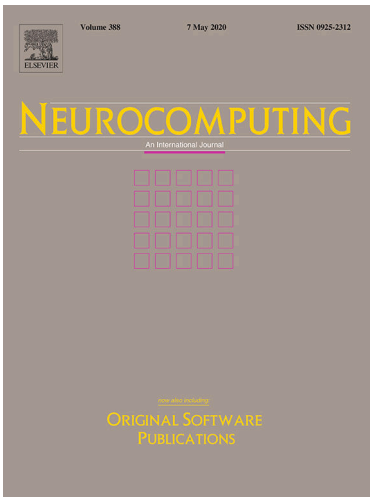
Journal Pre-proofs



Recognition of Grammatical Class of Imagined Words from EEG Signals using Convolutional Neural Network

Sahil Datta, Nikolaos V. Boulgouris

PII: S0925-2312(21)01219-4
DOI: https://doi.org/10.1016/j.neucom.2021.08.035
Reference: NEUCOM 24200

To appear in: *Neurocomputing*

Received Date: 26 January 2021
Accepted Date: 8 August 2021

Please cite this article as: S. Datta, N.V. Boulgouris, Recognition of Grammatical Class of Imagined Words from EEG Signals using Convolutional Neural Network, *Neurocomputing* (2021), doi: https://doi.org/10.1016/j.neucom.2021.08.035

Recognition of Grammatical Class of Imagined Words from EEG Signals using Convolutional Neural Network

Sahil Datta, Nikolaos V. Boulgouris*

Department of Electronic and Electrical Engineering, Brunel University London, UB8 3PH, U.K.

Abstract

In this paper we propose a framework using multi-channel convolutional neural network (MC-CNN) for recognizing the grammatical class (verb or noun) of covertly-spoken words from electroencephalogram (EEG) signals. Our proposed network extracts features by taking into account spatial, temporal, and spectral properties of the EEG signal. Further, sets of signals acquired from different regions of the brain are processed separately within the proposed framework and are subsequently combined at the classification stage. This approach enables the network to effectively learn discriminative features from the locations of the brain where imagined speech is processed. Our network was tested using challenging experiments, including cases where the test subject did not take part in system training. In our main application scenario, where no instance of a specific noun or verb was used during training, our method achieved 85.7% recognition. Further, our proposed method was evaluated on a publicly available EEG dataset and achieved recognition rate of 93.8% in binary classification. These results demonstrate the potential of our method.

Keywords: Electroencephalogram (EEG), Imagined speech, Covert speech, Multi-Channel Convolutional Neural Network (MC-CNN)

1. Introduction

Language plays an important role in human interaction and constitutes the most essential aspect of communication. For that reason, the development of brain computer interfaces (BCI) for imagined speech recognition has been of interest to researchers for more than two decades. BCI systems have enabled the recognition of imagined speech, and have led to possibilities of communication without speech production [1]. Research on recognition of overt and covert speech has gained attention and has been the focus of several studies [2]. These studies mainly focused on different elements of speech, such as phonemes, syllables and vowels [3]. Electroencephalogram (EEG) signals produced during of imagined speech have been used for recognition of vowels, and subject identification [4]. Recognition of covertly spoken (imagined) words has also been performed in binary classification tasks [2], [5]. Studies with covert speech have focused on recognition of each word individually.

It is known that the human brain interprets language by first interpreting the semantics and grammatical roles of the spoken words, such as the roles of nouns and verbs [6]. Therefore, in order to build a BCI for the transcription of imagined speech into text, it is important to recognise the grammatical classes of covertly spoken words. Linguistic interactions have a specific object and describe properties attributed to the object, this might be lexically reflected in nouns and verbs [7]. Nouns promote the primary concept, whereas verbs provide context to that concept [7]. Studies have investigated the distinction between nouns and verbs in the brain and explored brain areas associated with the processing of nouns and verbs [8]. The study in [9] investigated neural activity for nouns and verbs with magnetoencephalograph (MEG) signals. The study concluded that the largest amplitude changes during the processing of nouns and verbs occur in signals recorded from the temporal lobe and the frontal of left hemisphere. Another work with MEG signals [10] found more intensive activation in the frontal, Parietal and Temporal areas of the head during the processing of verbs rather

*Corresponding author:
Email address: Sahil.Datta@brunel.ac.uk, Nikolaos.Boulgouris@brunel.ac.uk (Sahil Datta, Nikolaos V. Boulgouris)

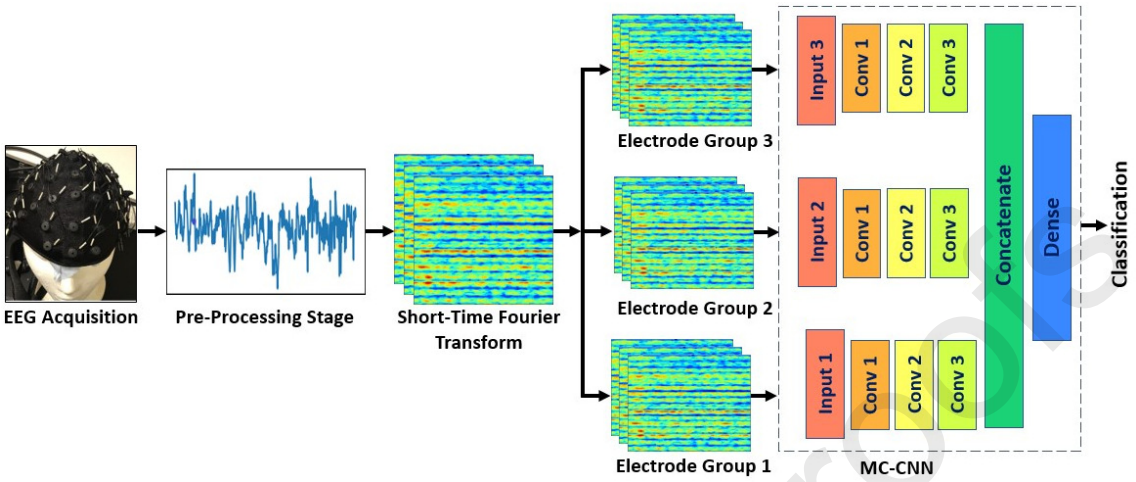


Figure 1: The proposed framework for recognition of grammatical class from EEG signals of covertly spoken words.

than nouns. Further, several studies have tried to distinguish between neural processing of nouns and verbs based on evoked potentials [10]. However, to the best of our knowledge no work so far has distinguished between nouns and verbs using machine learning techniques.

