Speech2EEG: Leveraging Pretrained Speech Model for EEG Signal Recognition

Jinzhao Zhou, Yiqun Duan, Yingying Zou, Yu-Cheng Chang, Yu-Kai Wang, *Member, IEEE*, and Chin-Teng Lin, *Fellow, IEEE*

***Abstract*— Identifying meaningful brain activities is critical in brain-computer interface (BCI) applications. Recently, an increasing number of neural network approaches have been proposed to recognize EEG signals. However, these approaches depend heavily on using complex network structures to improve the performance of EEG recognition and suffer from the deficit of training data. Inspired by the waveform characteristics and processing methods shared between EEG and speech signals, we propose Speech2EEG, a novel EEG recognition method that leverages pretrained speech features to improve the accuracy of EEG recognition. Specifically, a pretrained speech processing model is adapted to the EEG domain to extract multichannel temporal embeddings. Then, several aggregation methods, including the weighted average, channelwise aggregation, and channel-and-depthwise aggregation, are implemented to exploit and integrate the multichannel temporal embeddings. Finally, a classification network is used to predict EEG categories based on the integrated features. Our work is the first to explore the use of pretrained speech models for EEG signal analysis as well as the effective ways to integrate the multichannel temporal embeddings from the EEG signal. Extensive experimental results suggest that the proposed Speech2EEG method achieves state-of-the-art performance on two challenging motor imagery (MI) datasets, the BCI IV-2a and BCI IV-2b datasets, with accuracies of 89.5% and 84.07%, respectively. Visualization analysis of the multichannel temporal embeddings show that the Speech2EEG architecture can capture useful patterns related to MI categories, which can provide a novel solution for subsequent research under the constraints of a limited dataset scale.**

***Index Terms*— Transfer learning, motor imagery, electroencephalogram.**

Manuscript received 28 January 2023; revised 11 April 2023; accepted 14 April 2023. Date of publication 20 April 2023; date of current version 2 May 2023. This work was supported in part by the Australian Research Council (ARC) under Discovery DP210101093 and Discovery DP220100803, in part by the University of Technology Sydney (UTS) Human-Centric AI Centre by GrapheneX (2023–2031), in part by the Australia Defence Innovation Hub under Contract P18-650825, in part by the Australian Cooperative Research Centers Projects (CRC-P) Round 11 under Grant CRCPXI000007, in part by the U.S. Office of Naval Research Global Cooperative under Agreement ONRG-NICOP-N62909-19-1-2058, in part by US Air Force Office of Scientific Research and Defence Science and Technology (AFOSR–DST) Australian Autonomy Initiative under Agreement ID10134, and in part by the NSW Defence Innovation Network and NSW State Government of Australia under Grant DINPP2019 S1-03/09 and Grant PP21-22.03.02. *(Corresponding authors: Yu-Kai Wang; Chin-Teng Lin.)*

Jinzhao Zhou, Yiqun Duan, Yu-Cheng Chang, Yu-Kai Wang, and Chin-Teng Lin are with the Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: jinzhao.zhou@student.uts.edu.au; yiqun.duan@student.uts.edu.au; yu-cheng.chang@uts.edu.au; yukai.wang@uts.edu.au; chin-teng.lin@uts.edu.au).

Yingying Zou is with the School of Economics and Commerce, South China University of Technology, Guangzhou 510006, China (e-mail: yingying.zou.mse@gmail.com).

Digital Object Identifier 10.1109/TNSRE.2023.3268751

I. Introduction

A Brain-computer interface (BCI) allows direct communication between the human brain and computer devices by gathering central nervous system activities from the cerebral cortex [1]. Unlike invasive BCIs that require brain surgery prior to recording intracranial information, electroencephalography (EEG) has gained increasing popularity due to its ability to collect macroscopic brain signals without potential surgical risks and decreasing biosignal quality [2]. During the synaptic excitations of neuronal dendrites, electric currents from the cortex and the deep brain structure are measured using electrodes as EEG signals to control external devices [3]. Emerging BCI applications include artificial limb control [4], teleoperation [5], [6], gaming [7], healthcare [8], [9], and clinical diagnosis [10]. The foundation of brain-actuated teleoperation control is the recognition of a user's intent based on their brain signals [11], [12]. Despite the encouraging prospects and progress thus far, the extraction of discriminative features and accurate classification of EEG signals remains a challenge in the development of more advanced BCI applications.

In the pursuit of reliable EEG signal analysis, previous works have shown that traditional feature extraction and noise removal methods are useful not only for acoustic signals but also for EEG signals [13], [14], [15], [16], [17], [18], [19], [20]. The adaptation of traditional features such as independent component analysis (ICA) [21], fast Fourier transforms (FFTs) [22], wavelet transforms (WTs) [23], and Mel-frequency cepstral coefficients (MFCCs) [24] from speech processing to EEG analysis has been successful because both are mixture signals with similar characteristics [25], [26]. These methods can be summarized as EEG sonification, which treats an EEG signal as an audio signal and analyses it using speech recognition techniques [27], [28], [29]. More recently, deep learning models have the potential to learn better features that outperform traditional features and improve recognition accuracy [8], [30], [31], [32], [33], [34], [35], [36], [37]. These methods have sought to stack more complicated network structures into existing architectures because a shallow neural network can only learn to extract limited features. However, the lack of EEG data in the training stage and its poor quality are often overlooked, which can make deep neural network learning difficult [38].

This work is inspired by the shared characteristics between acoustic speech and EEG signals as well as recent works that draw the connection between noninvasive brain signals and automatic speech recognition [39], [40], [41], [42]. In this paper, we propose Speech2EEG, a novel EEG signal classification approach with better elasticity towards continuous EEG recognition in real-life scenario. The Speech2EEG approach exploits pretrained speech processing model in combination with EEG signal framing to improve the performance in capturing brain dynamics. To the best of our knowledge, we are the first to propose adopting structural feature extractors pretrained from massive speech datasets rather than training from scratch using the small and noisy EEG dataset. This is because the quality and scale of EEG data can have as much impact on the network's learning process as the EEG-specific network design [43]. Considering the low signal-to-noise ratio of EEG signals, adapting a well-trained feature extractor from a similar signal could be more efficient because the heavy noise in the training data often has a negative impact on the training process [44]. In addition, we propose to extract multichannel temporal embeddings from overlapping time frames for EEG signals rather than extracting features from the whole EEG sequence, which enables more fine-grained analysis of EEG signals for the recognition of brain dynamics as well as for potential real-life BCI applications. Last but not lease, the spatial dependencies among EEG channels are exploited by a variety of straightforward feature aggregation networks to combine channelwise information and yield the final representation for EEG classification.

Extensive experiments were conducted to illustrate the effectiveness of our proposed Speech2EEG model on two challenging motor imagery (MI) datasets, including the BCI IV-2a and BCI IV-2b datasets, as discussed in Section IV. We also provide visualization and analysis for the most relevant features in Section IV-I. The main contributions of this paper can be summarized as follows:

- The Speech2EEG model is the first approach proposed to leverage a pretrained speech processing model for effective EEG recognition and flexible network design. Experimental results show that Speech2EEG can achieve state-of-the-art accuracy on two challenging MI datasets.
- Multichannel temporal embeddings are introduced for EEG signal processing to enable fine-grained EEG analysis and more realistic applications in real-life scenarios.
- Visualization of the most influential features and topographic map suggest that the proposed approach extracts more plausible features from behavior-related brain activities.

II. Related Work

A. Pretrained Models for EEG Analysis

The reason for using pretrained models in the EEG domain is straightforward. First, due to collection difficulty and strict experimental protocol, data in the EEG domain are usually not large-scale. Therefore, a well-trained model is more crucial to provide rich and effective feature extractors and alleviate the dependence on the quality and size of the training dataset [45]. As a result, the practical benefit of achieving better EEG features from a pretrained neural network has provoked transfer learning development for EEG signal classification [46], [47], [48].

Typically, the pretrained neural networks used in existing EEG research are trained using other subjects or using other sessions with the same subject [49], [50], [51]. In [52] and [53], neural networks are first trained using existing data from source subjects with source sessions before being fine-tuned to the target subject and session using a small amount of target data. To achieve further improvement and learn more general features that can reveal or separate different factors of the phenomena entangled in the input data, unsupervised learning methods are introduced to make use of knowledge learned from a different EEG task [54], [55], [56], [57], [58]. For instance, autoencoders are first trained to reconstruct EEG time series before fine-tuning the encoder to a classification task [59], [60], [61], [62], [63]. These methods indicate that downstream EEG tasks can also benefit from more general feature extractors to a certain extent.

Recently, BENDR [38] trained a transformer model on the Temple University Hospital EEG Corpus speech processing domain dataset [64] to learn to increase the EEG representation generalization level. Although their pretrained model is not as competitive as more task-specific models, they show that useful features for EEG data can be captured using structures from language models. However, insufficient information in EEG signals can result in suboptimal feature extractors in the pretraining process. Different from their approach, this study explores the possibility of using a feature extraction network trained using the speech processing domain to provide general representations and avoid overfitting on a small dataset.

B. Effective Architectures for Raw EEG Signals

Deep learning methods for EEG analysis reduce the burden of designing feature extraction methods manually and allow end-to-end learning of task-related feature extractors automatically. Many works on deep learning models for EEG analysis focus on improving the network architecture.

