Decoding Imagined Speech using Wavelet Features and Deep Neural Networks

Jerrin Thomas Panachakel
Indian Institute of Science
Bangalore, India
jerrinp@iisc.ac.in

A.G. Ramakrishnan
Indian Institute of Science
Bangalore, India
agr@iisc.ac.in

T.V. Ananthapadmanabha
Voice and Speech Systems
Bangalore, India
tva.blr@gmail.com

***Abstract*—This paper proposes a novel approach that uses deep neural networks for classifying imagined speech, significantly increasing the classification accuracy. The proposed approach employs only the EEG channels over specific areas of the brain for classification, and derives distinct feature vectors from each of those channels. This gives us more data to train a classifier, enabling us to use deep learning approaches. Wavelet and temporal domain features are extracted from each channel. The final class label of each test trial is obtained by applying a majority voting on the classification results of the individual channels considered in the trial. This approach is used for classifying all the 11 prompts in the KaraOne dataset of imagined speech. The proposed architecture and the approach of treating the data have resulted in an average classification accuracy of 57.15%, which is an improvement of around 35% over the state-of-the-art results.**

***Index Terms*—imagined speech, brain-computer interaction, deep neural network, commone spatial pattern, EEG**

I. Introduction

Speech is one of the most basic and natural form of human communication. However, nearly 70 million people have speech disabilities around the world. Speech disability due to complete paralysis prevents people from communicating with other through any modality. It will greatly help such people, if by some means we are able to decode his/her thoughts, commonly referred to as "imagined speech" [1].

The interest in imagined speech dates back to the days of Hans Berger, who invented electroencephalogram (EEG) as a tool for synthetic telepathy [2]. Although it is almost a century since the first EEG recording, the success in decoding imagined speech from EEG signals is rather limited. One of the major reasons for the same is the very low signal-to-noise ratio (SNR) of EEG signals.

The potential of the recent developments in the field of machine learning, such as deep neural networks (DNN) has not been exploited fully in the field of decoding imagined speech, since such techniques require a huge amount of training data. In this paper, we selected 11 EEG channels that cover the cortical areas involved in speech imagery. For each imagined word, each of the EEG channels so selected is considered as an independent input signal, thus providing more training data. This is in contrast to the earlier approaches concatenating the features to form a single feature vector.

Our new approach has been validated using the KaraOne dataset [3] and we have obtained accuracy values better than the state-of-the-art results reported in the literature for the same dataset.

The rest of the paper is organized as follows: Section II describes prior work in the literature in the field of decoding imagined speech. Section III describes the dataset and the procedure for generating the feature vectors. Section IV describes the proposed DNN classifier in some detail. The results obtained are given in section V.

II. Related Work in the Literature

This section briefly describes the work in the field of decoding imagined speech, reported over the last decade.

C.S. DaSalla *et al.* developed a brain-computer interface (BCI) system based on vowel imagery [4] in the year 2009. The objective was to discriminate between the imagined vowels, */a/* and */u/*. The experimental paradigm consisted of three parts:

1) Imagined mouth opening and imagined vocalisation of vowel */a/*.
2) Imagined lip rounding and imagined vocalisation of vowel */u/*.
3) Control state with no action.

Using common spatial pattern (CSP) generated spatial filter vectors as features and nonlinear support vector machine (SVM) as the classifier, they achieved accuracies in the range of 56% to 72% for different subjects. As noted by Brigham *et.al* [1], the relatively higher accuracy obtained might have arisen because of the additional involvement of motor imagery.

Following a similar approach, Wang Li *et al.* in 2013 developed a system to distinguish between two monosyllabic Chinese characters meaning "left" and "one" [5]. Visual cue was provided to the subject to instruct him/her on the character to be imagined. When the cue disappeared, the subject had to repeatedly imagine the character in his/her mind as many times as possible for a duration of 4 sec. They obtained an accuracy of around 67%.

