IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 12, 2020, accepted August 10, 2020, date of publication August 14, 2020, date of current version August 25, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3016756

Neural Speech Decoding During Audition, Imagination and Production

RINI A. SHARON[1], (Student Member, IEEE), SHRIKANTH S. NARAYANAN[2],(Fellow, IEEE), MRIGANKA SUR[3], (Member, IEEE), AND A. HEMA MURTHY[1], (Senior Member, IEEE)

[1]Department of Computer Science and Engineering, IIT Madras, Chennai 600036, India
[2]Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90007, USA
[3]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Corresponding author: Rini A. Sharon (ee15d210@smail.iitm.ac.in)

ABSTRACT Interpretation of neural signals to a form that is as intelligible as speech facilitates the development of communication mediums for the otherwise speech/motor-impaired individuals. Speech perception, production, and imagination often constitute phases of human communication. The primary goal of this article is to analyze the similarity between these three phases by studying electroencephalogram(EEG) patterns across these modalities, in order to establish their usefulness for brain computer interfaces. Neural decoding of speech using such non-invasive techniques necessitates the optimal choice of signal analysis and translation protocols. By employing selection-by-exclusion based temporal modeling algorithms, we discover fundamental syllable-like units that reveal similar set of signal signatures across all the three phases. Significantly higher than chance accuracies are recorded for single trial multi-unit EEG classification using machine learning approaches over three datasets across 30 subjects. Repeatability and subject independence tests performed at every step of the analysis further strengthens the findings and holds promise for translating brain signals to speech non-invasively.

INDEX TERMS Assistive technology, brain computer interface, EEG, imagined speech, speech-EEG correlation, unit classification.

I. INTRODUCTION

The growth of modern computing technologies coupled with the understanding of the human brain have led to the evolution of promising Brain Computer Interfaces (BCIs) [1], [2]. BCIs provide a way to observe neural signals and convert the same to an actionable output. The potential to manipulate machinery with nothing more than a thought facilitates severely motor-disabled people to function independently [3], [4]. Given the limitless applications of converting neural signals to computerized actions, improving the quality and robustness of BCI systems to suit and serve the specific needs of each user is important.

Neuronal signals can be collected for experimentation by invasive (Neurosurgery), semi-invasive (Electrocorticogram (ECoG)) and non-invasive (Electroencephalogram (EEG), Magnetoencephalogram (MEG)) methods as reported extensively in the literature [5]. Since invasive electrodes are implanted directly in the brain, they are risky and expensive and also pose a hindrance to the process of data collection for the purpose of experimentation. Advances in accurate and easily deployable non-invasive EEG systems, including those that can communicate with smartphones via Bluetooth, and the relative ease of integration with other technologies suggests that EEG is likely to become a popular choice for building BCIs [6], [7]. Additionally, EEG signals offer high temporal resolution, while their poor spatial resolution is addressable by increasing the number of sensors [8], [9]. In this study, we describe use of the EEG to interpret electrical activity in the brain by means of recorded surface potentials on the scalp of the human subjects.

Although the fundamental principles behind BCIs using EEG are well established, there exist several drawbacks [10]. Among these, a few significant ones are discussed below. The EEG signal records the electrical potentials from the surface of the scalp which is distant from the source, and thus exhibits poor signal to noise ratio. Further, any involuntary muscular movement affects the EEG signal significantly. Albeit these cons, EEG-based BCIs still possess comforting prospects because of their noninvasive nature, convenience of

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wang.

recording, and potent applications as a communication medium for speech and motor-impaired individuals.

The organization of the paper is outlined below. The motivation behind the analysis of EEG signals related to speech, previous works on speech based neural signal analysis, and our contribution are summarized in the remainder of this section. The details of the methods used for classification are detailed in Section II. Section III describes the EEG data collection process, the different data-sets used for analysis, and their specifics. The general experimental setup and the features used are outlined in Section IV. The experimental results and discussions are presented in Section V, which is followed by Section VI which reviews some control checks to validate the various protocols employed in this work. Finally, Section VII concludes the work described in this article.

A. MOTIVATION

Most often, using BCIs to communicate with external devices requires the subject to perform conversation-irrelevant artificial tasks such as motor imagery [11], mental calculations [12], and so on. High accuracy BCI systems are also modeled based on P300 responses and steady state visually evoked potentials (SSVEP) [13]–[15]. Inconvenience of usage aside, the use of motor/mental actions for classification, severely restricts the number of events that can be represented. Research has shown that the brain has specific areas that are responsible for language understanding, speech comprehension, interpretation of meaning, and forming associations of sounds [16]–[18]. Although significant work has been done to model human speech comprehension abilities, it has been difficult to utilize this information for non-invasive computing purposes. In this work, we aim to investigate the reliability of speech-induced EEG signals in discriminating between distinct speech-like units in EEG. Three different phases of speech-EEG interaction, namely, speaking, listening, and imagining speech, are considered for the same. While speech imagination and production based interfaces possess high applicability for device control, inspecting the audition phase helps gain a better understanding of detectable neural processing units.

B. RELATED WORK: CO-SPEECH NEURAL SIGNALS

Despite disparities in the nature of Speech and EEG signals, there exist significant similarities of interest. Both speech and EEG signals are rich in temporal content and have high domain specificity in terms of speaker and subject influence respectively. In view of these similarities, numerous studies have outlined the existence of speech signatures in EEG by studying the correlation between speech envelopes and EEG envelopes [19], [20]. The high correlation reported between the audio speech envelopes and the reconstructed EEG-speech envelopes suggests that speech related signatures are present in EEG. Across different experimental perspectives/protocols such as followed in [19], [21], with differing input stimuli, good reconstruction accuracies are reported. These results thereby provide the primary support to explore the possibility of reconstructing speech from EEG. Many invasive electrode recording methods also suggest the dominant influence of speech on brain waves. The work presented in [22] specifically aims to reconstruct speech from neural ECoG recordings at a rate comparable to natural speech production rate. While [23], [24] study passive listening, [22] looks at speech production and miming. Non-invasive alternatives offering good temporal resolution such as EEG thus serve as better choices to model communication interfaces.

There have also been recent research advances aiming to classify real speech or imagined speech EEG units. Many works on the KARA ONE database attempt binary classification [25], [26] and 11-unit classification [27] with a maximum reported accuracy of 87% and 53% respectively. This database however, comprises isolated speech units played to the subjects or imagined by the subject as opposed to continuous conversational speech. Most of the works in the literature are based on a two-class problem [28], [29]. When a multi-class framework is considered, a significant decrease in performance is observed [30], [31]. The work proposed in this article not only addresses the multi-class problem but also considers continuous speech.

C. CONTRIBUTION

The objective of this article is to look for signatures of the fundamental units of speech (if any), in coherent-speech (co-speech) EEG signals. We collectively refer to the EEG signals collected when the subject is producing, listening, and imagining speech as co-speech EEG signals. The novel contributions of this study and how they are different from the work done previously in this domain are outlined below:

- Previous works have attempted imagined isolated speech-unit classification [28], [29], [31], spoken speech classification [30] and heard speech envelope correlation analysis [19], [20]. This work attempts to compare speech-unit classification across all three different phases of co-speech EEG, namely, imagination, perception, and production.
- Work so far considered either syllable-level, vowel-level, word-level, or phrase-level classification. Here a level-wise base unit selection is performed to best represent co-speech EEG classification in the three phases.
- A majority of the unit classification efforts so far have been defined as a binary isolated-unit classification problems. Our work presents a multi-class formulation for continuous conversational speech with a maximum of 54 classes in consideration.
- Features across different brain frequency bands and electrode cap regions are analyzed across five different types of Datasets.
- Unit classification and visualization of temporal unit structures are done across subjects and across sessions to investigate generalization.

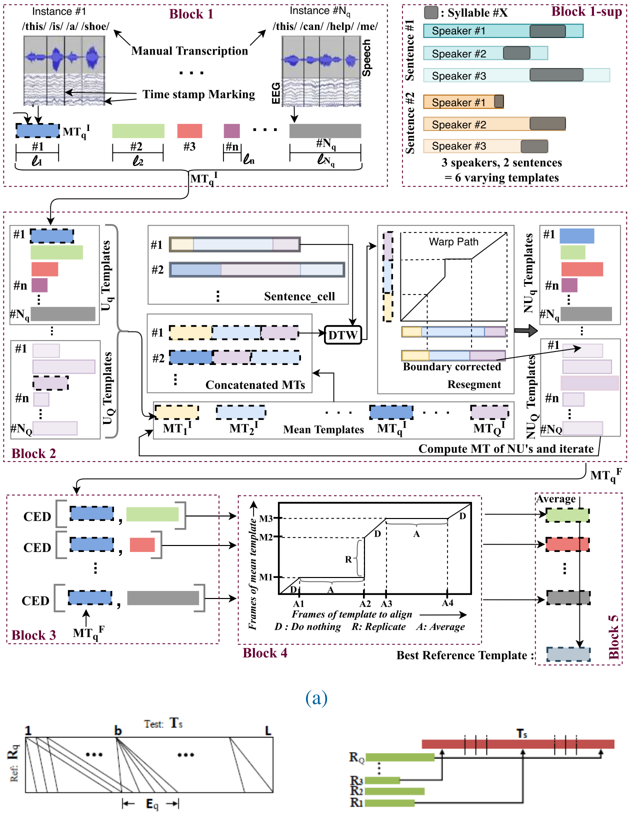
- In addition to result-based experimentation, a variety of control checks are performed to validate the implemented protocols.

II. METHODS

As EEG is a time-varying signal and has good temporal resolution, two different machine learning approaches, namely, two level dynamic programming and Gaussian mixture based hidden Markov models, were considered to match temporal patterns. These classification-oriented protocols are built with the intention of recognizing units of speech from recorded brain signals.

A. COMMON WORD REFERENCE TEMPLATE BASED TWO LEVEL DYNAMIC PROGRAMMING

The Common Word Reference Template matching algorithm along with a two level dynamic programming protocol (CWRT-2LDP) proposed by the authors in [32] was implemented. A series of functional blocks, as shown in Figure 1 were involved in implementing this protocol.



(a)

(b)

(c)

FIGURE 1. CWRT based 2LDP: a, Block 1,2 EEG data initial segmentation using manual markers obtained from the input/output speech signals, followed by iterative boundary correction. a, Block 1-sup Explanation of varying templates per class. a, Block 3,4 CED Algorithm to make all the templates equilength. a, Block 5 Average across the equilength templates to obtain the class-wise BRT. b, 2LDP distance score calculation c, Allocation of class labels.