Shallow ConvNet [65] uses a temporal convolution layer with a small 1 × 25 convolutional kernel and a spatial filter over all electrodes to aggregate temporal features from the previous layer. Further feature abstraction and downsampling are carried out using temporal convolutional layers with increasing kernel sizes and pooling layers. EEGNet [30] uses a channelwise spatial convolution layer as in Shallow ConvNet while improving the Shallow ConvNet structure by using a larger 1 × 64 temporal convolutional kernel and a

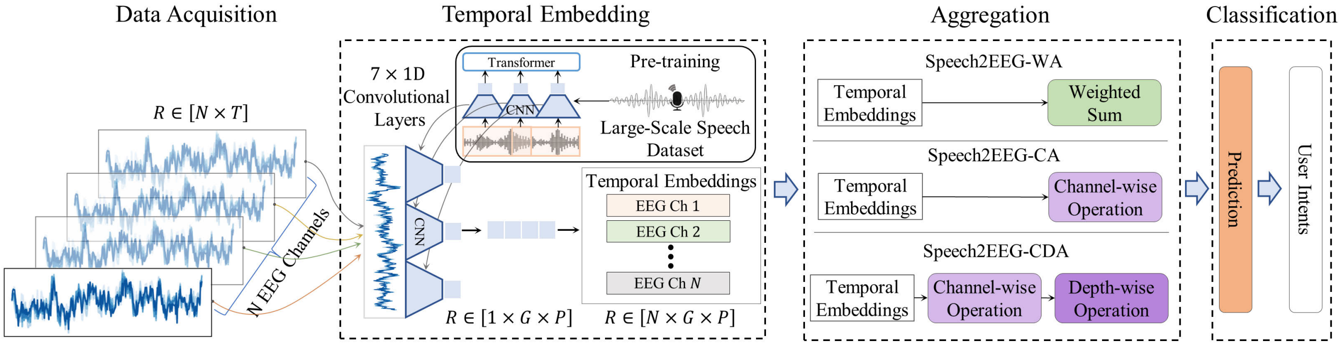


Fig. 1. The general framework of the proposed Speech2EEG method. After data acquisition from BCI caps, an embedding network pretrained on a large-scale speech dataset is adapted to the EEG domain to extract temporal embeddings from EEG signals within each time frame. Then, the generated temporal embeddings from each EEG channel are aggregated using a feature aggregation network. In this study, three types of feature aggregation schemas are designed to exploit temporal embeddings. Finally, the aggregated features are used to determine the category of the EEG signal and the user's intent.

separable convolutional layer. The latter allows direct information exchange among feature maps generated from different spatial filters. TCNet [37] further improves the EEGNet architecture by adding a temporal convolutional network that stacks two dilated causal convolutional layers and a residual skip connection that reorganizes the output temporal features from an EEGNet while increasing the depth of the neural network to allow more abstract feature summaries of previous EEG signals. Both the larger temporal kernel size in EEGNet and the application of dilated convolution enable a larger receptive field for feature aggregation across the temporal dimension. However, the overall receptive field of the network increases mainly by building a deeper network. The last convolutional layer in the network hierarchy has the largest receptive field but the lowest temporal resolution. To enable a global receptive field in the early network hierarchy and prevent the loss of temporal resolution, ATCNet [66] adapts a multihead attention block [67] after a Shallow-ConvNet-like convolutional block to extract a feature representation for each sliding window. The multihead attention block computes the attention score between features of every pair of sliding windows while keeping the number of sliding windows unchanged throughout the attention layers. Other architectures, such as graph embedding [68] and long-short term memory (LSTM) [69], have also been proposed to recognize raw EEG signals. To determine the optimal architecture, recent work has also tried to search for the optimal network architecture for each subject using neural network search (NAS) algorithms [70].

Our insight into these architectures is that they use channel-wise spatial filters to impose an inductive interchannel relation bias, which corresponds to the nature of EEG signals being a mixture of brain activities projected on the human scalp surface. As a result, brain activity of interest should have an impact on several neighboring electrodes. The identification of similar EEG signals among neighboring electrodes can thus be the key to identifying specific brain activity. Inspired by these developments, we adopted the spatial convolutional filtering mechanism that is widely used in these state-of-the-art neural networks to design the feature aggregation network, while the temporal convolution layer is replaced by a pretrained speech feature extraction network that extracts multichannel temporal embeddings from raw EEG signals.

III. Proposed Transfer Learning Method

This section provides technical details of the proposed Speech2EEG approach. The proposed Speech2EEG model is depicted in Fig.1. From left to right, Speech2EEG takes a raw EEG waveform as input and applies three modulated subnetworks, including a temporal embedding subnetwork, a feature aggregation subnetwork, and a classification subnetwork. After EEG data acquisition, the pretrained temporal embedding network extracts temporal embeddings for all EEG channels within a time frame. Then, the aggregation network is used for the integration and selection of features among different spatial or temporal locations. Finally, the classification network performs classification and yields a distributional score over the testing categories. The proposed Speech2EEG approach transfer knowledge learn from a large-scale waveform dataset to improve the performance of the EEG classification task. In the training stage, the fine-tuning method is utilized to adjust the pretrained temporal embedding subnetwork to the EEG classification with the help of the feature aggregation subnetwork and classification subnetwork. This paper describes the task settings in Section III-A. Afterwards, a detailed description of the temporal embeddings is introduced in Section III-B. Then, the aggregation networks are described in Section III-C. Finally, the training objective and settings will be described in Section III-D.

A. Preliminaries

We use $\mathcal{X} = \{X^{(i)}\}_{i=1}^{M}$ to denote the set of all M EEG samples available at the training phase, with $\mathcal{Y} = \{y^{(i)}\}_{i=1}^{M}$ denoting the categorical labels of these EEG sequences. Let C denote the number of EEG categories in a specific EEG classification task, i.e., $y^{(i)} \in \{1, \ldots, C\}$. Each sample of EEG signal $X^{(i)} \in \mathcal{R}^{N \times T}$ consists of N EEG channels of sequential EEG sample points with a length of T timesteps. T is fixed to a constant value after resampling or segmentation throughout the dataset. If not noted otherwise, the superscript in $X^{(i)}$ will be dropped for brevity.

B. Multichannel Temporal Embedding for Raw EEG Signals

The proposed Speech2EEG model utilizes a transformer-like network pretrained on a large-scale speech dataset to generate temporal embeddings over a small time frame for the EEG sequence from each channel. Modern transformer-like speech processing networks use self-supervised learning objectives for training to learn to discover meaningful speech units such as phonemes in a short time period [71]. We hypothesize that filters that are able to identify speech units can also generate useful features for event-induced brain activities such as event-related desynchronization (ERD) [72], [73], [74]. Thus, these features would contain more useful information than the noisy EEG signal for further analysis.

For a speech processing transformer-like architecture, one way to obtain feature representations is to use the output of the attention layers or fine-tune the whole transformer network model to a BCI dataset. However, features obtained from self-attention layers are biased toward the exploitation of long-term dependencies among speech units and have less information about the original input signal. Likewise, classification layers can be overly adapted to the original speech-related task, which is unlikely to be helpful. We validated this empirical assumption by building a additional classification layer on top of different components of a transformer-like architecture by an ablation study in Section IV-G. Therefore, we make use of the convolutional subnetwork before the attention layers as the network ϕ to extract temporal features.

We use the convolutional subnetwork from the Wav2Vec 2.0 [75] model to obtain multichannel temporal embeddings in this study, as depicted in Fig.1. The major reason for selecting this pretrained speech processing model is that it is trained on a large-scale waveform dataset in a self-supervised manner. The obtained general features allow this model to perform well after fine-tuning on a small-scale dataset. In addition, it has a modularized network architecture that can be exploited flexibly. Denote $\phi_j(X_i)$ as the i^{th} channel features obtained from the j^{th} layer of ϕ, $j \in \{1, \ldots, 7\}$. As discussed in [76] and [77], representations from earlier convolutional layers preserve more detailed local information, while those from later layers are more compact and abstract. We use the output from the last layer of ϕ ($j = 7$) to avoid local noise and enjoy a smaller embedding size for the temporal embeddings. ϕ consists of seven 1D convolutional layers with group normalization (GN) [78] used to normalize the features within each feature group. The Wav2Vec 2.0 model convolutional encoder has a receptive field of 400 time steps with a stride of 320 time steps. For an EEG signal with a sampling rate of 250 Hz, temporal embeddings are extracted from approximately 1.5 seconds duration of the EEG signal as a time frame with approximately 1 second overlapping. The resulting time frame for the temporal embedding is similar to the minimum time course of a typical ERD signal evoked during the MI process [79]. The overlapping window of the embeddings also allows more detailed scanning of the EEG signal for the ERD signal compared to networks that use the entire 3 to 4-second epoch [80], [81].

As depicted in the bottom of the temporal embedding block from Fig.1, we assume the output embedding of a single EEG channel $e_k = \phi_j(X_k) \in \mathcal{R}^{G\times P}$ to be a two-dimensional tensor of group number G and feature size P, where G is the number of time frames within the EEG time sequence and P is the output feature size for each EEG time frame. In our specific case, $P = 512$. We stack e_k from all N EEG channels to form a three-dimensional tensor $E = \left[\phi_j(X_1), \ldots, \phi_j(X_N)\right] = [e_1, \ldots, e_N]$, $E \in \mathcal{R}^{N\times G\times P}$ as the multichannel temporal embeddings for a complete EEG signal X. Ideally, we expect multichannel temporal embeddings to contain a useful description of brain dynamics that can be used to identify various MI categories while the differences between the EEG and speech signal distributions to be alleviated through a fine-tuning process.

C. Aggregation and Classification

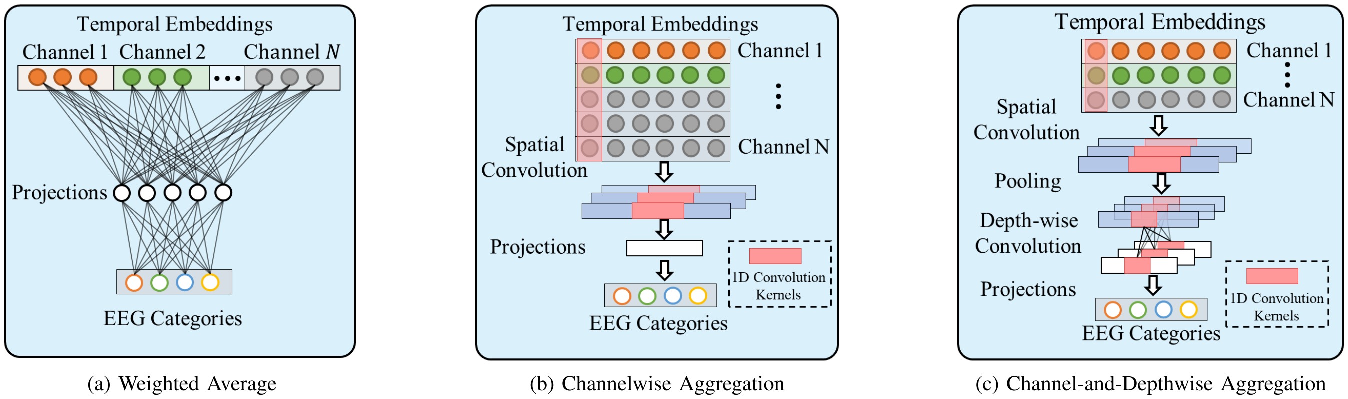
Nonetheless, temporal embeddings alone are not sufficient to make an EEG signal prediction. This is also supported in the experiment in Section IV-G. Given the temporal embeddings from each EEG channel, we employed different schemas to design the feature aggregation networks for the Speech2EEG model. A total of three aggregation methods are used with different levels of complexity. They are weighted average aggregation (WA), channelwise aggregation (CA), and channel-and-depthwise aggregation (CDA).

