Journal of Scientific & Industrial Research
Vol. 82, July 2023, pp. 756-766
DOI: 10.56042/jsir.v82i07.2899

Phoneme-based Imagined Vowel Identification from Electroencephalographic Sub-Band Oscillations during Speech Imagery Procedures

Anandha Sree Retnapandian* & Kavitha Anandan

Centre for Healthcare Technologies, Department of Biomedical Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai 603 110 Tamil Nadu, India

Received 09 October 2021; revised 11 May 2023; accepted 19 June 2023

Speech Imagery (SI) corresponds to imagining speaking an intended speech or a segment of speech. Decoding the SI process aids in building speech-based neural prosthetic devices. Though SI-based research has been carried out to decode imagined speech for more than a decade, there is a lag in achieving the naturalness of the spoken language. This is because the words are built as the combination of phonemes in any natural language, but the research so far has been involving the SI of vowels only. Hence, this work focuses on identifying the vowels from EEG signals acquired while imagining the corresponding phonemes. The acquisition process was repeated for multiple trials. The EEG signals were decomposed into five sub-band frequencies to analyze the activity during SI tasks. The energy coefficients extracted from the sub-band frequencies were employed in training the Recurrent Neural Network to classify the English vowels. Further, to emphasize the importance of training the classifier with multi-trial data, the results were compared with that of the single-trial data acquired from the same set of participants, and an accuracy of 84.5% and 88.9% were achieved for single and multi-trial protocols, respectively. The analysis using multi-trial data was able to achieve 4.4% higher accuracy when compared to single-trial data. Higher activations in the theta band during the speech imagery tasks and higher Classification accuracy while applying theta band features show the capability of using the theta band features in imagined speech decoding tasks.

Keywords: Electroencephalography, Imagined vowel identification, Phoneme, Recurrent neural network, Speech imagery

Introduction

Speech imagery is the act of visualizing speech without articulation. It is a type of mental imagery that uses covet speech. Electroencephalography (EEG) has always been effective in obtaining tangible information from brain activity with a potential for capturing neuronal activation during imagining tasks.[1] By training the participant using mental imagery processes, neural pathways for speech and cognition can be restored.[2] In speech-impaired individuals, identification of the intended speech is the basic requirement of any brain-computer interaction system.[3]

Decoding imagined speech with neural decoders using embedded electrodes on the cerebral cortex was reported by Guenther *et al.*[4] Later, the need for using the non-invasive technique to decode imagined speech arouse and the binary classification of the imagined vowels and phonemes was addressed.[5] As speech is a higher cognitive behavior of the brain, it generally involves the functioning of a wide range of cortical areas. Speech is mainly produced, processed, and comprehended in the left hemispheric brain regions namely the Broca's area (posterior-inferior frontal gyrus), Wernicke's area (posterior-superior temporal lobe), and the laSTG (left anterior superior temporal gyrus).[6] Hence, analysis of the brain activity in the above-mentioned brain regions during SI tasks will likely reveal the neural signatures related to the intended speech. This can provide a basis for building BCI systems to support people with neuronal speech disorders, with people who can process and comprehend speech but cannot produce proper or continuous speech. Most of the existing neural prosthesis systems are the "P300-speller" type.[7] More robust systems to detect the user thoughts can be developed by including the brain responses associated with SI.[8]

Imagined vowel decoding has been implemented by Min *et al.*[9] in 2016, using Ensemble Learning Machine-based binary classification, and overall classification accuracy of 70% was reported. Classification of the phonemes into five groups based on their phonological categories has also been reported by Zao *et al.* in 2015.[10] The same dataset was further explored and a slightly higher

*Author for Correspondence
E-mail: anandha.sree@gmail.com

classification accuracy of 74% was reported by employing Deep Belief Network.[11] A maximum accuracy of 87% was reported for all the phonemes in the dataset while using Convolutional Neural Network and Denoising Auto-Encoder[12] and a higher classification accuracy of 90% was reported while using DenseNet.[13] Three imagined action words were classified using Relevance Vector Machine and a classification accuracy of 70% and 95% was reported for multi-class and binary classifications, respectively.[14] A multi-class classification of five action words in the Spanish language was performed using Random Forest and wand the results showed an average classification accuracy of 58%.[15] Imagined /a/, /u/, and /rest/ classes were classified using Deep Capsule Neural Network, and an average accuracy of 93.32% was reported.[16] A Capsule Neural Network to categorize SI patterns /iy/ and /uw/ on the Karaone dataset was reported.[17] An improved classification accuracy on the classification of Chinese characters 'left' and 'one' using a Light GBM-based classification algorithm was reported.[18]

In our earlier studies, brain connectivity parameters have been estimated from the imagined (without articulation), spoken and unspoken (with articulation) Consonant-Vowel (CV) pairs, and a detailed region-specific analysis has been performed. Predominant activation of left frontal and temporal regions during speech imagery-related tasks was observed.[19] Statistical features were derived from the brain signals of left frontal and temporal regions acquired during speech imagery of vowels and the vowels were classified using multi-class Deep Belief Networks (DBN).[20] Inter-hemispheric, as well as intra-hemispheric analyses, were performed for protocols involving speaking and imagining speaking of Consonant-Vowel-Consonant (CVC) words. In this process, the brain connectivity parameters were extracted from the sub-band frequencies of the EEG signals.[21] The vowels were classified from the imagined CVC words using DBN and Recurrent Neural Networks (RNN).[22] RNN is a deep learning technique that is more adaptive for the classification of time-series data.[23,24] The imagined Vowels were identified from single-trial SI-based tasks involving imagining of vowels, consonant-vowel syllables, and consonant-vowel-consonant words. Even though these approaches were able to take the process close to the identification of words from imagined speech, they lag in naturalness since words are not just combinations of vowels and consonants but instead the combination of their corresponding phonemes.

