Classifying Imaginary Vowels from Frontal Lobe EEG via Deep Learning

Megha Parhi and Ahmed H. Tewfik
Department of Electrical and Computer Engineering
The University of Texas at Austin, Austin, Texas
Email: mparhi@utexas.edu, tewfik@austin.utexas.edu

***Abstract*—Brain-Computer Interface (BCI) is a promising technology for individuals who suffer from motor or speech disabilities due to the process of decoding brain signals. This paper uses a dataset for imagined speech to classify vowels based on the neurological areas of the brain. The normalized cross-correlation matrices between two electrodes are used as features. We demonstrate that by using the EEG from the frontal region of the brain, we obtain higher than 85 percent accuracy for correct vowel decoding by using two types of neural networks: convolutional neural network (CNN) and long short-term memory (LSTM). This accuracy is higher than previous studies that have classified the dataset using the entire brain region. This work shows great promise for task decoding where the physiological regions of the brain associated with specific tasks are exploited. The proposed approach has the potential to be deployed in future BCI applications.**

I. Introduction

Advances in brain-computer interface (BCI) hold great promise in improving the quality of life in clinical disorders associated with neurology and rehabilitation [1], [2]. BCI has been used for many different applications including entertainment (i.e., games such as ping pong and pacman) [3]–[6] and communication (i.e., internet searches and virtual keyboards) [7], [8]. BCI's main objective is to aid people to interact with their environment by decoding brain signals instead of relying on their muscle movement [9]. This technology could be very beneficial to people, for example, those who suffer from Locked-in syndrome. Locked-in syndrome is a rare neurological disease where a person is completely paralyzed and is unable to move any of their muscles [10]. BCI could help these individuals to communicate by using covert or silent speech.

Speech is a vital sense for communication. Several studies have investigated how to classify individual speech into categories like English vowels, short words, and long words [11]–[13]. The majority of this work has been carried out using classical machine learning (ML) models like support vector machines (SVMs). Many of the experiments for speech use data from all the electrodes and don't specifically pinpoint a certain location of the brain where activity occurs. The question lies in how do the neurological areas of the brain associate with the data and how a model can be learned that requires less computation leading to a low-cost speech-based BCI.

In this paper we show that by using the electrodes from the frontal lobe, i.e., the region responsible for speech in the brain, we can get the same, if not better, accuracy than using the measurements from all the electrodes. To the best of our knowledge, this is the first study of using a certain lobe to classify speech from the dataset of [12]. To show that the classification accuracy is as good, we analyze the data using the entire 64 electrodes and a subset of the electrodes from the frontal region of the brain where speech occurs. The classification process is modeled using two very well known and frequently used deep learning algorithms. These include the Long Short-Term Memory (LSTM) and the Convolutional Neural Network (CNN). Our results show that by using the frontal electrodes, the accuracy of the data is above 90 percent for each participant. This shows that speech-based BCI signals can be classified using only the active parts of the brain, which would help in enabling less computation time and less hardware. More details of this work are presented in [14].

A. Past Work

BCIs have been studied extensively for applications in speech and other tasks to understand what information is obtained from brain signals such as EEG signals. Nguyen *et al.* investigated an imaginary speech dataset using Riemannian manifold features classified by a Relevant Machine Vector [12]. A study by DaSalla *et al.* showed 68-79 percent accuracy when classifying a, u, and $rest$ using Common Spatial Patterns (CSP) [13]. The accuracy was found to be high as the CSPs from the discriminating channels Fz, C3, Cz, and C4 were used in the classification. These four channels are related to motor imagery and not speech imagery. Deng *et al.* used Huang-Hilbert transform to get an accuracy of 72.6 percent by classifying ba and ku imagined syllables [11]. These studies have primarily used signal processing algorithms and classical ML approaches to determine the classification for imagined speech in vowel data. These works have shown the promise of classification; however, neural networks have shown better accuracy results in classification.

Neural networks have shown great promise in classification tasks for many different applications including speech. Several papers have demonstrated speech recognition with recurrent neural networks (RNNs) [15], [16]. Long short-term memory models are a subgroup of recurrent neural networks that have been shown to provide very good accuracy in categorizing speech [17], [18]. Another work used frequency-following responses to project electrophysiological responses onto a low-

dimensional spectral feature for two vowels [19]. These studies investigate speech signals based on the entire brain's data. Our approach for this paper shows that the physiological aspects of the brain play a big part in understanding the BCI data and that less computation is sufficient when using the parts of the brain that are active during speech.