1) Weighted Average Aggregation (WA): is the simplest aggregated schema with minimum transformation to multichannel temporal embeddings (depicted in Fig. 2a). This can be treated as the baseline aggregation method. Assuming the multichannel temporal embeddings are already good enough for downstream EEG classification, a simple mechanism that allows feature selection and weighting will be sufficient to separate various EEG categories. Therefore, a dense layer is used in the Speech2EEG-WA to learn static weighting matrices that integrate temporal embeddings to a compact feature vector of size D. Denoting an aggregated feature as h, mathematically, each unit computes a weighted summation using multichannel temporal embeddings, which can be written as:

$$h_d = \sigma\left(\sum_{i,j,k} W^d_{ijk} \otimes E_{ijk}\right), \quad \forall d \in \{1, \ldots, D\} \tag{1}$$

where $\otimes$ denotes elementwise multiplication and σ is an elementwise nonlinearity operation or identical function. If an identical operation is used as σ, then the dense layer can also be viewed as a linear projection of the temporal embeddings. $W^d \in \mathcal{R}^{N\times G\times P}$ is the learnable weights to aggregate information for the d^{th} output hidden unit. An obvious drawback of the WA schema is that it cannot construct more complex and abstract features using embeddings from neighboring channels. Additionally, the scale of the Speech2EEG-WA architecture will increase drastically as the number of channels increases.

2) Channelwise Aggregation (CA): is depicted in Fig. 2b where the popular spatial convolution from multiple EEG network architectures is adopted. After converting raw EEG



(a) Weighted Average

(b) Channelwise Aggregation

(c) Channel-and-Depthwise Aggregation

Fig. 2. Aggregation strategies are used to exploit multichannel temporal embeddings in this study. (a) Weighted aggregation strategy: the multichannel temporal embeddings are flattened and concatenated before applying weight summation using hidden projection layer units. (b) Channelwise aggregation strategy: multichannel temporal embeddings are aggregated using a spatial convolutional layer where spatial dependencies among channels are exploited. (c) Channel-and-depthwise aggregation strategy: after the spatial aggregation of multichannel temporal embeddings, features from different feature maps are further integrated. The subsequent classification modules remain static throughout all aggregation strategies, which is an MLP network consisting of a dense layer.

signals to temporal embeddings using a pretrained speech feature extraction network, we intend to investigate whether the exploitation of spatial dependencies among EEG channels can improve recognition performance. A total of $F1$ convolutional filters with a kernel size of $N \times 1$ are trained so that each spatial filter performs channelwise aggregation. Assume we have a filter $\omega \in \mathcal{R}^{N\times 1}$ from a spatial convolutional layer and denote the output of the spatial convolutional as h. Each feature map h_f of the output h is computed by:

$$h_f = \sigma(\sum_{i=1}^{G}\sum_{j=1}^{N}\omega_j^f E_{ijk}), \quad \forall f \in \{1,\ldots,F_1\} \tag{2}$$

As discussed in [30], the use of spatial filters enables the extraction of features related to a specific frequency band. After spatial filtering, we apply an additional max pooling layer of pool size Kp and 1D convolutional layers of kernel size $K \times 1$ to reduce the feature size before using a dense layer with D hidden units for the final classification. The Speech2EEG-CA architecture constructs higher-level features from speech features when they are not immediately distinctive among EEG categories.

3) Channel-and-Depthwise Aggregation (CDA): uses an architecture similar to EEGNet after temporal embedding extraction, as depicted in Fig.2c. We apply a convolutional layer with $F1$ filters of filter size $K1 \times 1$, followed by a spatial convolutional layer with $F1 \times F2$ filters. After applying the spatial convolution, a pooling layer of pool size Kp is used to reduce feature size, while a separable convolutional layer of $F3$ filters and $K3 \times 1$ kernel size is used to enable depthwise interaction between convolutional feature maps. Another benefit of using a separable convolutional layer is the increase in the number of network layers for the learning of more complex features from multichannel temporal embeddings. To help regulate the model, batch normalization [82] layers are used in this architecture. Assuming that embeddings extracted from a speech model have limited discriminative capacity toward EEG recognition, we at least expect a mature network to perform basic filtering and transformation of raw EEG waveforms so that low-level noise can be reduced while similar EEG time frames remain close in the continuous embedding space. From this perspective, the Speech2EEG-CDA architecture uses less noisy embedding to learn good features for EEG recognition.

The detailed architecture settings for the three aggregation networks are listed in Table II. Despite using different feature aggregation subnetworks, we use a simple multilayer perceptron (MLP) as the classification network for all Speech2EEG architectures.

D. Training Objective and Search Space for the Hyperparameters

To train the Speech2EEG models, cross-entropy loss is applied as the objective function, which can be written as:

$$L(p^{(i)}) = -\sum_{i=1}^{M} y^{(i)}\log(p^{(i)}) \tag{3}$$

where $p^{(i)}$ denotes the prediction for an EEG sample in the training set $X^{(i)}$, and $y^{(i)}$ is the true label for the corresponding EEG sample. We use Adam [83] for the optimization algorithm with a weight decay rate of $1e^{-2}$. The batch size is set to 64. The network parameters as well as the learning rate are searched for each experiment using cross-validation on the training data similar to [37] and [84]. The search space for the hyperparameters is listed in Table I. To allow more efficient searching of the hyperparameters, we adopted the Tree-structured Parzen Estimator algorithm (TPE) algorithm [85] for choosing parameters within the parameter search space for a total of 100 rounds. To avoid excessive training and searching time, we adopted the Median Pruning algorithm to stop search rounds that has significantly worse performance.

IV. Experiments

To verify the effectiveness of the adapted features and the feature aggregation architectures, the datasets 2a and 2b from BCI Competition IV are utilized for extensive experiments.

TABLE I
SEARCH SPACE FOR THE HYPERPARAMETERS

Hyper-Parameters	Search Space
Speech2EEG-WA	
D	64,128,256,512,1024
Speech2EEG-CA	
K	10 to 25
D	64,128,256,512,1024
Speech2EEG-CDA	
F1	2 to 64
F2	2 to 64
F3	2 to 64
K1	30 to 128
K3	30 to 128
D	64,128,256,512,1024
Other Parameter	
Activation	None, RReLU, PRELU, GELU, ELU, Swish, HardSwish, Hardtanh, LeakyReLU, Sigmoid
Learning Rate	1e-6 to 1e-3

In particular, both internal ablation studies and external comparisons with state-of-the-art methods are conducted in this paper. For both datasets, subject-specific results are obtained using different the proposed Speech2EEG approaches. Since a number of works on the BCI IV-2a dataset are also interested in the cross-validation performance on the entire dataset, a mix-subject five-fold cross-validation experiment on the BCI IV-2a dataset is also carried out to better compare with external methods and evaluate the more general capacity of the proposed method. Additionally, ablation studies are carried out to investigate the performance of directly fine-tuning output from different modules of the pretrained speech model on both datasets to justify the rationale of using the feature extraction module for extracting temporal embeddings. Since the BCI IV-2a dataset contains more motor imagery categories and complete information from all channels compared to the BCI IV-2b dataset, an ablation study on the impact of different training set sizes. A brief description of the two datasets and the preprocessing steps are available in Section IV-B. Then, Section IV-C describes the evaluation metrics used in our experiments. The experimental outcome and the ablation studies are discussed in Section IV-D to Section IV-H. Finally, visualization of the proposed Speech2EEG model is presented and discussed in Section IV-I.

A. Implementation Details

The Speech2EEG method is implemented in PyTorch, a deep learning library based on Python, and is run on an Intel Xeon E-2274G 4.0 GHz CPU and an NVIDIA Quadro RTX 6000 with CUDA 10.2 GPU. The algorithms used for hyperparameter searching is implemented using the Optuna framework [86].

B. Datasets and Preprocessing

1) The BIC IV-2a dataset [87] is an oscillatory EEG dataset containing four types of imagined movements from 9 subjects (A01-A09). There are a total of 288 trials for each session. Each subject conducted two sessions (session T and session E) on different days. Hints were provided before the imagined movements of both hands, feet, and tongue in each trial. The EEG data were recorded using 25 electrodes (22 EEG channels and 3 EOG channels) with a sampling rate of 250 Hz. A bandpass filter between 0.5 and 100 Hz was applied to both the 2a and 2b datasets during the data collection phase. Although a number of studies suggest that additional preprocessing methods such as resampling [30], channel selection, exponential smoothing or dataset cleansing [65] are useful techniques for classification performance, we empirically find that these methods have a limited effect on the performance of the proposed method. Therefore, for the BCI IV-2a dataset, the raw EEG data from the 22 EEG channels are used as input to our Speech2EEG architectures.

2) The BCI IV-2b dataset [88] contains EEG data collected from 9 subjects (B01-B09) using 6 electrodes (3 bipolar EEG channels and 3 EOG channels) with a sampling rate of 250 Hz. Subjects were asked to imagine the movements of the left or right hand in each trial during 3 training sessions and 2 evaluation sessions. To minimize the overfitting issue and compensate for the smaller data size, we apply data augmentation processing to the BCI IV-2b dataset. Following [80], we utilize a Butterworth high-pass filter to cut off frequencies above 100 Hz as noise samples. Then, we combine the noise samples with other EEG data from the same session to obtain augmented samples. Generation of the augmented samples is shown in Eq. 4.

$$X_{aug}^{(i)} = X^{(i)} - \bar{X}^{(i)} + \bar{X}^{(j)}, \tag{4}$$

where $\bar{X}^{(i)}$ denotes the noise from the i^{th} sample in the training set, $\bar{X}^{(j)}$ denotes noise from another sample j within the same session, and $X_{aug}^{(i)}$ is the augmented sample using the original data sample $X^{(i)}$.

C. Evaluation Metrics

Two frequently used metrics in the EEG literature are used to evaluate the performance of the method proposed in this study. They are the accuracy and the kappa score. To quantitatively compare the effectiveness of each MI classification method, the accuracy Acc of the convergence model can be obtained by

$$Acc = \frac{N_{correct}}{N_l}, \tag{5}$$

where N_l and $N_{correct}$ are the total number of ground truth labels and the number of correct predictions, respectively. In addition to the accuracy score, the kappa value κ is computed to eliminate the impact of random guessing in the classification. The κ value is computed as follows [89]:

$$\kappa = \frac{Acc - p_r}{1 - p_r}, \tag{6}$$

where p_r is random classification accuracy in a dataset.

D. Subject-Specific Results on the BCI IV-2a Dataset

In this experiment, we train variants of the Speech2EEG model for each subject and compare the performance with existing methods. The subject-specific results are shown in Table IV. Table III summarizes the optimal hyperparameters of all Speech2EEG architectures.