Block 1(Initialization) - Boundary Segmentation:

Since we cannot determine the detailed ground truth boundaries for EEG, we followed a manual transcription based marking for segmentation. The speech input/output waveform and its corresponding EEG time aligned signal was imported in audacity and a label file was created to mark the unit-labels occurring in the speech waveform. The speech waveform is segmented unit-wise manually and the respective label files with their time stamps are saved. These time segments are then extracted from the 128 channel EEG data and assigned to their respective labels, thus giving us unit-level EEG segments. We segmented the hearing phase EEG data using markers from the corresponding input speech signal. The speaking phase EEG data for each trial was segmented using markers from the recorded speech output waveforms. The initial segmentation for the imagine-EEG data was done in a flat-start fashion by dividing the signal into equilength portions according to the number of constituent units (after verifying the manual mouse clicks by the subject to ensure that the subject had indeed imagined during the specific period).

Consider a dataset containing Q unique units. Distinct input cue sentences comprised of these units were formed in such a way that the units occurred in different contexts. Given that multiple speakers recorded the same set of cues, the segments of a specific unit occurring in the same sentence varied in duration depending on the speaker's rate of speech(Block 1-sup). Let's assume, the unit belonging to q^{th} class, where $1 \leq q \leq Q$, occurred in N_q instances (speakers × sentences, disregarding inter-sentence repetitions). Initial segmentation in this case, contributes N_q template segments of varying contexts and duration to the q^{th} class as shown in Block 1. These are clubbed together in the form of a cell array, U_q.

Block 2(Iteration) - Boundary Correction and Mean template:

Post speech-referenced segmentation, an iterative boundary correction algorithm is implemented as shown in Block 2. An initial Mean Template(MT_q^I) is chosen per class among the 'N_q' segments of cell U_q. This MT_q^I is the template with a length closest to the average length of all the 'N_q' templates. If l_n denoted the length of the n^{th} segment, where $1 \leq n \leq N_q$, then

$$l_\mu = (\sum_{n=1}^{N_q} l_n)/N_q \quad ; \quad d = \text{argmin}_{1\leq n\leq N_q} \left|l_\mu - l_n\right| \qquad (1)$$

$$MT_q^I = U_q\{d\} \qquad (2)$$

Once MT_q^Is were obtained, an iterative algorithm to self-adjust the boundaries was performed as described in Algorithm 1. The number of iterations was set empirically for each dataset. The inputs to the algorithm were the cell arrays(U_q) and the outputs were boundary adjusted(NU_q) cell arrays. Now, from the output NU_q cell arrays, we chose a final mean template(MT_q^F) per class, as the template with length closest to the average length of all templates in NU_q.

Block 3,4,5 - CWRT using Compress Expand Dynamic Time Warping(CED) Algorithm:

Algorithm 1 Iterative Boundary Adjustment

```
Input: Cell arrays $U_q$, with $N_q$ templates each, $MT_q^I$
Prior: Sentence_cell: A cell array of input cue sentences
Output: $NU_q$
For each t=1:max-iterations do
    For each s=1:total-sentences do
        sen=Sentence_cell{s};
        ind= sequential array of unit indices($\in$ 1 : $Q$) in sen
        concat: Concatenate $MT_k^I|_{k \in ind}$, in sequence order
        warp-path = DTW(concat, sen)
        For each j=1:length(ind) do
            S{j} = re-segment sen using warp-path
            $NU_{ind(j)}$\{count\} =S{j}; count ++
        end for
    end for
    $MT_q^I$= Compute mean templates of $NU_q$
end for
```

The cross-word reference template (CWRT) matching method as described in [33] was implemented to obtain one reference template from the many training templates across trials. To best represent every class, we need to unify the information contributed by every train instance to the particular class q. Owing to unequal lengths of segments in NU_q, we performed a CED method to make all segments equilength to facilitate averaging across trials. Consider the class q with N_q varying length templates and final mean template MT_q^F. A CED algorithm is performed between MT_q^F and every other (N_q-1) templates as shown in Block 3. When given a pair of input templates, the CED algorithm works as described in Block 4. The horizontal warp paths are collapsed by taking average of the frames in the path (avg($A1$ to $A2$)),(avg($A3$ to $A4$)). In vertical warp paths, the particular frame is replicated ($M2 - M1$) times. For diagonal warp path, the frames are kept as such. Hence, the CED algorithm yields N_q templates of the same length as its output. Now Block 5 computes the average across all the equilength templates of a particular unit class to obtain the class's best reference template(BRT).

Once the final templates post CWRT are obtained, a two level dynamic programming based classifier was devised as follows to classify test segments. Given the test signal, T_s and the final BRTs R_q of Q unit classes where $1 \leq q \leq Q$, a 2LDP as discussed in [32], [34], [35] is performed. Depending on the length L_q of each R_q, a range of end frames E_q is set as $b + \frac{L_q}{2} \leq e \leq b + 2L_q$, where b denotes the beginning frame and e the end frame in T_s. Then, a matrix of distance based scores $\bar{S}$, is formulated for every pair of beginning frame b and ending frame $e \in E_q$ as follows:

$$\widehat{S}(R_q, b, e) = \text{dtw-distance-measure}(R_q, T_s(b:e)) \tag{3}$$

$$\widetilde{S}(b, e) = \min_{1 \leq q \leq Q} [\widehat{S}(R_q, b, e)]\text{: retain best template match} \tag{4}$$

$$\widetilde{PI}(b, e) = \underset{1 \leq q \leq Q}{\text{argmin}}[\widehat{S}(R_q, b, e)]\text{: retain best path index} \tag{5}$$

$$\bar{S}(e) = \min_{1 \leq b < e} [\widetilde{S}(b, e) + \bar{S}(b-1)]\text{: recursive accumulation} \tag{6}$$

Keeping $\bar{S}$ as evidence of template match, backtracking of the path using $\widetilde{PI}$ is performed to obtain the best sequence of labels that transcribe the EEG corresponding to speech.

B. GMM HMM

A time tested and robust classification model for speech is the Gaussian Mixture Model - Hidden Markov Model (GMM-HMM) [36], more recently replaced by Deep Neural Network(DNN)-HMM models. HMM based models are gaining popularity in the domain of neural signals as well [37], [38]. Considering the data sparsity of EEG data, a GMM-HMM framework was considered for classification. Gaussian Mixture Model is a clustering algorithm that groups similar data points together based on their attributes or features and models the means and variances of these clusters. GMMs are probabilistic models and use soft clustering to assign data to each of the clusters. An Expectation Maximization(EM) algorithm is followed to estimate these parameters for each of the classes under consideration. Since each data point is considered independent of the others in a GMM, we incorporated an HMM based model to better capture the temporal statistics. HMMs provide a simple and efficient framework for modeling time-varying sequences by defining hidden states and their transition probabilities. Baum Welch training algorithm based on EM is used for training the HMMs for every class. They also facilitate context modelling that uses information about which unit is likely to occur in the context of another unit as opposed to the independent mono-unit modeling. Alongside mono-unit, tri-unit models which club left and right context units along with the unit under consideration to form standalone data instances can also be trained.

III. DATA COLLECTION AND DATABASES

This study is based on the data obtained from experiments conducted using three different elicitation protocols. The timeline of the experiments used to collect these datasets, referred as DS1, DS2, and DS3 respectively, are depicted in Figure 2. The Ethics Committee of the Indian Institute of Technology Madras approved this study (IEC/2018-03/HAM/09). All the subjects were informed about the aim and scope of the experiment, and a written consent was obtained to collect the data. Experiments were designed to collect EEG data in an acoustically isolated an-echoic chamber. A 128 channel HydroCel Geodesic Sensor net 130 was used with Cz as the reference. The sampling rate for obtaining EEG data was set at 250 Hz, and sensors' impedance were continuously monitored. Speech stimuli presented in the experiments were recorded in a sound-attenuated room by non-native male and female speakers of English. The speech was conversational-like and

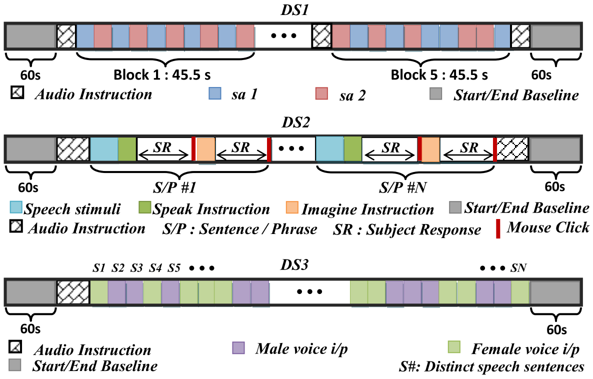


FIGURE 2. Datasets timeline description Inputs to DS1 are 2 sentences *sa*1: "she had your dark suit in greasy wash water all year" and *sa*2: "Dont ask me to carry an oily rag like that. These in totality constitute 25 syllabic units. The inputs for DS2 are formed from this set of syllables; i.e. "Dont rag me", "Carry your suit", "Ask like that" and so on. Inputs for DS3 are daily-use sentences like "Get me some food", "Thank you" and so on. Every experiment had a beginning and ending resting state of 60 seconds.

continuous, with no prolonged pauses. Each speech stimuli was presented multiple times in a session with a randomized presentation order using external speakers.

A non-overlapping set of 12, 9, and 9 subjects volunteered for the collection of dataset 1(DS1), 2(DS2), and 3(DS3) respectively. The total number of subjects in all the experiments is 30. All the subjects had good language proficiency in Indian English. The subjects (approximately 2:1 ratio of men and women) were healthy non-native English speakers between 25-34 years of age. The subjects were seated comfortably and were instructed to keep their eyes closed and minimize other voluntary movements throughout the experiment. Although this is not natural, the objective was to set up controls with minimal intervention due to artifacts in the EEG signals. Prior to the actual recording sessions, subjects were given an initial acclimatization session wherein few mock trials were performed. Once the subject was trained and comfortable with the elicitation protocol, he/she participated in one or two evaluation sessions where they were asked to complete the experimental task(s). Performance statistics are recorded and analysed individually for each subject-session pair and across sessions and subjects. Three experiments were designed to observe co-speech EEG influence, the details of which are summarized below. Table 1 briefly lists the purpose behind collecting each Dataset and the concerns it addresses.

TABLE 1. Purpose of collecting each Dataset.

Datasets	Aims to address
DS 1	• Do the EEG signals corresponding to speech units have distinct signatures? • Can syllable-like-units be used for classification? If not, what basic-unit best captures co-speech EEG characteristics? • Can a multi-unit classification framework yield promising results?
DS 2	• Are speech signatures detectable while imagining speech? (Special emphasis for imagination as this is likely to influence the design of BCIs) • Which phase best captures speech signatures from a classification perspective? • Are the signature of the speech units occurring in different contexts preserved in EEG?
DS 3	• Does replacing standard TIMIT words with everyday-use words affect system performance? • Do the algorithms scale when the number of units and their contexts are increased? • Does the classification protocol scale across subjects and across sessions?

A. DATASET 1

12 subjects volunteered for the collection of this dataset, out of which 6 subjects gave 2 sessions each on different days. Here two standard TIMIT sentences *sa*1 and *sa*2, spoken by 5 speakers(1 female and 4 males) were played to the subject multiple times in a randomized fashion while the subject was passively listening. These sentences together comprise of 25 syllabic units.

