his article has been accepted for publication in a future issue of this conference proceedings, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/SSCI50451.2021.9659960, 2021 IEEE Symposium Series on Computational Intelligence (SSCI)

Electrode Selection and Convolutional Attention Network for Recognition of Silently Spoken Words from EEG Signals

Sahil Datta
Department of Electronic and Electrical Engineering
College of Engineering, Design and Physical Sciences
Brunel University
London, United Kingdom
sahil.datta@brunel.ac.uk

Jorunn Jo Holmberg
Division of Psychology
Department of Life Sciences
College of Health, Medicine and Life Sciences
Brunel University
London, United Kingdom
jo.holmberg-hansson@brunel.ac.uk

Elena Antonova
Division of Psychology
Department of Life Sciences and Centre for Cognitive Neuroscience
College of Health, Medicine and Life Sciences
Brunel University
London, United Kingdom
elena.antonova@brunel.ac.uk

***Abstract*—Brain signals generated during silent speech have shown to be useful in designing a communication-based brain computer interface (BCI). However, brain signals are non-stationery and complex in nature, and therefore challenging to recognize. We propose a framework for recognizing imagined words using brain signals captured through electroencephalograph (EEG) sensors. Our method consists of two main components: (i) an electrode selection method, and (ii) a convolutional attention network. The electrode selection method provides the electrodes containing the most discriminative time-frequency information for imagined speech recognition. Further, spectrograms from selected electrodes are used as input to the convolutional attention network, that extracts time-frequency features and performs classification by ascribing higher importance to the time points with higher discriminatory capacity. Experimental results using EEG dataset shows that the proposed method efficiently recognizes mentally spoken words and exhibits performance superior to that of state-of-the-art methods.**

***Index Terms*—EEG, Brain Computer Interface, Convolutional Network, Attention, Inner Speech, Silent Speech, Electrode Selection, Time-Frequency**

I. Introduction

The use of electroencephalography (EEG) signals has grown in prominence for a wide array of applications during the last few decades [1]. Simultaneously, the analysis and interpretation of EEG signals generated during imagined speech has sparked considerable scientific attention [2]. As brain signals are non-stationary and challenging to analyze, previous techniques have concentrated on binary classification of silently spoken words [2], [3], [4]. Sereshkeh [3] employed multi-linear perceptron neural networks and attained a classification rate of 63% with two classes and 54% with three classes. In addition, a new method for binary classification of imagined words based on a covariance matrix descriptor was proposed by Nguyen [2], which achieved a classification rate of 50% for short words and 66% for long words. Further, a convolutional neural network-long short-term memory units (CNN-LSTM) and deep auto-encoders was used to recognize the presence of an articulation from silent speech EEG signals [5]. Similarly, different articulations were recognized from silent speech using wavelet features and deep learning classifier [6].

Although the above approaches show partial success, recognizing neural events associated with a given word remains a challenge due to the fleeting nature of a single event. Humans are capable of producing a word in 0.33-0.5 seconds [7], therefore, recognizing neural events associated with a given word becomes complicated. Furthermore, feature extraction methodologies have limitations; for example, Fast Fourier Transform (FFT) is most effective when the signals are stationary [8], while auto-regressive modelling can suffer from poor spectral estimation of signal [9]. The most popular features used in EEG research are time-frequency features, such as the spectrograms [10]. On the other hand, deep learning has been successful in feature learning for brain computer interface (BCI) applications [11]. In addition, attention-based neural networks have improved accuracy in several deep learning tasks, especially with sequential data [12]. To exploit the ability of neural networks to learn non-linear features, in this paper we introduce an application of the convolutional attention network for silent speech recognition.

In order to capture EEG signals from the different scalp regions multiple electrodes are needed, resulting in a high dimensionality of the EEG data. EEG signals from a large number of electrodes require extra computational resources and time to analyze [13]. For optimization, it is crucial to minimize the dimensionality of the data by appropriately choosing the electrodes. The standard procedure within the field is aggregating features from different electrodes into a vector; however, this technique does not exploit the spatial correlation between electrodes [11]. Therefore, in this paper we introduce an efficient electrode selection technique to reduce the EEG data dimensionality. Further, to explore the

his article has been accepted for publication in a future issue of this conference proceedings, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/SSCI50451.2021.9659960, 2021 IEEE Symposium Series on Computational Intelligence (SSCI)

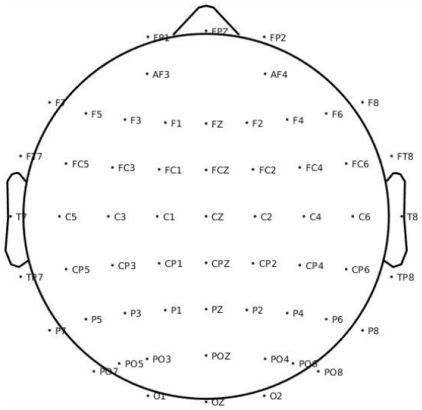


Fig. 1: Electrodes in the head-cap used for data acquisition.

