiments setup, Chapter 3 is Results including classification performance, feature visualization and ablation study, Chapter 4 is Discussion and Chapter 5 is Conclusion.

# CHAPTER 2. Materials and Methods

## 2.1 Dataset Description

In our project, the ASU imagined speech dataset [10] is employed. In this dataset, the data was collected using a 64-electrode BrainProducts ActiCHamp amplifier in accordance with the 10/20 international system. The dataset involved EEG imagined speech data from15 healthy subjects, consisting of 11 males and 4 females, with ages ranging from 22 to 32 years old.

During the data collection process, the participants were directed to imagine speaking one of eight words or phonemes: “cooperate”, “independent”, “in”, “out”, “up”, /a/, /i/, /u/. Each trial had a duration of 7 × T seconds (where T represents a time unit, with 1s for vowels and short words, and 1.4s for long words), followed by a 2 second rest stage. In the beginning 4T period, there was a periodic beep sound occurring within each T interval, and a visual cue was presented in the whole 7T periods. The participants were required to periodically imagine one of the words or phonemes in response to the visual cue, aided by the beep sound. In the last 3T, during which there was no beep sound, the EEG signals corresponding to the periodically imagined speech were recorded.

One session consisted of 100 trials (except for 2 subjects who only had 80 trials) for each word or phoneme within a subtask. The subtasks included: Long Words (“cooperate”, “independent”), Short Long Words (“cooperate”, “in”), Short Words (“in”, “out”, “up”), and Vowels (/a/, /i/, /u/). Given that none of the subjects completed all four subtasks, and some subjects only participated in one subtask, for the sake of convenience, we regarded the same subject in different subtasks as

different individuals. As a result, the collected data contains 1200 trials for the long words subtask, 1800 trials for the short words subtask, 1120 trials for the short long words subtask, and 2400 trials for the vowels subtask.

## 2.2 Data Preprocessing

After the collection of the imagined speech EEG data, a series of preprocessing steps were conducted. First, the data underwent a band-pass filtering within the frequency range of [8-70] Hz. This was achieved using a 5th-order Butterworth filter, which effectively attenuated frequencies outside this specified range. Subsequently, a Notch filter centered at 60Hz was applied. The purpose of this Notch filter was to eliminate the power-line interference that typically occurs at 60Hz in electrical systems.

Among the collected data, 4 channels were electrooculogram (EOG) data. These EOG data were recorded for the removal of EOG artifacts, which are electrical signals generated by eye movements and blinks that can contaminate the EEG recordings.

Finally, the processed EEG data, which had already been filtered and had EOG artifacts removed, was downsampled. The original sampling rate of 1000Hz was reduced to 256Hz. This downsampling process was carried out to reduce the computational load while still retaining the relevant information in the EEG signals for subsequent analysis in the project.

## 2.3 Model Architecture

In the original Attention-based Dual-scale Fusion Convolutional Neural Network (ADFCNN), the preprocessed EEG data is represented as $X \in R^{C \times T}$, where C denotes

the number of channels and T represents the number of recorded time points. In the ASU dataset, for each trial in each subject, C=60 and T=1280. According to the original study of ADFCNN [9], temporal convolutions with larger kernel sizes are better at extracting broader EEG frequency signals. On the contrary, smaller kernel sizes are better at higher frequency EEG information.

In ADFCNN, its Branch-I is designed to extract global and detailed spatial information on a large temporal scale. Meanwhile, Branch-II extracts abundant and detailed spatial information on a smaller temporal scale. By combining the advantages of both larger and smaller kernels, ADFCNN is capable of capturing multi-scale temporal and spatial features within the input EEG signals. Also, ADFCNN incorporates a self-attention feature fusion module to integrate the features extracted by the dual-scale convolutional neural network. The process of ADFCNN can be visualized in Fig. 1. First, the input EEG data is duplicated. Then, spectral and spatial convolutions are performed at two scales separately on the two duplicates. Subsequently, the data from the two scales are concatenated, and then self-attention is applied to the concatenated data. Finally, the data is classified into imagined speech classes using a dense layer.

The Attention-based Multi-scale Fusion Convolutional Neural Network (AMFCNN) is developed based on ADFCNN. In AMFCNN, an additional smaller scale, similar to Branch-II in ADFCNN (which is good at extracting abundant and detailed spatial information through standard spatial convolution), is added, which is able to extract a smaller scale of features which makes the model able to extract more information from a smaller scale that the model of ADFCNN missed. As a result, AMFCNN can extract information across more scales, rather than being limited to only two scales, making it more effective in extracting features from EEG data.

In the first branch of the temporal-spatial CNN (Branch-I), the extraction of large-scale features from imagined speech EEG data follows a similar approach to the first branch of ADFCNN. Given EEG data of the shape C×T, a temporal convolution is performed using large-scale kernels of shape [1×125], with the number of kernels being F. These kernels are designed to extract low-frequency information. Subsequently, F separable spatial convolution layers, each of size [C×1], are applied to extract global spatial features. After that, a point-wise convolution layer with a kernel size of [1×1] is conducted to integrate the separated spatial features. Finally, the processed data passes through a pooling layer of size [1×32].

The second branch of the temporal-spatial CNN (Branch-II) is responsible for extracting medium-scale features from EEG data. This is conceptually similar to the second branch in ADFCNN, which extracts small-scale features. The preprocessed EEG data first undergoes F medium-scale temporal convolutions with a size of [1×30] to obtain medium-frequency information. Then, the data is processed using F×F standard convolution kernels of shape [C×1]. Finally, an average pooling layer of size [1×75] is applied to the processed data.

The third branch of the temporal-spatial CNN (Branch-III) extracts small-scale features from EEG data, where the scale is even smaller than that of the second branch in ADFCNN. The preprocessed EEG data initially goes through F small-scale temporal convolutions with a size of [1×10] to acquire high-frequency information. Subsequently, the data is processed by F×F standard convolution kernels of shape [C×1]. Finally, an average pooling layer of size [1×10] is carried out on the data.

After the feature extraction from the three branches of the temporal-spatial CNN, the data from these three branches are concatenated to form a feature map of shape

[F×1× (T1 +T2 +T3)]. Then, self-attention is applied to this data to focus on specific extracted features. Finally, the data passes through a fully connected linear layer, which classifies the processed feature map into one of the imagined speech classes.

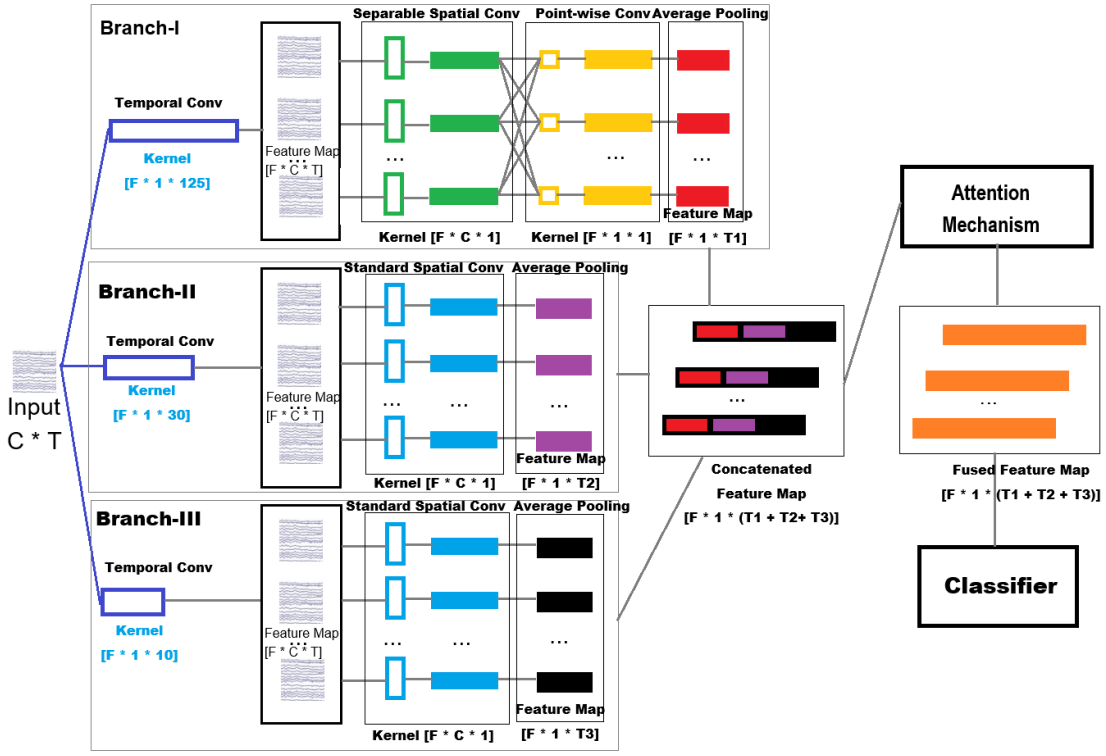



Fig. 1. Framework of the proposed AMFCNN.