TABLE II
ARCHITECTURES OF SPEECH2EEG-WA, SPEECH2EEG-CA, AND SPEECH2EEG-CDA

Architecture	**Speech2EEG-WA**			**Speech2EEG-CA**			**Speech2EEG-CDA**		
Module	Type	Filters/ Units	Kernel Size	Type	Filters/ Units	Kernel Size	Type	Filters/ Units	Kernel Size
Feature Extractor (Pretrained)	Conv1D-GN	512	10	Conv1D-GN	512	10	Conv1D-GN	512	10
	Conv1D-GN	512	3	Conv1D-GN	512	3	Conv1D-GN	512	3
	Conv1D-GN	512	3	Conv1D-GN	512	3	Conv1D-GN	512	3
	Conv1D-GN	512	3	Conv1D-GN	512	3	Conv1D-GN	512	3
	Conv1D-GN	512	3	Conv1D-GN	512	3	Conv1D-GN	512	3
	Conv1D-GN	512	2	Conv1D-GN	512	2	Conv1D-GN	512	2
	Conv1D-GN	512	2	Conv1D-GN	512	2	Conv1D-GN	512	2
Feature Aggregation	Dense	D	-	Sp.Conv1D	F1	N	Conv1D-BN	F1	K1
	-	-	-	MaxPool	-	Kp	Sp.Conv1D	F2 × F1	N
	-	-	-	Conv1D	F2	K	S.Conv1D-BN	F3	K3
	-	-	-	Dense	D	-	AvgPool	-	Kp
	-	-	-	-	-	-	Dense	D	-
Classification	Dense	C	-	Dense	C	-	Dense	C	-

[1] C denotes the number of EEG categories. D denotes the number of hidden units of the dense layer. N denotes the number of channels in the EEG signal.
[2] F1, F2, F3 denotes the number of filters while K1, K3, Kp denotes the kernel size of the corresponding convolutional layer.
[3] Sp.Conv1D denotes a 1D spatial convolutional layer.
[4] S.Conv1D denotes a 1D separable convolutional layer.
[5] Conv1D-BN denotes a 1D convolutional layer followed by a batch normalization layer, and Conv1D-GN denotes a 1D convolutional layer followed by a group normalization layer.

TABLE III
OPTIMAL NETWORK PARAMETERS FOR EACH SUBJECT IN THE BCI IV-2A DATASET

	Speech2EEG-WA		**Speech2EEG-CA**					**Speech2EEG-CDA**						
Subjects	Act.	D	Act.	F1	F2	K1	D	Act.	F1	F2	F3	K1	K3	D
A01	None	512	RReLU	40	20	20	128	GELU	2	30	10	128	30	64
A02	RReLU	1024	ELU	16	16	20	64	SELU	30	60	60	100	32	32
A03	RReLU	512	ELU	16	64	20	64	GELU	8	64	2	45	35	64
A04	PReLU	1024	L.ReLU	32	32	15	256	Tanh	64	60	40	90	128	200
A05	SELU	256	PReLU	16	20	20	128	H.tan	20	4	60	45	40	100
A06	ELU	128	ReLU	32	40	15	64	H.tan	32	10	32	70	128	64
A07	Sigmoid	256	H.tan	64	32	20	32	None	50	30	10	30	32	100
A08	Swish	256	Tanh	40	20	25	64	None	10	8	8	80	35	256
A09	Swish	1024	Tanh	32	32	25	32	Tanh	32	4	40	90	128	100

[1] Act. denotes activation function.
[2] L.ReLU denotes LeakyReLU, H.tan denotes HardTanh.

TABLE IV
RESULTS FOR 4-CLASS CLASSIFICATION ON THE BCI IV-2A DATSET

	EEG-TCNet [37]		MBEEGSE [90]		ATCNet [66]		TCNet-Fusion [91]		**Speech2EEG-WA**		**Speech2EEG-CA**		**Speech2EEG-CDA**	
	Acc.	kappa	Acc.	kappa	Acc.	kappa	Acc.	kappa	Acc.	kappa	Acc.	kappa	Acc.	kappa
A01	89.3	0.86	89.1	0.85	88.5	0.85	90.7	0.87	91.1	0.88	96.4	0.90	96.4	0.90
A02	72.4	0.63	69.7	0.59	70.5	0.61	70.7	0.60	71.4	0.61	76.4	0.93	89.3	0.85
A03	97.4	0.97	95.3	0.93	97.6	0.97	95.2	0.93	96.4	0.95	98.2	0.98	98.2	0.98
A04	75.9	0.68	81.4	0.75	81.0	0.75	76.8	0.68	85.7	0.80	91.8	0.89	87.8	0.83
A05	83.7	0.78	80.0	0.73	83.0	0.77	82.2	0.76	83.3	0.78	87.0	0.83	81.5	0.68
A06	70.7	0.61	63.3	0.51	73.0	0.65	68.8	0.58	75.0	0.66	75.0	0.66	76.4	0.68
A07	93.1	0.91	94.1	0.92	93.1	0.91	94.2	0.92	90.9	0.83	81.8	0.76	90.9	0.83
A08	86.7	0.82	89.6	0.86	90.3	0.87	88.9	0.85	90.7	0.80	92.6	0.88	92.6	0.9
A09	85.2	0.8	83.4	0.77	91.0	0.88	85.9	0.81	94.1	0.92	86.3	0.76	92.2	0.83
Avg.	83.8	0.78	82.9	0.77	85.3	0.81	83.7	0.78	86.5	0.80	87.3	0.84	**89.5**	0.82
Std.	9.2	0.12	10.8	0.14	9.2	0.12	9.6	9.8	8.5	0.11	8.3	0.10	6.9	0.09

[1] Avg. denotes the averaged accuracy among all subjects.
[2] Acc. denotes the classification accuracy.
[2] Std. denotes the standard deviation of the accuracies among subjects.

For all aggregation strategies, the Speech2EEG architectures achieve higher accuracy and kappa scores on average and consistently steady performance across all subjects thanks to the help of feature extractors adapted from a pretrained speech processing model. Compared to the results reported for EEG-TCNet [37] and TCNet-Fusion [91], which share a setting similar to our method, our architecture achieves an over 3% increase in average accuracy. Moreover, the standard deviation of the accuracy between subjects is only approximately 0.1, which is more consistent across different subjects. Note that the baseline Speech2EEG-WA and Speech2EEG-CA architectures use a simple aggregation method compared to MBEEGSE [90] and ATCNet [66], which use more sophisticated network architectures. This suggests that the quality of

TABLE V
Optimal Network Parameters for Each Subject in the BCI IV-2b Dataset

	Speech2EEG-WA		Speech2EEG-CA					Speech2EEG-CDA						
Subject	Act.	D	Act.	F1	F2	K1	D	Act.	F1	F2	F3	K1	K3	D
B01	RReLU	256	H.tan	40	32	15	32	SELU	60	2	60	50	35	32
B02	Sigmoid	1024	ELU	40	64	10	128	Tanh	16	8	8	32	50	128
B03	Sigmoid	1024	RReLU	16	32	10	32	Tanh	16	10	8	128	40	128
B04	SELU	128	SELU	64	32	5	128	L.ReLU	64	20	8	50	100	512
B05	None	512	RReLU	64	20	5	256	GELU	32	8	50	32	32	100
B06	SELU	1024	SELU	32	20	5	32	None	60	50	20	100	90	512
B07	Tanh	256	SELU	40	16	5	64	SELU	30	4	40	128	45	64
B08	Hardswish	1024	Tanh	64	64	10	32	H.tan	60	4	60	32	50	128
B09	Tanh	1024	H.tan	64	40	15	64	ELU	64	40	60	70	128	256

[1] Act. denotes activation function.
[2] L.ReLU denotes LeakyReLU, H.tan denotes HardTanh.

TABLE VI
Comparison of External Methods and the Proposed Method on the BCI Competition IV-2b Data

Method	B01	B02	B03	B04	B05	B06	B07	B08	B09	Avg.
TPP [81]	78.75	66.43	67.50	95.00	94.38	84.38	**85.31**	92.19	81.56	82.83
DRDA [95]	62.86	62.86	63.63	95.94	93.56	**88.19**	85.00	95.25	**90.00**	83.98
EMD-MI [93]	62.80	**67.10**	**98.70**	88.40	**96.30**	75.30	72.20	87.80	85.30	81.54
Bi-spectrum [96]	77.00	64.50	61.00	96.50	82.00	84.50	75.00	91.00	87.90	79.93
MI-CNN [94]	75.31	57.50	56.56	**96.88**	92.19	83.44	84.06	92.81	86.25	80.56
FBCSP [92]	70.00	60.36	60.94	97.50	93.12	80.63	78.13	92.50	86.88	80.00
ConvNet [65]	76.56	50.00	51.56	**96.88**	93.13	85.31	83.75	91.56	85.62	79.37
Speech2EEG-WA w/ pre-trained weights	78.07	58.78	64.35	95.44	96.34	82.87	80.17	96.52	83.67	81.80
Speech2EEG-CA w/ pre-trained weights	75.88	62.86	67.39	96.09	96.34	80.48	81.03	**96.52**	85.71	82.48
Speech2EEG-CDA w/ pre-trained weights	**80.70**	62.04	71.74	96.09	94.51	84.06	84.06	95.65	87.76	**84.07**
Speech2EEG-WA w/o pre-trained weights	66.67	57.89	62.60	92.64	81.88	68.18	78.20	84.55	81.95	74.95
Speech2EEG-CA w/o pre-trained weights	66.67	60.32	70.73	95.09	83.22	65.91	77.44	86.99	81.95	76.48
Speech2EEG-CDA w/o pre-trained weights	76.19	63.15	69.92	94.48	82.55	80.30	78.95	80.49	69.92	77.33

[1] Avg. denotes the averaged accuracy among all subjects.

training data could have a larger impact on model performance than the selection of model architecture in this case. Finally, the Speech2EEG-CDA architectures achieve a state-of-the-art MI classification accuracy of 89.5%, which shows that the Speech2EEG model can leverage knowledge learned from acoustic speech data to improve recognition performance for EEG signals.