In 2010, Brigham *et al.* came up with an algorithm based on autoregressive (AR) coefficients and k-nearest neighbor (k-NN) algorithm for classifying two imagined syllables */ba/* and */ku/* [1]. In this experiment, the subjects were given an auditory cue on the syllable to be imagined, followed by a series of click sounds. After the last click, the subjects were instructed

978-1-7281-2327-1/19/$31.00 ©2019 IEEE

to imagine the syllable once every 1.5 sec for a period of 6 sec. They reported an accuracy of around 61%.

In 2016, Min *et.al* used statistical features such as mean, variance, standard deviation, and skewness for pairwise classification of vowels (*/a/*, */e/*, */i/*, */o/*, and */u/*) using extreme learning machine (ELM) with radial basis function. In their experimental paradigm, auditory cue was provided at the beginning of the trial to inform the subject as to which vowel was to be imagined. After the auditory cue, two beeps were played, after which the subject had to imagine the vowel heard during the beginning of the trial. An average accuracy of about 72% was reported.

In 2017, Nguyen, Karavas and Artemiadis [6] came up with an approach based on Riemannian manifold features for classifying four different sets of prompts:

1) Vowels (*/a/*, */i/* and */u/*).
2) Short words ("*in*" and "*out*").
3) Long words ("*cooperate*" and "*independent*").
4) Short-long words ("*in*" and "*cooperate*").

The accuracy reported for the four sets of prompts are 49.2%, 50.1%, 66.2% and 80.1%, respectively. This dataset is one amongst the few imagined speech datasets that are available in the public domain and is referred to as the *"ASU dataset"*. More information about this dataset is given in Section III-A.

Balaji *et al.* in 2017 investigated the use of bilingual imaginary speech, namely English "*Yes"* & "*No*" and Hindi "*Haan*" (meaning "*yes*") & "*Na*" (meaning "*no*") for an imagined speech based BCI system [7]. Principal component analysis (PCA) was used for data reduction and an artificial neural network (ANN) was used as the classifier. Two specific sets of EEG channels corresponding to language comprehension and decision making were utilized. An interesting part of the experimental protocol was that there was no auditory or visual cue and the subjects were instructed to imagine the answer to a binary question posed either in English or Hindi. The study reported an accuracy of 75.4% for the combined English-Hindi task and quite a surprisingly high accuracy of 85.2% for classifying the decision.

In 2017, Sereshkeh *et al.* came up with an algorithm based on features extracted using discrete wavelet transform (DWT) and regularized neural networks for classifying the imagined decisions of "*yes*" and "*no*" [8], similar to the work by Balaji *et al.* They reported an accuracy of about 67%.

In 2018, Cooney *et al.* [9] used Mel frequency cepstral coefficients (MFCC) as features and SVM as classifier to classify all the 11 prompts in the KARAONE dataset [3]. The prompts consisted of seven phonemic/syllabic prompts (*/iy/*, */uw/*, */piy/*, */tiy/*, */diy/*, */m/*, */n/*) and four words ("pat", "pot", "knew" and "gnaw"). A maximum accuracy of only 33.33% was achieved. The lower accuracy might have arisen because of a larger number of choices, unlike the binary choice employed in the previous works.

In a recent work [10], Jerrin et al. used deep neural networks (DNN) for the first time to classify imagined speech. The specific task was to classify imagined words "in" and "cooperate". The features used were based on discrete wavelet transform and a DNN with three hidden layers was employed. The highest accuracy reported was around 86%.

III. Dataset Used for the Study and Methods

A. The KaraOne Dataset

The KaraOne dataset [3] has been used for our study. The KaraOne dataset consists of EEG data captured during the imagination and articulation of 11 prompts, which included 7 phonemic/syllabic prompts (iy, uw, piy, tiy, diy, m, n) and 4 words derived from Kent's list of phonetically-similar pairs (i.e., pat, pot, knew, and gnaw). The data was captured at 1 KHz sampling rate using SynAmps RT amplifier. The electrode placement was based on the 10/10 system [11].

Each data recording trial had four stages:

1) A 5-second rest state.
2) A stimulus state in which an auditory and a visual cue were provided to the participant.
3) A 5 seconds imagined speech state.
4) An articulation state.