B. DATASET 2

EEG data was collected from 9 subjects with 3 subjects providing 2 sessions each on different days. Each sentence and phrase used as input stimulus for data collection is drafted using a combination of a subset of the 25 syllables from the syllabic content of *sa*1 and *sa*2. A total of 25 such sentences were played randomly to the subject. The prompts were spoken by one female and one male volunteer. The 25 sentences played once constituted one trial and 3 such trials were recorded in one recording session. To differentiate between the different phases of speech-based cognition in this dataset, we define DS2a as the hearing phase, DS2b as the imagining phase, and DS2c as the speaking phase.

1) **Hearing Phase (H):** Here, the subject is required to perform passive listening when exposed to speech stimuli. The subject is requested to pay attention to what is being played.
2) **Imagination Phase (P):** Here, the subject is instructed to imagine speaking the prompt(subvocalize). Sub-vocalizing or silent speech is the act of saying words in your head without any articulatory or phonatory muscular movements. While this form of internal speech is typically common while reading, here we assume it to be a form of imagined speech.
3) **Speaking phase (S):** Here, the subject is required to verbally repeat the sentence/phrase that was played. The subject is specifically instructed to speak with clarity and poise while minimizing other head movements.

Since 'Imagining' and 'Speaking' duration vary depending on the linguistic proficiency of the subject, a mechanism of manual mouse click is adopted. Every subject is asked to indicate with a mouse click the end of his/her imagining/speaking response. The audio spoken by the subjects is also recorded using a recording device to be used for initial level segmentation and to playback and discard mistakenly spoken trials.

C. DATASET 3

EEG data was collected from 9 subjects, 2 sessions each on different days. A set of most commonly used English words were shortlisted and sentences based on these words were drafted to be played as the input speech stimuli. One male and one female speaker were chosen to record the audio inputs. The syllabic content of these spoken sentences contained 54 syllables, which were then combined to form distinct words. A set of 64 sentences so formed was repeated twice in a single session.

IV. GENERAL EXPERIMENTAL SET-UP

A. PRE-PROCESSING

After obtaining the EEG data, we band-passed the signal between 0.3 Hz and 60 Hz to retain the frequencies that contain relevant information and also applied a notch at 50 Hz to discard AC interference. Common average reference over all electrodes was applied offline post data collection. Trials containing subject induced artifacts(mean ≈5 trials per subject) were visually monitored and discarded for experimentation purposes. Other stereotypical artifacts including electromyographic ones were corrected using runica of EEGLAB, as applied in works dealing with similar objectives [39], [40]. Post this, data was segmented with the help of flags distinctly set to mark sentences' beginnings and ends as can be seen in Figure 3. Each flag, four characters long, is associated with a unique sentence ID. A mapping file is created apriori, which matched the flags with the sentence being played, imagined, or spoken. This mapping file is then used to match the segmented EEG time-sections with the corresponding sentence' text. The segments so obtained after mapping are considered as independent trials for training/classification.

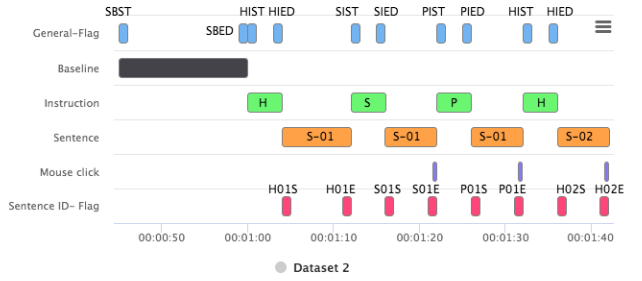


FIGURE 3. Flag Setting description for Dataset 2(an example time chunk): The first character of the flag indicates the action being performed; Hearing(H), Speaking(S), Imagining(P). The next 2 characters hold the unique sentence ID(01-25/64). The last character indicates the beginning(S) or end(E) of the action.

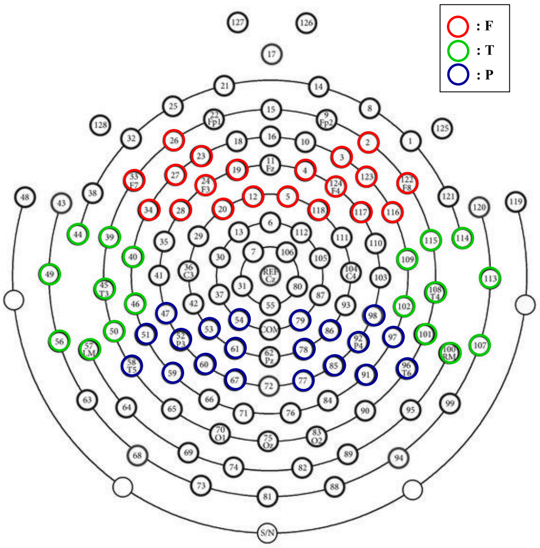


FIGURE 4. Region-wise mapping of electrodes- F: Frontal, T: Temporal, P: Parietal.

As the literature claims that different bands of frequencies correspond to different forms of cognition, the EEG signals were further bandpass filtered into delta(0.3-3Hz), theta(3-8Hz), alpha(8-13Hz), beta(13-30Hz) and gamma (30-50Hz) bands for analysis. A Beta-Gamma band (13-39Hz) is additionally bandpass filtered. The filters used for the above are second-order Butterworth bandpass filters. Channels corresponding to specific areas in the electrode cap, as depicted in Figure 4, are also extracted.

B. FEATURE EXTRACTION

EEG feature extraction and signal decomposition schemes characterize the signal as a function of time and/or frequency [41]–[43]. The classification objective and the nature of data involved highly influence the optimal choice of features. Based on the analysis results of speech-EEG classification presented in [32], temporal features based on short term energy are chosen for this work. This is further motivated by measurements on EEG that suggest that the resolution decreases as we move from temporal to spectral to spatial domains [44]–[46].

Short term processing is particularly meaningful while dealing with time-varying signals like speech and EEG where we assume fixed properties in a finite short-term temporal block [47], [48]. The Short term Energy (STE) is calculated as given in Equation 7, where "x_i" is the input EEG signal of length "N" samples, "h" is the Hamming window function of length 125 samples and "m" is the time shift of 1 sample.

$$E_m = \sum_{n=1}^{N} [x_i[n]h[m-n]]^2 \tag{7}$$

TABLE 2. Comparison of classification absolute accuracy in (%) for different BUs (36 channels, 0.3-50Hz, STE, 2LDP); CA: Chance Accuracy given by (1/(number of classes))*100; LU: like-unit; Avg Acc: Average Accuracy (%) across all subjects' data available for that dataset. (*): In Dataset 1, there is a sentence-level class bias. Since this dataset comprises of only 2 sentences, the sentence level classification module seems to perform the best. Contrary to this, in multi-sentence unbiased datasets, sentence-level modelling does not perform best. (**): Phrase-level and sentence-level accuracies are close for dataset 2 and 3 as these mostly comprise of single phrasal sentences.

	Dataset 1			Dataset 2a			Dataset 2b		Dataset 2c		Dataset 3		
	#cl	*Avg Acc(%)*	*Times CA*	*#cl*	*Avg Acc(%)*	*Times CA*	*Avg Acc(%)*	*Times CA*	*Avg Acc(%)*	*Times CA*	*#cl*	*Avg Acc(%)*	*Times CA*
Syllable LU	25	36.5	9.3	25	**24.9**	6.2	**24.5**	6.1	**21.1**	5.3	54	**17.9**	9.31
Word LU	21	31.8	6.8	21	21.3	4.4	20.8	4.3	17.1	3.6	59	15.1	8.9
Phrase LU	8	61.5	4.9	28	22.2**	6.1	21.4**	5.9	18.7**	5.1	62	13.4**	8.3
Sentence LU	2	**71.2***	1.4	25	21.7**	5.4	20.9**	5.2	18.0**	4.5	64	13.2**	8.4

The extracted STE at a particular time instant is a C-dimensional feature vector, where 'C' indicates the number of channels extracted during pre-processing. All the experiments discussed hereafter use this STE feature extraction module to generate EEG features for classification.

C. INTER-INTRA EVALUATION

High EEG-based subject classification accuracies [49], [50] imply the existence of significant variability in EEG signals while different subjects perform the same task. Since our datasets involve multiple sessions and subjects, the variability induced by them needs to be addressed. Hence, we define the following three testing strategies:

1) Intra-subject + Intra-session (Case A): Disjoint sets of train and test are taken from a single session of a single subject.
2) Intra-subject + Inter-session (Case B): Training is done on one session and testing on another session of the same subject.
3) Inter-subject (Case C): Training is done on data belonging to one subject and the model is tested on another subject's data.

The results of all the experiments reported unless otherwise specified are accuracies averaged over case A (for single-session subjects) and case B (for multi-session subjects).

D. PERFORMANCE METRIC AND DATA PREPARATION

The classification accuracy is taken to be $1 - UER$, where UER is the unit error rate. It is calculated by taking into account the number of Insertions(I), Substitutions(S), and Deletions(D) in the decoded output as compared to the ground truth. When an extra unit occurs in the decoded output between two existing in the ground truth, it is termed as an insertion. When a unit in the ground truth is replaced by another unit in the decoded output, it is termed a substitution. When a unit in the ground truth does not appear in the decoded output, it is counted as a deletion. The UER is hence calculated as given in Equation 8.

$$UER(\%) = \frac{(I + S + D)}{Total \quad number \quad of \quad units} \tag{8}$$

The central objective of this study is to investigate if EEG signals can enable the development of robust BCIs. To address this question, an extensive set of experiments were performed. All the intra-subject experiments henceforth considered a 4-fold cross validation approach(disjoint set of 75% for training and 25% for testing in each fold) and attempted single trial EEG decoding. For inter-subject(/session) reports, K-fold cross validation is performed, where K is the number of subjects(/sessions per subject) in that particular dataset.

V. EXPERIMENTAL RESULTS AND DISCUSSION

A. UNIT OF ANALYSIS - SELECTION OF BASIC UNIT (BU)

The primary step towards co-speech EEG recognition is to determine the units of speech processing in the brain, which are best recognized in a BCI set-up. For speech recognition, the basic unit of recognition is chosen as a phoneme, which is considered as the smallest single meaning-bearing unit of speech perceptible to the listener. Words, phrases, and sentences are then considered as a sequence of phonemes. Here we tried to classify EEG signals in syllable, word, phrase, and sentence levels to discern the best performing base unit(BU). Since the underlying cognitive decoding process is still not fully understood, we consider the possibility that the unit of auditory speech-stimuli might not be directly reflected in the measured EEG. Hence, if we assume $\mathbb{Y}$ to be a unit in speech under consideration, we propose a corresponding $\mathbb{Y}$-like-unit($\mathbb{Y}$LU) in EEG. Since the experimental framework published in [32] is established, a similar implementation was followed here considering the 36 channels belonging to the temporal and parietal regions, but without frequency band filtering. The classification results for all datasets using STE features in the 2LDP framework for all BUs under consideration are presented in Table 2. Results indicate that syllable-like-unit(SLU) classification yields the best performance across unbiased datasets. Taking note of this, we fixed SLU as the best BU for classification for the experiments that follow hereafter.