Table I

The Parameters of AMFCNN Architecture

| *MODULE* | *LAYER* | *SIZE* | *PARAMS* | *OUTPUT SHAPE* |
|---|---|---|---|---|
| ***BRANCH-I*** | Temporal Convolution | [F*1*125] | F*125 | [F*C*T] |
| | Separable Spatial Convolution | [F*C*1] | F*C | [F*1*T]] |
| | Point-wise Convolution | [F*1*1] | F*F | [F*1*T] |
| | Average Pooling | [1*32] | / | [F*1*T1] |

| | | | | |
|---|---|---|---|---|
| ***BRANCH-II*** | Temporal Convolution | [F*1*30] | F*30 | [F*C*T] |
| | Standard Spatial Convolution | [F*C*1] | C*F*F | [F*1*T] |
| | Average Pooling | [1*75] | / | [F*1*T2] |
| ***BRANCH-III*** | Temporal Convolution | [F*1*10] | F*10 | [F*C*T] |
| | Standard Spatial Convolution | [F*C*1] | C*F*F | [F*1*T] |
| | Average Pooling | [1*10] | / | [F*1*T3] |
| ***FUSION*** | Concatenation | / | / | [F*1*(T1+T2+T3)] |
| | Attention-based | / | C*C*3 | [F*1*(T1+T2+T3)] |
| ***CLASSIFIER*** | Linear | / | N*F*(T1+T2+T3) | [N*1] |

**T1=DIMENSION OF FEATURE EXTRACTED BY BRANCH I, T2= DIMENSION OF FEATURE EXTRACTED BY BRANCH II, T3=DIMENSION OF FEATURE EXTRACTED BY BRANCH III, N= NUMBER OF CLASSES**

## 2.4 Experiments Setup

In the imagined speech recognition task, the proposed method will be contrasted with four representative deep learning baselines: EEGNet [42], DeepConvNet [43], ShallowConvNet [43], FBCNet [44], and ADFCNN [9]. For all these models, the learning rate is set to 0.001, the number of epochs is 100, the batch size is 16, the window length is 3, and the weight decay is 0.075. 5-fold cross-validation is also used for the training of the models.

Adam is used as the optimizer, Cross-Entropy Loss is used as the loss function, PyTorch is employed as the deep learning framework, and an NVIDIA RTX 4050 Laptop GPU is used for the training of the models. The performances of these

models in the imagined speech task are evaluated using accuracy, considering both within-subject and cross-subject cases.

The code is publicly available at https://github.com/XQH139/AMFCNN-Imagined-Speech.

# CHAPTER 3. Results

## 3.1 Classification Performance

Table II presents the results of performance in the within-subject case, where the EEG data of each individual subject is independently trained using the aforementioned six deep learning methods.

AMFCNN demonstrates the highest average accuracy of 55.44% in within-subject training, marking an improvement of 4.09% over ADFCNN. Although, as depicted in Fig. 3, this improvement might not be significant, the enhancement achieved by AMFCNN is significantly greater than that of the DeepConvNet method.

In the long words subtask (involving Subjects 1-5 which is a 2-class classification of the words “independent” and “cooperate”), AMFCNN attains the highest accuracy of 64.12%, which is 8.25% higher than the second-best method, ADFCNN. As illustrated in Fig. 4(a), it exhibits a substantial improvement compared to all other methods except ADFCNN.

For the short long words subtask (comprising Subjects 6-11 which is a 2-class classification of the words “cooperate” and “in”), AMFCNN reaches an accuracy of 76.18%. This value is 7.12% higher than that of the second-highest method, ADFCNN. According to Fig. 4(b), it shows a significant improvement relative to ShallowConvNet and a highly significant improvement when compared to other methods, excluding ADFCNN.

In the short words subtask (with Subjects 12-16 which is a 3-class classification of the words “up”, “out”, and “in”), AMFCNN achieves an accuracy of 42.40%, which

is the same with the highest-performing method, ADFCNN. As shown in Fig. 4(c), it demonstrates a significant improvement over EEGNet and DeepConvNet.

Regarding the vowels subtask (involving Subjects 17-22 which is a 3-class classification of the phonemes /a/, /i/, and /u/), AMFCNN reaches an accuracy of 38.33%, ranking fourth among the six methods. Notably, it is still 1% higher than ADFCNN, and as indicated in Fig. 4(d), there is no significant improvement in this particular subtask.

Even though the performance improvement of AMFCNN compared to ADFCNN is not significant, the accuracy of AMFCNN is higher than that of ADFCNN. In the subtasks of long words, short long words, and short words, AMFCNN's performance is significantly superior to that of other methods, excluding ADFCNN.

As observed from Fig. 2, when compared to other methods, AMFCNN achieves the highest accuracy in the imagined speech classification of the EEG data on Subjects 1, 2, 5, 6, 8, 9, 10, 11, and 12. More specifically, Subjects 1 and 2 achieve an accuracy of 67.65%, and Subject 5 attains an accuracy of 76.47%, in the long words subtask. In the short long words subtask, Subject 6 reaches 76.47%, Subject 8 reaches 70.59%, Subject 9 reaches 67.65%, Subject 10 reaches 77.78%, and Subject 11 reaches 85.19%, respectively. In the short words subtask, Subject 12 achieves an accuracy of 52.00%.

In the within-subject case, these performance results indicate that AMFCNN has a stronger capability in handling the tasks of long words and short long words, while showing relatively lower performance in the subtasks of short words and vowels when compared with the other five methods.

In addition, we carried out cross-subject training for the six methods to assess the generalization ability of the models across different subjects. This involved combining the EEG data of all subjects to train a single model.

Table III presents a comparison of the performances of different deep learning methods in the cross-subject case. AMFCNN demonstrates the highest average accuracy of 48.67% during cross-subject training, representing an improvement of 3.92% over ADFCNN. It also outperforms the second-highest model, ShallowConvNet, by 2.79%. Although, as shown in Fig. 6, this improvement is not highly significant.

In the long words subtask (featuring Subjects 1-5 which is a 2-class classification of the words “independent” and “cooperate”), AMFCNN attains the second-highest accuracy of 54.71%. This value is 2.94% lower than that of the best performing method, ADFCNN. As illustrated in Fig. 7(a), it does not exhibit any significant improvement compared to other methods.

For the short long words subtask (comprising Subjects 6-11 which is a 2-class classification of the words “cooperate” and “in”), AMFCNN reaches an accuracy of 54.52%. This is 4.79% higher than ADFCNN, but it ranks only third among the models, trailing behind ShallowConvNet with 56.63% and EEGNet with 54.65%. According to Fig. 7(b), there is no significant improvement when compared to other methods.

In the short words subtask (involving Subjects 12-16 which is a 3-class classification of the words “up”, “out”, and “in”), AMFCNN achieves an accuracy of 44.80%. This is 5.6% higher than the second-highest method, ADFCNN. As shown in Fig. 7(c), it demonstrates a very significant improvement over EEGNet and DeepConvNet.

Regarding the vowels subtask (with Subjects 17-22 which is a 3-class classification of the phonemes /a/, /i/, and /u/), AMFCNN reaches an accuracy of 40.33%, which is the highest among the methods. It is 6.66% higher than ADFCNN and 5.66% higher than the second-highest method, EEGNet. As indicated in Fig. 7(d), there is a significant improvement compared to DeepConvNet and a very significant improvement compared to FBCNet.

Even though the performance improvement of AMFCNN compared to ADFCNN is not overly significant, the accuracy of AMFCNN is slightly higher than that of ADFCNN. In the subtasks of short words and vowels, AMFCNN's performance is significantly superior to that of other methods, excluding ADFCNN.

As observed from Fig. 5, when compared to other methods, AMFCNN achieves the highest accuracy in the imagined speech classification of the EEG data from Subjects 6, 8, 13, 14, 15, 18, 19, 21, and 22. More specifically, Subjects 6 and 8 achieve accuracies of 55.88% and 67.65%, respectively, in the short long words subtask. Subjects 13, 14, and 15 attain accuracies of 46.00%, 50.00%, and 52.00%, respectively, in the short words subtask. Subjects 18, 19, 21, and 22 achieve accuracies of 46.00%, 46.00%, 40.00%, and 48.00%, respectively, in the vowels subtask.

Contrary to the within-subject case, in the cross-subject case, the performance results indicate that AMFCNN has a stronger capability in handling the tasks of short words and vowels, while showing relatively lower performance in the subtasks of long words and short long words when compared to the other five methods. This suggests that AMFCNN has a better generalization ability across subjects compared to the other five methods.