We followed the same preprocessing steps as in [3], which included ocular artifact removal using blind source separation [12], band-pass filtering from 1 to 50 Hz and a Laplacian spatial filtering.

B. Wavelet Feature Extraction

In our work, instead of concatenating the features obtained from several channels, each channel is treated as a distinct input. This is possible because of the high correlation present between the signals of various channels [13]. The following 11 channels only have been chosen to be used in our work, based on the involvement of the underlying brain regions in the production of speech [14], [15]:

1) 'C4': postcentral gyrus
2) 'FC3': premotor cortex
3) 'FC1': premotor cortex
4) 'F5': inferior frontal gyrus, Broca's area
5) 'C3': postcentral gyrus
6) 'F7': Broca's area
7) 'FT7': inferior temporal gyrus
8) 'CZ': postcentral gyrus
9) 'P3': superior parietal lobule
10) 'T7': middle temporal gyrus, secondary auditory cortex
11) 'C5': Wernicke's area, primary auditory cortex

This choice of channels is also backed by the common spatial patterns (CSP) analysis on imagined speech v/s rest state EEG data [6].

Since each EEG channel is considered as an independent data vector, algorithms that extract a single feature vector from the entire set of EEG channels (such as Reimannian manifold features used by Nguyen *et.al* [6] and fuzzy entropy features [16]) cannot be used with the proposed architecture.

For each trial, only the first 3000 samples (3 seconds) of collected data have been used for feature extraction. For extracting the temporal features, we divided the first 3000

TABLE I
COMPARISON OF THE MEAN CROSS-VALIDATION ACCURACIES IN PERCENTAGE OBTAINED USING DIFFERENT METHODS (GIVEN IN EACH COLUMN) IN CLASSIFYING 11 IMAGINED PROMPTS IN THE KARAONE DATASET. "S01" TO "S08" ARE THE PARTICIPANT IDS.

	Method		
Subject	**SVM+MFCC [9]**	**DT+MFCC [9]**	**Proposed Method (DNN)**
s01	22.27	24.52	43.02
s02	33.33	31.06	60.91
s03	23.62	15.12	84.23
s04	15.31	21.14	45.78
s05	14.84	11.41	37.43
s06	20.86	21.17	60.81
s07	26.08	26.84	75.07
s08	23.15	18.37	49.98
Average:	22.43	21.20	57.15

samples into 4 equal blocks and extracted the following statistical features for each block:

1) Root-mean-square
2) Variance
3) Kurtosis
4) Skewness
5) 3rd order moment

Daubechies-4 (db4) wavelet is extensively used to extract features from EEG signals [17]. The 3 second-EEG signals are decomposed into 7 levels using db4 wavelet, for extracting the wavelet domain features. The above mentioned statistical features are extracted from the last approximation coefficients and for each of the last three detailed coefficients. This is performed to capture specific frequency bands that possess information on the cortical activity corresponding to the speech imagery. Hence, there are 20 temporal domain features and another 20 wavelet domain features adding up to feature vectors of dimension 40.

IV. DETAILS OF THE DNN CLASSIFIER

A DNN with two hidden layers is used as the primary classifier. Since the dimension of the feature vector is 40, the number of neurons in the input layer is also 40. Each dense hidden layer consists of 40 neurons. Also, dropout and batch normalization layers are added after each dense layer. The dropout ratio is 10% for the two hidden layers. The activation function of all the layers except the first hidden later is the rectified linear unit. The activation function of the first hidden layer is hyperbolic tangent. The activation function of the output layer is *softmax*. Loss function is *categorical cross-entropy*. This DNN architecture is adopted based on cross-validation performance of several DNN architectures.

Because of the availability of very limited data, only cross-validation is performed. Since we have derived 11 feature vectors (one per each chosen channel) per trial, 11 outputs are obtained for each trial, one each for each channel. The final decision for each trial is then based on majority or hard voting of the 11 outputs.