B. REGION-BAND ANALYSIS

Different regions of the human brain are responsible for specific cognitive functions [16], [18]. Since this work deals with non-invasive EEG signals, we intend to analyze specific regions in the electrode cap to determine which regions yield superior discriminative properties. Analogous to region-wise functional distinctions, specific frequency

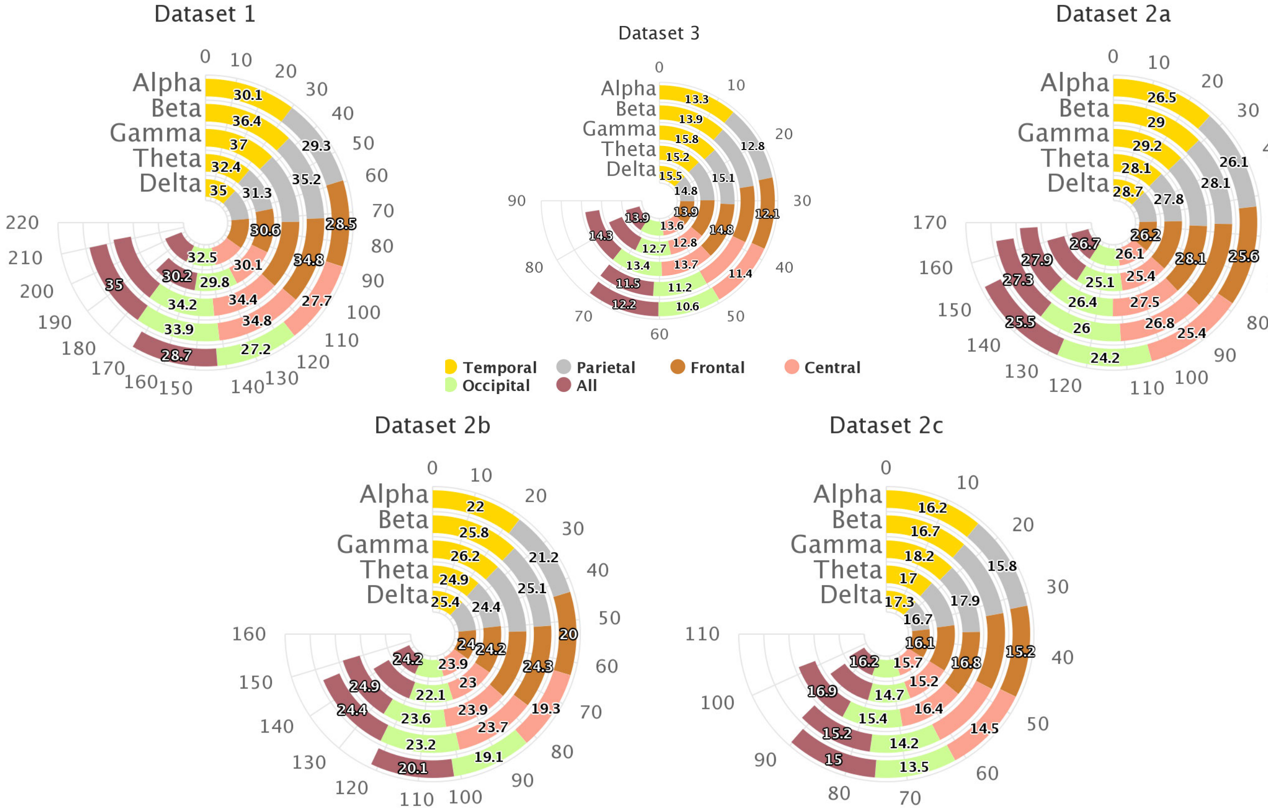


FIGURE 5. Region-Band Analysis: Pair-wise combinations of electrode cap regions and frequency bands are extracted and used to perform SLU unit classification for all Datasets using the 2LDP classification framework.

bands are seen to be dominant(in terms of energy) during certain cognitive states/actions. Many studies have shown that speech rhythms are captured as rhythmic modulations in the human auditory cortex [51], [52]. Hence, we focus on brain oscillations (bands) in subjects as they process continuous speech. Frequencies in the range 0.3 to 8Hz form the delta-theta band. This band is said to capture the perceived, non-speech-specific acoustic rhythm and syllabic rate [51], [52]. The beta band activity corresponds to 13 to 30Hz and is dominant when subjects are consciously reading, writing, and comprehending. In the literature, beta band also plays a part in high-level speech comprehension [53]. Frequencies in the range 30-50Hz indicate cognition, information processing, learning, and perception. Research shows gamma oscillations are engaged by speech and may have the potential to track its dynamics [51], [54].

In this section, EEG signals from different regions and bands are studied to understand their importance in speech-induced EEG. SLU classification results for all experimental data conditions corresponding to all region-band pairs across the five Datasets are depicted in Figure 5. Temporal and parietal region channels consistently perform better than channels extracted from other regions(average absolute gain(AAG) of 1.85% and 1.68% over other regions respectively). Concerning the frequency bands, the gamma band gives the best classification performance(AAG of 2.62% over other bands), closely followed by the beta, delta, and theta band. Given these observations, the two best performing regions and bands in combination were extracted and analysed across all datasets. As suggested in [32], the delta subtracted beta band characteristics were also analyzed, and their SLU classification accuracies were compared with the best performing frequency band in a similar experimental set-up for Dataset 1. We find that the Beta:Gamma band renders better classification performance than the Beta-Delta band (≈1.2% AAG across datasets).

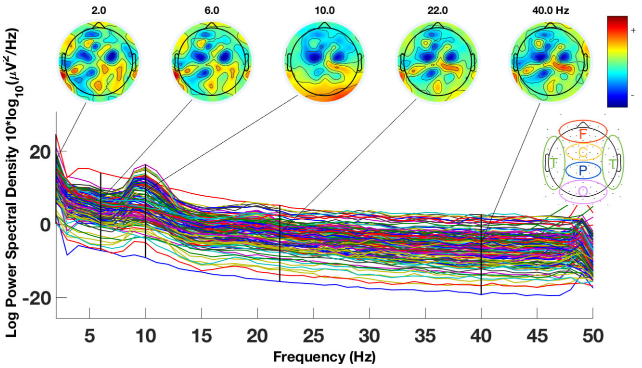


FIGURE 6. Channel-wise power spectra analysis.

To further visualize the region-band characteristics and to support the reported recognition scores, channel-wise frequency spectra and associated topographic maps of raw EEG signals are plotted for a session picked at random in Figure 6. A subset of 65% of the data is utilized to generate the plot. A single frequency representing every band considered is

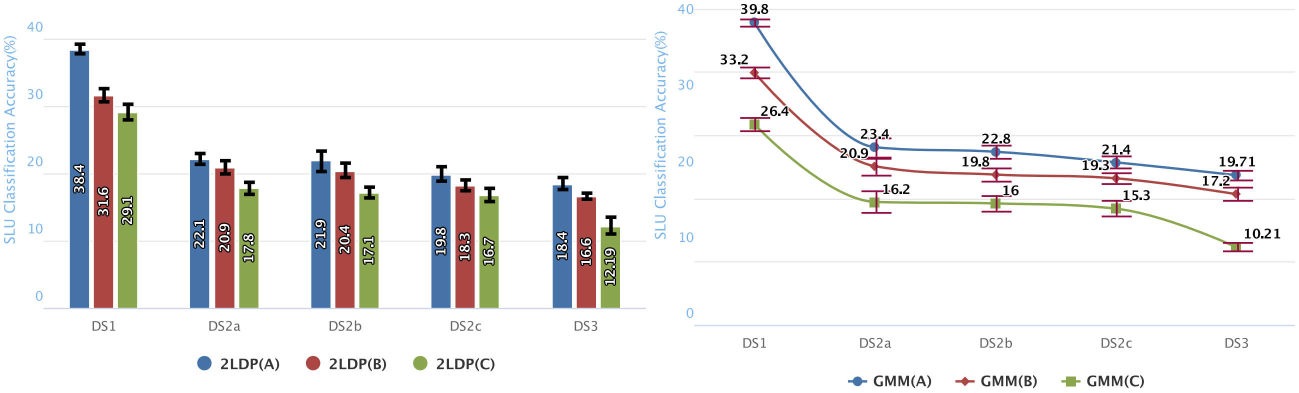


FIGURE 7. (a) 2LDP and (b) GMM-HMM performance is (plotted side by side for comparison) for Case-A, Case-B and Case-C for all Datasets.

picked at random for visualization(2Hz from delta, 6Hz from theta, 10Hz from alpha, 22Hz from beta, and 40Hz from gamma). The log power spectral density for each channel is calculated from the subset and is plotted across a frequency range of 0-50Hz. The scalp topographic maps show the scalp distribution of power in that particular frequency.

Comparing the topographic maps with the reference electrode cap region marking in the inset, it can be observed that specific frequency ranges have corresponding regions of data concentration. As is extensively reported in the literature, the alpha band frequency activity is concentrated in the occipital region. Similarly, high concentration can be observed in the temporal and parietal regions for the gamma and beta band frequencies. Relatively lesser temporal and parietal concentrations are observable in the delta and theta bands. These inferences are consistent with the classification values reported experimentally.

C. UNIT CLASSIFICATION

The analysis so far suggests that SLUs are the best BU for classification. Also, the temporal and parietal region channels filtered in the beta-gamma band capture maximum discrimination. Post analyzing these results, and the best performing feature set(determined using the selection by exclusion method above) was used for classification. All experiments henceforth report SLU classification results using temporal-parietal channels in the beta-gamma band.

Apart from the 2LDP classifier, a GMM-HMM classifier is also trained for SLU recognition. While the 2LDP model is trained using 36-dimensional STE features (from 36 temporal and parietal channels), the GMM-HMM model is trained by considering each channel to be an independent data instance. The implementation of 2LDP was done in MATLAB and GMM-HMM was done using the Kaldi toolkit [55] in this article. Hyper-parameters were tuned empirically. The number of states of the HMM used to model the temporal units were specifically hard-coded in the range 3-5 for distinct units depending on their average duration. The number of GMM components was set to 3 per state, and a 1.28 boost silence probability was assigned for the beginning and ending portions of the waveform.