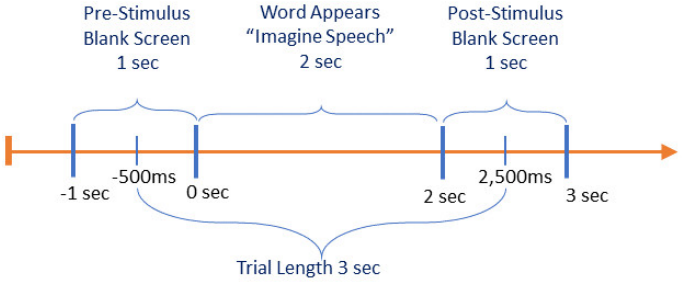


Fig. 2: The order of stimulus presentation for recording a single EEG trial. In our evaluation, we used only 3 sec of activity per trial to avoid overlapping of inter-trial activity.

spatial correlation between electrodes, spectrograms from the selected electrodes were processed as a multi-channel input by the proposed convolutional attention network.

To summarize, the proposed framework for recognition of silent speech from EEG signals presented here, combines the following novel features:

1) An electrode selection method that reduces the dimensionality of the EEG data providing the most discriminatory information for recognition of silently spoken words.
2) A convolutional attention network that treats the spectrograms as a time-varying input and uses convolutional layers to extract features from each time point separately. In addition, we used the self-attention layer for learning the most discriminatory temporal features associated with silently spoken word.
3) The evaluation of the proposed approach, combining electrode selection technique with the convolutional attention network obtained high accuracy in recognizing silently spoken words. In addition, our results demonstrate robustness of the proposed method in comparison to previous techniques.

II. Data Recording and Pre-Processing

A. EEG Head-Cap

To record EEG data, we used a Neuroscan 64 channel Quik cap (electrodes) with extended 10-20 system, which included the horizontal electrooculogram (HEOG) and vertical electrooculogram (VEOG) electrodes for eye blink measurements. The position of electrodes in the head-cap is shown in Figure 1. The cap was connected to a synamp amplifier operating at a 1kHz sampling rate. The amplifier was connected to the system where signals were being recorded in Neuroscan Curry 8.

B. Data Collection

An EEG dataset was acquired from 12 participants (Mean age 37, range 21-71). The inclusion criteria were: (i) fluency in English and (ii) no neurological or speech impairments. The recording was performed in a laboratory specifically designed for EEG-based experiments at Brunel University London, UK. Participants sat on a chair one meter away from the computer screen and were requested to stay stationary throughout the recording. Each participant performed four tasks; for the purpose of this study, only covert speech signals were used, i.e., when the participant was requested to silently speak the presented word. The participants were informed of their right to withdraw from the study at any time. The study has been approved by the College of Engineering, Design, and Physical Sciences Research Ethics Committee, Brunel University London, reference number 7361-LR-Sep/2017-8301-1.

Each recording session included words *Apple* and *Write*. The words were select from the list of most frequently used words in spoken and written English [14]. In addition, both words match phonetically, having the same number of letters and syllables, as well as being affectively neutral. To eliminate the temporal effects, words were displayed randomly on the screen [15]. E-prime-2 software was used to design the experimental paradigm. The words were presented on a white background, black in color with capital letters. The white background was used to reduce the potential caused by visual stimulus [16]. First, a blank screen appeared for 1 sec, where the participant did not perform any activity. This was followed by the word appearing for 2 sec, where the participant was told to mentally read the word as soon as it was presented. The word presentation was followed by another blank screen for 1 sec. Each trial lasted 4 sec; however, only 3 sec (500 ms before and 2,500 ms after stimulus onset) of trial was used to avoid overlapping of EEG activity between trials, as shown in Figure 2. Each participant performed ten trials for each word. The data recording was time-locked to ensure that the stimulus appeared on the screen at the correct time. To minimize fatigue, participants were given a break half way through the experiment.

C. Pre-Processing

Noise and artifacts such as eye blinks, eye movements, breathing, and muscle movement were removed from the EEG signals. The raw EEG data were filtered using a $0.01Hz$ high-pass filter to remove artefacts such as slow voltage shifts, which occur at frequencies less than 0.1 Hz. A notch filter was utilized to remove the $50Hz$ line noise. The

his article has been accepted for publication in a future issue of this conference proceedings, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/SSCI50451.2021.9659960, 2021 IEEE Symposium Series on Computational Intelligence (SSCI)

EMG electrode was used to eliminate the noise at higher frequencies, such as noise caused by muscle movement. To account for eye movement artifacts, the peak-to-peak voltage of the VEOG signal was measured in conjunction with the threshold voltage of $\pm 200\,\mu V$ [16]. Baseline raw data were corrected in real time and during offline processing.