E. Subject-Specific Results on the BCI IV-2b Dataset

To demonstrate the generalization of the proposed method, we train the Speech2EEG models using the first three sessions and evaluate the last two testing sessions for each subject. The accuracy on each subject as well as the average accuracy on all subjects using the proposed Speech2EEG models is compared with state-of-the-art algorithms on the BCI IV-2b dataset, as shown in Table VI. The optimal network parameters for each subject are presented in Table V. The results show that the proposed Speech2EEG model has superior average accuracy performance. Our method outperforms traditional methods such as filter bank common spatial pattern (FBCSP) [92] and the empirical mode decomposition based filtering method (EMD-MI) [93]. Compared to recent deep learning methods including the ConvNet [65], the MI-CNN [94], and temporal pyramid pooling (TPP) network [81], the Speech2EEG model achieves more compatetive performance. Although when compared to the deep representation-based domain adaptation (DRDA) method, the improvement on accuracy is insignificant. Our method is more data-efficient since the proposed Speech2EEG models only uses the data from the target subject while the DRDA method requires additional data from other subjects to learn domain-invariant features. Additionally, an internal comparison between Speech2EEG models with and without pretrained weights is listed in Table VI. The comparison illustrates that with the help of the pretrained feature extraction module, the proposed Speech2EEG models are able to obtain useful waveform information from the start and avoid overfitting the noisy training dataset. Compared to learning the whole network from scratch using the noisy EEG training data, the pretrained feature extraction module could have a positive impact on the classification performance.

F. Mixed-Subject Results on the BCI IV-2a Dataset

To verify the general capacity of the proposed Speech2EEG architecture as well as to compare with existing methods, we conduct a mixed-subject classification experiment on the BCI IV-2a dataset. We train Speech2EEG-WA, Speech2EEG-CA, and Speech2EEG-CDA models for all subjects in a 5-fold cross-validation setting. Table VIII shows the optimal performance for each architecture compared to existing methods and Table VII shows the statistical test between Speech2EEG models with and without pretrained weights using the top-3 results from each architecture.

As illustrated in Table VIII, for every Speech2EEG architecture, the use of a pretrained feature extractor improves the classification accuracy compared to training from scratch with raw EEG data while Table VII shows that the difference is significant. The Speech2EEG-CA and Speech2EEG-CDA models

TABLE VII

TOP-3 CLASSIFICATION RESULTS AND STUDENT-T TEST BETWEEN SPEECH2EEG MODELS WITH AND WITHOUT PRETRAINED WEIGHTS ON BCI IV-2A DATASET

Architecture	With pretrained weights			Without pretrained weight		
Speech2EEG-WA	85.10	84.92	84.81	80.00	77.42	79.68
Speech2EEG-CA	85.38	85.24	84.32	81.96	81.43	76.79
Speech2EEG-CDA	86.17	85.40	84.32	76.76	74.89	74.26
Student-t Test	p-value $= 0.000056 < 0.05$			Reject the Null Hypothesis		

[1] The Student-t test is performed with the Null Hypothesis as ”The mean of models with pretrained weights equals the mean of models without pretrained weights”.

[2] Numbers in the table denotes each of the Top-3 classification accuracy of the corresponding model and setting.

TABLE VIII

COMPARISON OF EXTERNAL METHODS AND THE PROPOSED METHOD ON THE BCI COMPETITION IV-2A DATASET

	Method	Accuracy
10-Fold	CWT-CNN (3 Classes) [97]	71.20
	C2CM [84]	74.46
	MB-3DCNN [90]	75.02
	Multi-view Feature Learning [98]	78.51
	M3DCNN [99]	81.22
	Spatio-Spectral CNN (2 Classes) [100]	87.15
5-fold	AX-LSTM [101]	76.90
	EEGNet (Variable Structure)*	77.00
	TSGL-EEGNet [102]	81.34
	Speech2EEG-WA w/o pre-trained weights	80.00
	Speech2EEG-CA w/o pre-trained weights	81.90
	Speech2EEG-CDA w/o pre-trained weights	76.70
	Speech2EEG-WA w/ pre-trained weights	85.10
	Speech2EEG-CA w/ pre-trained weights	85.38
	Speech2EEG-CDA w/ pre-trained weights	**86.17**

* denotes reproduced result.

TABLE IX

RESULTS ON FINE-TUNING DIFFERENT NETWORK COMPONENTS ON THE BCI IV-2A AND 2B DATASETS

Dataset	Whole Model	Transformer Module	Feature Extraction Module
BCI IV-2a	26.17	30.31	72.23
BCI IV-2b	50.08	56.05	69.32

achieve higher accuracy than the baseline Speech2EEG-WA method, indicating that the exploitation of spatial information remains effective for processing multichannel temporal embeddings. For the Speech2EEG-WA and Speech2EEG-CA architectures, the improvement from pretraining the speech feature extraction network is nearly 5%. When training the Speech2EEG-CDA architecture without pretrained feature extractors, the model's performance drops to only 76.7%, which is inferior to the reproduced EEGNet baseline (77.0%). This is consistent with empirical findings that shallower neural networks tend to be more effective than their deeper counterparts in BCI [38], [65], [103]. The Speech2EEG-CDA model has the deepest architecture, and training such a large network can be difficult, especially with a limited amount of training data and poor data quality. However, with the help of the pretrained feature extractor, Speech2EEG-CDA benefits from its deeper architecture while achieving the highest accuracy.

An external comparison between the proposed Speech2EEG architectures and the latest methods evaluated on the BCI IV-2a dataset is also summarized in Table VIII. The proposed Speech2EEG method achieves superior performance compared to existing methods. The Spatio-Spectral CNN [100] reported an accuracy of 87.15% but only used 2 MI categories. Different from their method, the proposed Speech2EEG method is evaluated on all 4 MI categories in the dataset. The C2CM [84], MB-3DCNN [90], and multiview feature learning [98] methods introduced structures to exploit certain EEG characteristics. Unlike these methods, the adaptation of structured feature extractors achieves better performance on this dataset than the introduction of a more complex network structure.

G. Ablation Study on Fine-Tuning the Whole Pretrained Network

To support the rationale of only using the feature extraction module from the pretrained speech model for extracting temporal embeddings, we conduct ablation experiments to study the performance of fine-tuning different components of the pretrained speech processing network for EEG signals classification. We are interested in the transferability of the logit output (the whole pretrained model), the output from the transformer module, and the output of the feature extraction module of the pretrained model. An additional dense layer with a softmax activation function is added on top of these network components to obtain the classification result for each EEG category. Table IX shows the performance of fine-tuning different network components to the BCI IV-2a and the BCI IV-2b datasets. It illustrates that using the transformer module as well as fine-tuning the whole model yields inferior prediction accuracies which are close to random guesses. However, fine-tuning the feature extraction subnetwork could reach 72.23% and 69.32% accuracies on the BCI IV-2a and 2b datasets respectively without using an additional aggregation network. This ablation experiment validates the previous assumption that certain low-level features in the pretrained speech model could also capture discriminative patterns from the EEG signal after a fine-tuning process. Therefore, the feature of the feature extraction module of the pretrained speech model could be the more reasonable candidate to use for extracting temporal embeddings in the proposed Speech2EEG approach.

H. Ablation Study on Different Training Data Sizes

We further perform an ablation study on the performance of the proposed Speech2EEG model on an EEG dataset with a smaller scale by reducing the training data while keeping the testing data unchanged. The pretrained Wav2Vec 2.0 model can achieve good performance when fine-tuned on small datasets. We intend to investigate how well Speech2EEG models both with and without pretrained weights perform in different training dataset scales.

On the BCI IV-2a and BCI IV-2b datasets, the performance of each Speech2EEG architecture in various training data percentages is displayed in Fig.3 and Fig.4 respectively. We implemented EEGNet with a variable network structure as in [37] as the benchmark method for comparison. According to Fig.3, on the BCI IV-2a dataset, all architectures can

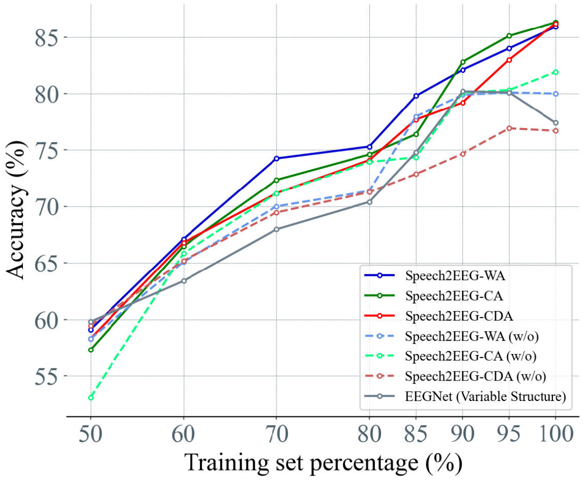


Fig. 3. Comparison of classification performance on different training dataset percentages on the BCI IV-2a dataset. (w/o) denotes the Speech2EEG model not using pretrained weights.

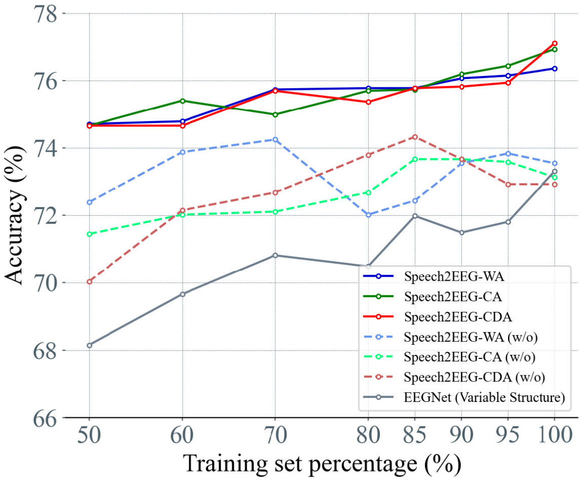
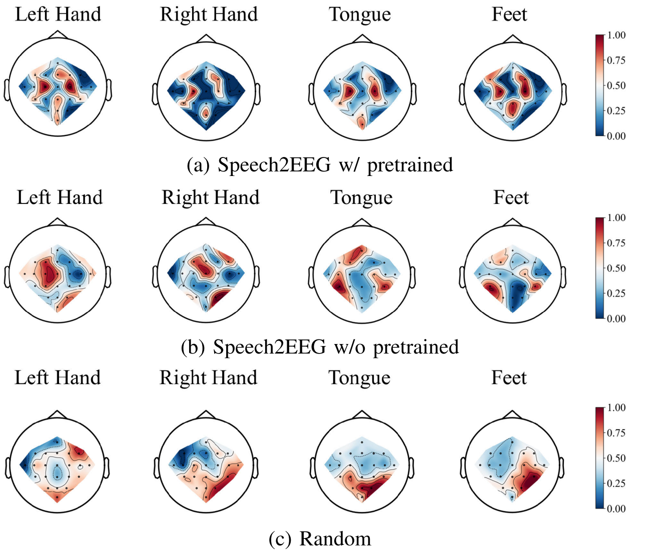


