Unleash Your Psyche : A Systematic Review on Imagined Speech Recognition

Aadarsh Gupta, Ashwani Pandey
Department of Electrical Engineering, Indian Institute of Technology, Delhi

***Abstract*— The recent explosion of speech-driven interfaces has had a significant impact on users' lifestyles, however there are still some constraints to the utility of such technology for the mute and those with speaking disabilities due to the requisite speech and physical gestures. The use of imagined speech using electroencephalographic (EEG) signals is an alluring domain of brain-computer interfaces (BCI) as it enables effective communication strategies for controlling devices using speech directives extracted from brain signals. This systematic review examines viable datasets and architectures for the purpose of finding opportunities for various imaging modalities to establish directions for recognition of speech from the brain signals. This study investigates novel EEG data processing techniques, examines three widely used databases: KARA ONE, BD1, and BD2 and evaluates effective modalities based on Convolutional Neural Networks (CNNs), LSTMs, and various deep learning architectures on a range of different of speech recognition tasks. Eventually, our work reinforces the validity of a brain imagery footprint that may be used to aid in the decoding of imagined speech in a range of real-world applications and devices.**

***Index Terms*— Brain-Computer Interfaces (BCI), Electroencephalographic (EEG) signals, Convolutional Neural Networks (CNNs), Deep Learning**

I. Introduction

Speech-driven interfaces have gained widespread acceptance and are currently used by a multitude of individuals in a vast variety of real-world applications and devices. These speech interfaces allow for natural interaction with electronic devices and enable fast input of texts [2]. However, the assumption of being able to generate coherent speech has a slew of drawbacks. Most of the current research is based on motor imagery-based control of external devices, which uses imagined hand, arm, or foot movements to deliver directional commands [4,5]. Daily human interaction is marked by effective verbal and nonverbal communication in the form of vocal speech (or sounds) and physical gestures. The corresponding requirement of higher functionalities and degrees of freedom restrict the possibilities of interaction for people with speaking disabilities, neuro-muscular disorders, and diseases as it involves intricate muscular hydrostat structure

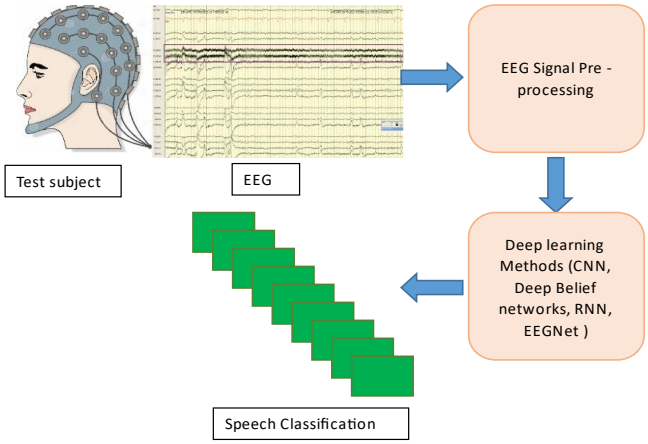


Figure 1: *General Framework for imagined speech recognition from EEG signals*

movement and vocal space involving labial, lingual, naso-pharyngeal and jaw motion [1].

Imagined speech is the internal pronunciation of phonemes, words, or sentences, without the movement of the phonatory apparatus or any audible output. The initial work showed that unspoken speech produces certain brain activities. However, studies hypothesized that the results are overestimated due to temporal correlated artifacts. Porbadnigk et al. showed that unspoken speech can be recognised using various deep learning mechanisms.

Consequently, the need for interfaces based on imagined speech would be beneficial not just to the mute but also to the disabled persons (i.e., locked-in syndrome) and can serve other applications of general interest. It was also proven that imagining a word/vowel has effect on EEG alpha, beta and theta bands. Initial work on imagined speech recognition focuses on identifying vowels ‘a’, ‘e’, ‘i’, ‘o’ and ‘u’ instead of entire word or sentence because they convey maximum variation in vocal articulation [16] and hypothesized to show similar variance in EEG waveform.

Speech-related Brain Computer Interface (BCI), also referred to as human-machine interfaces, are systems that use brain signals to control computers or hardware devices.

Aadarsh Gupta is with the Indian Institute of Technology, Delhi 110016 India (e-mail: ee1190451@iitd.ac.in).

Ashwani Pandey is with the Indian Institute of Technology, Delhi 110016 India (e-mail: ee1190451@iitd.ac.in).