This paper is organized in the following manner. Section II explains the data used for the experiments. Section III explains the modeling of the data. Section IV discusses and compares the results. Finally, Section V concludes the work and presents future work.

II. Data

The data used for this paper is from a publicly available imaginary speech dataset [12]. This dataset was collected at the Human-Oriented Robotics and Control Lab of the Arizona State University. The collection of the data was approved by appropriate institutional review board (IRB). There were 15 subjects in this study; however, data for seven subjects was made publicly available. The subjects for the dataset are all right-handed except for Subject 13. During the experiments, the subjects were instructed to pronounce words in their mind with cues from a computer monitor. For each trial, a beep would appear when the trial started and a visual cue would be prompted. The beeping sound was repeated with a period of time T. In this experiment, T is 1 second for the vowel data. The trials would then end with a rest period of approximately two seconds. The authors give the reasoning for this is to allow the areas of the brain where sounds occurs to remain inactive. This dataset contains data for short words, long words, and vowels for EEG that follows the international 10/20 system placement [20]. For this paper, we are interpreting solely the vowel data, which consists of a, i, and u. The placement of the electrodes for the experiments used the 10/20 internationally recognized system. Table I shows what each abbreviated letter in the 10/20 system corresponds to which lobe. The function of each region in the 10/20 system can be observed in the illustration in Fig. 1.

TABLE I
LOBES OF THE BRAIN PERTAINING TO ELECTRODE ABBREVIATION

Electrode abbreviation	Lobe in Brain
F	Frontal
T	Temporal
C	Central
P	Parietal
O	Occipital

A. Preprocessing

The data was preprocessed as listed below:

- Bandpass filter at 8-70 Hz using a 5th order Butterworth filter.
- Notch filter at 60 Hz to remove power line signal.
- EOG artifact removal by the ADJUST algorithm [22] to remove muscle and eye blinks.

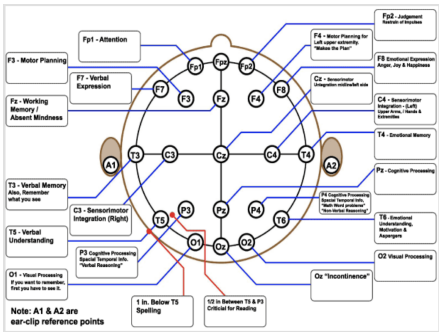

Fig. 1. Example of the neurological function of each electrode and placement using the 10-20 system with brain function [21].

- Downsample from 1000 Hz to 256 Hz.

B. Cross-Correlation matrix

The features for the deep learning models chosen are the cross-correlation matrices between the electrodes for each trial for each of the seven subjects. We calculate the *cross-correlation matrices* for each subject. Each cross-correlation matrix is calculated in the following manner with e_i representing electrode i and e_j representing electrode j. X represents the EEG signal, t represents the time, and τ represents the lag. The covariance between two electrodes is computed using:

$$Cov(X_{e_i}, X_{e_j})(\tau) = E[(X_{e_i}(t) - E(X_{e_i}(t))(X_{e_j}(t+\tau) - E(X_{e_j}(t+\tau)]$$

We make the assumption that the electrode signals have zero mean. For shorthand notation, we write $Cov(X_{e_i}, X_{e_j})(\tau)$ as $Cov_{e_i e_j}$. The cross-correlation matrix between each electrode is then calculated as follows using the covariance for each subject:

$$R_{e_i e_j} = \frac{Cov_{e_i e_j}}{\sqrt{Cov_{e_i e_i} Cov_{e_j e_j}}}$$

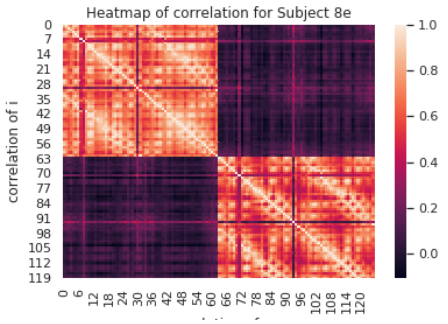

Fig. 2. Example of heatmap of cross-correlation matrix for Subject 8e between a and i. The first half of the electrodes correspond to a and the second half to i.