In order to recognise imagined speech based on EEG signals it is important to extract discriminative features representing different classes. Several techniques have been proposed for feature extraction in imagined speech recognition from EEG signals. The study in [2] proposed an feature extraction method using Riemannian manifold algorithm. In addition, the work in [11] used Riemannian distance of correntropy spectral density (CSD) matrices as feature for classification of imagined speech. An artificial neural networks (ANN) in combination with PCA have also been used to classify imagined speech from EEG signals [12]. However, most feature extraction methods are unable to adapt to variations within a given class. Adaptation to variations is particularly important in imagined speech recognition, where semantic variations lead to changes in the processing of words in the brain [13].

Feature extraction and classification can also take place using deep learning, which has exhibited excellent performance in various recognition tasks [14] including the recognition of imagined speech based on EEG signals. The work in [15] used deep learning to perform multi-class classification phonemes and words. Artificial neural network (ANN) was used to classify bilingual unspoken speech in [12]. In [16], a hybrid network consisting of CNN, recurrent neural network (RNN) and auto-encoder was used and achieved a recognition rate of 79.9% with long imagined words. However, none of these studies focused on the joint use of temporal, spatial, and spectral properties of the EEG signals. Further, the feature extraction methods used in these works focused on either using electrodes from a small region or using all the electrodes in the EEG cap.

Another conceivable approach for imagined speech recognition could be based on a recently proposed class of algorithms, termed Spiking Neural Networks [17], [18]. These approaches are biologically inspired and result in implementations with low power consumption. Similarly, digital neuromorphic computing [19], [20] has been proposed for designing effective artificial intelligence systems. While these approaches are valuable and hold great potential, the patterns represented in EEG spectrograms include rich spectro-spatio-temporal information that cannot be adequately modelled by the spiking patterns of a spiking network [21]. For this reason, in this work we focus our attention on comparisons with traditional machine learning methods that have been used for the interpretation of imagined speech using EEG signals.

In this paper, we present a framework for the recognition of the grammatical class of mentally-spoken words using EEG signals. The proposed framework constructs spectrograms that are classified by means of a multi-channel convolutional neural network (MC-CNN) [22]. The contributions of this paper are:

1. A framework (shown in figure 1) for the recognition of the grammatical class (noun, verb) from EEG signals produced during covertly (mentally)

spoken words. Each MC-CNN channel takes input from a different set of electrodes considering the spatial, spectral and temporal structure of the EEG signals separately for each area. To the best our knowledge this is the first work where an MC-CNN is used to classify EEG signals for the purpose of recognition of the grammatical category of covertly-spoken words.

2. A study of the brain areas that can be used as inputs to the MC-CNN architecture in order to improve overall accuracy. These brain areas are known to play significant role in the processing of nouns and verbs.

3. A thorough evaluation of the performance of the proposed framework, tested on our own database as well as on the publicly available Kara-One EEG database [3] for covertly-spoken nouns and verbs. We tried three different experimental protocols, which included testing on previously unseen subjects, and also on unseen nouns and verbs.

2. Data Acquisition

In order to analyse grammatical classes of mentally spoken words, we recorded EEG signals that were produced in response to 10 imagined nouns or verbs. This is in contrast to other databases, which recorded EEG signals for mentally spoken phonemes or syllables and contains either no words or only a few words [3]. In the paper we refer to our dataset as *CovertSpeech* Database.

2.1. Head Cap

EEG activity was recorded using a Neuroscan 64 channel Quik cap of extended 10-20 system. Two electrodes, VEOG and HEOG, were placed above and below the eye to record its horizontal and vertical movement. Two reference electrodes were placed at mastoid. To ensure good contact between the scalp and electrodes abrasive gel (conductive electrolyte) was used. Further, dead skin was removed using alcohol pads, which help reduce impedance. Data was sampled at 1kHz sampling rate using synamp amplifier. Neuroscan curry software was used to process the raw EEG signals.

2.2. Participants and Stimuli

EEG signals produced in response to imagined speech were recorded from 19 participants who were fluent in English and whose age was between 21 to 70 years. Ten words (stimuli), five nouns and five verbs, were presented randomly on a computer screen. This

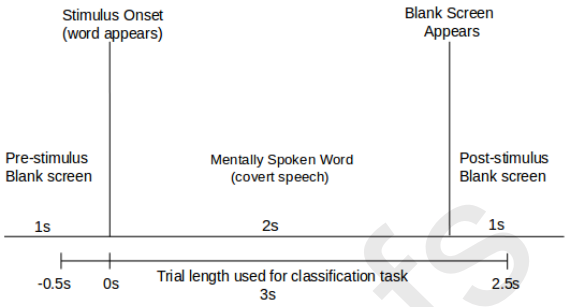


Figure 2: The sequence in which stimulus was presented for single trial of covertly spoken word. To avoid overlapping of inter-trial activity only three seconds of signal was used for recognition of grammatical class of mentally spoken words.

was done in order to avoid temporal effects [23]. A blank screen appeared for 1 sec before the stimulus onset, and the ensuing stimulus was presented for 2 sec. Subjects were asked to mentally read the word as soon as it appeared on the screen. The stimulus was followed by blank screen for 1 sec. Therefore, the total duration of a trial was four seconds. However, only three seconds (-0.5s to 2.5s) of the trial was used for classification task in order to avoid overlapping inter-trial activity. The sequence of EEG signal recording is shown in figure 2. The words appeared in capital letters, black in colour with white background presented on a computer screen 1 meter away from the subject. White background was chosen in order to avoid potential due to visual stimulus [24]. Subjects were asked to read the word presented on the computer screen covertly (speech imagery). Each word was presented 10 times, a total of 100 trials were recorded for a given subject. The experiment was designed using E-prime-2 software. The words (stimulus) used in the experiment are presented in Table 1. During recording all the subjects were instructed to refrain from any kind of movement. This research has been approved by College of Engineering, Design, and Physical Sciences Research Ethics Committee, Brunel University London, reference number 7361-LR-Sep/2017-8301-1.