Table II

Within-Subject Performance Comparison

| | CLASSIFICATION ACCURACY | | | | | |
|---|---|---|---|---|---|---|
| **SUBJECTS** | FBCNet | ShallowConvNet | DeepConvNet | EEGNet | ADFCNN | AMFCNN |
| **SUB1** | 58.82% | 50.00% | 52.94% | 47.06% | 58.82% | **67.65%** |
| **SUB2** | 61.76% | 52.94% | 44.12% | 47.06% | 50.00% | **67.65%** |
| **SUB3** | 50.00% | 55.88% | 52.94% | 41.18% | 47.00% | 52.94% |
| **SUB4** | 44.12% | 55.88% | 44.12% | 61.76% | **67.65%** | 55.88% |
| **SUB5** | 50.00% | 47.06% | 58.82% | 61.76% | 55.88% | **76.47%** |
| **SUB6** | 67.65% | 64.71% | 41.18% | 44.12% | 58.82% | **76.47%** |
| **SUB7** | 64.71% | 82.35% | 73.53% | 61.76% | **85.29%** | 79.41% |
| **SUB8** | 64.71% | 55.88% | 47.06% | 55.88% | 64.71% | **70.59%** |
| **SUB9** | 61.76% | 52.94% | 47.06% | 55.88% | 50.00% | **67.65%** |
| **SUB10** | 66.67% | 66.67% | 70.37% | 62.96% | 74.07% | **77.78%** |
| **SUB11** | 77.78% | 70.37% | 51.85% | 70.37% | 81.48% | **85.19%** |
| **SUB12** | 38.00% | 44.00% | 40.00% | 40.00% | 44.00% | **52.00%** |
| **SUB13** | 36.00% | 40.00% | 28.00% | 34.00% | **46.00%** | 42.00% |
| **SUB14** | 42.00% | 28.00% | 30.00% | 40.00% | **44.00%** | 38.00% |
| **SUB15** | 44.00% | 38.00% | 40.00% | 34.00% | 42.00% | 40.00% |
| **SUB16** | 38.00% | 38.00% | 36.00% | 32.00% | 36.00% | **40.00%** |
| **SUB17** | 40.00% | 40.00% | 42.00% | 44.00% | **44.00%** | 36.00% |
| **SUB18** | 42.00% | 38.00% | 44.00% | 46.00% | 38.00% | 44.00% |

| | | | | | | |
|---|---|---|---|---|---|---|
| **SUB19** | 48.00% | 28.00% | 40.00% | 38.00% | 40.00% | 42.00% |
| **SUB20** | 36.00% | 30.00% | 22.00% | 28.00% | 22.00% | 36.00% |
| **SUB21** | 40.00% | 36.00% | 36.00% | 48.00% | 40.00% | 42.00% |
| **SUB22** | 34.00% | 52.00% | 48.00% | 40.00% | 40.00% | 30.00% |
| **AVG_LONG** | 52.94% | 52.35% | 50.59% | 51.76% | **55.87%** | **64.12%** |
| **AVG_LONGSHORT** | 67.21% | 65.49% | 55.17% | 58.50% | **69.06%** | **76.18%** |
| **AVG_SHORT** | 39.60% | 37.60% | 34.80% | 36.00% | **42.40%** | **42.40%** |
| **AVG_VOWEL** | 40.00% | 37.33% | 38.67% | 40.67% | **37.33%** | **38.33%** |
| **AVG** | 50.27% | 48.49% | 45.00% | 46.99% | **51.35%** | **55.44%** |

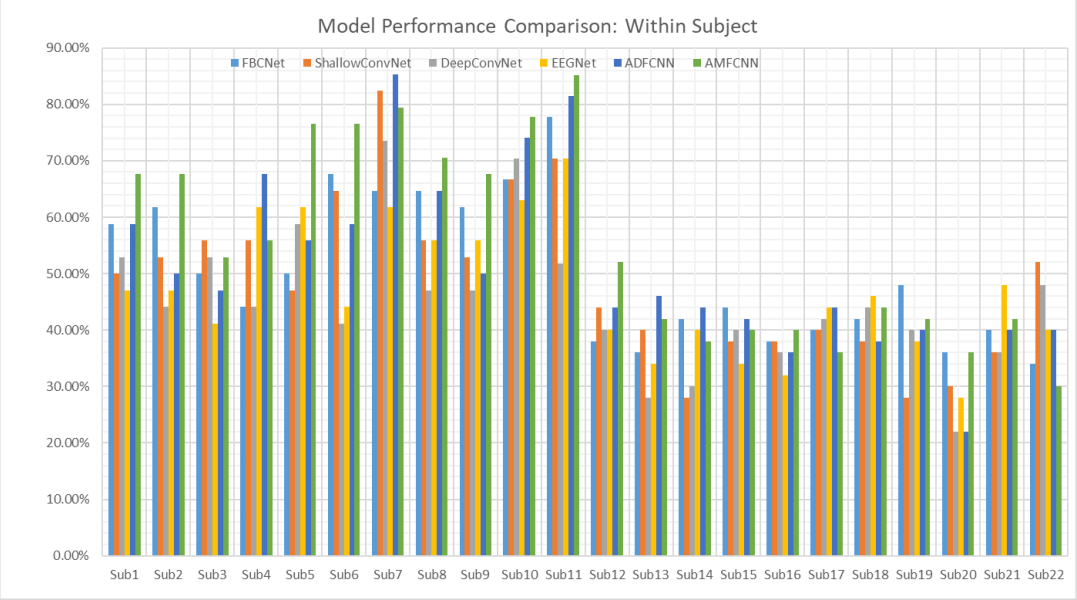



Fig. 2. Within-Subject Performance Comparison.

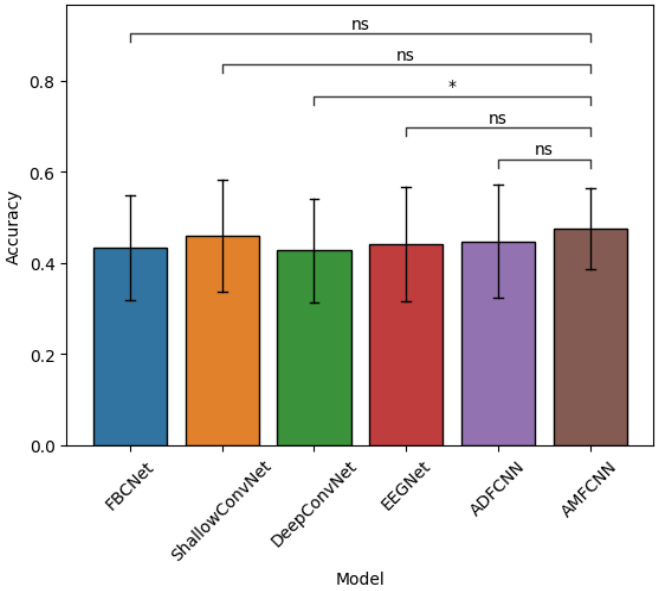
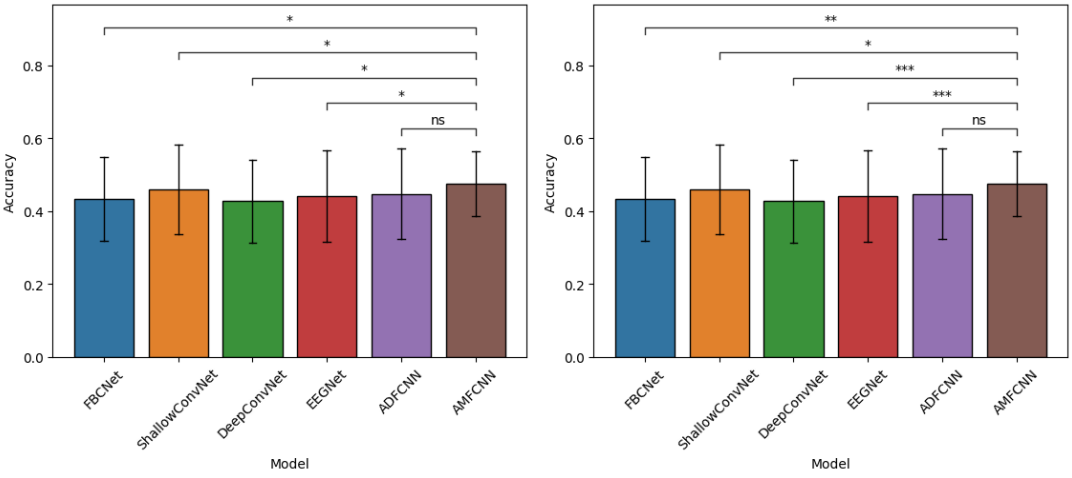
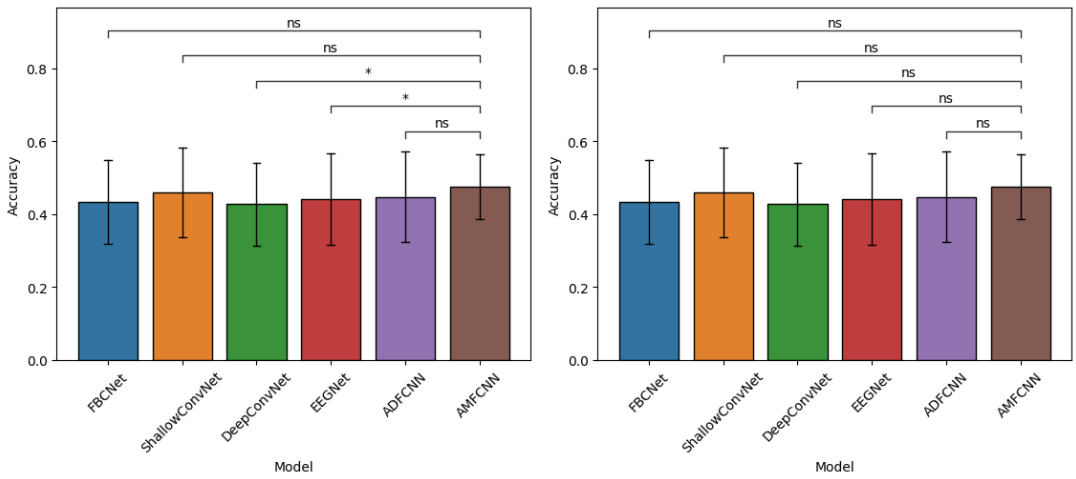



Fig. 3. Within-Subject Classification Performance Comparison, ns(Not Significant): 0.05 < p <= 1, *: 0.01 < p <= 0.05, **: 0.001 < p <= 0.01, ***: 0.0001 < p <= 0.001, ****: p <= 0.0001.