Therefore, such systems can equip the users with a medium to communicate and express their thoughts, thereby improving the quality of rehabilitation and clinical neurology. This technology is an enticing facet for the future because of the vast range of deployment and prospective uses ranging from motor and cognitive rehabilitation, assistance in the recovery of compromised communication and/or physical skills control of video games, augmentative assistance platforms and specialized communications, such as non-verbal interaction between army units for confidential message transfers. The number of alphanumeric characters under consideration can also be increased to check feasibility of making a device that can decode all 36 major alphanumeric characters.

In this paper we will undertake a review of existing research and a few innovative methods which are taking this domain towards realization. In the initial part we investigate the datasets which are available for the purpose, the experimental setup for collecting data will include subject preparation, stimulus design & EEG recording method. This work will serve as an analysis to identify the prospects of various imaging modalities to devise directions for identifying speech from the imagination of the human brain.

II. BACKGROUND & MATERIALS

There are different methodologies for the non-invasive capturing of brain signals such as magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI) and electroencephalography (EEG). The advantages that EEG has over the other methods are its low cost, portability, and high time resolution. The stages of a BCI processing system with EEG are: Experimental Design, Signal acquisition, Pre-processing, Feature extraction, Classification and Device control.

A. Kara One Database

The database combines three modalities i.e. EEG, face tracking & audio during imagined and vocalized phonemic and single-word prompts. The study was conducted in an office environment and an appropriately sized EEG cap was placed on the participant's head and a small amount of gel was applied to improve electrical conductance. A 64- channel Neuroscan Quick-cap was used where the electrode placement follows the standard 10-20 system. Over the duration of 30 to 40 minutes, individual prompts appeared on the screen one-at-a-time. 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) were used. These prompts were chosen to maintain a relatively even number of nasals, plosives, and vowels, as well as voiced and unvoiced phonemes. Each trial consisted of 4 successive states:

a) A 5-second rest state, where the participant was instructed to relax and clear their mind of any thoughts.
b) A stimulus state, where the prompt text would appear on the screen and its associated auditory utterance was played over the computer speakers. This was followed by a 2-second period in which the participant moved their articulators into position to begin pronouncing the prompt.

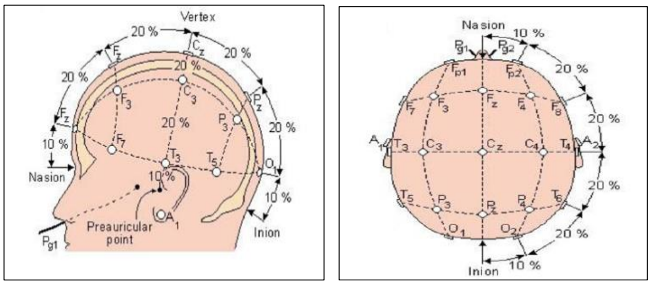


Figure 2: *International 10 – 20 system of Electrode placement by the American Electroencephalographic Society*

c) A 5-second imagined speech state, in which the participant imagined speaking the prompt without moving.
d) A speaking state, in which the participant spoke the prompt aloud. The Kinect sensor recorded both the audio and facial features during this stage.

B. Reference Database (BD1)

The reference database (BD1) [17] is an openly accessible database of EEG signals, developed by Coretto et al. which records imagined speech tasks with 5 vowels and 4 words. The experimental protocol for this database consisted in asking each subject to sit on a chair one meter away from an LCD screen.

Once seated, they were shown a message on the screen for two seconds warning them to get ready. Then, they were shown the vowel they had to imagine for two seconds. Next, they imagined the vowel continuously for four seconds.

Finally, they were shown a message on the screen indicating them to rest for four seconds. This procedure was repeated 40 times for each imagined vowel. In this database, the signals were recorded with an 18-electrode Grass device at a sampling frequency of 1024 Hz. The EEG electrodes were located according to the international 10–20 system and the database contains information from six electrodes F3, F4, C3, C5, P3, and P4.