Table 1: Words presented as stimulus during recording of EEG signals for covert speech.

Noun:	"*Apple*"	"*Bottle*"	"*Football*"	"*Laptop*"	"*Orange*"
Verb:	"*Carry*"	"*Run*"	"*Swim*"	"*Laugh*"	"*Write*"

3. Pre-Processing

EEG signals are contaminated by artifacts and noise, these could be physiological and/or environmental. Although subjects were told not to move during the recording, certain physiological artifacts cannot be avoided, especially those due to breathing, eye blinks and movement. To remove noise from the data, pre-processing was applied using Neuroscan curry 8 software. Low voltage shifts at lower frequencies were avoided by high-pass filtering at 0.01Hz. Most of the high frequency noise due to muscle movement was eliminated using the EMG electrode. Similarly, artifacts due to eye movement were removed using VEOG electrode, by measuring VEOG signal peak-to-peak voltage along with threshold voltage [24]. Baseline correction was done in real time and offline processing. Information above a threshold of $\pm 200\mu V$ was eliminated, and contaminated electrodes were interpolated using neighbouring electrodes. Also, 50Hz line noise was eliminated by applying notch filter.

4. Feature Extraction

EEG signals have discriminatory characteristics that appear in both temporal and frequency domain. Considering that, in this work we used time-frequency features as input to our neural network. We used Short Time Fourier Transform (STFT) for calculating spectrograms from EEG signals. The STFT [25] is defined as:

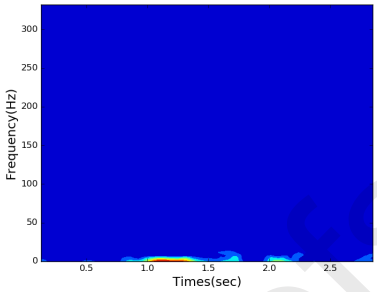
$$X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n} \quad (1)$$

where $x[n]$ is the EEG signal, $w[n]$ is the window function, and m represents the time index. Windowing was performed during STFT in order to avoid the discontinuities known as leakage. In our implementation of STFT we used the Hann window with a window length of 256 samples and temporal overlap of 87% between consecutive windows. We used a shorter window in order to improve temporal resolution, which can help detect temporal events. We did not include lower frequencies (below 5Hz). The spectrogram can be trivially obtained from (1) as

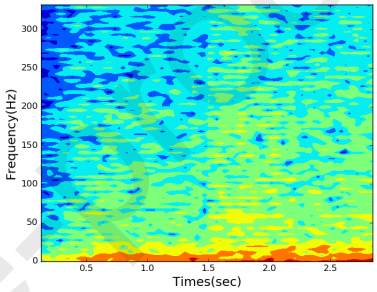
$$S_x = |X(m,\omega)|^2 \quad (2)$$

The spectrogram of (2) was the time-frequency feature that was input to our neural network.

Spectrograms were subject to baseline normalization. This was done because EEG signals suffer from 1/f phenomena (low power at high frequencies), which can cause uneven power representation at different frequency ranges. As a result, important events can be misinterpreted as background activity, making the comparison between classes difficult [25]. A short time period of 300 *msec* was chosen from the 500 *msec* pre-stimulus time period (when the blank screen before the stimulus appears) and was used for averaging over all training trials for all given classes from 64 electrodes. Although the pre-stimulus time period is considered to have no event related activity, due to the effects of windowing (overlapping pre-trial and post-trial time periods) a safer temporal window (t_1 = -400 to t_2 = -100 *msec*) was chosen. It was considered that the time before the onset of the stimulus does not provide any useful information with regards to the word being mentally spoken. A baseline vector $B(f)$ was calculated, comprising of frequencies averaged over the baseline time window (average column of the spectrogram). Therefore, the baseline vector is



(a)



(b)

Figure 3: (a) Before baseline normalization there is limited power representation at higher frequencies. (b) Baseline normalization amplifies higher frequency components and enables extraction of discriminatory features.

$$B(f) = \frac{1}{t_2 - t_1} \sum_{t_1}^{t_2} S_x(t, f)$$

and the normalized spectrogram (in decibels) is

$$S_{dB}(t, f) = 10log10(\frac{S_x(t, f)}{B(f)}) \quad (3)$$

Although baseline normalization is computed with respect to baseline (task unrelated activity) it also helps in discarding background activity. This leads to spectrograms where only the task-specific brain events are represented [25], [26].

5. Multi-Channel Model

For each subject, a total of 50 trials per class (noun, verb) were used. However, some subjects ended up having only 45 trials because certain trials had to be removed due to excessive noise. The spectrograms calculated using the acquired EEG signals were combined to form a multi-dimensional feature array that was input to our neural network architecture.

*5.1. **Brain Signal Selection***

Brain signals were selected from three different brain areas. The respective electrodes are presented in Table 2. These electrodes were selected based on certain factors. Firstly, electrodes in group 1 cover Broca's and Wernicke's area. Work in [27] suggests that Broca's area is activated during different stages of word production. However, Broca's area is suggested to be strongly activated during verb production in comparison to noun production [28]. Broca's area and Wernicke's areas are connected through a junction of nerves known as *arcuate fasciculus*. Wernicke's area is recruited in translating auditory input to overt and also covert speech [29], and in speech production and comprehension [30]. Although these two areas are considered the main language regions of the brain, the processing of nouns and verbs has been reported to involve areas outside these two regions [31]. To exploit the discriminatory potential of the signal acquired outside the main language processing areas of the brain, we used electrodes from two more areas.

Electrode group 2 covers a large part of the frontal lobe, which is known to play a major role in the processing of verbs [32]. The work in [33] showed frontal lobe activation during word processing. Further, the study found that word production results in activation of the inferior frontal region along with Wernicke's area and temporal gyrus.

Table 2: Groups of electrodes used as input channels to the MC-CNN network.