Therefore, in this work, imagined vowels using the EEG sub-band energy coefficients and recurrent neural networks have been identified. English vowels have been classified from the EEG signals acquired during multi-trial SI tasks involving imagining speaking the phonemes of vowels. Reliability analysis was carried out and furthermore, energy coefficients and relative power of each sub-band were used to train the RNN for retrieving the imagined vowels. Since words are built as a combination of phonemes and are the building blocks of sentences in natural speech, the process of decoding imagined speech through their phonemes seems meaningful in the Brain-computer interaction platforms. This can be implemented in building prosthetic devices for people with speech production disorders.

Materials and Methods

A 14-channel EEG signal acquisition system was used to acquire the data during Speech Imagery (SI) task. The raw EEG signals acquired while performing the task were pre-processed using conventional pre-processing routines to remove the noise. The activations irrelevant to the task were removed using Independent Component Analysis (ICA). The noise and artifact-free signals were decomposed into sub-band frequencies using Discrete Wavelet Transform (DWT). Features from each sub-band frequency were estimated to identify the imagined vowel using Machine Learning Technique. The pipeline of the techniques employed in this study is shown in Fig. 1. The aim of BCI applications along with higher classification accuracy is to minimize the complexity of holding the signal acquisition device during the daily usage of the BCIs. Hence, this work aimed at

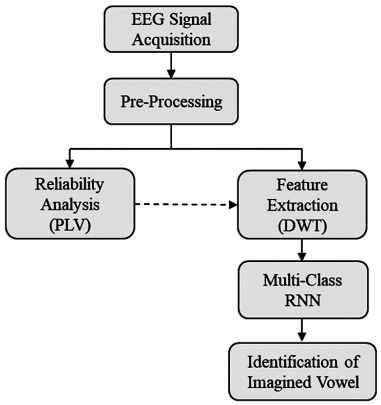

Fig. 1 — Pipeline of the methodology (an overview)

identifying the imagined vowels with a minimal set of electrodes, instead of using a high-density electrode system for EEG acquisition.

Data Acquisition and Pre-processing

Five male volunteers participated in the process of data collection. Only native right-handed Tamil (a south Indian language) speakers who had no prior history of neurological disorders took part in the study. The experiment was carried out in a soundproof room at the institution. The experimental protocol and the difference between speech and speech imagery-based tasks were explained clearly to all the participants ahead of the signal acquisition.

Wireless Emotiv EPOC+ Neuro-technology EEG acquisition system was used for recording the EEG signals. The EMOTIV EPOC+ is a portable 14-channel EEG system with two reference electrodes placed at the left and right mastoid labeled as CMS and DRL, respectively. Emotiv EPOC+ has been used as an EEG acquisition system in various human-computer interactive applications.[22] In Emotiv, built-in digital filters facilitate filtering during data acquisition with a resolution of 0.51μV. The signals are sampled at the rate of 128 samples per second and the impedance is maintained below 10kΩ using real-time contact quality measures supported by the Emotiv software. As Emotiv is an easy-to-use device and comes with user-friendly software support, initial assistance alone was obtained from a technician specialized in EEG device electrode positioning and data acquisition.

The experimental protocol involves a visual stimulus of phonemes of vowels shown to the participants using a display placed at a 1-meter distance. Though native Tamil speakers will have a different dialect and accent in speaking English, the commonly used sound forms which are more common for the accent of native Tamil speakers, as well as English speakers, have been chosen for the procedure. For instance '/a/' as in apple, '/e/' as in egg, '/i/' as in ink, '/o/' as in ostrich, and '/u/' as in umbrella have been chosen for data acquisition. The experimental protocol for data acquisition of the phoneme of vowel /i/ for subject #3 is shown in Fig. 2 and the same protocol was repeated for all the subjects for all the vowels. The experimental protocol used for this study has been verified and approved by the Institutional Human Ethical Committee (Ref. No.: IHEC/SSN CE/Pr. No. 02/26.10.2018).

The acquired signals were segregated into the rest, visual stimulus, and SI task states based on the occurrence of the events of the protocol. The phoneme-based vowel SI protocol consists of the following states:

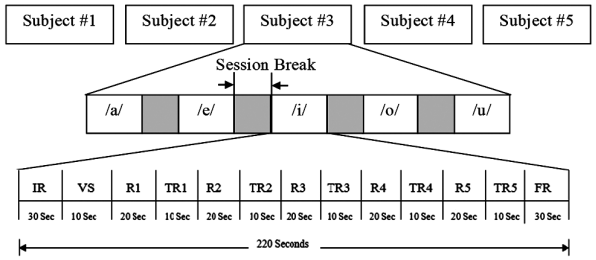

Fig. 2 — Experimental protocol for multi-trial phoneme-based vowel speech imagery; IR: Initial Rest, FR: Final Rest, R1-R5: Inter task Rests, TR #1-5: Trials 1-5; VS: Visual Stimuli

- **Rest:** a mentally and physically inactive state to get prepared for the following SI stage.
- **Visual Stimulus (VS):** a visual representation of the phoneme of one vowel at a time for 10 seconds.
- **Speech Imagery (Tr #1-5):** imagined saying the phoneme of the vowel shown in the visual stimulus without actual articulation.

The process was repeated five times with intermittent session breaks and the corresponding signals were acquired. The channels corresponding to the Frontal, Temporal, and Parietal regions were chosen for the analysis. This is mainly because the frontal and temporal regions are highly activated during the SI process. The frontal and temporal regions of the left and right hemispheric brain are active during speech production and comprehension tasks.[25] The brain wave rhythms are categorized into bands of different frequency ranges. Since the valuable information in the EEG signals is packed within the 0 to 40 Hz frequency range, the higher frequencies were considered noise. The acquired signals were band-passed between 0.1 Hz and 40 Hz using a Butterworth band-pass filter of order 8.[(26)] Filtered signals were baseline shifted by subtracting the mean of each signal and the mean-shifted signals were normalized using the min-max normalization technique. The normalized EEG signals were smoothened by dividing them into epochs using Hamming window.[27] After filtering and normalizing, the signals were free from sensor noise and line interference that commonly contaminate the EEG recordings.[28] ICA was implemented to extract multiple functionally independent sources of activation generated by specific cortical regions of the brain using EEGLAB as mentioned by Sandhya *et al*.[29] The MATLAB R2019b was used for signal processing and classification.