(a) (b)




(c) (d)

Fig. 4. Within-Subject Subtasks Classification Performance Comparison: (a) Long Words Classification Performance Comparison, (b) Short Long Words Classification Performance Comparison, (c) Short Words Classification Performance Comparison, (d) Vowels Classification Performance Comparison, ns(Not Significant): 0.05 < p <= 1, *: 0.01 < p <= 0.05, **: 0.001 < p <= 0.01, ***: 0.0001 < p <= 0.001, ****: p <= 0.0001.

Table III

Cross-Subject Performance Comparison

| | **CLASSIFICATION ACCURACY** | | | | | |
|---|---|---|---|---|---|---|
| **SUBJECTS** | FBCNet | ShallowConvNet | DeepConvNet | EEGNet | ADFCNN | AMFCNN |
| **SUB1** | 50.00% | 55.88% | 50.00% | 50.00% | 50.00% | 52.94% |
| **SUB2** | 61.76% | 67.65% | 61.76% | 64.71% | 61.76% | 64.71% |
| **SUB3** | 50.00% | 52.94% | 50.00% | 58.82% | 52.94% | 55.88% |
| **SUB4** | 52.94% | 47.06% | 50.00% | 47.06% | **58.82%** | 47.06% |
| **SUB5** | 52.94% | 50.00% | 44.12% | 52.94% | **64.71%** | 52.94% |
| **SUB6** | 41.18% | 50.00% | 52.94% | 47.06% | 52.94% | **55.88%** |
| **SUB7** | 52.94% | 67.65% | 52.94% | 58.82% | 50.00% | **64.71%** |

| | | | | | | |
|---|---|---|---|---|---|---|
| **SUB8** | 55.88% | 58.82% | 58.82% | 64.71% | 61.76% | **67.65%** |
| **SUB9** | 61.76% | 55.88% | 55.88% | 64.71% | 55.88% | 50.00% |
| **SUB10** | 40.74% | 48.15% | 51.85% | 44.44% | 29.63% | 44.44% |
| **SUB11** | 55.56% | 59.26% | 51.85% | 48.15% | 48.15% | 44.44% |
| **SUB12** | 40.00% | 46.00% | 36.00% | 26.00% | **46.00%** | 38.00% |
| **SUB13** | 38.00% | 44.00% | 32.00% | 40.00% | 38.00% | **46.00%** |
| **SUB14** | 34.00% | 32.00% | 28.00% | 32.00% | 26.00% | **50.00%** |
| **SUB15** | 42.00% | 40.00% | 36.00% | 32.00% | 44.00% | **52.00%** |
| **SUB16** | 44.00% | 30.00% | 34.00% | 32.00% | 42.00% | 42.00% |
| **SUB17** | 36.00% | 36.00% | 32.00% | 42.00% | **46.00%** | 34.00% |
| **SUB18** | 36.00% | 46.00% | 30.00% | 34.00% | 38.00% | **46.00%** |
| **SUB19** | 34.00% | 32.00% | 38.00% | 24.00% | 22.00% | **46.00%** |
| **SUB20** | 26.00% | 26.00% | 24.00% | 38.00% | **38.00%** | 28.00% |
| **SUB21** | 26.00% | 36.00% | 28.00% | 34.00% | 34.00% | **40.00%** |
| **SUB22** | 22.00% | 28.00% | 42.00% | 36.00% | 24.00% | **48.00%** |
| **AVG_LONG** | 53.53% | 54.71% | 51.18% | 54.71% | **57.65%** | **54.71%** |
| **AVG_LONGSHORT** | 51.34% | 56.63% | 54.05% | 54.65% | **49.73%** | **54.52%** |
| **AVG_SHORT** | 38.80% | 36.40% | 32.40% | 35.60% | **39.20%** | **44.80%** |
| **AVG_VOWEL** | 30.00% | 34.00% | 32.33% | 34.67% | **33.67%** | **40.33%** |
| **AVG** | 43.35% | 45.88% | 42.74% | 44.16% | **44.75%** | **48.67%** |

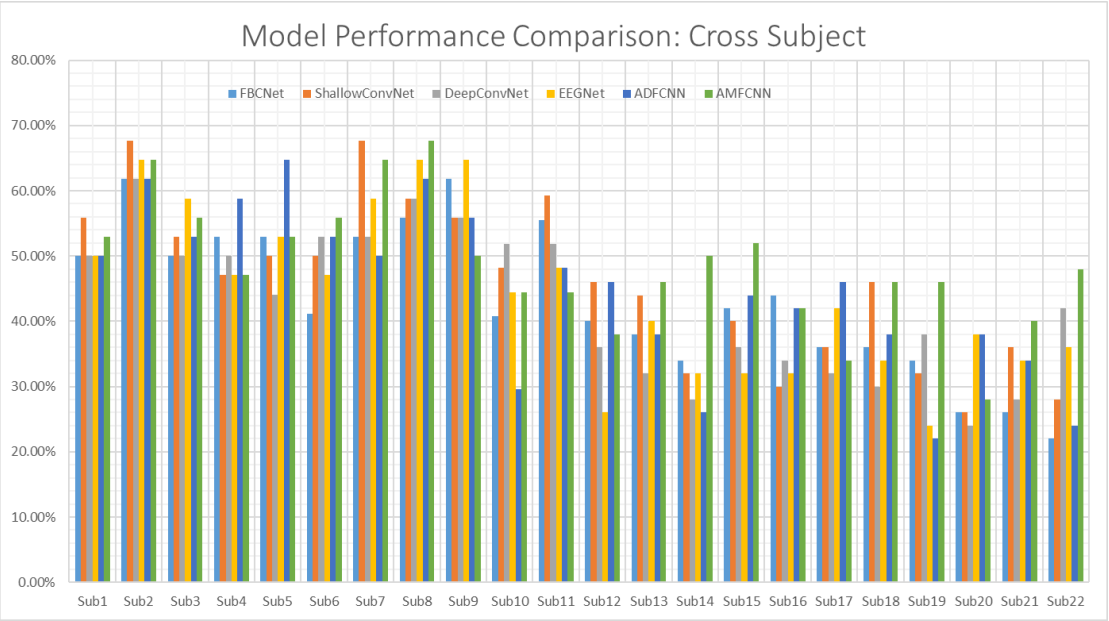



Fig. 5. Cross-Subject Performance Comparison.

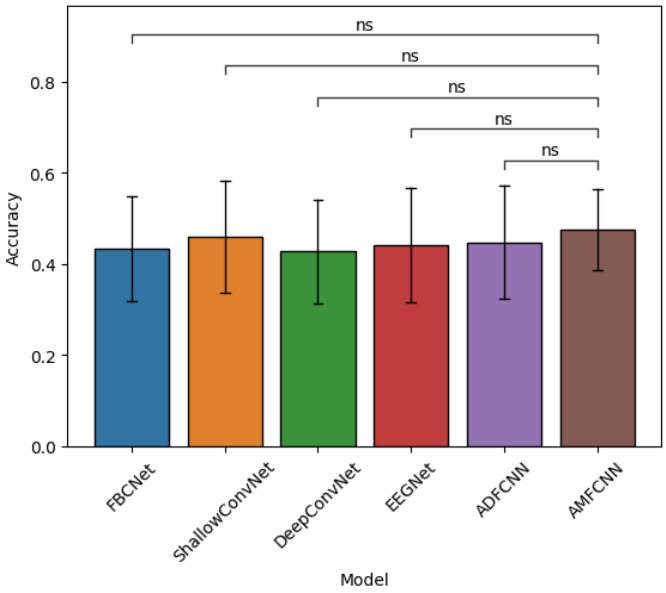
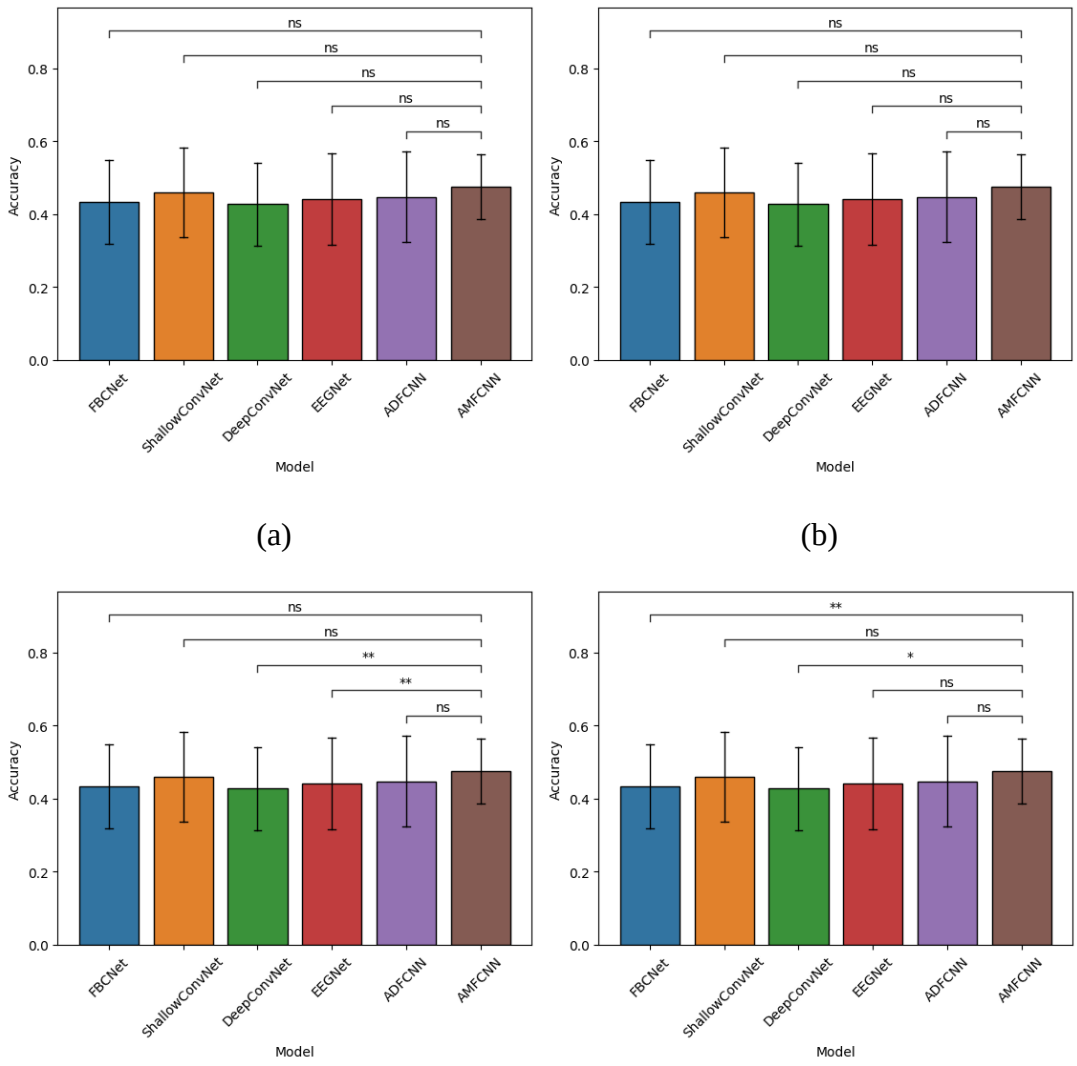