Electrode Group 1:	F5, F3, FC3, FC5, C5, CP5, P1
Electrode Group 2:	F1, FZ, F2, F4, F6, F7, FCZ, FC2, FPZ, FP2, AF3, AF4
Electrode Group 3:	P1, PZ, P2, P4, POZ, PO4, PO6

Electrode group 3 covers the Occipital lobe and the Parietal lobe. The Parietal lobe is known to be active during the silent reading task [34]. An important factor we considered when selecting this brain area was that Occipital and Parietal lobes are known to be recruited during word processing when written words are presented visually [35]. This is because visually perceived information is first processed by the Occipital lobe and subsequently communicated to the parieto-frontal region of the brain [36]. Further, the Occipital lobe is known to play a crucial role in the processing of nouns [31]. Therefore, electrodes in these areas are important for our study and are used in our system. In addition, considering that signals from neighbouring electrodes have high correlation [37], in each electrode group we used electrodes with spatial proximity to make it easier for the MC-CNN to learn features from the input.

5.2. Network Architecture

We trained a multi-channel convolutional neural network (MC-CNN) [22] to recognize the grammatical class of imagined words from EEG signals. As CNNs can adapt to variations in the input signal [38], they can learn discriminative features for covert speech recognition from the time-frequency information captured in spectrograms. The architecture of our MC-CNN is not the same as that of conventional CNN. Instead, it includes three channels, with each channel receiving a three dimensional tensor as input. For each channel, the input is of dimension $T \times f \times C$, where T is time, f is frequency, and C is the number of electrodes. The number of electrodes depends on the brain area, as detailed in Table 2. Feature learning is done separately on each channel and the resultant channel feature maps are converted into vectors. The vectors from the three channels are concatenated together and fed to the fully connected layer.

The proposed model architecture, shown in figure 4, consists of three channels, where each channel has three blocks and the network has three fully-connected layers

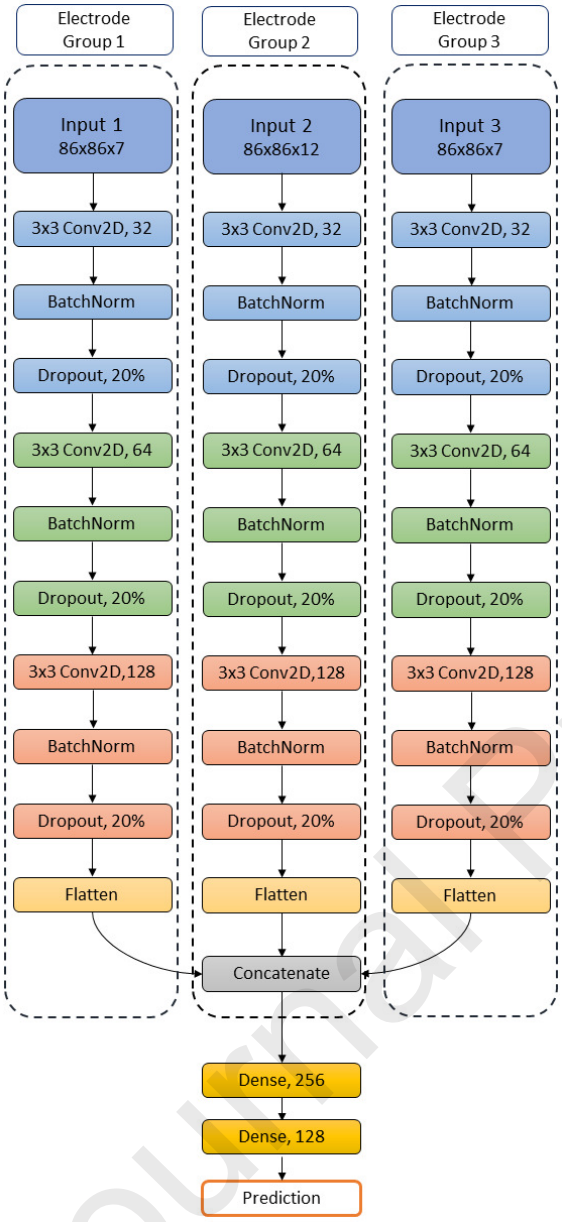


Figure 4: Architecture of the MC-CNN (best viewed in color). The network layers in blue, green and pink refers to block 1, 2, and 3. The CNN layer in blue used the sigmoid activation function, whereas the CNN in green and pink layer used the ELU activation function. All the CNNs used stride of size 2×2. The final layer used the sigmoid function to making the prediction (classification) for a given input.

in the classifier. All three channels have the same architecture. However, channels 1 and 3 take as input spectrograms from seven electrodes whereas channel 2 takes spectrograms form twelve electrodes. Each block contains two-dimensional convolutional layers followed by a batch-normalization layer [39], and a dropout layer.

A single convolutional layer was used in each block because of small data dimensions of 86×86. The convolutional layer in the first block filters the data using 32 kernels with a receptive field of size 3×3 and stride of size 2×2, a process that can capture high-level features from the spectrogram. The convolutional layers in the second and third block have 64 and 128 kernels of size 3×3 applied with a stride of size 2×2. These layers learn intricate features that are important for class discrimination. The performance of the architecture was assessed using filters of different sizes, among which the 3×3 filter performed best. The ability of the CNN to extract features from different time-frequency patches in the EEG spectrogram is particularly useful, as different feature maps can represent activity at different time-frequency windows. The feature map y at a given layer is obtained as [40]:

$$y_j = f(a) = f((W^n * x)_{ij} + b_n) \tag{4}$$

where x is the input spectrogram, $*$ is the convolutional operation, b_n is the bias value, y_j is the j^{th} output map and W^n is the weight matrix of filter n, with $n = 1, 2, ..., n_f$, n_f = 32, 64 or 128. For convolution we used "same" (zero) padding in order to preserve the spatial resolution of the input. This also results in better edge detection. The padding is defined as:

$$p = \frac{f_r - 1}{2} \tag{5}$$

where f_r is receptive field size. The used activation function f is the exponential linear unit (ELU) or the sigmoid. Specifically, the first convolutional layer used the sigmoid activation function:

$$sigmoid(a) = \frac{1}{1 + e^{-a_i}} \tag{6}$$

However, the sigmoid function is known to suffer from the vanishing gradient problem when used in the hidden layers [41]. Therefore, the rest of the network used the (ELU) activation function [42], which endows the network with the ability to learn non-linear features. The ELU is defined as:

$$ELU(a) = \begin{cases} a_i & \text{if } a_i \geq 0 \\ \alpha(e^{a_i} - 1), & \text{otherwise} \end{cases} \tag{7}$$

where a_i is the i^{th} value in the feature map. The ELU was chosen over the ReLU function because the ReLU

Table 3: Three experimental protocols: leave one subject out (LSO) is a subject-independent experiment; leave trial out (LTO) and leave one word out (LWO) are done on subject-by-subject basis, i.e., training and testing take place using different data from the same subject.