Fig. 4. Comparison of classification performance on different training dataset percentages on the BCI IV-2b dataset. (w/o) denotes the Speech2EEG model not using pretrained weights.

benefit from the increase in training EEG samples from the training dataset scale of 50% to 90%. The performance of the EEGNet architecture peaks at 90% of the training data but drops when provided more EEG samples in the training phase. We consider the main reason why EEGNet fails to benefit from a larger training dataset is its insufficient scale and depth of its architecture. In contrast, Speech2EEG-WA and Speech2EEG-CA models without the pretrained weights perform similarly to the baseline EEGNet model when sufficient training data is presented. However, the Speech2EEG-CDA model is consistently worse than the baseline EEGNet model. This is aligned with the empirical finding that a deeper and larger neural network will often be more difficult to train without a sufficiently large data corpus [38]. After all, without the pretrained weights, the Speech2EEG models are similar to a deep convolutional network with certain channel aggregation structures. Our proposed Speech2EEG method bypasses this problem through the introduction of pretrained feature extractors, which avoids the difficult training of a mature deep neural network using noisy EEG signals. With well-structured multichannel temporal embeddings, the subsequent feature aggregation and classification networks are more likely to construct useful features after fine-tuning. As a result, the Speech2EEG architectures benefit from the increase in the scale of the training dataset and the neural network. As for the BCI IV-2b dataset, Fig.4 shows that on this 2 classes EEG dataset, Speech2EEG models with pretrained weights perform substantially outperform the baseline EEGNet model as well as Speech2EEG models without pretrained weights. The result on BCI IV-2b further demonstrates the positive impact of pretrained weights on the classification performance.



(a) Speech2EEG w/ pretrained

(b) Speech2EEG w/o pretrained

(c) Random

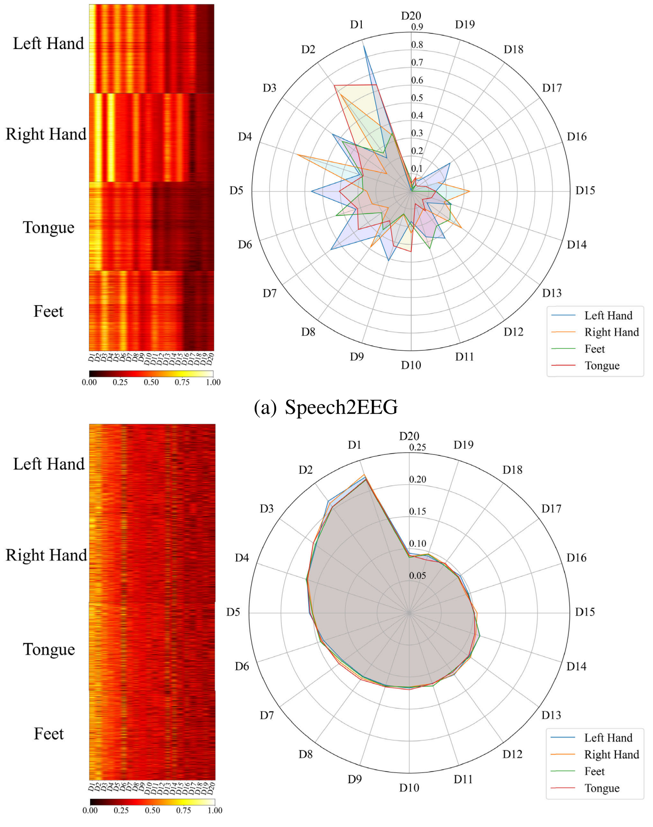
Fig. 5. Topology map of salient channels. The red regions indicate a higher importance during classification, while the blue regions indicate a lower importance.

I. Visualization

A major advantage of the proposed Speech2EEG method is that it effectively adapts a mature feature extraction network pretrained on large-scale speech data to extract useful temporal embeddings from the EEG data. In this section, we utilize the InputGradient method [104] to determine how our Speech2EEG model distinguishes one EEG category from another. Due to the fact that the BCI IV-2b dataset only utilizes 3 electrodes corresponding to the C3, Cz, and C4 locations while the BCI IV-2a dataset has more complete coverage of the brain area with a total of 22 electrodes, we consider the BCI IV-2b dataset has better visualization effects when demonstrating the performance of the proposed Speech2EEG model.

First, we visualize topology maps of the saliency of the EEG channels from a Speech2EEG model using pretrained speech embeddings, a Speech2EEG model trained from scratch without using pretrained speech embeddings, and a randomly initialized Speech2EEG model in Fig.5. Fig.5a shows that when using pretrained speech embeddings, Speech2EEG is more focused on channels located in the middle of the scalp, which is aligned with the occurrence of ERD signals in previous research [73], [79]. In Fig.5b, without the adaptation of a mature feature extraction network, although the model can still make use of channels located in the middle scalp, it pays too much attention to the channels at the edge of the scalp, which are more vulnerable to noise. In contrast, the topology map of the randomly initialized model (Fig.5c), does not respond to a specific brain region.

The saliency map is the most commonly used method to measure and visualize the spatial support of a particular class in each input signal. In EEG signal classification, since each channel represent a particular spatial location of the human scalp, the saliency map can provide information about which brain area are the most informative when the neural network tries to classify the input EEG signal. The topographical saliency map for the four motor imagery category from the proposed Speech2EEG model with and without pretrained weights is shown in Fig.5a and Fig.5b respectively. For better visualization, the normalized saliency maps of different samples are averaged to get the mean saliency map for each motor imagery category. Values are normalized into the range [0, 1]. The red color in the brain topology maps denotes high salient and informative brain area while the blue color denotes low salient and less informative brain area in the motor imagery task. The upper side of the saliency map is corresponding to the frontal cortex area while the lower side is corresponding to the posterior cortex area. Additionally, the saliency map of a randomly initialized Speech2EEG model is also shown in Fig.5c for better sanity check [105]. As can be seen from Fig.5a, for input signals from the four motor imagery categories, the saliency map shows that the central (near the C3 and C4 location) and the posterior (near the Pz location) of the brain are the most informative in the network's perspective. This is consistent with the previous literatures that study the responding distribution of brain signals of motor imagery [11], [73], [79], [106], [107], [108]. According to Brodmann brain function partition [109], brain area around the C3 and C4 location correspond to the primary motor cortex and is mainly associated with the sensorimotor functions. On the other hand, the brain area near the Pz location are correspond to visuo-motor coordination as well as brain functions for perception and processing of stimuli related to the senses [110]. In the view of deep learning model, by focusing on the key areas of ERD signal occurrence during motor imagery can avoid negative influence of the noise and other brain activities. Thus, it is beneficial to the recognition of different EEG categories. In Fig.5b, the saliency map for a Speech2EEG model without using the pretrained weights shows that the model finds the central area informative for the left hand, right hand, tongue and the feet categories. This is similar to the salient area from Fig.5a. The correct focus on the motor-related brain area could provide sanity support for how the Speech2EEG model without pretrained weights managed to correctly classify a certain amount of input EEG signals. However, it also considers frontal areas around the F7 and F8 locations as well as the posterior areas around the P7 and P8 locations. Since the functionality of these area is less relevant to the motor imagery activities, the model is more vulnerable to noise and influence from irrelevant brain activities. Finally, the saliency map of the randomly initialized model (Fig.5c) completely omit the motor imaginary regions and thus can only perform random guess towards an input EEG signal. The comparison between trained model and randomly initialized model shows that the saliency results are independent of the data and that the models indeed learn to exploit the spatial information from the EEG data reliable.



(a) Speech2EEG

(b) Random

Fig. 6. Visualization of the top-20 important embedding dimensions of the Speech2EEG model compared to a randomly initialized Speech2EEG network. On the left, the saliency map of the important temporal embedding dimensions for all MI categories is plotted from left to right. On the right is the normalized value of the top-20 important temporal embedding dimensions, where D1 denotes the top-1 important temporal embedding dimension and D20 denotes the top-20 important temporal embedding dimension.

In addition to the topology map, we visualize the top-20 most impactful dimensions of the multichannel temporal embeddings by the same saliency method (reindexed as D1 to D20). The saliency values for these dimensions are displayed using the saliency maps on the left side of in Fig.6a and Fig.6b while the value for each dimension is displayed using the radar maps located on the right. Values of the saliency maps are normalized into the range [0, 1] to facilitate display. For each motor imagery category, the x-axis of the saliency map represents the index of the embedding, while the y-axis represents each data sample in the test set. The displayed dimensions index and the order are consistent for each motor imagery category. As depicted in the saliency maps from Fig.6a, the Speech2EEG model focuses on a particular set of these embedding dimensions when predicting each brain activity. While the saliency maps from a randomly initialized model (Fig.6a) display identical saliency distribution among all categories. This result shows that after adapting the pretrained feature extraction module to the EEG dataset, the Speech2EEG model discovers a certain group of temporal information to be consistent and useful for this motor imagery classification task. Such that the model can rely on more diverse viewpoints of the EEG waveform signal to support its prediction. As for the radar map on the right side of Fig.6a, the resulting value for each dimension varies greatly among different brain activities. This indicates that the adapted speech feature extractors can capture discriminative patterns in unseen EEG data during the testing phase. Therefore, the adaptation process (i.e., fine-tuning) can be an effective way to introduce good feature extractors learned from speech data to EEG data. The adaptation preprocess is similar to the transfer of classical acoustic filters to EEG processing in existing research. In contrast, embeddings from a randomly initialized model exhibit no difference among MI categories and can only produce guess-level accuracy (Fig.6b).

V. Conclusion and Future Work

In this paper, we demonstrate that pretrained features from a large-scale speech processing model can be used to improve performance for EEG signal analysis. Using feature extraction networks pretrained using speech, we obtain multichannel temporal embeddings from raw EEG data. These embeddings are further processed using a feature aggregation network with a relatively simple structure. A total of 3 feature aggregation structures are designed in our study to utilize these adapted speech embeddings. Experimental results show that the proposed method can achieve state-of-the-art results on the BCI IV-2a and BCI IV-2b datasets. Our findings suggest the potential for using existing pretrained speech models to improve the performance of EEG signal classification with a more flexible network design. In particular, ablation studies on both BCI IV-2a and 2b datasets suggest that when the scale of training data is reduced by up to 40%, the proposed Speech2EEG method still achieves better performance compared to the popular EEGNet method. This research opens the door to building larger models for BCI systems. It is possible that we can utilize off-the-shelf pretrained speech processing models to improve the performance in a particular BCI task. In the future, we plan to investigate more advanced transfer learning methods to further improve the overall performance of other BCI tasks. Techniques such as knowledge distillation methods could be a potential way to reduce the volume of the deep learning model while attaining reasonable performance. Furthermore, the impact of speech models pretrained on other speech datasets on the performance of EEG signal classification will also be investigated in the future.