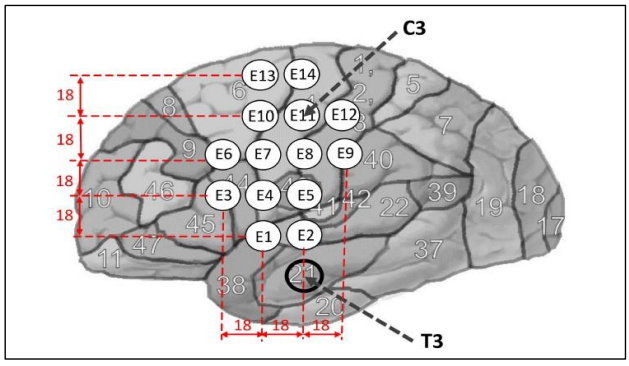


Figure 3: *Location of the neuroheadset, which contains 14 electrodes (E1, . . . ,E14) covering a section of the left hemisphere (language area). This is, the sensorimotor interface area and articulatory network of Hickok and Poeppel (Broca's area and motor cortex) related to Brodmann areas: 4, 6, 43, 44 and 45. C3 and T3 are reference points from the 10–20 positioning system*

Table 1: *Time intervals for the imagined vowels experiment.*

0	3	7	10	14
Relax	Imagined Vowel	Relax	Imagined Vowel	Relax...

C. *New Database (BD2)*

This relatively new database was created specifically for the study, held the information of 50 university students (20 women and 30 men) whose native language was Spanish (M = 24.76, SD = 7.66). The participants did not exhibit any medical or neurological conditions. First, each subject was asked to sit on a comfortable chair and an EEG neuroheadset was placed on their heads. The neuroheadset has 14 electrodes located on the left hemisphere, covering the language area. Two reference electrodes were located on the forehead. The electrodes were placed according to Hickok and Poeppel's neurological model of language related to the sensorimotor interface and articulatory network (Broca's area and motor cortex) related to Brodmann areas: 4, 6, 43, 44 and 45. The electrodes were placed on the neuroheadset in a matrix-like structure where the rows and columns of electrodes, were 18 mm apart.

To reference the neuroheadset on the head of each subject, the T3 and C3 positions were used according to the 10–20 system. Once the headset was secured, a light source, placed at one meter from the subject, was lit to indicate the moment when they should start or finish the task of thinking about a specific vowel with imagined speech.

To decrease blinking and eye movement artifacts, subjects were asked to keep their eyes closed. For the experiment, each subject was told to imagine a given vowel continuously and without pronouncing it while the light source was on. They were also told that, when the light source was turned off, they had to stop imagining the vowel and relax their body. During the experiment, the light source remained on for four seconds and then was turned off for three seconds. The procedure was repeated 25 times for each one of the imagined vowels. Upon completion of the 25 imagined speech tasks for each vowel, subjects rested for 5 min to continue with the next vowel. The imagined tasks were arranged in the following order as given in Table 1.

Each pre-processing block is mainly composed of a filtering stage using Adaptive-Projection Intrinsically Transformed MEMD (APIT-MEMD) and a signal transformation stage using spectral analysis. The signals were divided in trials with 64 samples and an overlap of 85%.

III. RELATED WORK

This work presents and evaluates two recent works in the domain, each of which provides comprehensive study of superlative architectures for the purpose.

A. *Speak Your Mind!* [1]

The paper attempts to detect speech tokens from speech

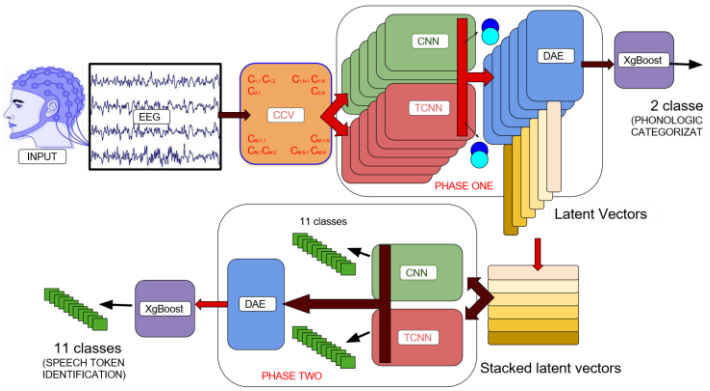


Figure 4: *Framework of proposed approach: CNN+TCNN+DAE*

imagery based on KARA ONE database [13] and achieves an accuracy of 83.42% across six different binary phonological classification and 53.36% for the individual token identification task. Electroencephalography (EEG) has been shown to be a promising signal for recognizing various brain activities among the numerous brain activity-monitoring modalities for use in BCI [6, 7]. On the other hand, these are multidimensional, with a low Signal-to-Noise ratio, low spatial resolution, and numerous artefacts. Furthermore, decoding the desired information from the high-dimensional raw EEG signals is not entirely evident.