Exp	Training			Testing		
	Subjects	Trials	Words	Subjects	Trials	Words
LSO	All but one	All	All	one	All	All
LTO	-	80%	All	-	20%	All
LWO	-	All	All but one	-	All	All but one

performs poorly when placed after the sigmoid function [43], and also because the ReLU suffers from the dying ReLU problem [44].

Batch-normalization was used in every block, which helped speed up system learning by centering the data [42]. In the classification block, the number of nodes in fully connected layers were 256, 128, and 2 in the last dense layer (classification layer) with sigmoid function (6) for classification. The use of the ELU function after dense layers makes the network capable of learning complex features passed from previous blocks.

We tried different variants of our MC-CNN architecture, by varying the number of input channels (i.e., using more electrodes from different areas), activation functions, the kernel size and the filter size. The additional signals were input to the MC-CNN by using an architecture with more than three inputs. This approach was meant to test whether the information contributed by a larger set of electrodes, from different areas of the brain, could improve the recognition rate. Another approach was to use a larger set of electrodes in inputs 2 and 3 of the MC-CNN network, covering more areas of the head at the Frontal and Parieto-Occipital lobe. However, in both cases the recognition rate was reduced. This highlights the importance of selecting electrodes from areas that play a role in the processing of nouns and verbs.

5.3. Network Training

The MC-CNN network was implemented in keras [45] with tensorflow backend [46]. Weight optimization was performed using Adam optimizer [47] with learning rate of 0.0001 in order to minimize the cross-entropy loss. A slower learning rate was used in order to avoid varying gradient at different layers [38]. Since unsuitable initialization can lead to unstable gradient [48], "He" initialization [49] was used to initialize weights. This initialization method was used as it works best with the ELU activation function [42]. The network was trained for 500 epochs, with batch gradient descent (i.e., a batch consisted of all the samples in the training data). In order to avoid overfitting, we used batch-normalization as well as dropout regularization [50] with dropout rate of 20%. These techniques were shown to be effective in [14].

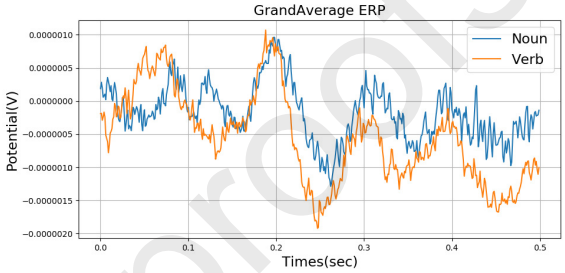


Figure 5: Grand averaged evoked potential for Nouns and Verbs, showing the four main components (best viewed in color). The negative deflection between 0.2s to 0.3s is the ERP component that is associated with imagined nouns and verbs.

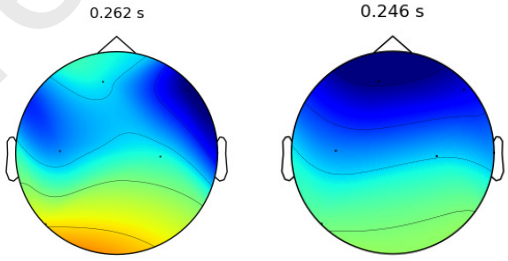


Figure 6: Topographical maps of grand averaged imagined speech evoked potential (ERP). (a) The map for the *noun* class shows power variation in temporal regions, (b) The map for the *verb* class shows reduced power in the frontal part of the brain.

5.4. Event Related Potential

We also investigated the evoked potential or event related potential (ERP) for two speech parts, i.e., nouns and verbs. ERPs are used in neuroscience to detect the onset of an event produced by a particular activity [24]. ERP was calculated by averaging trials for the same class (nouns and verbs) over 19 subjects. The grand average ERP is shown in figure 5. The ERPs from several

Table 4: Classification accuracy for EEG signals in our *CovertSpeech* dataset recorded during imagined speech of Nouns and Verbs. Results for three experimental protocols are shown.

Exp	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	Avg
LSO	85.4	62.2	88.3	52	49.25	82.45	87.8	69.2	73	88	85.5	84.1	86.85	90.45	74.8	87	92.4	91.3	62.1	78.5
LTO	93	91.6	86	77.3	63.6	91.4	82.6	73.9	84.3	86.5	77.8	78	89.9	94.3	80	91.9	93.7	81.8	89	84.6
LWO	91.2	94.2	90.5	66.8	69.8	90.8	88.3	84.5	82.3	89.3	72.7	80.6	88.3	94.6	84	93.4	93.1	87.1	88	85.7

sites were investigated, but the most informative were from electrode group 1 (Table 2). Therefore, we focus on the ERPs from electrode group 1. In our investigation of ERPs associated with nouns and verbs we found four components. The first component was a positive peak between 0.70*s* and 80ms post stimulus. The second component was a negative peak at around 0.110*s* which is also known as N100. The P70 and N100 components are manifestations of an early processing of the presented stimulus [31]. The third component was a very strong positive deflection at 200*ms* known as P200, which may reflect sensation-seeking behaviour of an individual [51]. The fourth component was noted to be most important as it is produced in response to imagined nouns and verbs in covert speech. A negative deflection around 0.250*s* was observed. This deflection was stronger for verbs in comparison to nouns. Past studies [10] have also observed similar temporal event (ERP) which validates the presence of distinctive activity produced by covertly-spoken nouns and verbs in imagined speech. These events were observed within 500*ms* of the onset of the stimulus. Therefore, only ERPs from that time range are presented. Topographical maps of the fourth event are shown in figure 6. The topographical maps show that covertly spoken nouns result in reduced power at the temporal regions of the head and slightly increased power at the Occipital area. In contrast, the processing of the covertly-spoken verbs results in reduced power in the frontal lobe. A similar observation was made by [52]. As can be seen in figure 5, the ERPs do not provide any discriminatory information about the two grammatical classes. However, the topographical maps (figure 6) indicate processing of nouns and verbs at different areas of the head.