Reliability Analysis

Reliability analysis was carried out to confirm that the acquired signals are purely confined to the Speech Imagery based tasks as well as rest. It has been reported in the literature that the alpha band relates to rest[30] and the theta band relates to mental activities.[31] Hence, the information from the alpha and theta bands was segregated by applying Empirical Mode Decomposition (EMD). EMD is a self-adaptive data-driven method is used as a time-space filter that generates localized time-frequency components by decomposing the signal into a number of smaller oscillating components called IMFs and residues. After decomposition, the IMFs were arranged in decreasing order of the frequency components. The segmented signals were then subjected to Hilbert-Huang Transform (HHT) from which the Phase Locking Value (PLV) was estimated for the corresponding Intrinsic Mode Functions (IMFs).[32] To quantify the zero crossings of the decomposed signal, the Hilbert-Huang transform was applied. HHT computes the instantaneous amplitude as well as the instantaneous phase of a signal.[33] Instantaneous phase (φ) of the signal was used for computing the Phase Locking Value (PLV). PLV is a synchronization measure that is used to analyze the phase difference between two brain signals. Here, the PLV has been calculated as given in Eq. (1).

$$PLV(t) = \left|E[e^{j\Delta\varphi_{xy}(t)}]\right| \quad \ldots(1)$$

where, $\Delta\varphi_{xy}(t) = \varphi_x(t) \sim \varphi_y(t)$, *E[]* is the statistical Expectation, and $\varphi_x(t)$, $\varphi_y(t)$ are the instantaneous phases of two time-varying signals.

Sub-band Energy Coefficients Extraction

For non-stationary signals such as EEG, a time-frequency method like Discrete Wavelet Transform (DWT) is the most appropriate method for feature extraction.[34] It has been proven that, amongst the various DWT techniques, Daubechies (Db) wavelet is more suitable for EEG-based imagery applications. Therefore, Daubechies (Db) wavelet has been applied for extracting the time domain features. A series of low-pass and high-pass filters were applied to the EEG signal to obtain the approximation coefficients and detail coefficients respectively.[35] For a sampling frequency of 128 Hz, EEG sub-bands frequencies were obtained at the fifth level decomposition. The Daubechies wavelet was used to decompose the signals acquired from the chosen Frontal, Temporal and Parietal electrodes ('F7', 'F3', 'T7', 'P7', 'P8', 'T8', 'F4' and 'F8') and extract the required features. The energy coefficients Root Mean Square (RMS), Mean Absolute Value (MAV), Integrated EEG (IEEG), Simple Square Integral (SSI), Variance of EEG (VAR), Average Amplitude Change (AAC) and the Relative Power (RP) are estimated as given in Eqs (2– 8).

$$RMS_i = \sqrt{\frac{1}{N}\sum_{n=1}^{N} D_i^2(n)} \quad \ldots(2)$$

$$MAV_i = \frac{1}{N}\sum_{n=1}^{N} |D_i(n)| \quad \ldots(3)$$

$$IEEG_i = \sum_{n=1}^{N} |D_i(n)| \quad \ldots(4)$$

$$SSI_i = \sum_{n=1}^{N} |D_i(n)^2| \quad \ldots(5)$$

$$VAR_i = \frac{1}{N-1}\sum_{n=1}^{N} D_i^2(n) \quad \ldots(6)$$

$$AAC_i = \frac{1}{N}\sum_{n=1}^{N} |D_i(n+1) - D_i(n)| \quad \ldots(7)$$

$$RP = \frac{Individual\ Band\ Power}{Total\ Power\ of\ EEG} \quad \ldots(8)$$

where, $D_i(n)$ is the n^{th} sample of a wavelet decomposed detail coefficient at level varying from $i=1,2,3,4,5$ and N is the length of the signal. The Energy coefficients such as RMS, MAV, IEEG, SSI, VAR, AAC and Relative power of each EEG sub-band were calculated from the detailed coefficients.[36] $F^{v*t*c*b}$ number of input vectors were generated, where *v* is the number of phonemes, *t* is the number of trials, *c* is the number of channels and *b* is the number of sub-band frequencies.

Vowel Classification using RNN

A multi-class Recurrent Neural Network (RNN) was built to classify the imagined vowels into five different classes. The network hierarchically learns categories through its hidden layer architecture. RNN is chosen as it works more effectively for the classification of time-series data. Each node within each network layer represents characteristics of the feature set and together they supply a complete representation of the corresponding feature vector. Each layer was allocated with a weight directly proportional to the previous output.[37]

In the RNN, a single-time step of the input x_t is supplied to the network. The vector x_t is created by concatenating the vector w_t that represents the extracted features of the current time step, and vector h_{t-1} represents output values of the hidden layer from the previous time step as given in Eq. (9). The steps are repeated as many times as the problem demands and the output y_t is calculated from the final current

state. The estimated output is then compared to the actual output and the error is calculated. The mathematical representation of the current state is given as

$$h_t = f\left(h_{(t-1)}, x_t, w_t\right) \quad \ldots(9)$$

where, h_t is the current state vector (at time t) and h_{t-1} is the previous state vector (at time t-1). The current state vector can be rewritten as,

$$h_t = \varphi(W_h \cdot h_{t-1} + W_x \cdot x_t) \quad \ldots(10)$$

where, φ is the sigmoid activation function, given as

$$\varphi(p_i) = \frac{1}{1+e^{-p_i}} \quad \ldots(11)$$

where, p_i is the sigmoid probability for the i^{th} class. The output vector y_t is obtained as

$$y_t = W_h \cdot h_t \quad \ldots(12)$$

Here, W_x, W_h *and* W_y represent the weight matrices corresponding to the edges connecting input to current state, previous state to current state and current state to output state, respectively. Relu based random weight initialization heuristics was used to initialize the weights. The initial weights at the input layer were calculated as given in Eq. (13). The weights are equal at each layer. The weights are updated using the gradients calculated by applying the chain rule.