The dimensionality of the input data was reduced by capturing joint variability of electrodes, given that raw multi-channel high-dimensional EEG data demands longer training times and resources. In contrast to the traditional technique of sampling a few channels [8,9], the study computes the channel cross-covariance (CCV), which results in positive, semi-definite matrices reflecting the electrode connection. CCV between two electrodes c_1 and c_2 is defined as:

$$Cov\left(X_t^{c_1}, X_{t+\tau}^{c_2}\right) = \mathbb{E}[X^{c_1}(t) - \mu x^{c_1}(t)][X^{c_1}(t+\tau) - \mu x^{c_1}(t+\tau)] \quad (1)$$

A 4-layered 2D Convolution Neural Networks (CNNs) [10] has been used to extract spatial features from the covariance matrix to predict phononical categories, while 6-layered temporal CNN (TCNN) [11, 12] has been used in parallel on the channel covariance matrices to investigate long-term dependencies and temporal correlations of the signal. To lower the dimensionality of the spatial temporal recordings and remove background noise effects, an unsupervised deep autoencoder (DAE) [15] on the fused heterogeneous features

Table 2: *Comparison of accuracy for 10% test data for speech token prediction task*

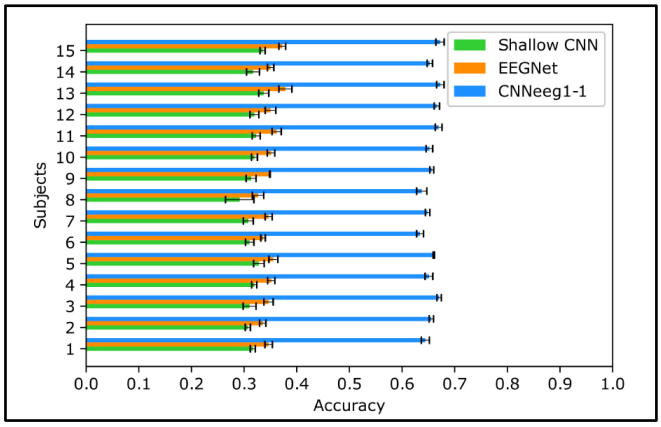
Method	**EEG Data**	**Phonological Features**
LSTM	8.45	15.83
CNN	8.88	16.02
CNN+LSTM	12.44	22.10
CNN+LSTM+DAE	23.45	49.19
Proposed model	**28.08**	**53.36**

obtained by the CNN and TCNN was employed. The bottleneck characteristics of the autoencoders are stacked into a 6x256 matrix that helps predict 11 unique speech tokens in the EEG dataset. The two baselines used are based on an individual LSTM and individual CNN architectures and difference in accuracy are tabulated in Table 2.

B. CNN architecture for CNNeeg1-1 Net

The input layer for each CNN receives the information of the images obtained from the EEG imagined vowels. It consists of a tensor of size 32 x 15 x 1 for database BD1 and 32 x 91 x 1 for database BD2 . Next comes the dropout layer with 0.25 probability the aim being minimizing the overfitting in the training process. Layer 3 is a 2D convolutional layer that applies a sliding convolution filter on the input. For this layer, 50 filters are configured with a size of 5 x 5, a stride of 1, and a padding of 0; thus, the output has a size of 28 x 57x 50.

Figure 5: *Subject wise accuracy of CNNeeg -1 Net on BD1 database. Evaluated mAP for BD1 is 0.6562*



Note: The evaluated mean and standard deviation accuracy of the models on BD1 are:

a) Shallow-CNN: $\mu = 0.3171, \sigma = 0.0114$
b) EEGNet: $\mu = 0.3506, \sigma = 0.0133$
c) CNNeeg1-1: $\mu = 0.6562, \sigma = 0.0123$

Figure 6: *Comparison of estimated marginal means of Shallow CNN, EEGNet & CNNeeg-1 on BD1 and BD2 datasets.*

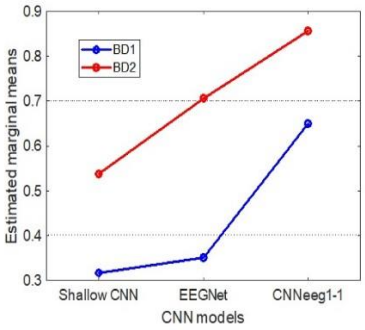


Table 3: *Machine learning classifiers used for imagined vowel recognition*

Classifiers	Accuracy	Subjects	No of Electrodes
SVM-G	77%	5	19
RVM-G	79%	5	19
RVM-L	50%	5	19
Bipolar Neural Network	44%	13	19
SVM	21.94%	15	6
Random Forest (RF)	22.72%	15	6

Table 4: *Classifiers with Deep Learning Networks used for imagined vowel recognition*

Architecture	Accuracy	Subjects	No of Electrodes
Deep Belief Networks	80%	6	19
Deep Belief Networks	87.96%	3	32
RNN	70%	6	19
CNN	32.75%	15	6
EEGNet	30.08%	15	6

C. Imagined Vowels Using Deep Learning Methods [16]

Classifiers that have been used for imagined vowel recognition ‘a’,’e’,’i’,’o’ & ‘u’ are summarized in Table 3. Deep Learning architectures that have been utilized for vowel recognition with EEG are summarized in Table 4.