III. Electrode Selection using Time-Frequency Information

A. Short Time Fourier Transform

Temporal characteristics alone are incapable of capturing the properties of EEG data. For instance, signals generated by two distinct triggers or events (different words in this case) may be comparable in terms of head mappings and neural activity, but different in terms of frequency characteristics [17]. As a result, we used the Short Time Fourier Transform (STFT) to analyze the brain signals in this study. Windowing was used during the STFT calculation to avoid discontinuities referred to as leakage. We employed the Hann window in our STFT implementation [8]. The length of the window was taken to be 256 and a temporal overlap of 87% was used between consecutive windows. The short window provides excellent temporal resolution and aids in detecting the events in the signal. Additionally, we performed baseline normalization to avoid the reduced power representation at high frequencies [8]. Baseline normalization also helps in highlighting task-discriminative activity from background activity [4].

B. Electrode Selection using Spectrograms

EEG signals recorded from multiple electrodes result in increased computational complexity and processing time. Further, EEG signals from multiple electrodes can lead to overfitting and poor performance of a BCI system [13]. Therefore, an efficient electrode selection technique is essential for improving speed and recognition rate of a BCI system. We propose an electrode selection method that provides the most task-discriminative electrodes for recognition of imagined speech.

The proposed method (Algorithm 1) takes input spectrograms from all electrodes and selects the top-K electrodes that contribute most towards recognition of silently spoken word. Where K is a user defined parameter. Input to the electrode selection algorithm was the training data $X \in \mathbb{R}^{n \times C \times T \times F}$, where n denotes the total number of training trials, T denotes the total number of time points in the spectrogram, F denotes the total number of frequency points in the spectrogram, and C is the total number of electrodes.

The first step is to average the spectrograms $S_{t,f}$ along the time axis. This is performed on spectrograms belonging to all trials and electrodes. This generates a frequency vector S_f for each electrode in a trial. In the next step, S_f from each electrode is divided into j overlapping vectors of varying length. A j^{th} overlapping vector $S_{f'_j}$ is obtained from the vector S_f, where $f - m_j$ refers to the length of the new vector $S_{f'_j}$. In our analysis, we created three new vectors (j=3); $S_{f'_1}$, $S_{f'_2}$, $S_{f'_3}$. All the vectors are of varying length, such that the length of vector $S_{f'_1}$ is less than $S_{f'_2}$, and $S_{f'_3}$ being the longest. For each vector $S_{f'_j}$ mean was estimated, with a_j being the mean of j^{th} vector. This was done in order to capture the magnitude of activity from different frequency bands. The proposed method regards an electrode as informative if the mean a_j of the vector $S_{f'j}$ is above zero. Mean was calculated to provide a measure of overall power within the vector $S_{f'j}$. A matrix X_f of size $n \times C$ is obtained, which contains information about C electrodes from all n trials. The C^{th} position of the matrix contained 1 or 0 value, depending on the mean value a_j of a vector (Algorithm 1, step 8). All values in $X_f(n, C)$ were added along n axis, to obtain an array. From the array, indices of K electrodes with highest values were retrieved. These were regarded as the K most informative electrodes. The top-K electrodes were estimated using training data and same electrodes were used in the test data for classification purposes.

Algorithm 1 Select Top-K Electrodes

Requires: A matrix $X \in \mathbb{R}^{n \times C \times T \times F}$, where $T \times F$ are dimensions of a spectrogram $S_{t,f}$ belonging to an electrode C of the n^{th} trial. K the no of electrodes to be selected, and m_j the vector length.

1: Calculate mean across T for each spectrogram
2: create an empty array X_f of size $n \times C$
LOOP through all trials n and electrodes C
3: **for** $n \in 1, ..., n$ **do**
4: **for** $C \in 1, ..., C$ **do**
5: S_f = Mean($S_{t,f}$, axis=0) {Mean across time axis}
Create vector of varying length from the frequency vector S_f.
6: $S_{f'_1}$, $S_{f'_j}$,..., $S_{f'_j}$ = S_{f-m_1}, S_{f-m_2},..., S_{f-m_j} {where length of $S_{f'_j}$ > $S_{f'_2}$ > $S_{f'_1}$}
7: a_1, a_2,..., a_j = Mean($S_{f'_1}$), Mean($S_{f'_2}$),..., Mean($S_{f'_j}$)
8: **if** ($a_1 > 0$ **or** $a_2 > 0$ **or** $a_j > 0$) **then**
9: $X_f(n, C)$ = 1 {C^{th} position in $X_f(n, C)$ is 1}
10: **else**
11: $X_f(n, C)$ = 0 {C^{th} position in $X_f(n, C)$ is 0}
12: **end if**
13: **end for**
14: **end for**
Create an array containing the number of times an electrode C was 1
15: L = Sum($X_f(n, C)$, axis=0) {sum across n trials}
16: L = Argsort(L) {Arrange indices of values in descending order}
17: $topK = L(1 : K)$ {Retrieve first K electrodes from the list L}
18: **return** $topK$