$$W_x = Random\ weight * \sqrt{\frac{2}{size^{(x-1)}}} \quad \ldots(13)$$

The gradient for updating W_x is calculated as,

$$\frac{\partial E_t}{\partial W_x} = \frac{\partial E_t}{\partial h_t} \cdot \frac{\partial h_t}{\partial W_x} \quad \ldots(14)$$

The gradient for updating W_h is calculated as,

$$\frac{\partial E_t}{\partial W_h} = \sum_{i=1}^{n} gradients\ calculated\ from\ states\ h_n \quad \ldots(15)$$

Similarly, the gradient for updating W_y is calculated as,

$$\frac{\partial E_t}{\partial W_y} = \frac{\partial E_t}{\partial y_t} \cdot \frac{\partial y_t}{\partial W_y} \quad \ldots(16)$$

In general, the gradient for updating weights is given as,

$$\frac{\partial y}{\partial W} = \sum_{i=t+1}^{t+N} \frac{\partial y}{\partial h_{t+N}} \cdot \frac{\partial h_{t+N}}{\partial h_i} \cdot \frac{\partial h_i}{\partial W_h} \quad \ldots(17)$$

The logistic sigmoid activation function of the RNNs hidden layer units is given in Eq. (11). The weights are adjusted as given in Eqs. (14) – (17). The truncated back propagation through time (t-BPTT) was applied to propagate error and update weights as explained in Algorithm 1.[38] The network was optimized using categorical cross entropy loss function as given in Eq. (19). The objective is to minimize the loss since, lower the loss the better the performance. The output error at time t is the difference between the actual output and the predicted output.

$$L(E_t) = (o_t - y_t)^2 \quad \ldots(18)$$

where, o_t is the actual output, y_t is the predicted output and $L()$ is the loss function. Here, the loss function used is the categorical loss entropy, which is given as,

$$L_{CE} = -\sum_{i=1}^{n} t_i \log(p_i) \quad \ldots(19)$$

where, n is the number of classes, t_i is the truth label and p_i is the sigmoid probability for the i^{th} class. p_i is estimated as given in Eq. (11).

The RNN built for this work consisted of 11 input units and 5 output units. The network learns the features from every 2 second data, windowed with 50% overlap, resulting in 9 timesteps from each signal. The RNN classifier output performance was measured in terms of following parameters: Classification Accuracy, Sensitivity, Specificity and F-score as given in Eqs. (20) – (23).

$$Classification\ Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \ldots(20)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad \ldots(21)$$

$$Specificity = \frac{TN}{TN+FP} \quad \ldots(22)$$

$$f - score = \frac{2TP}{2TP+FP+FN} \quad \ldots(23)$$

where, *TP* is true positive, *TN* is true negative, *FP* is false positive, and *FN* is false negative.

Algorithm : *t-BPTT (t1, t2) [where t1<t2<t]*
for 1 : timestep t : T do
run the RNN for one step, computing h_t (equation (9))
if t divides t1 then
unroll the network
calculate and accumulate errors across each timestep (equation (18))
roll up the network and update weights
update t to t − t2
end if
end for
t1: number of forward-pass timesteps between updates
t2: number of timesteps to which to apply BPTT

Results and Discussion

The acquired EEG signals were band passed between 0.1 to 40 Hz, normalized using min-max normalization technique, smoothened, and subjected

to ICA to identify the task relevant components using source localization principles. Preprocessed signals were segmented in to five sub-band frequencies using EMD and the PLV was estimated from the alpha and theta bands.

The averaged alpha and theta band activities during the SI task and rest of a representative subject are shown in Fig. 3. The representative topographic maps shown in Fig. 3 correspond to the eighth component of the averaged alpha and theta band frequencies. Red color represents higher activity and dark blue color represents lower activity in brain. It can be observed from the figure that, Alpha band showed a higher activity during rest and lower activity during speech imagery task. Contrarily, a higher activity was observed in the theta band during SI task with rest reporting a lesser activity. It was also observed that the theta band activity during the SI task is more in the left frontal regions. All the participants exhibited similar activations during the SI and rest tasks.

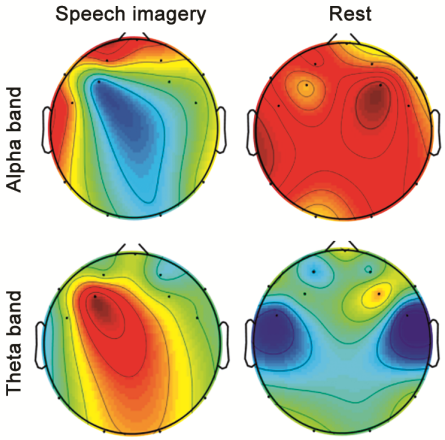

Fig. 3 — Averaged alpha and theta band activity during Task (Speech imagery of vowel 'a') and rest

Single subject's PLV strengths estimated during the speech imagery task from the alpha and theta bands are represented in Fig. 4(a) and 4(b), respectively. The PLV is a measure that estimates the level of co-ordination between the signals of each electrode. This PLV analysis revealed that, electrodes of the left hemisphere (F3, F7, T7 and P7) coordinated well with each other in theta band when compared with the right hemispheric electrodes. To expand the understanding, the band power of the two sub-band frequencies were estimated and analyzed individually. Similar PLV strength variations were observed from all the participants.