References

[1] R. Abiri, S. Borhani, E. W. Sellers, Y. Jiang, and X. Zhao, “A comprehensive review of EEG-based brain–computer interface paradigms,” *J. Neural Eng.*, vol. 16, no. 1, Feb. 2019, Art. no. 011001.

[2] X. Wang, G. Gong, N. Li, and Y. Ma, “A survey of the BCI and its application prospect,” in *Theory, Methodology, Tools and Applications for Modeling and Simulation of Complex Systems*. Berlin, Germany: Springer, 2016, pp. 102–111.

[3] L. F. Nicolas-Alonso and J. Gomez-Gil, “Brain computer interfaces, a review,” *Sensors*, vol. 12, no. 2, pp. 1211–1279, 2012.

[4] A. B. Schwartz, X. T. Cui, D. J. Weber, and D. W. Moran, “Brain-controlled interfaces: Movement restoration with neural prosthetics,” *Neuron*, vol. 52, no. 1, pp. 205–220, Oct. 2006.

[5] M. J. Khan and K.-S. Hong, “Hybrid EEG–fNIRS-based eight-command decoding for BCI: Application to quadcopter control,” *Frontiers Neurorobotics*, vol. 11, p. 6, Feb. 2017.

[6] C. Gorman and Y.-K. Wang, “A closed-loop AR-based BCI for real-world system control,” in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2021, pp. 1–7.

[7] A. Nijholt, “BCI for games: A ‘state of the art’ survey,” in *Proc. Int. Conf. Entertainment Comput.* Cham, Switzerland: Springer, 2008, pp. 225–228.

[8] M. Tanveer et al., “Deep learning for brain age estimation: A systematic review,” *Inf. Fusion*, vol. 96, pp. 130–143, Aug. 2023.

[9] A. Rafiei and Y.-K. Wang, “Automated major depressive disorder classification using deep convolutional neural networks and Choquet fuzzy integral fusion,” in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2022, pp. 186–192.

[10] K. Li et al., “Feature extraction and identification of Alzheimer’s disease based on latent factor of multi-channel EEG,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1557–1567, 2021.

[11] G. Pfurtscheller, C. Brunner, A. Schlögl, and F. H. Lopes da Silva, “Mu rhythm (de) synchronization and EEG single-trial classification of different motor imagery tasks,” *Neuroimage*, vol. 31, no. 1, pp. 153–159, May 2006.

[12] H.-I. Suk and S.-W. Lee, “Subject and class specific frequency bands selection for multiclass motor imagery classification,” *Int. J. Imag. Syst. Technol.*, vol. 21, no. 2, pp. 123–130, 2011.

[13] M. S. Pogach, N. M. Punjabi, N. Thomas, and R. J. Thomas, “Electrocardiogram-based sleep spectrogram measures of sleep stability and glucose disposal in sleep disordered breathing,” *Sleep*, vol. 35, no. 1, pp. 139–148, Jan. 2012.

[14] R. Khosrowabadi and A. W. B. A. Rahman, “Classification of EEG correlates on emotion using features from Gaussian mixtures of EEG spectrogram,” in *Proc. 3rd Int. Conf. Inf. Commun. Technol. Moslem World (ICTM)*, Dec. 2010, pp. 102–107.

[15] A. Mazaheri and T. W. Picton, “EEG spectral dynamics during discrimination of auditory and visual targets,” *Cogn. Brain Res.*, vol. 24, pp. 81–96, Jun. 2005.

[16] G. J. Mysore and P. Smaragdis, “A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 17–20.

[17] S. Chachada and C.-C.-J. Kuo, “Environmental sound recognition: A survey,” *APSIPA Trans. Signal Inf. Process.*, vol. 3, no. 1, p. e14, 2014.

[18] Y. Wang, T.-P. Jung, and C.-T. Lin, “EEG-based attention tracking during distracted driving,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 23, no. 6, pp. 1085–1094, Nov. 2015.

[19] Y.-C. Chang, Y.-K. Wang, N. R. Pal, and C.-T. Lin, “Exploring covert states of brain dynamics via fuzzy inference encoding,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2464–2473, 2021.

[20] J. Fumanal-Idocin, Y.-K. Wang, C.-T. Lin, J. Fernandez, J. A. Sanz, and H. Bustince, “Motor-imagery-based brain–computer interface using signal derivation and aggregation functions,” *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 7944–7955, Aug. 2022.

[21] W. Zhou and J. Gotman, “Removal of EMG and ECG artifacts from EEG based on wavelet transform and ICA,” in *Proc. 26th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Sep. 2004, pp. 392–395.

[22] J. Dennis, Q. Yu, H. Tang, H. D. Tran, and H. Li, “Temporal coding of local spectrogram features for robust sound recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 803–807.

[23] Y. Shao and C.-H. Chang, “Bayesian separation with sparsity promotion in perceptual wavelet domain for speech enhancement and hybrid speech recognition,” *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 41, no. 2, pp. 284–293, Mar. 2011.

[24] C. Cooney, R. Folli, and D. Coyle, “Mel frequency cepstral coefficients enhance imagined speech decoding accuracy from EEG,” in *Proc. 29th Irish Signals Syst. Conf. (ISSC)*, Jun. 2018, pp. 1–7.

[25] A. Kandaswamy, C. S. Kumar, R. P. Ramanathan, S. Jayaraman, and N. Malmurugan, "Neural classification of lung sounds using wavelet coefficients," *Comput. Biol. Med.*, vol. 34, no. 6, pp. 523–537, 2004.

[26] Z. Tufekci and J. N. Gowdy, "Feature extraction using discrete wavelet transform for speech recognition," in *Proc. IEEE SoutheastCon, Preparing New Millennium*, Apr. 2000, pp. 116–123.

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

[28] C. H. Nguyen, G. K. Karavas, and P. Artemiadis, "Inferring imagined speech using EEG signals: A new approach using Riemannian manifold features," *J. Neural Eng.*, vol. 15, no. 1, Feb. 2018, Art. no. 016002.

[29] T. Hermann and A. Hunt, "The importance of interaction in sonification," in *Proc. DBLP*, 2004, pp. 1–8.

[30] 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.

[31] C.-T. Lin, J. Liu, C.-N. Fang, S.-Y. Hsiao, Y.-C. Chang, and Y.-K. Wang, "Multistream 3-D convolution neural network with parameter sharing for human state estimation," *IEEE Trans. Cognit. Develop. Syst.*, vol. 15, no. 1, pp. 261–271, Mar. 2023.

[32] C.-T. Lin, C.-H. Chuang, Y.-C. Hung, C.-N. Fang, D. Wu, and Y.-K. Wang, "A driving performance forecasting system based on brain dynamic state analysis using 4-D convolutional neural networks," *IEEE Trans. Cybern.*, vol. 51, no. 10, pp. 4959–4967, Oct. 2021.

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

[34] S. Lemm, G. Curio, Y. Hlushchuk, and K.-R. Muller, "Enhancing the signal-to-noise ratio of ICA-based extracted ERPs," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 4, pp. 601–607, Apr. 2006.

[35] J. Wang and M. Wang, "Review of the emotional feature extraction and classification using EEG signals," *Cognit. Robot.*, vol. 1, pp. 29–40, 2021.

[36] H. Altaheri et al., "Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: A review," *Neural Comput. Appl.*, pp. 1–42, Aug. 2021.

[37] T. M. Ingolfsson, M. Hersche, X. Wang, N. Kobayashi, L. Cavigelli, and L. Benini, "EEG-TCNet: An accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2020, pp. 2958–2965.

[38] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data," *Frontiers Hum. Neurosci.*, vol. 15, p. 253, Jun. 2021.

[39] C. Herff and T. Schultz, "Automatic speech recognition from neural signals: A focused review," *Frontiers Neurosci.*, vol. 10, p. 429, Sep. 2016.

[40] A. Défossez, C. Caucheteux, J. Rapin, O. Kabeli, and J.-R. King, "Decoding speech from non-invasive brain recordings," 2022, *arXiv:2208.12266*.

[41] J. Millet et al., "Toward a realistic model of speech processing in the brain with self-supervised learning," 2022, *arXiv:2206.01685*.

[42] M. J. Monesi, B. Accou, T. Francart, and H. Van Hamme, "Extracting different levels of speech information from EEG using an LSTM-based model," 2021, *arXiv:2106.09622*.

[43] V. Gudivada, A. Apon, and J. Ding, "Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations," *Int. J. Adv. Softw.*, vol. 10, no. 1, pp. 1–20, 2017.

[44] Z. Wan, R. Yang, M. Huang, N. Zeng, and X. Liu, "A review on transfer learning in EEG signal analysis," *Neurocomputing*, vol. 421, pp. 1–14, Jan. 2021.

[45] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[46] S. Bhattacharyya, A. Konar, D. N. Tibarewala, and M. Hayashibe, "A generic transferable EEG decoder for online detection of error potential in target selection," *Frontiers Neurosci.*, vol. 11, p. 226, May 2017.

[47] Z. Lan, O. Sourina, L. Wang, R. Scherer, and G. R. Müller-Putz, "Domain adaptation techniques for EEG-based emotion recognition: A comparative study on two public datasets," *IEEE Trans. Cogn. Devel. Syst.*, vol. 11, no. 1, pp. 85–94, Mar. 2019.

[48] J. Li, S. Qiu, Y.-Y. Shen, C.-L. Liu, and H. He, "Multisource transfer learning for cross-subject EEG emotion recognition," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3281–3293, Jul. 2020.

[49] Y.-P. Lin and T.-P. Jung, "Improving EEG-based emotion classification using conditional transfer learning," *Frontiers Hum. Neurosci.*, vol. 11, p. 334, Jun. 2017.

[50] H. Daoud and M. A. Bayoumi, "Efficient epileptic seizure prediction based on deep learning," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 5, pp. 804–813, Oct. 2019.

[51] S. Raghu, N. Sriraam, Y. Temel, S. V. Rao, and P. L. Kubben, "EEG based multi-class seizure type classification using convolutional neural network and transfer learning," *Neural Netw.*, vol. 124, pp. 202–212, Apr. 2020.

[52] C.-S. Wei, T. Koike-Akino, and Y. Wang, "Spatial component-wise convolutional network (SCCNet) for motor-imagery EEG classification," in *Proc. 9th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, Mar. 2019, pp. 328–331.