his article has been accepted for publication in a future issue of this conference proceedings, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/SSCI50451.2021.9659960, 2021 IEEE Symposium Series on Computational Intelligence (SSCI)

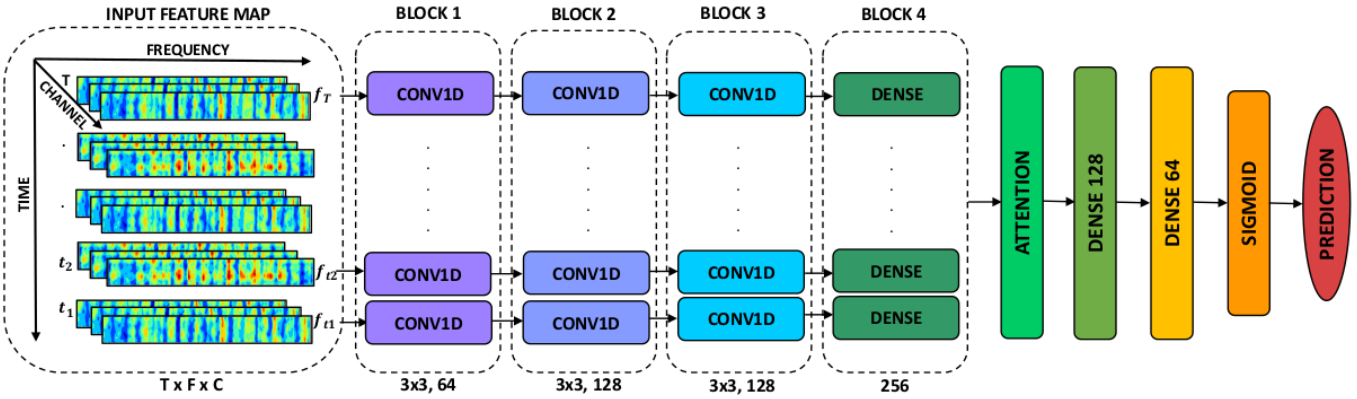


Fig. 3: Illustration of the proposed convolutional attention network. Input to the network is a multi-dimensional tensor, which is processed by a separate convolutional layer at each time point. At each time point t chain of three blocks is used to extract features from C channels, where each channel is a vector of length F = 86, which is the length of each frame input to the convolutional attention network. Two blocks containing one-dimensional convolutional and batch normalization layers other containing residual module. The output of these blocks is processed by a dense block followed by the self-attention layer for learning of important features.

IV. Classification using the Convolutional Attention Network

The proposed network architecture is shown in Figure 3. It uses the convolutional and attention layers to learn spectral and temporal patterns from the input spectrograms. At the first stage, the network uses convolutional layers to extract the important frequency components from each time point in the spectrograms. This reduces the input size and allows our network to exploit spatial correlations between neighboring electrodes. Further, the parallel dense layers perform dimensionality reduction on features extracted by the convolution blocks. We used a self-attention mechanism to learn important temporal points from the features extracted by the convolution blocks. In addition, the self-attention layer emphasizes the time points that provide most discriminative features. This makes the network capable of learning important spectro-temporal components belonging to the spectrograms of the selected electrodes.

Our proposed network consists of three convolution blocks and a dense block, proceeded by the self-attention mechanism and three dense layers, with the final dense layer performing binary classification using the $sigmoid$ function. Each block contains T parallel one-dimensional (1-D) convolution layers and batch normalization layers, where T denotes time points in the input spectrograms. 1-D convolution was used to extract the features at each time point (frequency vectors) in the spectrogram, two convolution blocks have single convolutional layer, and one block contains residual module with two convolutional layers [18]. The first block's convolutional layer filters the data using 64 kernels with a receptive field of size 3 and a stride of size 2. This procedure can extract high-level information from the spectrogram's frequency vectors. The second block in the network contains a residual module with two convolutional layers [18] with 128 kernels of size 3. The third block's convolutional layer uses 128 kernels of size 3 that are applied with a stride of size 2. The feature extracted by the convolution blocks are transferred to the dense layers for dimensionality reduction. Strategically constructed layers endowed with non-linearities can help deep learning models [19]. Thus, the exponential linear unit (ELU) [20] activation was utilized in the network to learn non-linear patterns from EEG spectrograms. The convolutional and dense layers used the ELU function.