Fig. 6. Cross-Subject Classification Performance Comparison, ns(Not Significant): 0.05 < p <= 1, *: 0.01 < p <= 0.05, **: 0.001 < p <= 0.01, ***: 0.0001 < p <= 0.001, ****: p <= 0.0001.




(a) (b)




(c) (d)

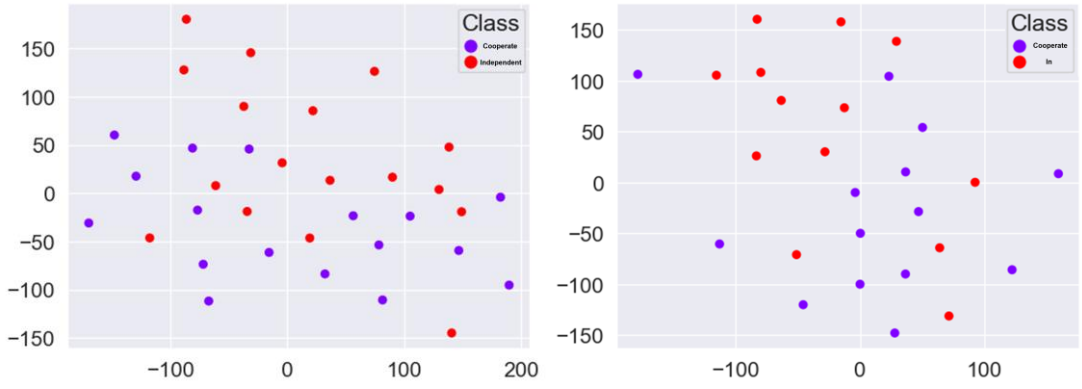
Fig. 7. Cross-Subject Classification Performance Comparison: (a) Long Words Classification Performance Comparison, (b) Short Long Words Classification Performance Comparison, (c) Short Words Classification Performance Comparison, (d) Vowels Classification Performance Comparison, ns(Not Significant): 0.05 < p <= 1, *: 0.01 < p <= 0.05, **: 0.001 < p <= 0.01, ***: 0.0001 < p <= 0.001, ****: p <= 0.0001.

## 3.2 Feature Visualization

To more comprehensively demonstrate the effectiveness of the three branches of the temporal-spatial Convolutional Neural Network (CNN) within the Attention-based Multi-scale Fusion Convolutional Neural Network (AMFCNN), the t-distributed

Stochastic Neighbor Embedding (t - SNE) technique is employed. This method can visualize the high-dimensional features extracted by these three branches. Specifically, features extracted by the three branches from four subjects who exhibit the best performance in the subtasks are used to generate t-SNE representation scatter plots, considering both the cases of individual branches and within-subject/cross-subject.

Notably, with the exception of the two plots corresponding to the long words subtask in the within-subject case (Fig. 8(a)) and the short long words subtask in the within-subject case (Fig. 8(b)), none of the other result plots display a clearly divided representation. This observation proves the advantage of incorporating Branch-III into the original Attention-based Dual-scale Fusion Convolutional Neural Network (ADFCNN) to form AMFCNN, as compared to relying solely on the original two branches.




(a) (b)

Fig. 8. t-SNE Visualization of Feature Extracted from Branch-III (a) t-SNE Visualization of Feature Extracted from Branch-III in the case of Within-Subject of long words subtask, (b) t-SNE Visualization of Feature Extracted from Branch-III in the case of Within-Subject of short long words subtask.

## 3.3 Ablation Experiments

To test the effectiveness of the different branches in the proposed model in the task of imagined speech recognition, we performed ablation experiments on the ASU dataset and designed the variant models in Table IV:

- Branch-I: This variant model only consists of one Branch-I rather than 3 branches of AMFCNN described in Table I, which means we deleted the Branch-II, Branch-III and the concatenation layer from the AMFCNN to make this variant model.

- Branch-II: This variant model only consists of one Branch-II rather than 3 branches of AMFCNN described in Table I, which means we deleted the Branch-I, Branch-III and the concatenation layer from the AMFCNN to make this variant model.

- Branch-III: This variant model only consists of one Branch-III rather than 3 branches of AMFCNN described in Table I, which means we deleted the Branch-I, Branch-II and the concatenation layer from the AMFCNN to make this variant model.

- Branch-I, II: This variant model only consists of Branch-I and Branch II rather than 3 branches of AMFCNN described in Table I, which means we deleted the Branch-III from the AMFCNN to make this variant model.

- Branch-I, III: This variant model only consists of Branch-I and Branch III rather than 3 branches of AMFCNN described in Table I, which

means we deleted the Branch-II from the AMFCNN to make this variant model.

- Branch-II, III: This variant model only consists of Branch-II and Branch III rather than 3 branches of AMFCNN described in Table I, which means we deleted the Branch-I from the AMFCNN to make this variant model.

Fig. 9 shows the results of the within-subject ablation experiments. As presented in Table V, although the average performance of the variant model with only Branch-III is 49.07%, which is lower than the 52.62% performance achieved by the variant model with only Branch-II, and the performance of the variant model consisting of Branch-I and Branch-III is 51.53%, slightly lower than the 51.74% accuracy of the variant model with Branch-I and Branch-II, the performance of the Attention-based Multi-scale Fusion Convolutional Neural Network (AMFCNN) still surpasses that of the Branch-I, II variant model.

Fig. 10 shows the results of the cross-subject ablation experiments. Contrary to the within-subject case, the average performance of the variant model with only Branch-III is 46.78%, which exceeds the 44.23% performance of the variant model with only Branch-II. However, the performance of the variant model composed of Branch-I and Branch-III is 46.03%, lower than the 48.26% accuracy of the variant model with Branch-I and Branch-II. As shown in Table VI, the performance of AMFCNN is still higher than that of the Branch-I, II variant model.

These results indicate that Branch-III indeed extracts features that are distinct from those extracted by Branch-I and Branch-II, rather than merely a replication of Branch-II. This proves the unique contribution of Branch-III to the overall feature

extraction of the AMFCNN model in both within-subject and cross-subject imagined speech recognition tasks.