Fig. 2 represents an example of a correlation matrix from the data between the vowel sounds a and i for subject 8e. The white diagonal represents the correlation between two of the same electrodes. The correlation matrices are symmetric. We

can observe that there are some highly correlated electrodes and a few electrodes that have low correlation. This correlation matrix was calculated for data with 64 electrodes, which creates a 128×128 matrix. The top right and bottom left 64×64 matrices or the heavily purple section of the heat map represent the cross-correlation of a and i. The top left and bottom right 64×64 matrices represent the correlation of the vowel and itself. A total of 191 such matrices are calculated for each subject and used as features for all electrodes. The less correlated electrodes show that those electrodes may provide more discrimination than the electrodes that are highly correlated with each other. A subset of all the 64 electrodes is the 20 frontal electrodes where speech occurs. The cross-correlation matrices for the frontal electrodes are calculated in the same manner as the electrodes. The difference between the frontal electrodes correlation matrices and all the electrodes is the number of channels and the dimension of the frontal cross-correlation matrices. The frontal electrodes consist of 20 channels. The dimension of the cross-correlation matrices is 40×40. A total of 191 cross-correlation matrices are calculated corresponding to the 20 frontal electrodes for the labels a, e, and i.

C. Labeling

The matrices are labeled by taking the first trial vowel with whatever vowel it is correlated with. For example, corr0 is labeled as a due to the fact that we take the first trial of a and then correlate it with i. This method aids in understanding what part of the data we are correlating. All the other matrices are calculated in a similar manner for each subject. The subset frontal electrodes are labelled in the exact same manner as described above for all the electrodes. Table II shows which correlation matrices (corr) pertain to which vowel for both subsets of electrodes. There are 75 matrices labeled as a, 61 labeled as i, and 55 labeled as u.

TABLE II
LABELING OF CORRELATION MATRICES

Vowel label	Correlation Matrix
a	corr0-18,corr55-70,corr100-112,corr136-145,corr163-169 corr181-190
i	corr19-37,corr71-85,corr113-124,corr146-154,corr170-175
u	corr38-54,corr86-99,corr125-135,corr155-162,corr176-180

III. MODELS

The two models that are used to investigate and understand the BCI EEG data are: LSTM and CNN. These two neural network models were chosen due to their promising results in classification tasks [23], [24]. The data input has 191 correlation matrices representing the features of the input to the classifiers for each subset of the electrode. Each correlation matrix is a 128×128 matrix for each electrode pair. For the subset of all the electrodes, the frontal electrodes consist of the same number of 191 correlation matrices of dimension 40×40 for each subject. Out of the 191, 151 matrices are used as the training data and 40 as the testing data for each subject.

For the experiments considered, we take all the electrodes in the brain, and the twenty electrodes from the frontal region for both the left and right hemisphere. We calculate the accuracy for each of the seven subjects. The vowels are *one–hot encoded*.

The loss is calculated by categorical cross entropy where the true class is represented as a one-hot encoded vector. The outputs are compared to the one-hot encoded vector, which will then determine the loss. This can be measured in the following manner where $\hat{y}$ is the predicted output, which is the output from the softmax:

$$L(y, \hat{y}) = -\sum_{j=0}^{M}\sum_{i=0}^{N}(y_{ij}\log(\hat{y}_{ij}))$$

A. LSTM

Long short-term memory (LSTM) has shown promising results in speech systems. LSTMs have four gates that interact in a specific way. A stacked LSTMs is one type of LSTM that was proposed by Hinton *et al.* for speech [15]. An example layer of the LSTM network used is illustrated in Fig. 3.

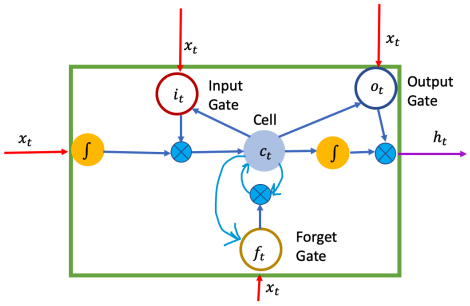

Fig. 3. One layer of LSTM architecture.