It is known that the neural events associated with a cognitive activity of a short duration last only a few milliseconds [21]. As a result, not all EEG time points are helpful for detecting covertly spoken words. Therefore, we used the *self-attention mechanism* to highlight the most informative temporal aspects in the output generated by the convolution-dense blocks. To integrate attention in our network, we created a self-attention layer, which used the output of parallel dense layers as input to construct a more informative global feature map g. In order to implement the self-attention mechanism, we initially calculated the normalized importance vector $\alpha_\mathbf{t}$ using two successive fully connected layers: first layer (FC1) with the $tanh$ and second layer (FC2) with the $softmax$ activation function. Both layers had only one neuron and $\alpha_\mathbf{t}$ were calculated as follows:

$$\alpha_t = \frac{\exp\left(\mathbf{W_2} \cdot \tanh(\mathbf{W_1} \cdot \mathbf{h_t})\right)}{\sum_{t=1}^{T} \exp\left(\mathbf{W_2} \cdot \tanh(\mathbf{W_1} \cdot \mathbf{d_t})\right)} \tag{1}$$

where $\mathbf{d_t}$ is the output from the dense layers at each time point and $\mathbf{W_1}$, $\mathbf{W_2}$ are the weights of first and second layer. Subsequently, the global feature vector $\mathbf{g}$ [22] was calculated as:

$$\mathbf{g} = \sum_{t=1}^{T} \alpha_t \mathbf{h_t} \tag{2}$$

his article has been accepted for publication in a future issue of this conference proceedings, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/SSCI50451.2021.9659960, 2021 IEEE Symposium Series on Computational Intelligence (SSCI)

TABLE I: Subject-dependent (SD) evaluation of accuracy achieved by electrodes selected using the proposed method in section III-B. The baseline is C=64, i.e., when spectrograms from all the electrodes are used for training and testing the network, where C is the number of electrodes. The results are presented in a subject-by-subject manner.

C	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	Avg Accuracy	Avg Precision
64	55.0	54.3	67.5	71.1	100	77.7	83.8	69.0	78.7	63.5	88.8	76.5	73.8	64.8
32	51.5	73.5	68.8	73.8	99.4	76.1	88.3	64.0	71.3	66.5	92.7	83.5	75.7	67.7
9	70.0	84.5	73.3	78.3	100	81.1	88.3	85.0	70.6	67.5	87.2	75.0	80.0	69.5

where g represents output of the attention layer known as the global feature vector. The self-attention layer is trained using back-propagation and the gradient is used to learn significant temporal points [22]. Following the self-attention layer, the network contains two dense layers of 128 and 64 neurons each. The dense layers employed in the ELU as an activation function, enabling the network to transfer non-linear information from the previous layers. For binary classification, the final layer utilizes the $sigmoid$ function.

The networks were implemented using the Keras library [23] with Tensorflow backend [24]. The network was trained on NVIDIA Tesla P100 for 200 epochs, the Adam optimizer [25] was used for weight optimization, and the cross-entropy loss was minimized with the learning rate of 0.0001. Due to weight sharing in convolutional networks, the gradient at different layers can vary widely [11]; therefore, we used a slower learning rate. The model was trained using a mini-batch gradient descent of size 5 for subject-dependent (SD) and 64 for subject-independent (SI) evaluation. Further, in order to avoid the problem of unstable gradient, our network used the He weight initialization method [26].

V. Results

For the experimental evaluation of our method, we used EEG signals acquired during silent speech of two words, i.e., *Apple*, and *Write*. We tested our system for the binary classification of silently spoken words formed from the above two words. The electrodes used in the recognition task were obtained using our electrode selection method. The proposed electrode selection method was evaluated by selecting different number of electrodes. All calculated spectrograms were of dimension 50×86. Spectrograms from the selected electrodes were combined to form a multi-channel input of shape $T \times F \times C$. In the multi-channel input, C refers to the number of electrodes (channels), $T = 50$ refers to the number of time points, and $F = 86$ is the number of frequency points in the spectrograms. Therefore, input to the convolutional attention network was a three-dimensional tensor, of shape $T \times F \times C$. Electrode selection was performed on training data and electrodes for the test data were selected based on the training electrodes. The effectiveness of the proposed convolutional attention network and electrode selection method was evaluated using two evaluation methods, presented in Table II.

TABLE II: Two evaluation methods: subject-dependent (SD) and subject-independent (SI). SD was performed on a subject-by-subject basis, i.e., training and testing used different data from the same subject, and SI used data from all subjects.