The theta band power of the electrodes considered for analysis is shown in Fig. 5. It was observed that, electrodes F3 and P7, which are nearer to the motor cortex, showed lesser theta band power during the imagery task due to lack of motor activity. Whereas the electrodes located nearer to the Wernicke's (posterior - superior temporal lobe) and Broca's (posterior - inferior frontal lobe) areas showed very

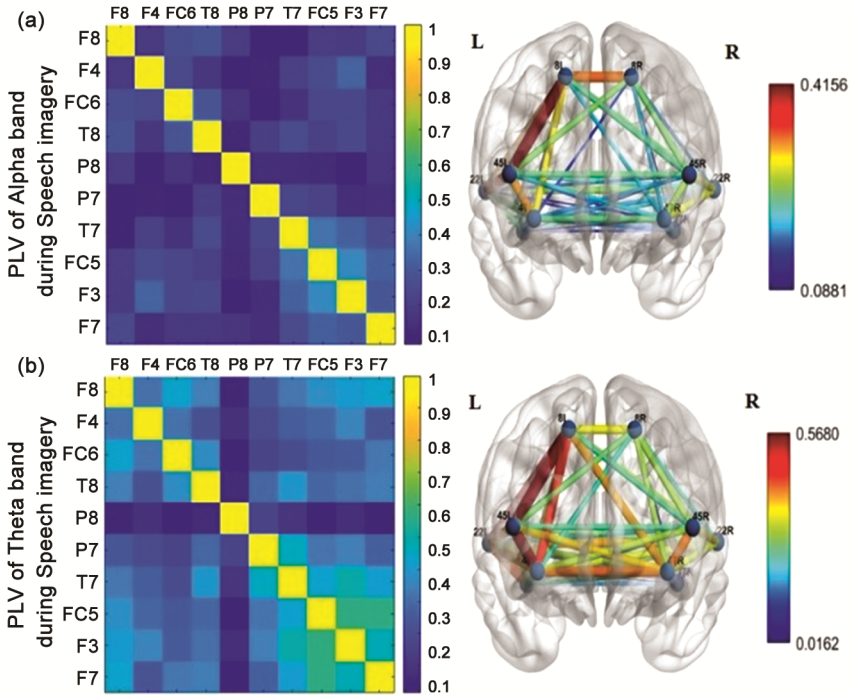

Fig. 4 — (a) PLV of Alpha band during speech imagery task of vowel 'a'; (b) PLV of Theta band during speech imagery task of vowel 'a'

high theta band power during SI task when compared with rest.

Other electrode regions showed an increased theta power with smaller differences during task and rest. Results also showed that all the participants exhibited similar activations in both alpha and theta bands during both the rest and SI tasks. The features extracted from the energy coefficients were normalized before fed to the network. The normalized features extracted from a single EEG signal are represented in Fig. 6.

The variations in Classification Accuracy of different vowels using the multi-trial features extracted from each sub-bands are shown in Fig. 7(a). The average accuracy and the standard deviation corresponding to each band have also been represented. The confusion matrices obtained for the EEG sub-bands is shown in Fig. 7(b). The higher true positive and true negative to false positive and false negative ratios of the theta and beta band proves the individual better performance of the respective bands. Even though other sub-band frequencies like beta band showed slightly higher performance measures than the theta band, the reliability analysis reveals higher activations in the theta band activations during speech imagery tasks. Hence, theta band oscillations alone can be considered for the imagined speech decoding.

The comparison of the Classification Accuracies of vowels identified from the features extracted from theta and beta bands using single trial and multiple trial protocols are shown in Fig. 8(a). Overall Classification Accuracy of 84.48% and 88.89% were observed for single-trial and multi-trial protocols, respectively. The averaged performance measures mapping (Sensitivity and Specificity) of RNN for identifying phoneme of vowels with the features extracted from the different sub-bands' for the single and multi-trial protocols are shown in Fig. 8(b). Observations showed that the Classification Accuracy (Table 1) in identifying the imagined vowels using multi-trial protocol was improved by 4.41%. A comparison with state-of-the-art works is tabulated in Table 2 to denote the superiority of the protocol implemented in this work.

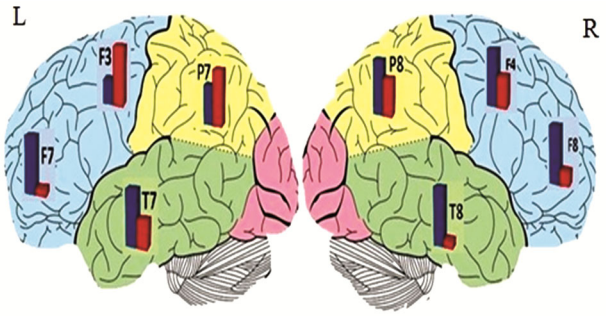

Fig. 5 — Theta band power obtained during task – blue (speech imagery of vowel 'a') and rest – red

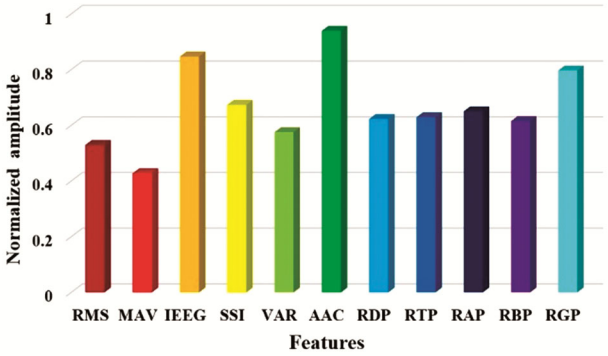

Fig. 6 — Features extracted from a single EEG Signal and normalized. Root Mean Square (RMS), Mean Absolute Value (MAV), Integrated EEG (IEEG), Simple Square Integral (SSI), Variance of EEG (VAR), Average Amplitude Change (AAC) Relative Delta Power (RDP), Relative Theta Power (RTP), Relative Alpha Power (RAP), Relative Beta Power (RBP) and Relative Gamma Power (RGP)

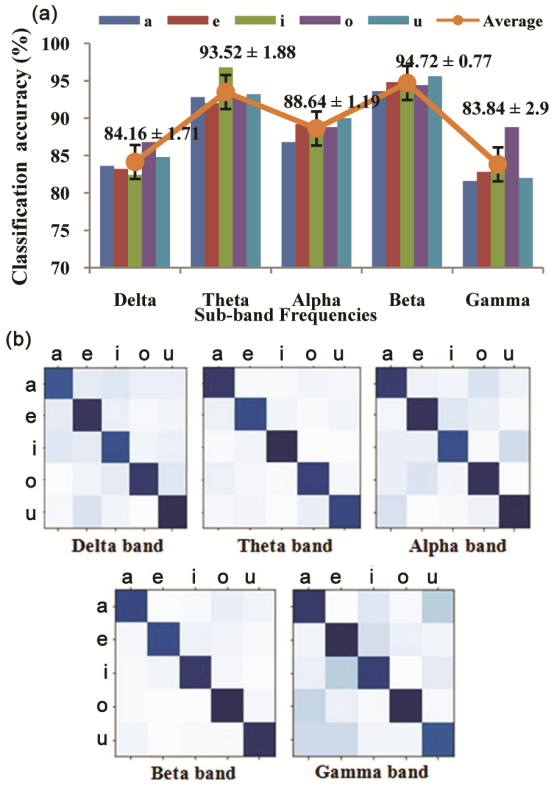

Fig. 7 — (a) Variations in classification accuracy of different vowels from sub-band frequency (x-axis) with average accuracy and standard deviation; (b) Confusion Matrices for each sub-bands