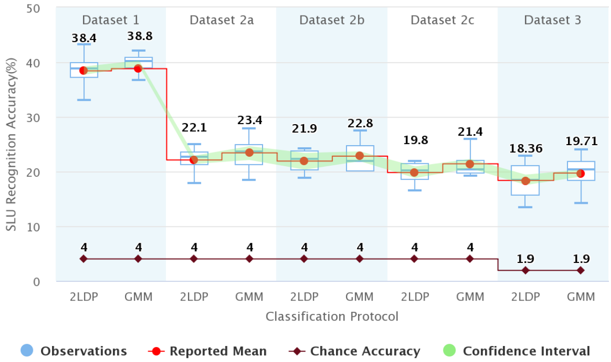


FIGURE 8. Boxplot for all Datasets.

Statistical Analysis of Inter-Intra Sessions and Subjects is performed to establish generalizability. Figure 7 compares the SLU classification accuracies across the two classification methods, 2LDP and GMM-HMM for the three cases of testing(Case-A(Intra session), Case-B(Inter session) and Case-C(Inter subject)). Confidence intervals across all subjects are marked for each instance. Chance accuracy is 4% for Dataset 1 and 2 and 1.9% for Dataset 3. Boxplot analysis using cross-validation results is also done for all Datasets to inspect the distribution of data. The following values of the accuracy distribution are plotted in Figure 8: minimum, maximum, median, first and third quartile. Leave-one-subject(LOS) out accuracy(Case C) is plotted for all 30 subjects for the audition experiment in Figure 9. The reported mean and chance accuracies for every dataset are also marked for comparison.

A comparison of the performance of the two models across datasets reveals that the EEG features classified using a GMM-HMM model give better accuracy than 2LDP for cases A and B. Since DS3 has 2 sessions for all subjects, the performance variation across cases A, B, and C form an unbiased evaluation. In DS3, it is observed that 2LDP consistently gives better classification accuracy for the inter-subject case as compared to the GMM-HMM model variant.

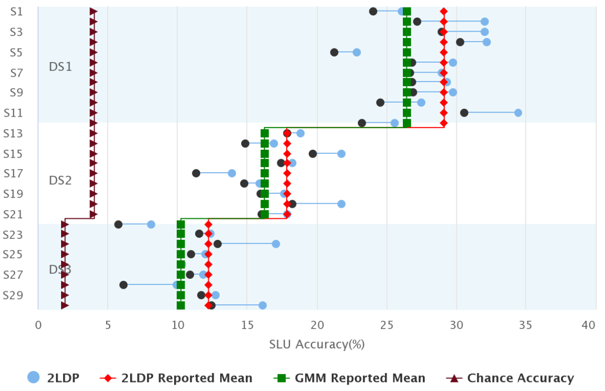


FIGURE 9. LOS accuracy for all 30 subjects(audition).

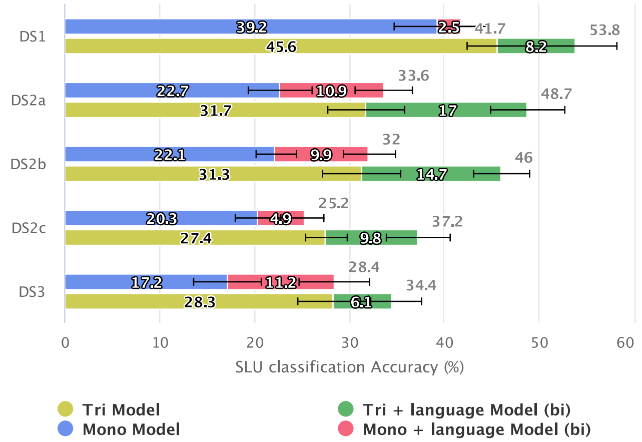


FIGURE 10. Effect of acoustic level and transcription level context modelling: Mono-SLU and Tri-SLU classification performance with and without bigram-SLU context modelling is plotted along with their confidence intervals to highlight the improvement in performance contributed by each stage of context modelling.

D. CONTEXT MODELLING

All the results reported above were obtained without explicitly using context information, i.e., assuming that the probability of an SLU is independent of the location at which it occurs. As there is a one-to-one temporal correspondence between the EEG signal and speech, this section investigates the effect of context modeling approaches in SLU classification performance. Two kinds of context modelling can be incorporated in the GMM-HMM model;

1) **Acoustic level**: These are the tri-unit models that capture the different contexts in which an EEG unit can occur and models them explicitly.
2) **Transcription level**: These utilize the data transcriptions to determine which linguistic paths are more probable than the others and help improve the confidence and correctness of the decoded output. Since our datasets have limited transcription vocabulary, the transcription level context modelling was built from a random set of 10,000 sentences taken from the text of wall street journal(WSJ0) database [56].

It is observed that incorporating a bi-SLU context model built using the text vocabulary greatly improved the performance. This boost in performance due to context modelling is seen to be significant across databases(Figure 10) for the best performing feature-model pair considered in this work. In order to comment on limited vocabulary applications, a transcription-level language model built using the data-specific text vocabulary was used for decoding. This further improved the decoding accuracy by 8.8 ± 1.7% across subjects as outlined by examples given in Table 3.

Continuous speech decoding from neural signals, as performed in [22], can be considered an invasive counterpart of our classification framework. They report 40% accuracy in a 25 syllable-pool set-up as compared to an accuracy of 37.2% in our work using non-invasive EEG. Further, the transcribed outputs from the decoded EEG signals can be synthesized as speech using a trained text to speech synthesizer, thus making it a functional communication interface.

TABLE 3. Single trial decoding of perception phase EEG signals for the Dataset 3(54 syllables). o: Original sentence, wd: decoded sentence with WSJ language model, od: decoded sentence with own data-specific language model. Substitutions are highlighted, deletions are striked out and insertions are in red.

Syllable error rate	Original sentences Vs decoded sentences
wd∼30% od∼15%	o: give me wa ter and food wd: there is wa ter and food od: give me wa ter and meet
wd∼50% od∼30%	o: nice to meet and know you wd: nice to meet ~~and know~~ you there od: nice to meet ~~and know~~ you
wd∼60% od∼40%	o: what is wrong with that wd: that is wrong ~~with that~~ od: there is wrong with him
wd∼80% od∼70%	o: peo ple help him with some thing wd: please ~~ple~~ help need ~~with~~ some wa ter od: ~~peo ple~~ help him ~~with~~ some thank you
wd∼80% od∼80%	o: I am sor ry dont fi nish the food wd: I am fine thank you ~~fi nish the food~~ od: I am fine ~~dont fi nish the food~~

E. METHODOLOGICAL DESIGN ADVANTAGES

Performance accuracies aside, the proposed approaches also offer design-level benefits. In comparison with popular speech-EEG decoding protocols, the advantages of the proposed framework are three-fold.

1) **Large-set Decoding:** Majority of works classify a closed-set vocabulary of units such as words [57], [58] and phrasal blocks [30]. This makes the scalability of the protocol to newer unseen test instances difficult. In the proposed approach 54 syllables are used as the fundamental units for recognition, therefore the supported vocabulary can be very large.
2) **Syllable recognition in continuous speech:** Existing syllable and vowel based classifiers disregard

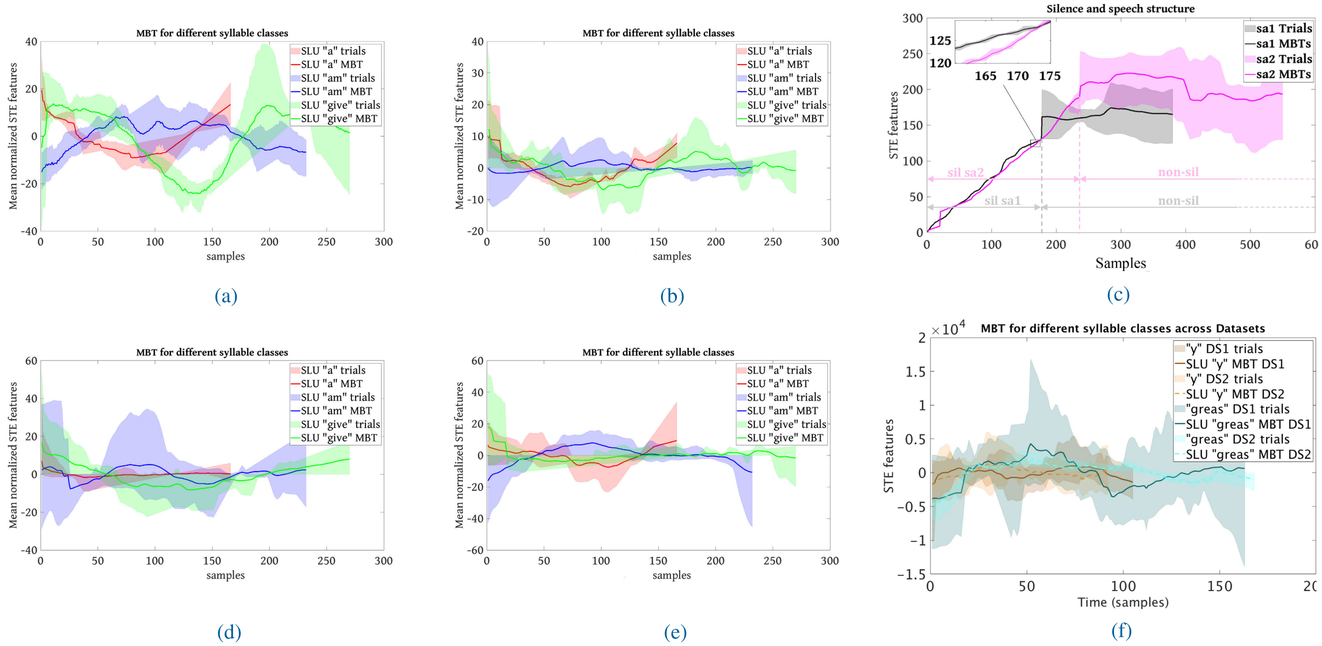


FIGURE 11. Distinct structures of three SLU classes ('a','am' and 'give') across Subjects across Sessions a, Subject A, Session 1 MBT, b, Subject A, Session 2 MBT, c, The STE features of a segment of sa1(black) and sa2(pink) EEG signals containing the beginning silence portion followed by the three non-silence SLUs occurring in a single trial are plotted, d, Subject B, Session 1 MBT, e, Subject B, Session 2 MBT, f, SLU templates across different Datasets with varying contextual occurrences.

contextual dependencies by training and testing models on isolated units rather than continuous speech [28], [29], [59]. The proposed method performs context-independent decoding of units in continuous speech-EEG signals across mismatched sentences.

3) **Model Generalization:** Most neural decoding approaches perform binary classification [25], [26], [29]. Although there have been few successful multi-class attempts, they do not consider subject and session independence [22], [27]. Addressing the concerns of variability due to these factors [60], the proposed approach provides generalization across multiple subjects and sessions while performing multi-class SLU decoding.

Summarising, the proposed approach is the first attempt to perform multi-class fundamental unit classification in continuous speech EEG generalizing over multiple subjects and sessions.