Exp	Training			Testing	
	Subjects	Trials	Batch Size	Subjects	Trials
SI	All	90%	64	All	10%
SD	-	90%	5	-	10%

A. Subject-Dependent (SD)

The first evaluation method was subject-dependent (SD). In this evaluation only trials from one participant were used for training and testing of the proposed system. In general, 10 trials for each class were recorded from each participant. However, after artifact rejection some participants contributed only 9 trials. Results for each participant were obtained using leave-one-out cross validation, where the dataset was divided into 90% training and 10% testing data. To circumvent variation in the network parameters caused by the stochastic nature of deep learning algorithms [27], the convolutional attention network was trained and tested ten times for each trial. The results for each outcome were averaged per participant.

We validated the performance of the proposed electrode selection method with a baseline, i.e., when spectrograms from all the electrodes (C = 64) were used as input to the convolutional attention network. For electrode selection, we used C=9 & C=32, i.e., spectrograms were selected from the C most informative electrodes using the proposed electrode selection method to train and test the network. Training data were used for obtaining the most discriminative electrodes. Three sets of results were obtained from each participant. As shown in Table I, the results obtained with C=9 achieves the highest recognition rate of 80%, outperforming the baseline. This means that our electrode selection method reduces the dimensionality and rejects noise in the EEG data to provide more task-discriminative information. Despite the high recognition rate the precision of the network is low.

B. Subject-Independent (SI)

The second experiment evaluated the convolutional attention network and electrode selection method in a SI manner, where the network was trained and tested on EEG data from all participants. In total, 113 trials were available for each class, where 90% were used for training and 10% for testing

his article has been accepted for publication in a future issue of this conference proceedings, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/SSCI50451.2021.9659960, 2021 IEEE Symposium Series on Computational Intelligence (SSCI)

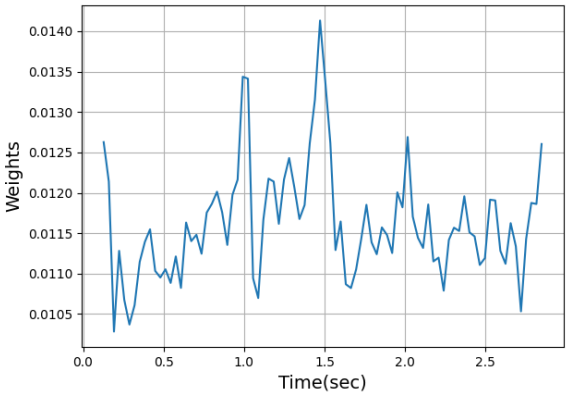
TABLE III: Classification accuracy of the proposed convolutional attention network and electrode selection technique in a subject-independent (SI) manner.

SI	C=64	C=32	C=9
Accuracy	74.0	74.2	71.1
Precision	74.9	75.3	71.9

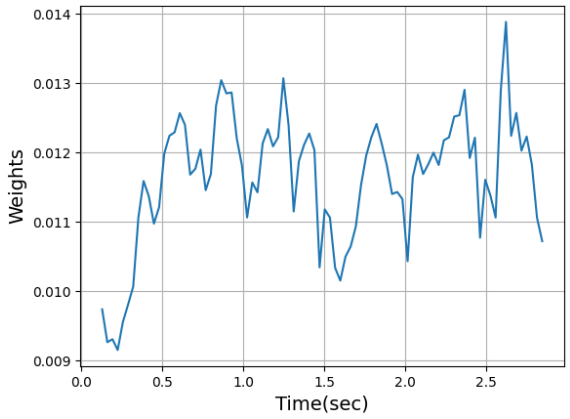
data. We validated the performance of the convolutional attention network in a leave-one-out cross validation manner. For electrode selection, we tested the system for three sets of electrodes with C = 9, 32, and 64. As shown in Table III the highest accuracy was achieved for C = 32. In SI evaluation, recognition rate is low with C=9, resulting from inter-subject variation in the EEG data; therefore, a larger set of electrodes performed better compared to fewer electrodes. This shows, the proposed electrode selection method can help reduce inter-subject variability by using fewer electrodes. In addition, our results show robustness of the proposed convolution-attention network in dealing with inter-subject variations.

C. Attention Weights Visualization

Figure 4 shows the attention weights for the two words. The self-attention mechanism assigned highest weights to the time points with most discriminative information about the silently spoken word. The attention weights for all the networks trained in SI and SD evaluation were averaged to create the global attention weights, shown in Figure 4. As can be seen in Figure 4, EEG signals for the word *Apple* exhibits most distinct characteristic at 1 sec, whereas EEG signals for the word *Write* contain important features after 2.5 sec. Interestingly, the weights at the pre-stimulus time period for the two conditions vary, which can be attributed to the trial-by-trial fluctuations in participants focus during the task which give rise to new features[28]. However, the convolutional attention network is designed to extract features associated with the silently spoken words, hence significant weights are given to the features after the stimulus onset.