Summarizing the observations made so far, a higher theta band activity during speech imagery tasks is clearly evident. The Phase Locking Value analysis revealed higher activations in theta band during the speech imagery tasks. The band power analysis confirmed the same. Finally, the performance measures prove that the features extracted from the theta band were capable of classifying the imagined phonemes of vowels more accurately. Hence, the theta band segmented from the complete signal can be used to speech decoding processes to minimize the computational complexity.

Speech Imagery is a typical process involving hearing and articulation imagery tasks. Assessment of network functional connectivity by EEG signals proves to enhance the training in such cases. In most neuro-developmental disorders like autism, most children have the understanding about what they are taught, see, and hear. The information about those understandings is processed accurately, but the lack of

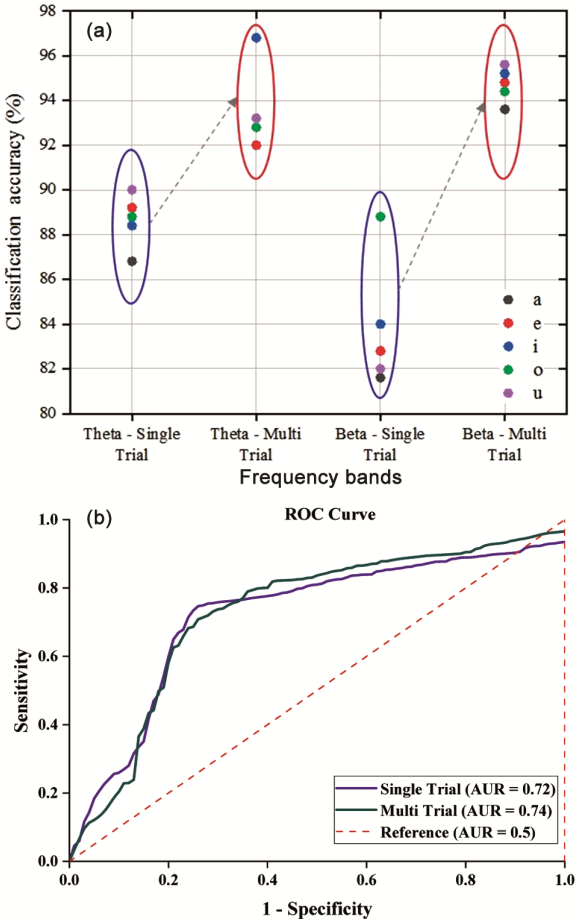

Fig. 8 — (a) Comparison of the classification accuracies between single trial and multiple trials for each vowel identified using the features extracted from the theta and beta bands; (b) Sensitivity and specificity mapping for the Multi-Trial Protocol Single-Trial Protocol

Table 1 — Performance measures, CA: Classification Accuracy, Sensitivity, Specificity and F - score, for classification of phoneme of vowels identified by Recurrent Neural Network for different sub-bands

Performance Measures		a	e	i	o	u
Delta	CA	83.6	83.2	82.4	86.8	84.8
	Sensitivity	60	57.1	56	68.8	61.1
	Specificity	88.7	90.7	89	90.73	91.3
	F-score	7.4	6.6	7.1	6.1	6.3
Theta	CA	**92.8**	**92**	**96.8**	**92.8**	**93.2**
	Sensitivity	79.3	81.2	92	82	85.1
	Specificity	96.3	94.5	98	95.5	95.7
	F-score	4.8	5.3	4.3	4.8	4.8
Alpha	CA	86.8	89.2	88.4	88.8	90
	Sensitivity	66.4	72.5	74.4	71.1	74.5
	Specificity	92.9	93.4	91.3	93.4	93.7
	F-score	5.8	5.4	5.8	5.5	5.6
Beta	CA	93.6	94.8	95.2	94.4	95.6
	Sensitivity	85.2	93.0	88	81.3	88.4
	Specificity	95.4	95.7	97	98.4	97.9
	F-score	4.7	4.5	4.5	4.6	4.5
Gamma	CA	81.6	82.8	84	88.8	82
	Sensitivity	53.7	56.1	60.2	76.9	55.2
	Specificity	89.9	90.6	89.6	91.5	88.8

Table 2 — Comparison between the current method and Other state-of-the-art works

Dataset	Classifier	Classification type	Max. %
/n/, /tiy/, /piy/, /m/, /diy/, /uw/, /iy/, gnaw, pot, and pat	DBN[10]	Binary Classification	90%
	DBN[11]	Binary Classification	74%
		Multi-class classification	41.5%
	CNN, TCNN, DAE[12]	Binary Classification	87%
	DenseNet[13]	Binary Classification	90.68%
in, out, up	Relevance Vector Machines[14]	Binary Classification	95%
		Multi-class classification	70%
up, down, left, right, and select (in spainsh)	Random Forest[15]	Multi-class classification	58.41%
/a/, /e/, /i/, /o/, and /u/	ELM, ELM-L, ELM-R, SVM-R, and LDA[9]	Binary Classification	73%
/a/, /e/, /i/, /o/, /u/	**RNN**	**Multi-class classification**	**88.9%**

information being transferred between the cortical regions may vary depending upon the level of disorder. Several connectivity parameters are available for studying the details like what and how much information is being transferred during a task. Whereas analysing the statistical changes occurring at the individual brain locations can clarify the degree of impact generated during any task.