We train a model using the correlation matrices via a stacked LSTM with two layers. In a typical RNN with an input sequence $\mathbf{x}=(x_1, ...x_T)$, $\mathbf{h}=(h_1, ..., h_T)$ is the hidden vector, $\mathbf{y}=(y_1, y_2, ...y_T)$ is the output vector, and the time series is from $t = 1, 2, ..., T$. The RNN is formulated as follows:

$$h_t = \mathcal{H}(W_{xh}x_t + W_{hh}h_{t-1} + b_h)$$
$$y_t = W_{yh}h_t + b_y$$

where W represents the weight matrices, b represents the bias vector, and $\mathcal{H}$ denoted the hidden layer function. The LSTM comprises of four gates: input gate, forget gate, output gate, and cell activation vector. These are all the same size as h from RNN. The LSTM module is described by:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i)$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f)$$
$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o)$$
$$h_t = o_t \tanh(c_t)$$

where σ represents the logistic sigmoid function, and i, f, o, and c represent the input gate, forget gate, output gate, and cell activation, respectively. The weight matrices are denoted by W_{ij} and the bias vectors are represented by b_i.

B. CNN

A Convolutional Neural Network is a multi-layer neural network with several convolution-pooling layer pairs and fully-connected layers at the output. It can take an input image and be able to differentiate from other images. For this paper, the images correspond to the correlation matrices. Preprocessing steps in CNNs are known to be much lower than that for other classification algorithms, which makes this model easier to work with. First, the input to a CNN is a tensor. For our data for all the 64 electrodes, the input tensor size is 191 by 128 by 128 corresponding to the cross-correlation matrices. An individual image is of size 128 by 128. The convolution layer or kernel is run with certain image dimensions. In our case, we take a 3*3*1 image. The kernel then shifts through the entire image. This is known as the convolution step, which is used to extract high-level features. We use two convolutional layers in our experiment. After convolving, a pooling layer is run, which is used to decrease the computation needed. We use a max pooling layer in our case. After the convolutional layers, an activation function is applied. This is followed by a softmax layer. The weights are trained using cross-entropy loss.

IV. Results

To test whether the hypothesis that the selected physiological regions of the brain lead to higher accuracy than using the entire brain, we ran experiments using all 64 electrodes and a subset of the 64 electrodes containing 20 electrodes that correspond to the frontal electrodes. These are known as physiological areas of the brain where speech occurs. We ran all our experiments using Google CoLab.

A. Model Parameters

Parameter tuning is an art when it comes to training models. For the experiments that are run on the correlation matrices, Table III summarizes the parameters for all 64 electrodes (LSTMa and CNNa) and the subset containing 20 frontal electrodes (LSTMf and CNNf).

TABLE III
Parameters of Models

Model Parameter	CNNa	LSTMa	CNNf	LSTMf
Epochs	100	50	100	50
Batch Size	150	100	150	100
Total Layers	2	2	2	2
Number of Hidden Layers	1	1	1	1
Activation	ReLu	ReLu	ReLu	ReLU
Optimizer	Adam	Adam	Adam	Adam

B. Model Results

Table IV shows the accuracy of the LSTM and CNN based on all 64 electrodes (LSTMa and CNNa) and the subset containing 20 frontal electrodes (LSTMf and CNNf) for each Subject (Sub). We can observe that utilizing the brain signals in the frontal electrodes shows higher accuracy than learning from the entire 64 electrodes for these eight subjects. Using all the electrodes, the accuracy using LSTM never reaches above 80 percent, while when just the frontal electrodes are used, the accuracy significantly improves to above 90 percent for each subject. This shows the promise of using EEG data in certain regions of the brain based on activities. This also shows that by removing about 67% of the data we are able to obtain higher accuracy. The 67% of the data removed pertains to the electrodes that are not measuring the frontal lobe. The reason behind this is because speech occurs in the frontal region, which is our region of interest.