Table IV

The Variant Models of Ablation Experiment

| | **BRANCH-I** | **BRANCH-II** | **BRANCH-III** | **BRANCH-I, II** | **BRANCH-I, III** | **BRANCH-II, III** | **AMFCNN** |
|---|---|---|---|---|---|---|---|
| **TEMPORAL CONVOLUTION LAYER** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **BRANCH-I (SEPARABLE SPATIAL CONVOLUTION LAYER)** | ✓ | × | × | ✓ | ✓ | × | ✓ |
| **BRANCH-II (STANDARD SPATIAL CONVOLUTION LAYER)** | × | ✓ | × | ✓ | × | ✓ | ✓ |
| **BRANCH-III (STANDARD SPATIAL CONVOLUTION LAYER)** | × | × | ✓ | × | ✓ | ✓ | ✓ |
| **AVERAGE POOLING LAYER** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **CONCATENATION LAYER** | × | × | × | ✓ | ✓ | ✓ | ✓ |
| **SELF-ATTENTION LAYER** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Table V

Within-Subject Ablation Experimental Results

| SUBJECTS | BRANCH-I | BRANCH-II | BRANCH-III | BRANCH-I,II | BRANCH-I,III | BRANCH-II,III | AMFCNN |
|---|---|---|---|---|---|---|---|
| **SUB1** | 64.71% | 38.24% | 58.82% | 41.18% | 52.94% | 44.12% | 67.65% |
| **SUB2** | 58.82% | 73.53% | 67.65% | 58.82% | 61.76% | 64.71% | 67.65% |
| **SUB3** | 50.00% | 52.94% | 52.94% | 55.88% | 64.71% | 47.06% | 52.94% |
| **SUB4** | 35.29% | 47.06% | 50.00% | 58.82% | 44.12% | 52.94% | 55.88% |
| **SUB5** | 67.65% | 61.76% | 50.00% | 64.71% | 44.12% | 50.00% | 76.47% |
| **SUB6** | 50.00% | 82.35% | 64.71% | 64.71% | 70.59% | 91.18% | 76.47% |
| **SUB7** | 76.47% | 76.47% | 73.53% | 82.35% | 67.65% | 88.24% | 79.41% |
| **SUB8** | 64.71% | 58.82% | 58.82% | 52.94% | 61.76% | 67.65% | 70.59% |
| **SUB9** | 64.71% | 64.71% | 61.76% | 73.53% | 55.88% | 70.59% | 67.65% |
| **SUB10** | 62.96% | 70.37% | 70.37% | 70.37% | 77.78% | 81.48% | 77.78% |
| **SUB11** | 62.96% | 81.48% | 62.96% | 88.89% | 70.37% | 85.19% | 85.19% |
| **SUB12** | 32.00% | 56.00% | 46.00% | 52.00% | 46.00% | 48.00% | 52.00% |
| **SUB13** | 32.00% | 40.00% | 38.00% | 38.00% | 40.00% | 24.00% | 42.00% |
| **SUB14** | 36.00% | 44.00% | 40.00% | 42.00% | 48.00% | 44.00% | 38.00% |
| **SUB15** | 44.00% | 42.00% | 36.00% | 42.00% | 44.00% | 46.00% | 40.00% |
| **SUB16** | 34.00% | 40.00% | 40.00% | 42.00% | 46.00% | 42.00% | 40.00% |
| **SUB17** | 28.00% | 36.00% | 30.00% | 30.00% | 50.00% | 34.00% | 36.00% |
| **SUB18** | 38.00% | 38.00% | 44.00% | 50.00% | 42.00% | 38.00% | 44.00% |
| **SUB19** | 36.00% | 34.00% | 36.00% | 30.00% | 40.00% | 26.00% | 42.00% |
| **SUB20** | 30.00% | 38.00% | 42.00% | 24.00% | 36.00% | 42.00% | 36.00% |
| **SUB21** | 44.00% | 42.00% | 26.00% | 38.00% | 40.00% | 46.00% | 42.00% |
| **SUB22** | 32.00% | 40.00% | 30.00% | 38.00% | 30.00% | 30.00% | 30.00% |
| **AVG_LONG** | 55.29% | **54.71%** | **55.88%** | **55.88%** | **53.53%** | 51.76% | **64.12%** |
| **AVG_LONGSHORT** | 63.63% | **72.37%** | **65.36%** | **72.13%** | **67.34%** | 80.72% | **76.18%** |
| **AVG_SHORT** | 35.60% | **44.40%** | **40.00%** | **43.20%** | **44.80%** | 40.80% | **42.40%** |
| **AVG_VOWEL** | 34.67% | **38.00%** | **34.67%** | **35.00%** | **39.67%** | 36.00% | **38.33%** |
| **AVG** | 47.47% | **52.62%** | **49.07%** | **51.74%** | **51.53%** | 52.87% | **55.44%** |

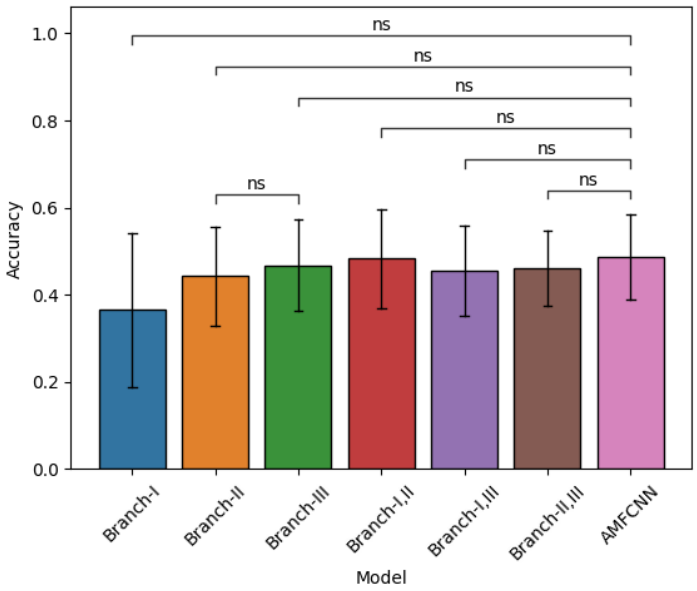



Fig. 9. Within-Subject Ablation Experimental Results. ns(Not Significant): 0.05 < p <= 1, *: 0.01 < p <= 0.05, **: 0.001 < p <= 0.01, ***: 0.0001 < p <= 0.001, ****: p <= 0.0001.

Table VI

Cross-Subject Ablation Experimental Results

| SUBJECTS | BRANCH-I | BRANCH-II | BRANCH-III | BRANCH-I, II | BRANCH-I, III | BRANCH-II, III | AMFCNN |
|---|---|---|---|---|---|---|---|
| **SUB1** | 50.00% | 52.94% | 58.82% | 50.00% | 52.94% | 47.06% | 52.94% |
| **SUB2** | 50.00% | 61.76% | 55.88% | 67.65% | 67.65% | 55.88% | 64.71% |
| **SUB3** | 50.00% | 41.18% | 52.94% | 50.00% | 58.82% | 55.88% | 55.88% |
| **SUB4** | 47.06% | 44.12% | 52.94% | 50.00% | 47.06% | 41.18% | 47.06% |
| **SUB5** | 50.00% | 58.82% | 47.06% | 50.00% | 50.00% | 52.94% | 52.94% |
| **SUB6** | 55.88% | 44.12% | 52.94% | 64.71% | 47.06% | 58.82% | 55.88% |
| **SUB7** | 47.06% | 64.71% | 67.65% | 70.59% | 52.94% | 52.94% | 64.71% |
| **SUB8** | 55.88% | 64.71% | 58.82% | 50.00% | 55.88% | 52.94% | 67.65% |
| **SUB9** | 61.76% | 52.94% | 61.76% | 67.65% | 55.88% | 52.94% | 50.00% |
| **SUB10** | 51.85% | 40.74% | 55.56% | 44.44% | 44.44% | 40.74% | 44.44% |
| **SUB11** | 44.44% | 37.04% | 40.74% | 40.74% | 48.15% | 59.26% | 44.44% |
| **SUB12** | 30.00% | 44.00% | 44.00% | 44.00% | 38.00% | 38.00% | 38.00% |
| **SUB13** | 34.00% | 42.00% | 46.00% | 44.00% | 42.00% | 52.00% | 46.00% |
| **SUB14** | 34.00% | 54.00% | 44.00% | 50.00% | 50.00% | 52.00% | 50.00% |
| **SUB15** | 32.00% | 30.00% | 40.00% | 38.00% | 38.00% | 44.00% | 52.00% |
| **SUB16** | 26.00% | 42.00% | 36.00% | 42.00% | 30.00% | 36.00% | 42.00% |
| **SUB17** | 22.00% | 36.00% | 34.00% | 46.00% | 48.00% | 40.00% | 34.00% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **SUB18** | 30.00% | 36.00% | 40.00% | 52.00% | 48.00% | 40.00% | 46.00% |
| **SUB19** | 8.00% | 36.00% | 34.00% | 36.00% | 26.00% | 36.00% | 46.00% |
| **SUB20** | 2.00% | 34.00% | 34.00% | 42.00% | 28.00% | 32.00% | 28.00% |
| **SUB21** | 2.00% | 30.00% | 44.00% | 40.00% | 36.00% | 36.00% | 40.00% |
| **SUB22** | 20.00% | 26.00% | 28.00% | 22.00% | 36.00% | 36.00% | 48.00% |
| **AVG_LONG** | 49.41% | **51.76%** | **53.53%** | **53.53%** | **55.29%** | 50.59% | **54.71%** |
| **AVG_LONG SHORT** | 52.81% | **50.71%** | **56.25%** | **56.35%** | **50.73%** | 52.94% | **54.52%** |
| **AVG_SHORT** | 31.20% | **42.40%** | **42.00%** | **43.60%** | **39.60%** | 44.40% | **44.80%** |
| **AVG_VOWEL** | 14.00% | **33.00%** | **35.67%** | **39.67%** | **37.00%** | 36.67% | **40.33%** |
| **AVG** | 36.54% | **44.23%** | **46.78%** | **48.26%** | **45.49%** | 46.03% | **48.67%** |