6. Results

In order to distinguish mentally spoken nouns from verbs, we used EEG signals produced during covert (imagined) speech of ten words, i.e., five nouns and five verbs. In general, 50 trials of a mentally-spoken word for each class (noun, verb) were recorded from each subject. However, some subjects ended up having only 45 trials because some of the recorded trials had to be removed due to excessive noise.

Three different approaches were used for the evaluation of the effectiveness of the proposed network. In particular, two sets of results were obtained in a subject-dependent manner, i.e., the network was trained and tested for each subject separately. In addition, a third experiment was conducted, in which results were obtained in a subject-independent manner, i.e., the network was trained on data from several subjects and was tested on data from a different subject. The three experimental protocols are summarized in Table 3.

On Nvidia Tesla K40, one iteration of subject-dependent training needs about 2 minutes, while subject-independent training needs about 15 minutes. As multiple iterations were run, the overall training time was a few hours. For a single testing trial, the response time of the algorithm is less than 1 sec. Therefore, in practical situations, the response of the system would be extremely fast.

6.1. Leave One Subject Out (LSO)

The first experimental protocol evaluated the performance of our network in a subject independent manner, where the MC-CNN was trained on EEG data from 18 subjects and tested using EEG trials from a different subject. This training and testing approach was repeated for all the subjects in a Leave one Subject Out (LSO) cross validation manner. Due to the large dataset and the limited GPU memory, trials from each subject were divided into two different sets that were used separately for the training and testing of our network. Results from both sets were averaged for each subject. In this experimental protocol the network was trained for 100 epochs with mini-batch gradient descent of size 64. Classification results, in terms of recognizing whether a trial corresponds to an imagined noun or verb, are shown in Table 4. As can be seen, the average classification rate is 78.9%, which shows that our system can accurately classify EEG signals from subjects that have not been used in the training of our network.

Table 5: Classification accuracy for EEG signals in our *CovertSpeech* dataset when network was trained on 64 and 26 electrodes together.

Exp	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	Avg
64	47.7	59	51.6	57.2	49.7	46.5	51.7	52.8	54.3	51	55.4	53.4	52.3	56.6	57.4	58.7	61.3	64.2	54.8	54.7
26	64.8	63.8	78	55.2	49.3	68.6	63.3	58.2	65.6	65.3	60	57.5	63.4	79.6	78.2	69.7	62.2	56.3	55.2	63.8

Table 6: Classification accuracy of our proposed model when training and testing took place using the *Kara-one* dataset [3].

Exp	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg
LSO	76.7	69.4	86	93.8	83.3	61	88.1	89.8	91.5	81.9	82.2
LTO	97.7	96.8	87	99.6	97	82.3	99.3	98.1	100	79.8	93.8

6.2. Leave Trial Out (LTO)

In the LTO experimental protocol, 80% of the trials were used for training the network, while 20% were used for testing. EEG data from all covertly-spoken words were used for both training and testing. Training and testing took place on a subject by subject basis and results from all the test trials were averaged together for each subject. To avoid variations in network parameters due to the stochastic nature of learning algorithms [53], the MC-CNN network was trained and tested 10 times for each trial. Test results were averaged. Subject-by-subject results are shown in Table 4. As seen, an average recognition rate of 84.6% was achieved. This demonstrates that it is possible to infer whether a subject has thought of a noun or verb by observing a subject's EEG signals.

6.3. Leave One Word Out (LWO)

Another experimental protocol for the evaluation of network performance was to train the MC-CNN network using EEG signals from four words and test it using the EEG signals from the remaining (fifth) word in the noun or verb class. In this case, 80% of the data was used for training and 20% data was used for testing. Henceforth, this approach will be referred to as Leave one Word Out (LWO) cross validation. The network was tested on EEG data from each word separately and the classification rates were averaged for each subject. These results are shown in Table 4. As seen, the mean classification rate is 85.7%, which shows that the system has the ability to recognize the grammatical class of previously unseen nouns or verbs.

6.4. Comparison with Single Channel Processing

In order to evaluate the effectiveness of forming electrode groups that are processed using separate channels, we compared our architecture with a system where all spectrograms were fed to a single-channel CNN. Two sets of results were obtained, first when spectrograms from all 64 electrodes were input simultaneously to single-channel CNN. Another set of results were obtained by providing to the network spectrograms from three brain areas (26 electrodes in total), which were fed simultaneously to the single-channel CNN. The results were evaluated in leave trial out (LTO) fashion and on a subject-by-subject basis. The results are shown in Table 5. As can be seen, the single-channel results are inferior to those of multidimensional method reported in Table 4. Specifically, our grouped approach, in conjunction with using three separate channels of MC-CNN, outperforms the best-performing single channel architecture by at least 25% in terms of recognition performance.

6.5. Kara-One Data-set

In addition to the assessment we conducted on our own data, we validated our model on the publicly available Kara-one [3] EEG dataset of covertly spoken words. In our analysis we used EEG data from ten subjects recorded during covert speech of four words, containing two nouns ("*Pat*", "*Pot*") and two verbs ("*Gnaw*", "*Knew*"). The raw signals were band pass filtered between 0.01Hz to 475Hz and 60Hz line was removed using a notch filter, other artifacts and noise were removed as described in section 3. Both classes (noun, verb) had 24 trials each, where 22 trials were used for the training of the network and two trials were used for testing. The sets of the training and testing trials were changed in a Leave One Trial Out (LTO) cross validation manner, with results for all test trials being calculated separately and later averaged for each subject. Classification rates for ten subjects are presented in Table 6. As seen, a mean accuracy of 93.8% was achieved by our model. In addition, we used the Kara-one database in order to test the performance of our network in a subject-independent manner. The system was trained on EEG data from nine subjects and tested on data from one (different) subject, i.e., the experiment was performed in a Leave One Subject out (LSO) cross

Table 7: Classification accuracy when the network was trained on our *CovertSpeech* database and tested on *Kara-one* dataset. Subject-dependent (SD) and subject-independent (SD) testing was performed.