TABLE IV
Test Accuracy of LSTM and CNN

Subject	LSTMa	LSTMf	CNNa	CNNf	Nguyen [12]	Saha [25]
Sub 8	74.4	**96.2**	68.5	**92.1**	51	73
Sub 8e	76.9	**100**	73.4	**95.1**	NA	NA
Sub 9	33.3	**92.1**	28.2	**89.5**	NA	NA
Sub 11	12.8	**90**	15.2	**85.0**	53	75
Sub 12	51.3	**99.2**	53.6	**94.5**	51	79
Sub 13	33.3	**100**	27.4	**98.4**	46.7	69
Sub 15	69.2	**100**	73.4	**100**	48	84

Similar to LSTM, we observe that there is significant improvement by just using the frontal electrodes. The CNN accuracy is slightly worse than LSTM, but it shows significant improvement.

C. Comparison of Results with Past Work

We compare the results from Table IV with the results from [25] and [12]. We compare with these two prior papers since these use the same dataset with different methods. Saha *et al.* proposed a LSTM and CNN hybrid model. They calculated the channel cross-covariance of the electrodes to determine the accuracy using the Nguyen *et al.* dataset, which we used in our experiments as well. Unlike Saha *et al.*, we calculate the cross-correlation matrices between the two electrodes for our features. Another difference between the two studies is that we pinpoint the frontal electrodes as the region where the most activity in the brain occurs during speech. Both Saha and Nguyen use all 64 electrodes of the brain region. Table IV summarizes the results from [25], which is compared with [12]. Nguyen *et al.* compute the covariance matrix as the feature vector and use a Relevance Vector Machine to classify the vowel data using Riemannian Manifold features [12].

Compared to past work, we observe that the results from [12] are very low in accuracy compared with Saha *et al.* [25]. The accuracy for [12] is approximately 50 percent for all the subjects using all the electrodes. For our Subjects 9 and 13, we get worse accuracy using all the electrodes using

neural network models than [12] and [25]. This could be due to the way the data has been measured. These results also demonstrates the difference between traditional ML methods and the accuracy that can be obtained with neural networks for classification.

The approach in Saha *et al.* [25] achieves accuracy less than 85% for all the subjects. This is slightly better than our results for all electrodes since we achieve less than 80%. For all 64 electrodes our results have similar accuracy to [25] for the Subject 8. When we use just the frontal electrode, we have accuracy above 85% for all subjects using both LSTM and CNN. These results are better than the hierarchical model using simple LSTM and CNN architectures proposed by [25]. This shows that by understanding the physiological aspects of the brain, we can better understand and classify brain signals to aid in computation time and accuracy.

Overall, we were able to show that by pinpointing the neurological area of the brain that is active, one is able to obtain higher accuracy than using the data from the entire brain. This creates better understanding of brain signals and shows the need to understand the physiology of the active brain.

V. Conclusion

We demonstrate in this paper that by using the correlation data from the frontal region of the brain we are able to obtain a vowel decoding accuracy that is above 90 percent, while using the entire brain region the accuracy tends to be below 80 percent using LSTM and CNN. This demonstrates that the neurological parts of the brain where the brain is active can aid in understanding the data. This would significantly reduce hardware as well as computational time in BCI experiments.

One limitation of the proposed work is that the sample size is small. Future work will investigate similar analysis using a larger sample size to validate the method proposed in the paper. If validated, the proposed approach can help patients who are locked in. The overall objective of this work is to develop a non-stationary system to aid in understanding the algorithms needed to create an online system that could be potentially used for people who suffer from speech disorders based on the specific areas of the brain. Another goal is to design a real-time system where EEG signals from the selected brain regions can be decoded in real time.

References

[1] J.S. Brumberg, A. Nieto-Castanon, P.R. Kennedy, and F.H. Guenther. Brain–computer interfaces for speech communication. *Speech communication*, 52(4):367–379, 2010.

[2] C. Herff and T. Schultz. Automatic speech recognition from neural signals: a focused review. *Frontiers in neuroscience*, 10:429, 2016.

[3] K. LaFleur, K. Cassady, A. Doud, K. Shades, E. Rogin, and B. He. Quadcopter control in three-dimensional space using a noninvasive motor imagery-based brain–computer interface. *Journal of neural engineering*, 10(4):046003, 2013.