VI. PROTOCOL CONTROL CHECKS

Specific control checks at different stages of the classification protocol were carried out to further establish the validity of these results. Since these are just for inspection purposes, the experiments contained in this section are not reported for all datasets.

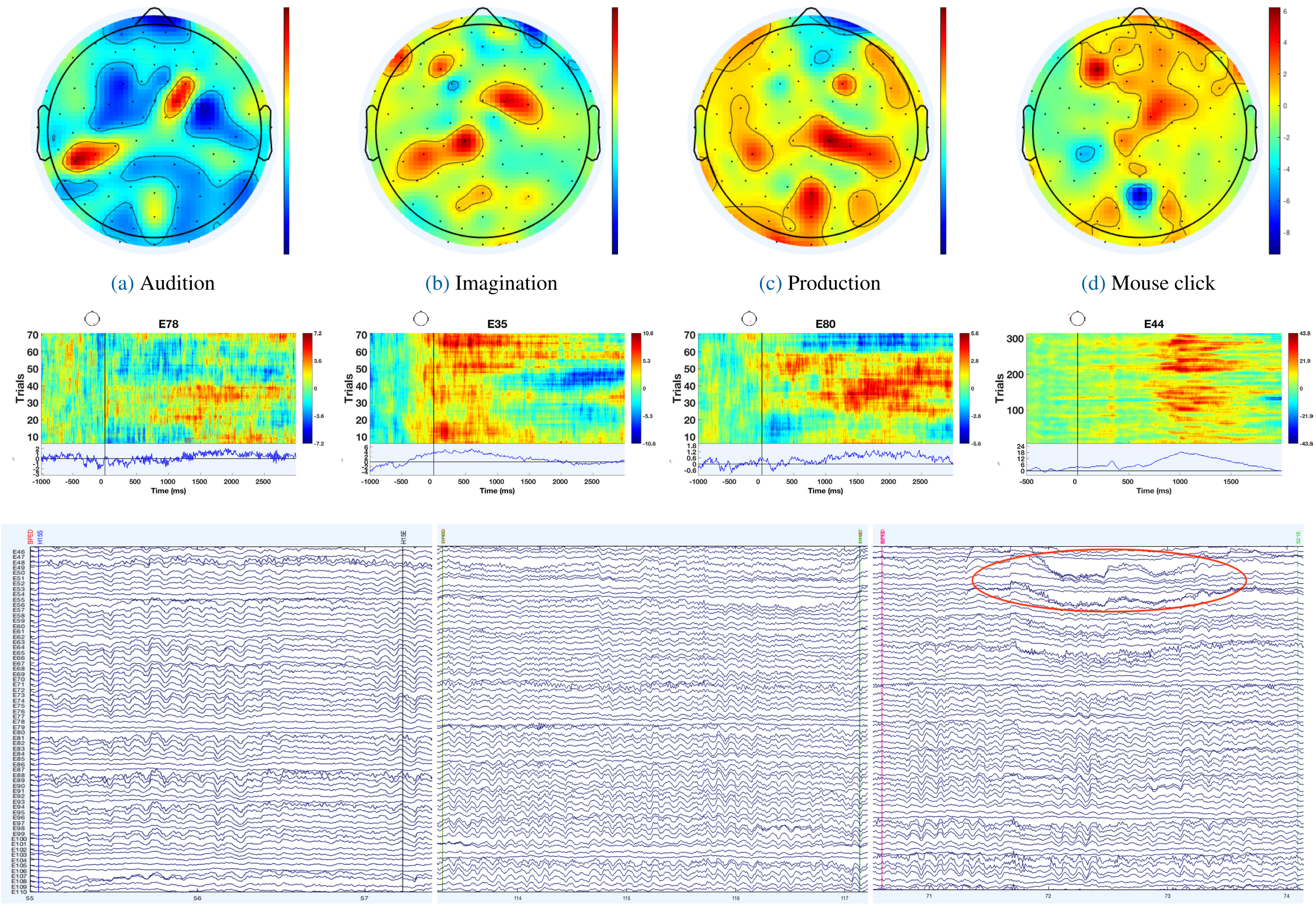
A. DTW WITHIN SEGMENTS

During segmentation, we obtained SLU segment boundaries which were iteratively corrected following the speech-marked manual boundary representations. Since we intend to capture the level of similarity in these segmented SLUs, we perform DTW across the SLUs and report the results. 50% of the segmented trials were taken to obtain a single train CWRT template per SLU class. The rest of the segments were tested against them by performing a DTW and assigning them to the class with minimum distance. An average classification accuracy of 31% was observed over 25 SLU classes across 12 subjects in Dataset 1. This establishes that there exist unique signatures specific to a particular SLU class.

B. VISUALIZATION OF SLU TEMPLATES

In order to visualize the signatures unique to an SLU class, we plot the EEG signal features corresponding to specific SLU classes and make the following observations:

1) Figure 11a, 11b, 11c and 11d aim to visually distinguish the short term EEG energy features of distinct SLUs. All trials are mean normalized before segmentation. The pre-averaging step of CWRT technique applies dynamic time warping based scaling to yield templates of constant length from all trials for a particular SLU class. The shaded portions in the figure represent the variance of these equilength STE feature templates across all trials, and the solid line corresponds to the mean best template (MBT) of that class which is chosen to perform 2LDP matching. This graphical visualization of templates belonging to three different SLU classes provides substantial evidence for discrimination in terms of the temporal shape and structure of distinct SLUs. Three SLUs, namely, 'a','am', and



(a) Audition (b) Imagination (c) Production (d) Mouse click

(e) Audition (f) Imagination (g) Production

FIGURE 12. Topographic map, ERP images, and Raw EEG signals plotted for EEG data recorded during different tasks, namely, Audition, Imagination, Production(artifact highlighted) and Mouse click followed by beep.

'give' are chosen and their templates are plotted across sessions and across subjects.

- The SLU templates seem to possess similar temporal structure while comparing different sessions from the same subject(left to right).
- The similarity in structure degrades when we compare across subjects(top to bottom).

This visualization thus provides the justification for the performances reported for inter-intra subjects and sessions.

2) Figure 11c is plotted to compare and analyze the distinct structures of the silence and non-silence portions of EEG. The STE features of a segment of sa1(red) and sa2(blue) EEG signals containing the beginning silence portion followed by the three non-silence SLUs occurring in a single trial are plotted. Two significant observations of this visualization are

- The silence portion appears similar irrespective of the sentence being played.
- The variance in the silence portion across different templates is small as compared to the non-silence SLUs. The zoomed inset shows a narrow shaded area confirming the above observations.

C. CROSS-DATASET TEMPLATES

Although Dataset 1 and Dataset 2 have different sets of input speech sentences, their constituent syllabic content is the same. In order to verify the SLU level signatures, we obtain CWRT MBT train templates from Dataset 2 and perform testing on EEG trials from Dataset 1 by following the 2LDP classification protocol. We get an SLU classification accuracy of 31.2% while testing against cross-dataset reference templates as opposed to 37.1% in self-dataset reference templates. Figure 11f temporally visualizes the MBTs of two specific syllables from datasets 1 and 2, and reveals template-level similarity across datasets. The good cross-dataset accuracy, despite the occurrence of SLUs in varying contexts, further supports the claim of the existence of SLU-specific signatures in EEG.

D. SIGNAL LEVEL VISUALIZATION OF THREE PHASES OF CO-SPEECH EEG DATA

As our experiments deal with multiple tasks being performed by the subjects, a signal level analysis of EEG data to locate visual differences between the tasks is performed. Here four tasks are chosen, namely hearing task, speaking task, imagination task, and mouse left-click. All

TABLE 4. Experimental Results Based Inferences.

Experimental Base Configuration	
Best Base Unit	Syllable-like-unit (SLU)
Most discriminating Regions	Temporal and Parietal (36 channels)
Most informative frequency band	Beta-Gamma (13-39 Hz)
Channel Handling (36)	
Concatenating to form 36 dimensional feature vector	Works best for 2LDP, as sequential patterns are matched
Averaging across 36 dimensions	Performance varies with a standard deviation of 1-2% for all models *Remark*: In speaking phase, accuracy always degrades while averaging
Taking each channel to be an independent trial	Utilized in GMM-HMM classifier to create more data instances that can be used to better model the states, mixtures and parameters
Intra-Inter Analysis	
Intra Subject - Intra Session	This gives best accuracy as it need not deal with other induced variations. However, practical feasibility is less as only a single session is considered to get disjoint train and test data
Intra Subject - Inter Session	As outlined previously, basic attempts to build a subject specific BCI falls here. Performance slightly degrades due to inter session variability caused by factors such as placement of electrode cap, familiarity with the experiment, subject's mental state and so on.
Inter Subject	Performance further degrades in this case due to subject induced variability. Subject ID is a widely studied EEG area where subject specific signatures are prominent enough to classify subjects with high accuracy. This subject specificity could restrict the task-level classification.
Protocol Performance	
2LDP	Average computational cost is highest for this. Given a test signal of length L, for every possible combination of beginning and end frames, 2LDP computes distance for all possible train templates. For experiments in this paper, average decode time is around 15mins$\pm\sigma$(8mins) per test trial .
GMM-HMM	This relatively faster decode protocol is ideal for online decoding of speech EEG for the convenience of BCI users. Average decode time is around 10sec$\pm\sigma$(3sec) per test trial .
Dataset-wise Performance Analysis, Best Accuracy Reported (BAR)	
Dataset 1 - (25 syl) , BAR: 53.8% (Perception)	The classification accuracy is very high in this dataset, mainly because there exist only 2 sentences spoken by 10 speakers played on repeat. This provides a significant amount of data per unit for modelling parameters. This is a control experiment, which shows that EEG signals do have some signatures of speech.
Dataset 2a- (25 syl), BAR: 48.7% (Perception)	Although this dataset has the same number of classes as DS1, 25 sentences with varying contexts exist. Due to this contextual diversity, overall accuracy reduces. Among the 3 phases, the perception phase best captures speech signatures.
Dataset 2b- (25 syl), BAR: 46% (Imagination)	Transitioning from Perception experiments to Imagination experiments, we find a slight dip in classification accuracy. The content of perception and imagination remains the same.
Dataset 2c- (25 syl), BAR: 37.2% (Speaking)	Protocols that work well for perception and imagination do not seem to scale the same way for production. Despite artifact removal techniques, residual effects of the speech production induced artifacts(which are relatively more prominent than perception or imagination artifacts), could be the reason for the same.
Dataset 3- (54 syl), BAR: 34.4% (Perception)	The experimental protocols followed for 25 class paradigm scales well for a 54 class problem as well. This dataset follows a similar pattern as DS2a in terms of contextual content and performance.

events of each task, irrespective of the sentence it involves, were segmented, segregated, and averaged across time. 2D topographic maps(TPM) of the average of all these instances and one-dimensional trial level event related potential images(ERPI) are plotted for raw EEG signals for each of the above tasks in Figure 12. The trials belonging to each task are stacked together to obtain a 2D average ERPI(values at trials-by-time). One second prior to the event onset and 3 seconds after the event onset is the time duration considered. Trials are sorted according to latency and adjacent trials are smoothed to obtain the final ERPI. Alongside the topographic maps and ERP images, the raw EEG signals are also plotted for the three phases, namely, audition, imagination, and production.