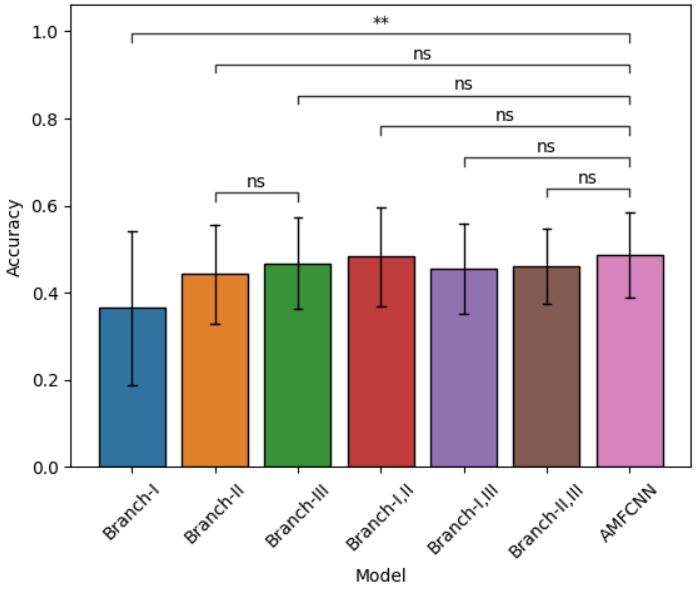



Fig. 10. Cross-Subject Ablation Experimental Results. ns(Not Significant): 0.05 < p <= 1, *: 0.01 < p <= 0.05, **: 0.001 < p <= 0.01, ***: 0.0001 < p <= 0.001, ****: p <= 0.0001.

# CHAPTER 4. Discussion

In recent years, several brain-computer interface (BCI) methods have predominantly revolved around Convolutional Neural Networks (CNNs), which have been empirically demonstrated to be highly effective in the domain of BCIs. However, the majority of these CNNs are single-scale architectures, capable of extracting information solely from a single scale of electroencephalogram (EEG) data. Consequently, the concept of multi-scale CNNs has increasingly captured the attention of researchers in the current era.

One such multi-scale CNN, the Attention-based Dual-scale Fusion Convolutional Neural Network (ADFCNN) [9], has exhibited remarkable performance in the motor imagery task. It extracts spectral and spatial features from EEG signals using CNNs at two distinct scales and subsequently fuses these features through a self-attention mechanism for the purpose of classification. Theoretically, this approach is also applicable to EEG signals related to imagined speech, given the similarities in the extraction of spatial and temporal features for classification between motor imagery and imagined speech and also the similarities of Primary Motor Cortex and Premotor Cortex activation during the tasks [13].

In our specific imagined speech recognition task, the classification accuracy of ADFCNN achieved an overall average of 51.35% in the within-subject case and 44.75% in the cross-subject case. When compared with benchmark methods such as FBCNet [42], ShallowConvNet [43], DeepConvNet [43], EEGNet [44], and our newly proposed Attention-based Multi-scale Fusion Convolutional Neural Network (AMFCNN), these accuracies ranked ADFCNN second in the within-subject case and third in the cross-subject case.

The experimental outcomes not only show the effectiveness of ADFCNN in motor imagery tasks but also in imagined speech recognition tasks. These results provide substantial evidence for the similarity between motor imagery tasks and imagined speech tasks. Also, they validate the feasibility of transferring BCI methods from one application area to another, demonstrating the potential for cross-task applicability within the broader field of BCIs.

Our newly proposed method for imagined speech recognition, the Attention-based Multi-scale Fusion Convolutional Neural Network (AMFCNN), incorporates an additional branch, namely Branch - III. This branch is designed with a standard spatial convolution structure similar to that of Branch-II in the original ADFCNN. The parameters of Branch-III are set to be smaller, enabling them to extract features at a smaller scale that the original two branches of ADFCNN failed to capture.

The improved performance of AMFCNN compared to ADFCNN is a strong evidence for the effectiveness of Branch-III. This additional branch can extract diverse and complementary features compared with the other two branches. The reason behind this is its ability to extract temporal features at a smaller scale using a temporal convolution layer and average pooling with smaller kernels. In this approach, there will be less information loss compared to the larger kernels used in the other two branches.

It is well-known that the convolution and average pooling processes will inevitably lead to information loss. During these operations, multiple data points within a kernel are aggregated into a single output data point, and there is no perfect inverse transformation to restore the original input data from the output. These operations are

non-reversible, and their application is premised on the assumption that sufficient information can be retained in the output.

Ideally, a method that retains more information should achieve higher accuracy. However, in reality, the data often contains noise, which can degrade performance if not properly processed by CNN, pooling operations or other signal processing methods. A neural network with a larger number of parameters requires more data to converge. In the imagined speech recognition or other brain-computer interface (BCI) applications, the available EEG data is typically limited, such a network demands more computational resources and time. Therefore, a balance between performance and resources is essential.

For the imagined speech recognition task, we can prioritize performance due to the low device burden, short time cost, and limited data requirements. This is precisely what Branch-III accomplishes. It focuses on preserving the original temporal information, which is reflected in its superior performance in t-SNE visualization compared to the other two branches. Simultaneously, the presence of more noise in the data does not offset the benefits brought about by the additional information retained by Branch-III.

However, in the ablation study, the performances of the variant model Branch-III do not outperform the similar variant model Branch-II. In the within-subject case, the average performance of variant Branch-II is higher than the average performance of variant Branch-III. In both the within-subject and cross-subject cases, the performances of the variant model Branch-I, III are lower than the performances of the variant model Branch-I, II, respectively. But the AMFCNN, which combines all three branches, attains a higher performance than any of the individual variant

models. The fact that AMFCNN's performance surpasses that of all other variants indicates that the impact of noise introduced by the more abundant data and additional parameters does not override the positive effect of increased information.

One plausible explanation for the lower performances of variant Branch-III and variant Branch-I, III compared with variant Branch-II and variant Branch-I, II is their different proportions of contributions from each scale to the whole AMFCNN. Among the three branches, Branch-II contributes the largest proportion of spatial-temporal features, followed by Branch-III and Branch-I. The distinction among the three branches is their scales of their convolution and pooling operations within the temporal-spatial domain. This implies that EEG signals in different time scales consist of distinct features, and these signals from various time scales are integrated together to form the EEG signals we collect. This explanation further validates the correctness and advancement of multi-scale CNNs, as it proves the significance of considering different scales for feature extraction from EEG data.

An interesting phenomenon has been observed. Although the overall performance of AMFCNN in the cross-subject case is lower than in the within-subject case, its performances in the subtasks of short words and vowels in the cross-subject case are higher than those in the within-subject case. This is a unique aspect not exhibited by ADFCNN. Typically, across all subtasks, the performance in the within-subject case is expected to be higher than in the cross-subject case.

The most likely cause of this result can be attributed to the presence of Branch-III. However, the exact mechanism behind this remains to be elucidated. Since ADFCNN does not show this phenomenon in the cross-subject case, the cause cannot be solely related to the self-attention mechanism or the linear classifier. Given that

AMFCNN's performance in the short words and vowels subtasks is higher in the cross-subject case than in the within-subject case, it must be related to the cross-subject training process itself.

It is plausible that Branch-III extracts additional features specific to the classes within these two subtasks, features that are distinct from those extracted by the other two branches. During cross-subject training, these features combine to integrate a significantly larger amount of information compared to either the within-subject or cross-subject training without Branch-III, thereby resulting in a notably higher performance. Another contributing factor could be that the limited number of trials used in the test stage might make small performance improvements indistinguishable, which make the impact of Branch-III in the cross-subject case for these particular subtasks unseen. Due to the limited data, neither of the data used for training or testing are enough for the imagined speech BCI. So, future researchers should focus more on data collection on imagined speech datasets because of the general problem of data shortage.

Though the objectives of the project are accomplished, we can also see that there are many limitations of AMFCNN. First is that the improvement of AMFCNN compared with ADFCNN is not so significant. Second is that the performances in the subtask of short words and vowels are still too low. Multi-scale CNNs with more scales should be studied, better branch architectures need to be experimented, and the parameters can also be improved to extract more information from the raw EEG data. Third is that more state-of-the-art methods for imagined speech recognition or other BCI methods should be considered in comparison. Fourth is that more imagined speech recognition datasets need to be tested to validate the generalization ability of the AMFCNN.

# CHAPTER 5. Conclusion

In conclusion, we evaluated the effectiveness of motor imagery BCI method, the Attention-based Dual-scale Fusion Convolutional Neural Network (ADFCNN) in the task of imagined speech recognition. We also proposed a new method adding a branch in smaller scale extracting smaller-scale features to ADFCNN called Attention-based Multi-scale Fusion Convolutional Neural Network (AMFCNN) which achieves higher performance compared with ADFCNN in both within-subject case and cross-subject case on the ASU imagined speech EEG dataset. Ablation experiments and t-SNE visualization are conducted and show the superiority of adding a smaller scale branch. The result shows the effectiveness of ADFCNN and AMFCNN in imagined speech recognition task, multi-scale CNNs have the potential to be widely used in future imagined speech even other fields of BCI.

# Acknowledgement

I’d like to thank my supervisor Prof. Feng Wan for his support in my study at the University of Macau.

I’d also like to show gratitude to Mr. Wei Tao for his patient guidance and advice in coding and writing of my project.

I acknowledge the use of AI tool Doubao to help me polish my writing in my project report. I declare that no content generated by AI has been presented and submitted as my own work.