Exp	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg
SD	94	50	63.5	54	76	93.7	83.3	62.5	92.7	50	71.9
SI	61.7	76.7	76	71	61.2	70.4	72.5	53.1	76.5	68.1	68

Table 8: Comparison of our method with past studies in a binary classification task.

Method	Class Type	Dataset	Recognition Rate
[54]	Words	[54]	63%
[55]	Words	[2]	79.9%
[2]	Words	[2]	80.1%
[11]	Words	[3]	90.2%
Ours	Words	Ours	84.6%
Ours	Words	[3]	93.8%

validation manner. An average accuracy of 82.2% was achieved.

6.6. Transfer Learning

In order to evaluate the robustness of the proposed method, the network was trained on spectrograms from our *Noun-Verb* EEG Database and was tested on *Kara One* database as a transfer learning model. For transfer learning, the last three dense layers of the pre-trained network were fine-tuned [56]. The weights of all layers except the last three dense layers were frozen. The network was fine-tuned with a slow learning rate equal to 10^{-4} to avoid over-fitting [41]. The network was evaluated for subject-dependent (SD) classification, where the last layers were fine-tuned and tested for each subject separately. Due to limited availability of data in subject-dependent evaluation, the network was fine-tuned and tested using leave-one-trial-out cross validation, followed by averaging the results for each trial. In the second approach, i.e., subject-independent (SI) classification, the network was fine-tuned on data from one subject and tested on data from all the other subjects. The results are shown in Table 7. As can be seen, our method achieves excellent recognition on a previously unseen dataset. This shows the robustness of the proposed method in recognizing grammatical class of imagined words.

6.7. Comparison

Although past works (e.g., [6], [7], [9]) have studied distinctive brain activity associated with the processing of words of different grammatical classes, to the best of our knowledge, no work in the past has performed recognition of the grammatical class (noun or verb) of mentally spoken words. For that reason, we compared our results with existing techniques that perform binary classification of imagined speech. Due to the different datasets used, a direct comparison may not be perfectly conclusive. We included in our comparison the method in [55], which performed classification of long words: "*cooperate*" and "*independent*", as well as the system in [54], which classified between two classes: "*yes*", "*no*". We also compared our results with the state-of-the-art method in [2]. As seen in Table 8, in the two-class scenario, our method outperforms the other methods in the comparison despite the fact that we used data from 19 subjects, a population that is substantially wider than that of other studies which included only few subjects [24]. This increases our confidence in the proposed system. Further, although our method used multiple nouns and multiple verbs, which increases intra-class variation and makes recognition more difficult, our system reached excellent results using three different experimental protocols and achieved a maximum classification rate of 85.7%.

7. Conclusion

We propose a framework for recognizing grammatical class (noun, verb) from EEG signals produced during covertly spoken words. Our proposed multi-channel convolutional neural network (MC-CNN) uses EEG signals captured from different areas of the head that were most appropriate for our application. The proposed method uses MC-CNN to extract time-frequency features from spectrograms belonging to different electrode groups and combine them to achieve high classification rate. Experimental results with the proposed framework showed that the proposed method can recognize the grammatical class of imagined nouns and verbs.

References

[1] L. Wang, X. Zhang, X. Zhong, Y. Zhang, Analysis and classification of speech imagery EEG for bci, Biomedical signal processing and control 8 (2013) 901–908.

[2] C. H. Nguyen, G. K. Karavas, P. Artemiadis, Inferring imagined speech using EEG signals: a new approach using riemannian manifold features, Journal of neural engineering 15 (2017) 016002.

[3] S. Zhao, F. Rudzicz, Classifying phonological categories in imagined and articulated speech, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, pp. 992–996.

[4] K. Brigham, B. V. Kumar, Subject identification from electroencephalogram (EEG) signals during imagined speech, in: Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on, IEEE, 2010, pp. 1–8.

[5] H. J. Hwang, H. Choi, J. Y. Kim, W.-D. Chang, D. W. Kim, K. Kim, S. Jo, C. H. Im, Toward more intuitive brain–computer interfacing: classification of binary covert intentions using functional near-infrared spectroscopy, Journal of biomedical optics 21 (2016) 091303.

[6] M. Bierwisch, Words in the brain are not just labelled concepts, Behavioral and Brain Sciences 22 (1999) 280–282.

[7] D. Crepaldi, M. Berlingeri, E. Paulesu, C. Luzzatti, A place for nouns and a place for verbs? a critical review of neurocognitive data on grammatical-class effects, Brain and language 116 (2011) 33–49.

[8] M. Popp, N. M. Trumpp, E.-J. Sim, M. Kiefer, Brain activation during conceptual processing of action and sound verbs, Advances in Cognitive Psychology 15 (2019) 236.

[9] A. Schilling, R. Tomasello, M. R. Henningsen-Schomers, A. Zankl, K. Surendra, M. Haller, V. Karl, P. Uhrig, A. Maier, P. Krauss, Analysis of continuous neuronal activity evoked by natural speech with computational corpus linguistics methods, Language, Cognition and Neuroscience (2020) 1–20.

[10] S. Tsigka, C. Papadelis, C. Braun, G. Miceli, Distinguishable neural correlates of verbs and nouns: A MEG study on homonyms, Neuropsychologia 54 (2014) 87–97.

[11] M. A. Bakhshali, M. Khademi, A. Ebrahimi-Moghadam, S. Moghimi, EEG signal classification of imagined speech based on riemannian distance of correntropy spectral density, Biomedical Signal Processing and Control 59 (2020) 101899.

[12] A. Balaji, A. Haldar, K. Patil, T. S. Ruthvik, C. Valliappan, M. Jartarkar, V. Baths, EEG-based classification of bilingual unspoken speech using ANN, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2017, pp. 1022–1025.

[13] G. Vigliocco, D. P. Vinson, J. Druks, H. Barber, S. F. Cappa, Nouns and verbs in the brain: a review of behavioural, electrophysiological, neuropsychological and imaging studies, Neuroscience & Biobehavioral Reviews 35 (2011) 407–426.

[14] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[15] J. T. Panachakel, A. Ramakrishnan, 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.