Since speech is a highly complex physiological signal, retrieval from unspoken, imagined speech is a major goal and needs to be progressed gradually. In the various attempts to retrieve imagined speech, vowels have been identified from EEG signals acquired from various forms of speech such as silent, imagined, and loud. Vowels have also been identified from single-trial SI-based tasks involving imagining vowels, consonant-vowel syllables, and consonant-vowel-consonant words. Though these approaches tend to take the process close to speech retrieval from imagined speech, they lag in naturalness since words are combinations of phonemes of vowels. Thus, this work leading to the retrieval of vowels from the phonemes of imagined vowels is a step closer to verbal information retrieval with better naturalness. It can be noted that the protocol was designed to acquire multi-trial EEG signals for phonemes of each vowel. Speech is the higher cognitive behavior of the brain, involving a wide range of distributed network of cortical areas. Therefore, the complexity of extracting information from EEG signals while imagining speaking is higher. This includes different brain areas reacting in a different manner when compared to speaking. Thus, assimilating the brain activity during different trials seemed to be more imperative than working with a single trial data.

Speech is produced, processed, and comprehended by Wernicke’s and Broca’s areas located in the left hemispheric brain region. And the right hemisphere of the brain oversees imagination tasks. The EEG signals acquired in and around the left and right, Frontal, Parietal and Temporal regions have been processed to estimate features for vowel identification. The acquired signals subjected to various signal processing routines were band separated and band-wise reliability analysis was carried out to make sure the subject's brain activations were in sync with the protocol. The reliability analysis performed by the extracted PLV measures showed less activity in alpha band and more activity in theta band during SI task. Energy based statistical features were extracted using level Db5 Wavelet transform. It was observed that the electrodes nearer to the motor cortex, F3 and P7, showed lesser band power during the imagery task due to lack of motor activity during the imagery task. Whereas the electrodes located nearer to the Wernicke's (posterior - superior temporal lobe) and Broca's (posterior - inferior frontal lobe) showed very high band power during SI task when compared with rest. Higher activations in theta band during the speech imagery tasks and higher Classification accuracy while applying theta band features show the capability of using the theta band features alone in imagined speech decoding tasks.

The extracted features were employed in training and testing the multi-trial RNN for identification of vowels and the RNN‘s performance was evaluated. The classification accuracy for the vowels classified with the features extracted from theta and beta band were found to be higher with average classification accuracy greater than 93%. The classification accuracy of the multi-trial RNN was found to be improved by 4.41% compared to the single trial RNN. This is because of the more number of training instances that improve the classifier’s performance and the classifier’s ability in picking up the significant signatures of each class with the repeated trials. Though higher classification accuracy was observed while classification using the features from the theta and beta sub-bands, since theta band activity is coherent to the speech imagery tasks, the theta band features alone can be used to classify the imagined vowels. Extension of the work in identifying imagined words seemed to be a tough process due to the variations in phonetic representations and the complexity of the signals being handled. The substantial outcome of this process is vowel identification from EEG signals obtained during speech imagery. Though there are similar works available in the literature, this work has a significant outcome as it identifies vowels not from EEG signals of imagined vowels, but imagined phonemes of vowels. This takes the work closer to the identification of imagined words, as words are combinations of phonemes. The possibility of identifying the vowels from imagined phonemes with lesser computational complexity and better performance measures was also achieved.

Conclusions

This work can be further extended for reconstructing the interrupted or broken speech with the EEG of imagined speech, which can be a major

finding in Speech retrieval processes. The work progresses by identifying the imagined CV and CVC words from the imagined phonemes of the Corresponding Consonants and Vowels. Speech is a higher cognitive behavior of the brain. It is a complex task that induces continuous activations in the cortical regions involved in the particular task. A similar type of speech imagery signals was acquired from a subject from different geographical regions whose mother tongue is different but can speak English, in which similar types of brain activations in the speech-related regions were observed. Though literature support for such acquaintances is limited, practical implementation proves the case. Since this paper limits to the identification of imagined phonemes alone, the inclusion of other phonemes and addressing the co-articulation problems are considered for future work.

References

1 Curley W H, Forgacs P B, Voss H U, Conte M M & Schiff N D, Characterization of EEG signals revealing covert cognition in the injured brain, *Brain*, **141(5)** (2018) 1404–421.

2 Tian X & Poeppel D, Mental imagery of speech: linking motor and perceptual systems through internal simulation and estimation, *Front Hum Neurosci*, **6** (2012) 314.

3 Conti E, Calderoni S, Marchi V, Muratori F, Cioni G & Guzzetta A, The first 1000 days of the autistic brain: a systematic review of diffusion imaging studies, *Front Hum Neurosci*, **9** (2015) 159.

4 Guenther F H, Brumberg J S, Wright E J, Nieto-Castanon A, Tourville J A, Panko M, Law R, Siebert S A, Bartels J L, Andreasen D S, Ehirim P, Hui Mao & Philip RK, A wireless brain-machine interface for real-time speech synthesis, *PloS one*, **4(12)** (2009) 8218.

5 DaSalla C S, Kambara H, Sato M & Koike Y, Single-trial classification of vowel speech imagery using common spatial patterns, *Neural Netw*, **22(9)** (2009) 1334–1339.

6 Nuñez A I R, Yue Q, Pasalar S & Martin R C, The role of left vs. right superior temporal gyrus in speech perception: An fMRI-guided TMS study, *Brain Lang*, **209** (2020) 104838.

7 Guy V, Soriani M H, Bruno M, Papadopoulo T, Desnuelle C & Clerc M, Brain computer interface with the P300 speller: usability for disabled people with amyotrophic lateral sclerosis, *Annals of physical and rehabilitation medicine*, **61(1)** (2018) 5–11.

8 Uzawa S, Takiguchi T, Ariki Y & Nakagawa S, Spatiotemporal properties of magnetic fields induced by auditory speech sound imagery and perception, in *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (IEEE) 2017, 2542–2545.

9 Min B, Kim J, Park H J & Lee B, Vowel imagery decoding toward silent speech BCI using extreme learning machine with electroencephalogram, *BiMed Research International*, 2016.