# References


[1] Dong-Yeon Lee, M.L., Seong-Whan Lee, Classification of Imagined Speech Using Siamese Neural Network. IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2020.

[2] Datta, S. and N. Boulgouris, Recognition of grammatical class of imagined words from EEG signals using convolutional neural network. Neurocomputing, 2021. 465: p. 301-309.

[3] Panachakel, J.T. and R.A. G, Classification of Phonological Categories in Imagined Speech using Phase Synchronization Measure. Annu Int Conf IEEE Eng Med Biol Soc, 2021. 2021: p. 2226-2229.

[4] Ahn, H.J., et al., Multiscale Convolutional Transformer for EEG Classification of Mental Imagery in Different Modalities. IEEE Trans Neural Syst Rehabil Eng, 2022. PP.

[5] Kamble, A., P.H. Ghare, and V. Kumar, Deep-Learning-Based BCI for Automatic Imagined Speech Recognition Using SPWVD. Ieee Transactions on Instrumentation and Measurement, 2023. 72: p. 1-10.

[6] Kamble, A., P.H. Ghare, and V. Kumar, Optimized Rational Dilation Wavelet Transform for Automatic Imagined Speech Recognition. Ieee Transactions on Instrumentation and Measurement, 2023. 72: p. 1-10.

[7] Kamble, A., et al., Spectral Analysis of EEG Signals for Automatic Imagined Speech Recognition. Ieee Transactions on Instrumentation and Measurement, 2023. 72: p. 1-9.

[8] Macias-Macias, J.M., et al., Interpretation of a deep analysis of speech imagery features extracted by a capsule neural network. Comput Biol Med, 2023. 159: p. 106909.

[9] Tao, W., et al., ADFCNN: Attention-Based Dual-Scale Fusion Convolutional Neural Network for Motor Imagery Brain-Computer Interface. IEEE Trans Neural Syst Rehabil Eng, 2024. 32: p. 154-165.

[10] Nguyen, C.H., G.K. Karavas, and P. Artemiadis, Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features. Journal of Neural Engineering, 2018. 15(1): p. 016002.

[11] Zhang, Liying, et al. "Speech imagery decoding using EEG signals and deep learning: A survey." IEEE Transactions on Cognitive and Developmental Systems, 2024.

[12] S. Martin, I. Iturrate, J. d. R. Millán, R. T. Knight, and B. N. Pasley “Decoding inner speech using electrocorticography: rogress and challenges toward a speech prosthesis,” Front. Neurosci., vol. 12, p.422, Jun. 2018.

[13] Kunz, E.M., et al., Representation of verbal thought in motor cortex and implications for speech neuroprostheses. bioRxiv, 2024: p. 2024.10.04.616375.

[14] A. J. Casson, D. C. Yates, S. J. M. Smith, J. S. Duncan, and E. Rodriguez-Villegas, "Wearable electroencephalography," IEEE Eng. Med. Biol. Mag., vol. 29, no. 3, pp. 44–56, May 2010.

[15] P. L. Nunez and R. Srinivasan, Electric Fields of the Brain: The Neurophysics of EEG. London, U.K.: Oxford Univ. Press, 2006.

[16] J. T. Panachakel, R. A. Ganesan, and T. Ananthapadmanabha, "Common Spatial attern Based Data Augmentation Technique for Decoding magined Speech," in Proc. IEEE Int. Conf. Electron., Comput. Commun. Technol. (CONECCT), Jul. 2021, pp. 1-5.

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

[18] J. S. García-Salinas, A. A. Torres-García, C. A. Reyes-Garćia, and L. Villaseñor-Pineda, "Intra-subject class-incremental deep learning approach for EEG-based imagined speech recognition," Biomed. Signal Process. Control, vol. 81, Mar. 2023, Art. no. 104433.

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

[20] C. Cooney, A. Korik, R. Folli, and D. Coyle, "Evaluation of hyperpara- meter optimization in machine and deep learning methods for decodingimagined speech EEG," Sensors, vol. 20, no. 16, p. 4629, 2020.

[21] L. C. Sarmiento, S. Villamizar, O. López, A. C. Collazos, J. Sarmiento, and J. B. Rodríguez, "Recognition of EEG signals from imagined vowels using deep learning methods," Sensors, vol. 21, no. 19, p. 6503, 2021.

[22] F. Gasparini, E. Cazzaniga, and A. Saibene, "Inner speech recognition through electroencephalographic signals,". [Online]. Available: arXiv: 2210.06472.

[23] J. M. Macías-Macías, J. A. Ramírez-Quintana, G. Ramírez-Alonso, and M. I. Chacón-Murguía, "Deep learning networks for vowel speech imagery," in Proc. 17th Int. Conf. Electr. Eng., Comput. Sci. Autom. Control (CCE), Nov. 2020, pp. 1-6.

[24] P. Agarwal and S. Kumar, "Electroencephalography-based imagined speech recognition using deep long short - term memory network," ETRI J., vol. 44, no. 4, pp. 672-685, Jun. 2022.

[25] M. M. Islam and M. M. Shuvo, “DenseNet based speech imagery EEG signal classification using Gramian Angular Field,” in Proc. 5th Int. Conf. Adv. Electr. Eng. (ICAEE), Sep. 2019, pp. 149-154.

[26] Q. eting and G. Nuo, “Research on the Classification Algorithm of maginary Speech EEG Signals Based on Twin Neural Network,” in Proc. IEEE 7th Int. Conf. Signal Image Process. (ICSIP), Jul. 2022, pp. 211-216.

[27] S. Datta and N. V. Boulgouris, “Recognition of grammatical class of imagined words from EEG signals using convolutional neural network,” Neurocomputing, vol. 465, pp. 301–309, Nov. 2021.

[28] M. P. P., T. Thomas, and R. Gopikakumari, “Wavelet feature selection of audio and imagined/vocalized EEG signals for ANN based multimodal ASR system,” Biomed. Signal Process. Control, vol. 63, Jan. 2021, Art. no. 102218.

[29] J. T. Panachakel and A. Ramakrishnan, “DCLL—A deep network for possible real-time decoding of imagined words,” in Proc. Int. Symp. Intell. Informat., 2022, pp. 3–12.

[30] A. Siva Sankar Reddy and R. Bilas Pachori, “Multivariate dynamic mode decomposition for automatic imagined speech recognition using multichannel EEG signals,” IEEE Sensors Lett., vol. 8, no. 2, pp. 1–4, Feb. 2024.

[31] W. Ko, E. Jeon, and H.-I. Suk, “Spectro-spatio-temporal EEG representation learning for imagined speech recognition,” in Proc. Asian Conf. Pattern Recognit., 2021, pp. 335–346.

[32] B.-H. Lee, B.-H. Kwon, D.-Y. Lee, and J.-H. Jeong, “Speech imagery classification using length-wise training based on deep learning,” in Proc. 9th Int. Winter Conf. Brain-Comput. Interface (BCI), Feb. 2021, pp. 1–5.

[33] N. C. Mahapatra and P. Bhuyan, “Multiclass classification of imagined speech vowels and words of electroencephalography signals using deep learning,” Adv. Hum.-Comput. Interact., vol. 2022, pp. 1–10, Jul. 2022.

[34] L. C. Sarmiento, S. Villamizar, O. López, A. C. Collazos, J. Sarmiento, and J. B. Rodríguez, “Recognition of EEG signals from imagined vowels using deep learning methods,” Sensors, vol. 21, no. 19, p. 6503, Sep. 2021.

[35] O. Banerjee, D. Govind, A. K. Dubey, and S. V. Gangashetty, “Significance of dimensionality reduction in CNN-based vowel classification from imagined speech using electroencephalogram signals,” in Proc. Int. Conf. Speech Comput., 2022, pp. 44–55.

[36] K. Tyrrell and M. H. Kapourchali, “Unsupervised learning for exploring hidden structures in self-talk,” in Proc. IEEE 19th Int. Conf. Body Sensor Netw. (BSN), Oct. 2023, pp. 1–4.

[37] J. Clayton, S. Wellington, C. Valentini-Botinhao, and O. Watts, “Decoding imagined, heard, and spoken speech: Classification and regression of EEG

using a 14-channel dry-contact mobile headset," in Proc. Interspeech, Oct. 2020, pp. 4886–4890.

[38] G. Dai, J. Zhou, J. Huang, and N. Wang, "HS-CNN: A CNN with hybrid convolution scale for EEG motor imagery classification," J. Neural Eng., vol. 17, no. 1, Jan. 2020, Art. no. 016025.

[39] W. Ko, E. Jeon, S. Jeong, and H.-I. Suk, "Multi-scale neural network for EEG representation learning in BCI," IEEE Comput. Intell. Mag., vol. 16, no. 2, pp. 31–45, May 2021.

[40] X. Zhao, H. Zhang, G. Zhu, F. You, S. Kuang, and L. Sun, "A multibranch 3D convolutional neural network for EEG-based motor imagery classification," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 27, no. 10, pp. 2164–2177, Oct. 2019.

[41] 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, Oct. 2018, Art. no. 056013.

[42] 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, Nov. 2017.

[43] R. Mane, N. Robinson, A. P. Vinod, S.-W. Lee, and C. Guan, "A multiview CNN with novel variance layer for motor imagery brain computer interface," in Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2020, pp. 2950–2953.