[16] P. Saha, M. Abdul-Mageed, S. Fels, Speak your mind! towards imagined speech recognition with hierarchical deep learning, arXiv preprint arXiv:1904.05746 (2019).

[17] S. Yang, T. Gao, J. Wang, B. Deng, B. Lansdell, B. Linares-Barranco, Efficient spike-driven learning with dendritic event-based processing, Frontiers in Neuroscience 15 (2021) 97.

[18] S. A. Lobov, A. N. Mikhaylov, M. Shamshin, V. A. Makarov, V. B. Kazantsev, Spatial properties of stdp in a self-learning spiking neural network enable controlling a mobile robot, Frontiers in neuroscience 14 (2020) 88.

[19] S. Yang, B. Deng, J. Wang, H. Li, M. Lu, Y. Che, X. Wei, K. A. Loparo, Scalable digital neuromorphic architecture for large-scale biophysically meaningful neural network with multi-compartment neurons, IEEE transactions on neural networks and learning systems 31 (2019) 148–162.

[20] S. Yang, J. Wang, B. Deng, C. Liu, H. Li, C. Fietkiewicz, K. A. Loparo, Real-time neuromorphic system for large-scale conductance-based spiking neural networks, IEEE transactions on cybernetics 49 (2018) 2490–2503.

[21] A. Kugele, T. Pfeil, M. Pfeiffer, E. Chicca, Efficient processing of spatio-temporal data streams with spiking neural networks, Frontiers in Neuroscience 14 (2020) 439.

[22] Y. Kim, Convolutional neural networks for sentence classification, arXiv preprint arXiv:1408.5882 (2014).

[23] A. Porbadnigk, M. Wester, T. S. Jan-p Calliess, EEG-based speech recognition impact of temporal effects (2009).

[24] S. J. Luck, An introduction to the event-related potential technique mit press, Cambridge, Ma (2005) 45–64.

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

[26] C. Demanuele, C. J. James, E. J. Sonuga-Barke, Distinguishing low frequency oscillations within the 1/f spectral behaviour of electromagnetic brain signals, Behavioral and Brain Functions 3 (2007) 62.

[27] A. Flinker, A. Korzeniewska, A. Y. Shestyuk, P. J. Franaszczuk, N. F. Dronkers, R. T. Knight, N. E. Crone, Redefining the role of Broca's area in speech, Proceedings of the National Academy of Sciences 112 (2015) 2871–2875.

[28] C. Weiller, C. Isensee, M. Rijntjes, W. Huber, S. Müller, D. Bier, K. Dutschka, R. P. Woods, J. Noth, H. C. Diener, Recovery from wernicke's aphasia: a positron emission tomographic study, Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society 37 (1995) 723–732.

[29] X. Pei, E. C. Leuthardt, C. M. Gaona, P. Brunner, J. R. Wolpaw, G. Schalk, Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition, Neuroimage 54 (2011) 2960–2972.

[30] X. Pei, J. Hill, G. Schalk, Silent communication: toward using brain signals, IEEE pulse 3 (2012) 43–46.

[31] H. Preissl, F. Pulvermüller, W. Lutzenberger, N. Birbaumer, Evoked potentials distinguish between nouns and verbs, Neuroscience Letters 197 (1995) 81–83.

[32] K. A. Shapiro, A. Pascual-Leone, F. M. Mottaghy, M. Gangitano, A. Caramazza, Grammatical distinctions in the left frontal cortex, Journal of cognitive neuroscience 13 (2001) 713–720.

[33] M. A. Gernsbacher, M. P. Kaschak, Neuroimaging studies of language production and comprehension, Annual review of psychology 54 (2003) 91–114.

[34] H. Petsche, D. Lacroix, K. Lindner, P. Rappelsberger, E. Schmidt-Henrich, Thinking with images or thinking with language: a pilot EEG probability mapping study, International Journal of Psychophysiology 12 (1992) 31–39.

[35] A. von Stein, P. Rappelsberger, J. Sarnthein, H. Petsche, Synchronization between temporal and parietal cortex during multimodal object processing in man, Cerebral Cortex 9 (1999) 137–150.

[36] D. Dentico, B. L. Cheung, J. Chang, J. Guokas, M. Boly, G. Tononi, B. Van Veen, Reversal of cortical information flow during visual imagery as compared to visual perception, Neuroimage 100 (2014) 237–243.

[37] A. Ramakrishnan, J. Satyanarayana, Reconstruction of EEG from limited channel acquisition using estimated signal correlation, Biomedical Signal Processing and Control 27 (2016) 164–173.

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

[39] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167 (2015).

[40] Y. R. Tabar, U. Halici, A novel deep learning approach for classification of EEG motor imagery signals, Journal of neural engineering 14 (2016) 016003.

[41] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT press, 2016.

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

[43] N. Nikhil, is ReLU after Sigmoid bad?, 2018. [Online] Available: https://towardsdatascience.com/is-relu-after-sigmoid-bad-661fda45f7a2.

[44] L. Lu, Y. Shin, Y. Su, G. E. Karniadakis, Dying relu and initialization: Theory and numerical examples, arXiv preprint arXiv:1903.06733 (2019).

[45] F. Chollet, keras, [Online] Available: https://github.com/fchollet/keras, 2015.

[46] 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).

[47] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[48] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[49] K. He, X. Zhang, S. Ren, 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.

[50] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, The journal of machine learning research 15 (2014) 1929–1958.

[51] S. Sur, V. Sinha, Event-related potential: An overview, Industrial psychiatry journal 18 (2009) 70.

[52] P. Khader, F. Rösler, EEG power and coherence analysis of visually presented nouns and verbs reveals left frontal processing differences, Neuroscience Letters 354 (2004) 111–114.

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

[54] A. R. Sereshkeh, R. Trott, A. Bricout, T. Chau, EEG classification of covert speech using regularized neural networks, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (2017) 2292–2300.

[55] P. Saha, S. Fels, Hierarchical deep feature learning for decoding imagined speech from EEG, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 10019–10020.

[56] M. Long, Y. Cao, J. Wang, M. Jordan, Learning transferable features with deep adaptation networks, in: International conference on machine learning, PMLR, 2015, pp. 97–105.