IV. DISCUSSIONS

In this work, we present a schematic review for superlative algorithms for the purpose and evaluate their performance of popular databases available. The proposed model [1] comprising of CNN+TCNN+DAE outperforms baselines on KARA ONE database for the speech token prediction task, when cross-covariance (CCV) of the EEG signals is evaluated. The architecture CNNeeg1-1 based on deep learning architectures for EEG imagined vowel signal recognition using two different databases: BD1, with 15 subjects and BD2, with 50 subjects. Statistical results were presented with a mixed analysis of variance of repeated measures for intra-subject and inter-subject training. The results show that CNNeeg1-1 outperforms both Shallow CNN and EEGNet for EEG imagined vowel classification in intra-subject and inter-subject training analysis with both databases. Thus, the work shows evidence for possibility of classifying imagined vowel with the novel CNNeeg1-1 algorithm.

REFERENCES

[1] Saha, P., Abdul-Mageed, M. and Fels, S., 2022. *SPEAK YOUR MIND! Towards Imagined Speech Recognition With Hierarchical Deep Learning*. [online] arXiv.org. Available at: <https://arxiv.org/abs/1904.05746> [Accessed 15 April 2022].

[2] Herff, C. and Schultz, T., 2022. *Automatic Speech Recognition from Neural Signals: A Focused Review*.

[3] Bakhshali, M., Khademi, M., Ebrahimi-Moghadam, A. and Moghimi, S., 2022. *EEG signal classification of imagined speech based on Riemannian distance of correntropy spectral density*.

[4] G. Pfurtscheller and C. Neuper, "Motor imagery and direct braincomputer communication," Proceedings of the IEEE, vol. 89, no. 7, pp. 1123–1134, 2001.

[5] P. Herman, G. Prasad, T. M. McGinnity, and D. Coyle, "Comparative analysis of spectral approaches to feature extraction for eeg-based motor imagery classification," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 16, no. 4, pp. 317–326, 2008.

[6] S. Machado, F. Araujo, F. Paes, B. Velasques, M. Cunha, ´ H. Budde, L. F. Basile, R. Anghinah, O. Arias-Carrion, M. Cagy ´ et al., "Eeg-based brain-computer interfaces: an overview of basic concepts and clinical applications in neurorehabilitation," Reviews in the Neurosciences, vol. 21, no. 6, pp. 451–468, 2010.

[7] F. Lotte, M. Congedo, A. Lecuyer, F. Lamarche, and B. Ar- ´ naldi, "A review of classification algorithms for eeg-based brain– computer interfaces," Journal of neural engineering, vol. 4, no. 2, p. R1, 2007.

[8] S. Zhao and F. Rudzicz, "Classifying phonological categories in imagined and articulated speech," in ICASSP, 2015. IEEE, 2015, pp. 992–996.

[9] P. Sun and J. Qin, "Neural networks based eeg-speech models," arXiv:1612.05369, 2016.

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

[11] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio." in SSW, 2016, p. 125.

[12] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv preprint arXiv:1803.01271, 2018.

[13] S. Zhao and F. Rudzicz, "Classifying phonological categories in imagined and articulated speech," in ICASSP, 2015. IEEE, 2015, pp. 992–996

[14] X. Zhang, L. Yao, Q. Z. Sheng, S. S. Kanhere, T. Gu, and D. Zhang, "Converting your thoughts to texts: Enabling brain typing via deep feature learning of eeg signals," in 2018 PerCom. IEEE, 2018, pp. 1–10.

[15] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT Press, 2016.

[16] R. A. Mitchell and A. Shaw, "Vowel recognition with time-delay neural network," IEEE International Conference on Systems Engineering, pp. 637– 640, 1990.

[17] Fich.unl.edu.ar. 2022. [online] Available at: <http://fich.unl.edu.ar/sinc/downloads/imagined_speech/> [Accessed 15 April 2022].