(a)



(b)

Fig. 4: Attention weights for the two words: (a) Apple; (b) Write.

D. Comparison with the State-of-the-art Optimization Methods

TABLE IV: Comparison of performance achieved by the proposed electrode selection methods with the state-of-the-art optimization methods in selecting electrodes for recognition of silently spoken words. SD: Subject-Dependent; SI: Subject-Independent.

Method	Accuracy		Precision		Processing Time	
	SI	SD	SI	SD	SI	SD
PSO	72.8	71.1	73.7	62.5	3.6 min	2.6 min
GA	72.5	71.6	74.1	63.3	2.4 min	1.1 min
Proposed	74.2	75.7	75.3	67.7	41 sec	8 sec

We compared the proposed electrode selection method with the state-of-the-art optimization methods such as the genetic algorithm (GA) and particle swarm optimization (PSO) [29]. The GA and PSO are meta-heuristic algorithms, which have been effective in solving complex engineering optimization problems [30]. Therefore, we applied the GA and PSO to select electrodes, which were then used to train the convolutional attention network for recognizing silently spoken words. For the GA, we used a population size of 500 and the algorithm converged after 20^{th} generation, similar to [31]. A logistic regression classifier was employed to produce the measure of fitness. For the PSO, we used 20 particles in a swarm with an inertial weight of 0.3. The objective function provided the measure of accuracy using the logistic regression classifier and the algorithm converged after 500 iterations. Input to the GA and PSO was a two-dimensional matrix $X \in \mathbb{R}^{n \times C}$, where n is the number of training trials and $C = 64$ is the number of electrodes. The input was estimated by averaging spectrograms across time and frequency axis for each electrode. As can be seen from Table IV reporting the comparison of performance achieved by the three methods, the proposed electrode selection method outperformed the PSO and GA optimization methods. In addition, our method is much faster, i.e., requires less processing time, making it more suitable for a BCI application.

E. Comparison Against a Baseline Network without Attention Mechanism

TABLE V: Comparison of performance of the proposed convolutional attention network with the baseline. Where the baseline is the network without attention mechanism. Performance of the two networks was compared using t-test.

C	Without Attention		Conv Attention		p-value	
	SI	SD	SI	SD	SI	SD
64	70.4	64.9	74.0	73.8	<0.001	0.045
32	70.1	66.9	74.2	75.7	0.58	0.056
9	67.9	70.2	71.1	80.0	0.02	0.015

To validate the contribution of attention mechanism in the proposed network, we compared the performance of the convolutional attention network with a baseline network. The baseline network had similar architecture to the proposed network without attention mechanism. As can be seen from Table V, the convolutional attention network performed significantly better than the network without attention mechanism for all comparisons for both subject-independent and dependent evaluations, except for C=32 for subject-independent evaluation (see Table V for p-values). This shows effectiveness of attention mechanism in highlighting important features for the recognition of silently spoken words.

F. Comparison with Existing Methods

TABLE VI: Comparison of accuracy achieved by different methods on our EEG dataset. Each method was evaluated in subject-dependent (SD) and subject-independent (SI) manner.

Method	C	SD	SI
Sereshkeh [3]	64	52.4%	68.8%
Panachakel [6]	64	51.7%	50.0%
Bashivan [11]	64	50.5%	65.0%
Proposed	64	73.8%	74.0%
Proposed	32	75.7%	74.2%
Proposed	9	80.0%	71.1%

We also compared our results with methods proposed in previous studies for the recognition of imagined words using EEG signals [3], [6]. Further, our comparison included the method proposed in [11] that processes EEG data as a time-varying input using convolutional recurrent network. The parameters used for the three methods were as described in [3], [11], [6]. We report the performance of previous methods on our EEG dataset in Table VI. The results demonstrate that more accurate recognition of words can be achieved using our proposed framework with fewer electrodes.

VI. Limitations and Future Work

One limitation of the proposed method is that it uses Short Time Fourier Transform (STFT) to extract time-frequency information which is computationally expensive. Therefore, in future work we intend to analyze EEG signals in time domain. Further, the electrode selection method failed to achieve high recognition rate with C=9 in the subject-independent evaluation. In addition, the number of electrodes selected has to be decided in advance. Therefore, it would be interesting to implement an electrode selection mechanism within the convolutional attention network which can select important electrodes in an end-to-end fashion. It would also be interesting to investigate a fully attention-based network for spectral and temporal pattern learning. Our future work will involve collecting a comprehensive dataset and EEG signals for a larger vocabulary of words (classes) including "no word" scenario.