Few significant points of observation arise from these plots. The artifacts during the production phase are prominently visible in the raw EEG signal plot. The production phase also shows higher energy/intensity spread in the TPM and EPRI. Across all three phases, fairly high temporal and parietal electrode cap region activity can be observed from the topo maps. Passive audition TPM and ERPI show lower energy intensity as compared to production and imagination. We also see that in both the production and imagination ERPI, the 1 second duration segment before the event onset is similar and closely resembles the ERPI of the audition phase. The reason for this could be the fact that an auditory instruction cue is played right before the subject starts to imagine or speak. Otherwise stated, the subject performs passive audition when the instruction is being played, hence supporting the similarity in the pre-onset ERPI segments. The most consistent localized energy activity across trials is observed in the mouse click event, which is followed by a beep.

VII. CONCLUSION

Inferences drawn from the results discussed so far are jointly summarized in Table 4. These observations also address the concerns mentioned in Table 1. This work analyzes non-invasive EEG signals and aims to relate physical aspects of the speech signal to its corresponding EEG modality by computationally showing the existence of speech-specific signatures in co-speech EEG. The paper adopts a ''Selection by exclusion'' method to design an optimal experimental set-up to classify SLUs from continuous co-speech EEG in a multi-class scenario with much higher than chance accuracy. In addition to result-based experimentation, a variety of control checks are performed to validate the implemented protocols. In conclusion, given a limited vocabulary and a strict language model, there is a growing possibility of modelling naturalistic interfaces by capturing significant co-speech EEG signatures.

VIII. DATA AVAILABILITY

The data that support the findings of this study are freely available in `https://www.iitm.ac.in/donlab/cbr/cospeech_eeg_dataset/`

IX. CODE AVAILABILITY

All codes may be obtained for non-commercial use by contacting the corresponding author.

REFERENCES

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, ''Brain–computer interfaces for communication and control,'' *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, 2002.

[2] J. Wolpaw and E. W. Wolpaw, *Brain-Computer Interfaces: Principles and Practice*. Oxford, U.K.: OUP, 2012.

[3] J. J. Shih, D. J. Krusienski, and J. R. Wolpaw, ''Brain-computer interfaces in medicine,'' *Mayo Clinic Proc.*, vol. 87, no. 3, pp. 268–279, Mar. 2012.

[4] R. Leeb, S. Perdikis, L. Tonin, A. Biasiucci, M. Tavella, M. Creatura, A. Molina, A. Al-Khodairy, T. Carlson, and J. D. R. Millán, ''Transferring brain–computer interfaces beyond the laboratory: Successful application control for motor-disabled users,'' *Artif. Intell. Med.*, vol. 59, no. 2, pp. 121–132, Oct. 2013.

[5] S. Waldert, ''Invasive vs. non-invasive neuronal signals for brain-machine interfaces: Will one prevail?'' *Frontiers Neurosci.*, vol. 10, p. 295, Jun. 2016.

[6] F. Cincotti, D. Mattia, F. Aloise, S. Bufalari, G. Schalk, G. Oriolo, A. Cherubini, M. G. Marciani, and F. Babiloni, ''Non-invasive brain–computer interface system: Towards its application as assistive technology,'' *Brain Res. Bull.*, vol. 75, no. 6, pp. 796–803, Apr. 2008.

[7] J. R. Del Millán, P. W. Ferrez, F. Galán, E. Lew, and R. Chavarriaga, ''Non-invasive brain-machine interaction,'' *Int. J. Pattern Recognit. Artif. Intell.*, vol. 22, no. 05, pp. 959–972, 2008.

[8] C. J. Chu, ''High density EEG' what do we have to lose?'' *Clin. Neurophysiol., J. Int. Fed. Clin. Neurophysiol.*, vol. 126, no. 3, p. 433, 2015.

[9] V. J. Acharya and J. N. Acharya, ''Localization with high-density EEG: Complexity of analysis versus accuracy,'' *Clin. Neurophysiol. Practice*, vol. 5, p. 10, 2020.

[10] S. N. Abdulkader, A. Atia, and M.-S. M. Mostafa, ''Brain computer interfacing: Applications and challenges,'' *Egyptian Inform. J.*, vol. 16, pp. 213–230, Jul. 2015.

[11] L. Ming-Ai, W. Rui, H. Dong-Mei, and Y. Jin-Fu, ''Feature extraction and classification of mental EEG for motor imagery,'' in *Proc. 5th Int. Conf. Natural Comput.*, vol. 2, 2009, pp. 139–143.

[12] T. Fernández, T. Harmony, M. Rodríguez, J. Bernal, J. Silva, A. Reyes, and E. Marosi, ''EEG activation patterns during the performance of tasks involving different components of mental calculation,'' *Electroencephalogr. Clin. Neurophysiol.*, vol. 94, no. 3, pp. 175–182, Mar. 1995.

[13] R. C. Panicker, S. Puthusserypady, and Y. Sun, ''An asynchronous P300 BCI with SSVEP-based control state detection,'' *IEEE Trans. Biomed. Eng.*, vol. 58, no. 6, pp. 1781–1788, Jun. 2011.

[14] Y. Wang and T.-P. Jung, ''Visual stimulus design for high-rate SSVEP BCI,'' *Electron. Lett.*, vol. 46, no. 15, pp. 1057–1058, Oct. 2010.

[15] J. Jin, B. Z. Allison, T. Kaufmann, A. Kübler, Y. Zhang, X. Wang, and A. Cichocki, ''The changing face of P300 BCIs: A comparison of stimulus changes in a P300 BCI involving faces, emotion, and movement,'' *PLoS ONE*, vol. 7, no. 11, Nov. 2012, Art. no. e49688.

[16] J. A. Kiernan, ''Anatomy of the temporal lobe,'' *Epilepsy Res. Treatment*, vol. 2012, pp. 1–12, 2012.

[17] A. Ardila, B. Bernal, and M. Rosselli, ''The role of wernicke's area in language comprehension,'' *Psychol. Neurosci.*, vol. 9, no. 3, pp. 340–348, 2016.

[18] S. L. E. Brownsett and R. J. S. Wise, ''The contribution of the parietal lobes to speaking and writing,'' *Cerebral Cortex*, vol. 20, no. 3, pp. 517–523, Mar. 2010.

[19] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, ''Attentional selection in a cocktail party environment can be decoded from single-trial EEG,'' *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, Jul. 2015.

[20] W. Biesmans, J. Vanthornhout, J. Wouters, M. Moonen, T. Francart, and A. Bertrand, ''Comparison of speech envelope extraction methods for EEG-based auditory attention detection in a cocktail party scenario,'' in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 5155–5158.

[21] M. J. Crosse, J. S. Butler, and E. C. Lalor, ''Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions,'' *J. Neurosci.*, vol. 35, no. 42, pp. 14195–14204, 2015.

[22] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, ''Speech synthesis from neural decoding of spoken sentences,'' *Nature*, vol. 568, no. 7753, p. 493, 2019.

[23] M. Yang, S. A. Sheth, C. A. Schevon, G. M. M. Ii, and N. Mesgarani, ''Speech reconstruction from human auditory cortex with deep neural networks,'' in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 1121–1125.

[24] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang, ''Reconstructing speech from human auditory cortex,'' *PLoS Biol.*, vol. 10, no. 1, Jan. 2012, Art. no. e1001251.

[25] S. Zhao and F. Rudzicz, ''Classifying phonological categories in imagined and articulated speech,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 992–996.

[26] P. Sun and J. Qin, ''Neural networks based EEG-speech models,'' 2016, *arXiv:1612.05369*. [Online]. Available: http://arxiv.org/abs/1612.05369

[27] P. Saha, M. Abdul-Mageed, and S. Fels, ''Speak your mind! Towards imagined speech recognition with hierarchical deep learning,'' 2019, *arXiv:1904.05746*. [Online]. Available: http://arxiv.org/abs/1904.05746

[28] K. Brigham and B. V. K. V. Kumar, ''Imagined speech classification with EEG signals for silent communication: A preliminary investigation into synthetic telepathy,'' in *Proc. 4th Int. Conf. Bioinf. Biomed. Eng.*, Jun. 2010, pp. 1–4.

[29] B. Min, J. Kim, H.-J. Park, and B. Lee, ''Vowel imagery decoding toward silent speech BCI using extreme learning machine with electroencephalogram,'' *BioMed Res. Int.*, vol. 2016, Oct. 2016, Art. no. 2618265.

[30] M. Rosinova, M. Lojka, J. Stas, and J. Juhar, ''Voice command recognition using EEG signals,'' in *Proc. Int. Symp.*, Sep. 2017, pp. 153–156.

[31] A. Porbadnigk, M. Wester, J. Calliess, and T. Schultz, ''EEG-based speech recognition—Impact of temporal effects,'' in *Proc. Int. Conf. Bio-Inspired Syst. Signal Process. (BIOSIGNALS)*, Jan. 2009.

[32] R. A. Sharon, S. Narayanan, M. Sur, and H. A. Murthy, ''An empirical study of speech processing in the brain by analyzing the temporal syllable structure in speech-input induced EEG,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 1–8.

[33] W. H. Abdulla, D. Chow, and G. Sin, ''Cross-words reference template for dtw-based speech recognition systems,'' in *Proc. Conf. Convergent Technol. Asia–Pacific Region*, 2003, pp. 1576–1579.

[34] H. Sakoe, ''Two-level DP-matching–A dynamic programming-based pattern matching algorithm for connected word recognition,'' *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 6, pp. 588–595, Dec. 1979.

[35] R. A. Sharon and H. A. Murthy, ''Comparison of feature-model variants for coSpeech-EEG classification,'' in *Proc. Nat. Conf. Commun. (NCC)*, Feb. 2020, pp. 1–6.

[36] M. Gales and S. Young, ''The application of hidden Markov models in speech recognition,'' *Found. Trends Signal Process.*, vol. 1, no. 3, pp. 195–304, 2008.

[37] D. A. Moses, M. K. Leonard, J. G. Makin, and E. F. Chang, "Real-time decoding of question-and-answer speech dialogue using human cortical activity," *Nature Commun.*, vol. 10, no. 1, pp. 1–14, Dec. 2019.

[38] C. Herff, L. Diener, M. Angrick, E. Mugler, M. C. Tate, M. A. Goldrick, D. J. Krusienski, M. W. Slutzky, and T. Schultz, "Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices," *Frontiers Neurosci.*, vol. 13, p. 1267, Nov. 2019.