10 Zhao S & Rudzicz F, Classifying phonological categories in imagined and articulated speech, in *IEEE Int Conf Acoust Speech Signal Process* (IEEE) 2015, 992–996.

11 Sun P & Qin J, Neural networks based EEG-speech models, *arXiv preprint arXiv:1612.05369 (*2016).

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

13 Islam M M & Shuvo M M H, DenseNet based speech imagery EEG signal classification using Gramian angular field, in *5th Int Conf Adv Electr Eng* (IEEE) 2019, 149–154.

14 Nguyen C H, Karavas G K & Artemiadis P, Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features, *J Neural Eng*, **15(1)** (2019) 016002.

15 González-Castañeda E F, Torres-García A A, Reyes-García C A & Villaseñor-Pineda L, Sonification and textification: Proposing methods for classifying unspoken words from EEG signals, *Biomed Signal Process Control*, **37** (2017) 82–91.

16 Ramirez-Quintana J A, Macias-Macias J M, Ramirez-Alonso G, Chacon-Murguia M I, & Corral-Martinez L F, A novel deep capsule neural network for vowel imagery patterns from EEG signals, *Biomed Signal Process Control*, **81** (2023) 104500.

17 Macías-Macías J M, Ramírez-Quintana J A, Chacón-Murguía M I, Torres-García A A, & Corral-Martínez L F, Interpretation of a deep analysis of speech imagery features extracted by a capsule neural network, *Comput Biol Med*, (2023) 106909.

18 Pan H, Li Z, Tian C, Wang L, Fu Y, Qin X, & Liu F, The LightGBM-based classification algorithm for Chinese characters speech imagery BCI system, *Cogn Neurodyn*, (2022) 1–12.

19 Sandhya C, Sree R A & Kavitha A, Analysis of speech imagery using consonant-vowel syllable speech pairs and brain connectivity estimators, in *2nd Int Conf Biomed Syst Signals Images*, February 2015.

20 Sree R A & Kavitha A, Vowel classification from imagined speech using sub-band EEG frequencies and deep belief networks, in *4th Int Conf Signal Process Commun Network* (IEEE) 2017, 1–4.

21 Sandhya C & Kavitha A, Analysis of speech imagery using brain connectivity estimators on consonant-vowel-consonant words, *Int J Biomed Eng Technol*, **30(4)** (2019) 329–343.

22 Chengaiyan S, Retnapandian A S & Anandan K, Identification of vowels in consonant–vowel–consonant words from speech imagery based EEG signals, *Cogn Neurodyn*, **14(1)** (2020) 1–19.

23 Chen J, Jiang D, Zhang Y & Zhang P, Emotion recognition from spatiotemporal EEG representations with hybrid convolutional recurrent neural networks via wearable multi-channel headset, *Comput Commun*, **154** (2020) 58—65.

24 Yu W, Kim I Y & Mechefske C, Analysis of different RNN autoencoder variants for time series classification and machine prognostics, *Mech Syst Signal Process*, **149** (2021) 107322.

25 Browarska N, Kawala-Sterniuk A, Zygarlicki J, Podpora M, Pelc M, Martinek R & Gorzelańczyk E J, Comparison of smoothing filters' influence on quality of data recorded with the emotiv EPOC flex brain–computer

interface headset during audio stimulation, *Brain Sci*, **11(1)** (2021) 98.

26 Reichert C, Dürschmid S, Bartsch M V, Hopf J M, Heinze H J & Hinrichs H, Decoding the covert shift of spatial attention from electroencephalographic signals permits reliable control of a brain-computer interface, *J Neural Eng*, **17(5)** (2020) 056012.

27 Xiong Q, Zhang X, Wang W F & Gu Y, A parallel algorithm framework for feature extraction of EEG signals on MPI, *Comput Math Methods Med*, 2020.

28 Shajil N, Mohan S, Srinivasan P, Arivudaiyanambi J & Murrugesan A A, Multiclass classification of spatially filtered motor imagery EEG signals using convolutional neural network for bci based applications, *J Med Biol Eng*, **40(5)** (2020) 663–672.

29 Sandhya C, Srinidhi G, Vaishali R, Visali M & Kavitha A, Analysis of speech imagery using brain connectivity estimators, in *IEEE 14th Int Conf Cognit Info Cognit Comput* (IEEE) 2015, 352–359.

30 Van Dijk H, Schoffelen J M, Oostenveld R & Jensen O, Prestimulus oscillatory activity in the alpha band predicts visual discrimination ability, *J Neurosci*, **28(8)** (2008) 1816–1823.

31 Summerfield C & Mangels J A, Coherent theta-band EEG activity predicts item-context binding during encoding, *Neuroimage*, **24(3)** (2005) 692–703.

32 Wang Z, Tong Y & Heng X, Phase-locking value based graph convolutional neural networks for emotion recognition, *IEEE Access*, **7** (2019) 93711–93722.

33 Tsai F F, Fan S Z, Lin Y S, Huang N E & Yeh J R, Investigating power density and the degree of nonlinearity in intrinsic components of anesthesia EEG by the hilbert-huang transform: an example using ketamine and alfentanil, *PloS one*, **11(12)** (2016) e0168108.

34 Yi W, Qiu S, Wang K, Qi H, Zhang L, Zhou P & Ming D, Evaluation of EEG oscillatory patterns and cognitive process during simple and compound limb motor imagery, *PloS one*, **9(12)** (2014) e114853.

35 Daubechies I, The wavelet transform, time-frequency localization and signal analysis, *IEEE Trans Inf Theory*, **36(5)** (1990) 961—1005.

36 Alomari M H, Awada E A, Samaha A & Alkamha K, Wavelet-based feature extraction for the analysis of EEG signals associated with imagined fists and feet movements, *Comput Inf Sci*, **7(2)** (2014) 17.

37 Elman J L, Finding structure in time, *Cogn Sci*, **14(2)** (1990) 179–211.

38 Hoshi I, Shimobaba T, Kakue T & Ito T, Single-pixel imaging using a recurrent neural network combined with convolutional layers, *Opt Express*, **28(23)** (2020) 34069–34078.