[53] H. Wu et al., "A parallel multiscale filter bank convolutional neural networks for motor imagery EEG classification," *Frontiers Neurosci.*, vol. 13, p. 1275, Nov. 2019.

[54] D. Wu, Y. Xu, and B.-L. Lu, "Transfer learning for EEG-based brain–computer interfaces: A review of progress made since 2016," *IEEE Trans. Cogn. Developmental Syst.*, vol. 14, no. 1, pp. 4–19, Mar. 2020.

[55] X. Chai, Q. Wang, Y. Zhao, X. Liu, O. Bai, and Y. Li, "Unsupervised domain adaptation techniques based on auto-encoder for non-stationary EEG-based emotion recognition," *Comput. Biol. Med.*, vol. 79, pp. 205–214, Dec. 2016.

[56] Z. Yin and J. Zhang, "Cross-session classification of mental workload levels using EEG and an adaptive deep learning model," *Biomed. Signal Process. Control*, vol. 33, pp. 30–47, Mar. 2017.

[57] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "EmotionMeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2019.

[58] H. Banville, I. Albuquerque, A. Hyvarinen, G. Moffat, D.-A. Engemann, and A. Gramfort, "Self-supervised representation learning from electroencephalography signals," in *Proc. IEEE 29th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Oct. 2019, pp. 1–6.

[59] D. Bethge et al., "EEG2 Vec: Learning affective EEG representations via variational autoencoders," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2022, pp. 3150–3157.

[60] T.-P. Jung and T. J. Sejnowski, "Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 96–107, Jan. 2022.

[61] X.-Z. Zhang, W.-L. Zheng, and B.-L. Lu, "EEG-based sleep quality evaluation with deep transfer learning," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2017, pp. 543–552.

[62] Y.-M. Jin, Y.-D. Luo, W.-L. Zheng, and B.-L. Lu, "EEG-based emotion recognition using domain adaptation network," in *Proc. Int. Conf. Orange Technol. (ICOT)*, Dec. 2017, pp. 222–225.

[63] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 4835–4839.

[64] I. Obeid and J. Picone, "The temple university hospital EEG data corpus," *Frontiers Neurosci.*, vol. 10, p. 196, May 2016.

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

[66] H. Altaheri, G. Muhammad, and M. Alsulaiman, "Physics-informed attention temporal convolutional network for EEG-based motor imagery classification," *IEEE Trans. Ind. Informat.*, vol. 19, no. 2, pp. 2249–2258, Feb. 2023.

[67] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[68] H. Wang, H. Yu, and H. Wang, "EEG_GENet: A feature-level graph embedding method for motor imagery classification based on EEG signals," *Biocybernetics Biomed. Eng.*, vol. 42, no. 3, pp. 1023–1040, Jul. 2022.

[69] S. U. Amin, H. Altaheri, G. Muhammad, W. Abdul, and M. Alsulaiman, "Attention-inception and long short-term memorybased electroencephalography classification for motor imagery tasks in rehabilitation," *IEEE Trans. Ind. Informat.*, vol. 18, no. 8, pp. 5412–5421, Aug. 2022.

[70] Y. Duan, Z. Wang, Y. Li, J. Tang, Y.-K. Wang, and C.-T. Lin, "Cross task neural architecture search for EEG signal classifications," 2022, *arXiv:2210.06298*.

[71] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," 2019, *arXiv:1910.05453*.

[72] G. Pfurtscheller, C. Neuper, and W. Mohl, "Event-related desynchronization (ERD) during visual processing," *Int. J. Psychophysiol.*, vol. 16, nos. 2–3, pp. 147–153, May 1994.

[73] S. Afrakhteh and M. R. Mosavi, "Applying an efficient evolutionary algorithm for EEG signal feature selection and classification in decision-based systems," in *Energy Efficiency of Medical Devices and Healthcare Applications*. Amsterdam, The Netherlands: Elsevier, 2020, pp. 25–52.

[74] C. Zich, S. Debener, A.-K. Thoene, L.-C. Chen, and C. Kranczioch, "Simultaneous EEG-fNIRS reveals how age and feedback affect motor imagery signatures," *Neurobiol. Aging*, vol. 49, pp. 183–197, Jan. 2017.

[75] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12449–12460.

[76] L. Bergman, N. Cohen, and Y. Hoshen, "Deep nearest neighbor anomaly detection," 2020, *arXiv:2002.10445*.

[77] K. Roth, L. Pemula, J. Zepeda, B. Scholkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14318–14328.

[78] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[79] K. Nakayashiki, M. Saeki, Y. Takata, Y. Hayashi, and T. Kondo, "Modulation of event-related desynchronization during kinematic and kinetic hand movements," *J. NeuroEng. Rehabil.*, vol. 11, no. 1, pp. 1–9, Dec. 2014.

[80] C. Zhang, Y.-K. Kim, and A. Eskandarian, "EEG-inception: An accurate and robust end-to-end neural network for EEG-based motor imagery classification," *J. Neural Eng.*, vol. 18, no. 4, Aug. 2021, Art. no. 046014.

[81] M. Riyad, M. Khalil, and A. Adib, "MI-EEGNET: A novel convolutional neural network for motor imagery classification," *J. Neurosci. Methods*, vol. 353, Apr. 2021, Art. no. 109037.

[82] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

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

[84] S. Sakhavi, C. Guan, and S. Yan, "Learning temporal information for brain–computer interface using convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5619–5629, Nov. 2018.

[85] Y. Ozaki, Y. Tanigaki, S. Watanabe, and M. Onishi, "Multiobjective tree-structured Parzen estimator for computationally expensive optimization problems," in *Proc. Genetic Evol. Comput. Conf.*, Jun. 2020, pp. 533–541.

[86] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 2623–2631.

[87] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "BCI competition 2008—Graz data set A," Lab. Brain-Comput. Interfaces, Inst. Knowl. Discovery, Graz Univ. Technol., 2008, pp. 1–6, vol. 1.

[88] R. Leeb, C. Brunner, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "BCI competition 2008—Graz data set B," Graz Univ. Technol., Graz, Austria, 2008, pp. 1–6.

[89] F. Li, F. He, F. Wang, D. Zhang, Y. Xia, and X. Li, "A novel simplified convolutional neural network classification algorithm of motor imagery EEG signals based on deep learning," *Appl. Sci.*, vol. 10, no. 5, p. 1605, Feb. 2020.

[90] G. A. Altuwaijri, G. Muhammad, H. Altaheri, and M. Alsulaiman, "A multi-branch convolutional neural network with squeeze-and-excitation attention blocks for EEG-based motor imagery signals classification," *Diagnostics*, vol. 12, no. 4, p. 995, Apr. 2022.

[91] Y. K. Musallam et al., "Electroencephalography-based motor imagery classification using temporal convolutional network fusion," *Biomed. Signal Process. Control*, vol. 69, Aug. 2021, Art. no. 102826.

[92] K. Keng Ang, Z. Yang Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (FBCSP) in brain–computer interface," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jun. 2008, pp. 2390–2397.

[93] P. Gaur, R. B. Pachori, H. Wang, and G. Prasad, "An empirical mode decomposition based filtering method for classification of motor-imagery EEG signals for enhancing brain–computer interface," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2015, pp. 1–7.

[94] H. Dose, J. S. Møller, H. K. Iversen, and S. Puthusserypady, "An end-to-end deep learning approach to MI-EEG signal classification for BCIs," *Exp. Syst. Appl.*, vol. 114, pp. 532–542, Dec. 2018.

[95] H. Zhao, Q. Zheng, K. Ma, H. Li, and Y. Zheng, "Deep representation-based domain adaptation for nonstationary EEG classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 535–545, Feb. 2021.

[96] S. Shahid and G. Prasad, "Bispectrum-based feature extraction technique for devising a practical brain–computer interface," *J. Neural Eng.*, vol. 8, no. 2, Apr. 2011, Art. no. 025014.

[97] D. F. Collazos-Huertas, A. M. Álvarez-Meza, C. D. Acosta-Medina, G. A. Castaño-Duque, and G. Castellanos-Dominguez, "CNN-based framework using spatial dropping for enhanced interpretation of neural activity in motor imagery classification," *Brain Informat.*, vol. 7, no. 1, pp. 1–13, Dec. 2020.

[98] J. Xu, H. Zheng, J. Wang, D. Li, and X. Fang, "Recognition of EEG signal motor imagery intention based on deep multi-view feature learning," *Sensors*, vol. 20, no. 12, p. 3496, Jun. 2020.

[99] T. Liu and D. Yang, "A densely connected multi-branch 3D convolutional neural network for motor imagery EEG decoding," *Brain Sci.*, vol. 11, no. 2, p. 197, Feb. 2021.

[100] J.-S. Bang, M.-H. Lee, S. Fazli, C. Guan, and S.-W. Lee, "Spatio-spectral feature representation for motor imagery classification using convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 7, pp. 3038–3049, Jul. 2022.

[101] P. Wang, A. Jiang, X. Liu, J. Shang, and L. Zhang, "LSTM-based EEG classification in motor imagery tasks," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 11, pp. 2086–2095, Nov. 2018.

[102] X. Deng, B. Zhang, N. Yu, K. Liu, and K. Sun, "Advanced TSGL-EEGNet for motor imagery EEG-based brain–computer interfaces," *IEEE Access*, vol. 9, pp. 25118–25130, 2021.

[103] D. Kostas and F. Rudzicz, "Thinker invariance: Enabling deep neural networks for BCI across more people," *J. Neural Eng.*, vol. 17, no. 5, Oct. 2020, Art. no. 056008.

[104] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.

[105] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[106] S. Lacey and R. Lawson, *Multisensory Imagery*. Berlin, Germany: Springer, 2013.

[107] A. Hassanpour, M. Moradikia, H. Adeli, S. R. Khayami, and P. Shamsinejadbabaki, "A novel end-to-end deep learning scheme for classifying multi-class motor imagery electroencephalography signals," *Exp. Syst.*, vol. 36, no. 6, Dec. 2019, Art. no. e12494.

[108] D. J. McFarland and J. R. Wolpaw, "EEG-based brain–computer interfaces," *Current Opinion Biomed. Eng.*, vol. 4, pp. 194–200, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S246845111730082X, doi: 10.1016/j.cobme.2017.11.004.

[109] K. Amunts and K. Zilles, "Architectonic mapping of the human brain beyond Brodmann," *Neuron*, vol. 88, no. 6, pp. 1086–1107, 2015.

[110] Y. Iwamura, "Somatosensory association cortices," in *International Congress Series*, vol. 1250. Amsterdam, The Netherlands: Elsevier, 2003, pp. 3–14.