[39] J. M. Correia, B. Jansma, L. Hausfeld, S. Kikkert, and M. Bonte, "EEG decoding of spoken words in bilingual listeners: From words to language invariant semantic-conceptual representations," *Frontiers Psychol.*, vol. 6, p. 71, Feb. 2015.

[40] M. A. Bakhshali, M. Khademi, A. Ebrahimi-Moghadam, and S. Moghimi, "EEG signal classification of imagined speech based on Riemannian distance of correntropy spectral density," *Biomed. Signal Process. Control*, vol. 59, May 2020, Art. no. 101899.

[41] A. S. Al-Fahoum and A. A. Al-Fraihat, "Methods of EEG signal features extraction using linear analysis in frequency and time-frequency domains," *ISRN Neurosci.*, vol. 2014, pp. 1–7, 2014.

[42] J. Kevric and A. Subasi, "Comparison of signal decomposition methods in classification of EEG signals for motor-imagery BCI system," *Biomed. Signal Process. Control*, vol. 31, pp. 398–406, Jan. 2017.

[43] M. T. Sadiq, X. Yu, Z. Yuan, F. Zeming, A. U. Rehman, I. Ullah, G. Li, and G. Xiao, "Motor imagery EEG signals decoding by multivariate empirical wavelet transform-based framework for robust brain–computer interfaces," *IEEE Access*, vol. 7, pp. 171431–171451, 2019.

[44] B. Burle, L. Spieser, C. Roger, L. Casini, T. Hasbroucq, and F. Vidal, "Spatial and temporal resolutions of EEG: Is it really black and white? A scalp current density view," *Int. J. Psychophysiol.*, vol. 97, no. 3, pp. 210–220, Sep. 2015.

[45] N. Salansky, A. Fedotchev, and A. Bondar, "High-frequency resolution EEG: Results and opportunities," *Amer. J. EEG Technol.*, vol. 35, no. 2, pp. 98–112, Jun. 1995.

[46] M. Vázquez Marrufo, E. Vaquero, M. J. Cardoso, and C. M. Gómez, "Temporal evolution of α and β bands during visual spatial attention," *Cognit. Brain Res.*, vol. 12, no. 2, pp. 315–320, Oct. 2001.

[47] M. Jalil, F. A. Butt, and A. Malik, "Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals," in *Proc. Int. Conf. Technol. Adv. Electr., Electron. Comput. Eng. (TAEECE)*, May 2013, pp. 208–212.

[48] J. Li and S. Sun, "Energy feature extraction of EEG signals and a case study," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jun. 2008, pp. 2366–2370.

[49] K. Brigham and B. V. K. V. Kumar, "Subject identification from electroencephalogram (EEG) signals during imagined speech," in *Proc. 4th IEEE Int. Conf. Biometrics: Theory, Appl. Syst. (BTAS)*, Sep. 2010, pp. 1–8.

[50] L. A. Moctezuma, A. A. Torres-García, L. Villaseñor-Pineda, and M. Carrillo, "Subjects identification using EEG-recorded imagined speech," *Expert Syst. Appl.*, vol. 118, pp. 201–208, Mar. 2019.

[51] J. Gross, N. Hoogenboom, G. Thut, P. Schyns, S. Panzeri, P. Belin, and S. Garrod, "Speech rhythms and multiplexed oscillatory sensory coding in the human brain," *PLoS Biol.*, vol. 11, no. 12, Dec. 2013, Art. no. e1001752.

[52] N. Ding and J. Z. Simon, "Cortical entrainment to continuous speech: Functional roles and interpretations," *Frontiers Human Neurosci.*, vol. 8, p. 311, May 2014.

[53] M. D'Zmura, S. Deng, T. Lappas, S. Thorpe, and R. Srinivasan, "Toward EEG sensing of imagined speech," in *Int. Conf. Hum.-Comput. Interact.*, 2009, pp. 40–48.

[54] A.-L. Giraud and D. Poeppel, "Cortical oscillations and speech processing: Emerging computational principles and operations," *Nature Neurosci.*, vol. 15, no. 4, p. 511, 2012.

[55] D. Povey, "The kaldi speech recognition toolkit," in *Proc. Workshop Autom. Speech Recognit. Understand.*, 2011, pp. 1–7. [Online]. Available: https://kaldi-asr.org/

[56] J. Garofalo, D. Graff, D. Paul, and D. Pallett, *Csr-i (wsj0) Complete*. Philadelphia, PA, USA: Linguistic Data Consortium, 2007.

[57] M. N. I. Qureshi, B. Min, H.-J. Park, D. Cho, W. Choi, and B. Lee, "Multiclass classification of word imagination speech with hybrid connectivity features," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 10, pp. 2168–2177, Oct. 2018.

[58] C. Cooney, A. Korik, F. Raffaella, and D. Coyle, "Classification of imagined spoken word-pairs using convolutional neural networks," in *Proc. 8th Graz BCI Conf.*, 2019, pp. 338–343.

[59] B. M. Idrees and O. Farooq, "EEG based vowel classification during speech imagery," in *Proc. 3rd Int. Conf. Comput. Sustain. Global Develop.*, 2016, pp. 1130–1134.

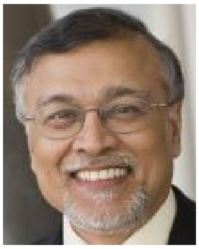
[60] A. Melnik, P. Legkov, K. Izdebski, S. M. Kärcher, W. D. Hairston, D. P. Ferris, and P. König, "Systems, subjects, sessions: To what extent do these factors influence EEG data?" *Frontiers Hum. Neurosci.*, vol. 11, p. 150, Mar. 2017.

RINI A. SHARON (Student Member, IEEE) received the bachelor's degree in electronics and communication engineering from the Vellore Institute of Technology, Vellore, India, in 2015. She is currently pursuing the Ph.D. degree with the Electrical Department, IIT Madras, Chennai, India. From 2015 to 2016, she worked as an Associate Engineer at Caterpillar Pvt., Ltd., Chennai. Since 2016, she has been serving as a Teaching Assistant with IIT Madras. She has published eight articles to date and is also an Active Science Communicator. Her research interests include speech processing, computational brain research, and other areas of machine learning and signal processing.

SHRIKANTH S. NARAYANAN (Fellow, IEEE) received the B.E. degree in electrical engineering from the College of Engineering, Guindy, Chennai, India, in 1988, and the M.S. and Ph.D. degrees in electrical engineering from UCLA, in 1990 and 1995, respectively.

He was an Engineer at UCLA, in 1992. From 1995 to 2000, he was with AT&T Labs-Research, Florham Park, AT&T Bell Labs, and Murray Hill, first as a Senior Member and later as a Principal Member of Technical Staff. He was the Research Area Director of the Integrated Media Systems Center and the NSF Engineering Research Center, University of Southern California (USC), and the Research Principal of the Pratt and Whitney Institute for Collaborative Engineering, USC, a unique partnership between academia and industry, from 2003 to 2007. He is currently a University Professor and the holder of the Niki and Max Nikias Chair in engineering with USC and a Professor with the Ming Hsieh Electrical and Computer Engineering Department, Signal and Image Processing Institute, USC, with joint appointments as a Professor in computer science, linguistics, psychology, neuroscience, pediatrics, and otolaryngology-head and neck surgery. He is also the Inaugural Director of the Ming Hsieh Institute and the Research Director of the Information Sciences Institute, USC. His laboratory was supported by federal (NSF, NIH, DARPA, IARPA, ONR, Army, and DHS) and industry grants. He has published over 800 articles and has 17 granted U.S. patents. His research interest includes signals and systems modeling with an interdisciplinary emphasis on speech, audio, language, multimodal, and biomedical problems and applications with direct societal relevance. His research and inventions have led to technology commercialization, including through startups he co-founded behavioral signals technologies focused on the telecommunication services and AI-based conversational assistance industry and Lyssn focused on mental health care delivery, treatment, and quality assurance.

Dr. Narayanan is a Fellow of the National Academy of Inventors (NAI), the Acoustical Society of America (ASA), the International Speech Communication Association (ISCA), the Association for Psychological Science (APS), the American Association for the Advancement of Science (AAAS), and the American Institute for Medical and Biological Engineering (AIMBE). He was a recipient of the NSF CAREER Award; the USC Associates Award for Creativity in Research and Scholarship; the USC Engineering Junior and Senior Research Awards and Use-Inspired Research Award; the USC Electrical Engineering Northrop-Grumman Research Award; the Mellon Award for Mentoring Excellence; the USC Distinguished Faculty Service Award; the Okawa Research Award; the IBM Faculty Awards, in 2008 and 2010; the Google Faculty Research Award, in 2016; the 2011 UCLA Engineering Alumni Professional Achievement Award; the 2019 Distinguished Alumnus Award from the College of Engineering, Guindy; and the Faculty Fellowship from the USC Center for Interdisciplinary Research.



MRIGANKA SUR (Member, IEEE) received the B.Tech. degree in electrical engineering from IIT Kanpur, Kanpur, and the Ph.D. degree in electrical engineering from Vanderbilt University, Nashville, TN, USA.

He is currently the Newton Professor of neuroscience and the Director of the Simons Center for the Social Brain, MIT, which he founded after 15 years as the Head of the Department of Brain and Cognitive Sciences, MIT. He studies the organization, plasticity, and dynamics of the cerebral cortex of the brain using experimental and theoretical approaches. He has discovered fundamental principles by which networks of the cerebral cortex are wired during development and change dynamically during learning. He has published over 400 articles and has five granted patents.

Dr. Sur is an elected Fellow of The Royal Society, U.K., the National Academy of Medicine, the American Academy of Arts and Sciences, the American Association for the Advancement of Science, the World Academy of Sciences, and the Indian National Science Academy. He has received numerous awards and honors, most recently the Cortical Discoverer Prize of the Cajal Club, and delivered distinguished lectures world-wide. He has trained over 75 doctoral students and postdoctoral fellows, and received awards for outstanding teaching and mentoring. At MIT, he has been recognized with the Sherman Fairchild and Newton Chairs.

A. HEMA MURTHY (Senior Member, IEEE) received the bachelor's degree from Osmania University, Hyderabad, India, in 1980, the master's degree from McMaster University, Hamilton, ON, Canada, in 1986, and the Ph.D. degree from IIT Madras, Chennai, India, in 1992.

From 1980 to 1983, she worked as a Scientific Officer (SO/SC) with the Speech and Digital Systems Group, TIFR, Mumbai. She is currently a Professor with the Department of Computer Science and Engineering, IIT Madras. She has over 39 journal publications, two book publications, and over 220 articles. Her research interests include speech processing, computer networks, music information retrieval, computational brain research, and other areas of machine learning and signal processing.

Dr. Murthy was a recipient of the Manthan Award, in 2012, and the IBM Faculty Award, in 2006. Her awards include being elected as a Fellow of the Indian National Academy of Engineering, in November 2017, and to the board of the International Speech Communication Association for the duration of 2017–2021.