VII. Conclusion

In this work, we proposed a framework for recognition of silently spoken word from EEG signals. The proposed methods reduce the dimensionality of EEG data and provide the most task discriminative information by using electrode selection method. Further, the convolutional attention network is used to learn frequency features and important temporal information of EEG signals. Experimental evaluation showed that combination of our electrode selection method and convolutional attention network can recognize silently spoken words more accurately than previously proposed state-of-the-art methods.

References

[1] G. Li and W. Y. Chung, "Combined EEG-gyroscope-tDCS brain machine interface system for early management of driver drowsiness," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 1, pp. 50–62, 2017.

[2] C. H. Nguyen, G. K. Karavas, and P. Artemiadis, "Inferring imagined speech using EEG signals: A new approach using riemannian manifold features," *Journal of Neural Engineering*, vol. 15, no. 1, p. 016 002, 2017.

[3] A. R. Sereshkeh, R. Trott, A. Bricout, and T. Chau, "EEG classification of covert speech using regularized neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2292–2300, 2017.

[4] S. Datta and N. V. Boulgouris, "Recognition of Grammatical Class of Imagined Words from EEG Signals using Convolutional Neural Network," *Neurocomputing*, in press 2021.

[5] P. Saha, S. Fels, and M. Abdul-Mageed, "Deep learning the EEG manifold for phonological categorization from active thoughts," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 2762–2766.

[6] J. T. Panachakel, A. Ramakrishnan, and T. Ananthapadmanabha, "Decoding imagined speech using wavelet features and deep neural networks," in *2019 IEEE 16th India Council International Conference (INDICON)*, IEEE, 2019, pp. 1–4.

[7] M. Alsaleh, "Toward an imagined speech-based brain computer interface using EEG signals," Ph.D. dissertation, University of Sheffield, 2019.

[8] M. X. Cohen, *Analyzing neural time series data: theory and practice*. MIT press, 2014.

[9] A. S. Al-Fahoum and A. A. Al-Fraihat, "Methods of EEG signal features extraction using linear analysis in frequency and time-frequency domains," *ISRN Neuroscience*, vol. 2014, 2014.

[10] S. Martin, P. Brunner, C. Holdgraf, H. J. Heinze, N. E. Crone, J. Rieger, G. Schalk, R. T. Knight, and B. N. Pasley, "Decoding spectrotemporal features of overt and covert speech from the human cortex," *Frontiers in Neuroengineering*, vol. 7, p. 14, 2014.

[11] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," *arXiv preprint arXiv:1511.06448*, 2015.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[13] T. Alotaiby, F. E. Abd El-Samie, S. A. Alshebeili, and I. Ahmad, "A review of channel selection algorithms for EEG signal processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–21, 2015.

[14] G. Leech, P. Rayson, *et al.*, *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge, 2014.

[15] A. Porbadnigk, M. Wester, and T. S. Jan-p Calliess, "EEG-based speech recognition impact of temporal effects," 2009.

[16] S. J. Luck, *An introduction to the Event-Related Potential Technique*. MIT Press Cambridge, MA, 2005.

[17] D. Roehm, M. Schlesewsky, I. Bornkessel, S. Frisch, and H. Haider, "Fractionating language comprehension via frequency characteristics of the human EEG," *Neuroreport*, vol. 15, no. 3, pp. 409–412, 2004.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] J. Donahue, L. Anne H., S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[20] D. A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," *arXiv preprint arXiv:1511.07289*, 2015.

[21] D. A. Butts, C. Weng, J. Jin, C. I. Yeh, N. A. Lesica, J.-M. Alonso, and G. B. Stanley, "Temporal precision in the neural code and the timescales of natural vision," *Nature*, vol. 449, no. 7158, pp. 92–95, 2007.

[22] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[23] F. Chollet *et al.*, *Keras*, https://keras.io, 2015.

[24] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

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

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[27] J. Brownlee, *Deep learning with Python: develop deep learning models on Theano and TensorFlow using Keras*. Machine Learning Mastery, 2016.

[28] J. S. P. Macdonald, S. Mathan, and N. Yeung, "Trial-by-trial variations in subjective attentional state are reflected in ongoing prestimulus EEG alpha oscillations," *Frontiers in psychology*, vol. 2, p. 82, 2011.

[29] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-international conference on neural networks*, IEEE, vol. 4, 1995, pp. 1942–1948.

[30] A. Konak, D. W. Coit, and A. E. Smith, "Multi-objective optimization using genetic algorithms: A tutorial," *Reliability engineering & system safety*, vol. 91, no. 9, pp. 992–1007, 2006.

[31] A. Albasri, F. Abdali-Mohammadi, and A. Fathi, "EEG electrode selection for person identification thru a genetic-algorithm method," *Journal of medical systems*, vol. 43, no. 9, pp. 1–12, 2019.