V. RESULTS AND COMPARISON WITH THE LITERATURE

Five-fold cross-validation is performed on the pre-processed data of each participant. During cross-validation, it is ensured that all the channels corresponding to a trial are either in the training set or in the test set. This is important, since the presence of a couple of channels from the test trials in the training set can lead to high spurious accuracy due to data leakage. The cross-validation results obtained are listed in Table I, along with other results reported in the literature.

VI. CONCLUSION

The present work shows that it is possible to treat each EEG channel as an independent data vector in order to increase the size of the training set for the purpose of decoding imagined speech using deep learning techniques. The proposed method gives around 35% improvement in accuracy on an average over the state-of-the-art results.

VII. ACKNOWLEDGEMENT

The authors thank Dr. Frank Rudzicz, University of Toronto, Mr. Ciaran Cooney, Ulster University, Dr. Kanishka Sharma, Mr. Pradeep Kumar G. and Ms. Ritika Jain, Indian Institute of Science for the support extended to this work.

REFERENCES

[1] K. Brigham and B. V. Kumar, "Imagined speech classification with EEG signals for silent communication: a preliminary investigation into synthetic telepathy," in *Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference on*. IEEE, 2010, pp. 1–4.

[2] T. La Vaque, "The history of EEG Hans Berger: psychophysiologist. A historical vignette," *Journal of Neurotherapy*, vol. 3, no. 2, pp. 1–9, 1999.

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

[4] C. S. DaSalla, H. Kambara, M. Sato, and Y. Koike, "Single-trial classification of vowel speech imagery using common spatial patterns," *Neural networks*, vol. 22, no. 9, pp. 1334–1339, 2009.

[5] L. Wang, X. Zhang, X. Zhong, and Y. Zhang, "Analysis and classification of speech imagery EEG for BCI," *Biomedical signal processing and control*, vol. 8, no. 6, pp. 901–908, 2013.

[6] 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. 016002, 2017.

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

[8] 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.

[9] C. Cooney, R. Folli, and D. Coyle, "Mel frequency cepstral coefficients enhance imagined speech decoding accuracy from EEG," in *2018 29th Irish Signals and Systems Conference (ISSC)*. IEEE, 2018, pp. 1–7.

[10] J. T. Panachakel, A. G. Ramakrishnan, and T. V. Ananthapadmanabha, "A novel deep learning architecture for decoding imagined speech from eeg," in *Proceedings of the IEEE Austria International Biomedical Engineering Conference (AIBEC2019)*. IEEE, 2019.

[11] G. Chatrian, E. Lettich, and P. Nelson, "Ten percent electrode system for topographic studies of spontaneous and evoked eeg activities," *American Journal of EEG technology*, vol. 25, no. 2, pp. 83–92, 1985.

[12] G. Gómez-Herrero, W. De Clercq, H. Anwar, O. Kara, K. Egiazarian, S. Van Huffel, and W. Van Paesschen, "Automatic removal of ocular artifacts in the EEG without an EOG reference channel," in *Proceedings of the 7th Nordic Signal Processing Symposium-NORSIG 2006*. IEEE, 2006, pp. 130–133.

[13] A. G. Ramakrishnan and J. V. Satyanarayana, "Reconstruction of EEG from limited channel acquisition using estimated signal correlation," *Biomedical Signal Processing and Control*, vol. 27, pp. 164–173, 2016.

[14] W. D. Marslen-Wilson and L. K. Tyler, "Morphology, language and the brain: the decompositional substrate for language comprehension," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 362, no. 1481, pp. 823–836, 2007.

[15] B. Alderson-Day, S. Weis, S. McCarthy-Jones, P. Moseley, D. Smailes, and C. Fernyhough, "The brain's conversation with itself: neural substrates of dialogic inner speech," *Social cognitive and affective neuroscience*, vol. 11, no. 1, pp. 110–120, 2015.

[16] S. Raghu, N. Sriraam, G. P. Kumar, and A. S. Hegde, "A novel approach for real-time recognition of epileptic seizures using minimum variance modified fuzzy entropy," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 11, pp. 2612–2621, 2018.

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