[4] C.G. Coogan and B. He. Brain-computer interface control in a virtual reality environment and applications for the internet of things. *IEEE Access*, 6:10840–10849, 2018.

[5] R. Krepki, B. Blankertz, G. Curio, and K. Müller. The berlin brain-computer interface (bbci)–towards a new communication channel for online control in gaming applications. *Multimedia Tools and Applications*, 33(1):73–90, 2007.

[6] M. W. Tangermann, M. Krauledat, K. Grzeska, M. Sagebaum, C. Vidaurre, B. Blankertz, and K. Müller. Playing pinball with non-invasive bci. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, pages 1641–1648. Citeseer, 2008.

[7] N. Birbaumer, N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kübler, J. Perelmouter, E. Taub, and H. Flor. A spelling device for the paralysed. *Nature*, 398(6725):297, 1999.

[8] B. Jarosiewicz, A.A. Sarma, D. Bacher, N.Y. Masse, J.D. Simeral, B. Sorice, E.M. Oakley, C. Blabe, C. Pandarinath, V. Gilja, et al. Virtual typing by people with tetraplegia using a self-calibrating intracortical brain-computer interface. *Science translational medicine*, 7(313):313ra179–313ra179, 2015.

[9] J.R. Wolpaw, N. Birbaumer, D.J. McFarland, G. Pfurtscheller, and T.M. Vaughan. Brain–computer interfaces for communication and control. *Clinical neurophysiology*, 113(6):767–791, 2002.

[10] S. Laureys, F. Pellas, P. Van Eeckhout, S. Ghorbel, C. Schnakers, F. Perrin, J. Berre, M. Faymonville, K. Pantke, and F. Damas. The locked-in syndrome: what is it like to be conscious but paralyzed and voiceless? *Progress in brain research*, 150:495–611, 2005.

[11] S. Deng, R. Srinivasan, T. Lappas, and M. D'Zmura. Eeg classification of imagined syllable rhythm using hilbert spectrum methods. *Journal of neural engineering*, 7(4):046006, 2010.

[12] C.H. Nguyen and P. Artemiadis. Eeg feature descriptors and discriminant analysis under riemannian manifold perspective. *Neurocomputing*, 275:1871–1883, 2018.

[13] C.S. DaSalla, H. Kambara, M. Sato, and Y. Koike. Single-trial classification of vowel speech imagery using common spatial patterns. *Neural networks*, 22(9):1334–1339, 2009.

[14] Megha Parhi. Classifying imaginary vowels from frontal lobe eeg via deep learning, 2020.

[15] A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6645–6649. IEEE, 2013.

[16] A. Graves. Supervised sequence labelling. In *Supervised sequence labelling with recurrent neural networks*, pages 5–13. Springer, 2012.

[17] G. Krishna, C. Tran, J. Yu, and A.H. Tewfik. Speech recognition with no speech or with noisy speech. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1090–1094. IEEE, May 2019.

[18] M. Sakthi, A. Tewfik, and B. Chandrasekaran. Native language and stimuli signal prediction from eeg. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3902–3906. IEEE, 2019.

[19] H.G. Yi, Z. Xie, R. Reetzke, A.G. Dimakis, and B. Chandrasekaran. Vowel decoding from single-trial speech-evoked electrophysiological responses: A feature-based machine learning approach. *Brain and behavior*, 7(6):e00665, 2017.

[20] F. Sharbrough, G.E. Chatrian, Ronald Lesser, H. Luders, M. Nuwer, and T. Picton. American electroencephalographic society guidelines for standard electrode position nomenclature. *Clinical Neurophysiology*, 8:200–202, 01 1991.

[21] *Edmonton Neurotherapy: QEEG Brain Mapping*. https://www.edmontonneurotherapy.com/edmonton-neurotherapy-qeeg.

[22] A. Mognon, J. Jovicich, L. Bruzzone, and M. Buiatti. Adjust: An automatic eeg artifact detector based on the joint use of spatial and temporal features. *Psychophysiology*, 48(2):229–240, 2011.

[23] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005.

[24] A. Krizhevsky, I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[25] P. Saha and S. Fels. Hierarchical deep feature learning for decoding imagined speech from eeg. *arXiv preprint arXiv:1904.04352*, 2019.