AN EMPIRICAL STUDY OF SPEECH PROCESSING IN THE BRAIN BY ANALYZING THE TEMPORAL SYLLABLE STRUCTURE IN SPEECH-INPUT INDUCED EEG

Rini A. Sharon* *Shrikanth Narayanan*† *Mriganka Sur*** *Hema A. Murthy**

* Department of Computer Science and Engineering, Indian Institute of Technology Madras
†Viterbi School of Engineering, University of Southern California, Los Angeles
** Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge

ABSTRACT

Clinical applicability of electroencephalography (EEG) is well established, however the use of EEG as a choice for constructing brain computer interfaces to develop communication platforms is relatively recent. To provide more natural means of communication, there is an increasing focus on bringing together speech and EEG signal processing. Quantifying the way our brain processes speech is one way of approaching the problem of speech recognition using brain waves. This paper analyses the feasibility of recognizing syllable level units by studying the temporal structure of speech reflected in the EEG signals. The slowly varying component of the delta band EEG(0.3-3Hz) is present in all other EEG frequency bands. Analysis shows that removing the delta trend in EEG signals results in signals that reveals syllable like structure. Using a 25 syllable framework, classification of EEG data obtained from 13 subjects yields promising results, underscoring the potential of revealing speech related temporal structure in EEG.

Index Terms— Speech, EEG, syllable, Multitaper, delta

1. INTRODUCTION

Inspired by the idea that machines can be controlled by one's thoughts, various brain computer interfaces (BCI) have been developed to interpret cognitive activity [1, 2]. Given the concurrent advances in neuroscience and engineering, BCI principles are being implemented notably to provide controls to the motor and speech impaired users [3, 4]. Albeit the popularity that BCIs have gained, there exist cognitive challenges in operating these systems [5, 6]. Moreover, using motor movements or auditory/visually evoked potentials to control the BCI devices restricts the semantic event classification to be binary in most cases. In order to exploit the multi-class scope and to facilitate ease of use of these interfaces, we need to design a better communication protocol at the user's end. Therefore designing a natural speech-like communication medium using BCIs could prove to be highly effective.

Over the years, researchers have aimed to understand how human linguistic and cognitive abilities are interwoven in the production and perception of intelligible speech [7,8]. The human auditory system is intricately designed to perform acoustic analysis, including the extraction of meaningful units that facilitate its interpretation. Numerous studies suggest examining the mechanisms of speech processing in order to enable speech reconstruction from brain signals. However, due to the challenges and limitations of obtaining invasive brain recordings of cognitive processes, such neural studies of human speech processing still remains ambitious. A step towards understanding speech processing in the brain is to focus on some basic units that constitute temporal structuring in speech. Syllables and phonemes have long been used for analyzing audio in many speech processing applications and provide a basic operational processing unit [9–11]. In this paper, we propose a novel protocol of using an iterative segmentation algorithm coupled with a two level dynamic programming classifier, to study the temporal structure of EEG at the syllable level in order to establish that distinguishable syllable like entities are indeed present in the EEG signal.

The rest of the paper is organized as follows. Section 2 discusses the motivation behind considering this problem statement. Section 3 details the EEG data-collection process and the pre-processing approaches followed. Section 4 highlights the proposed methods for feature extraction, analyzes the influence of removal of the delta trend in EEG and discusses the syllable classification protocol implemented in this paper. Section 5 outlines the result analysis followed by the conclusions in Section 6.

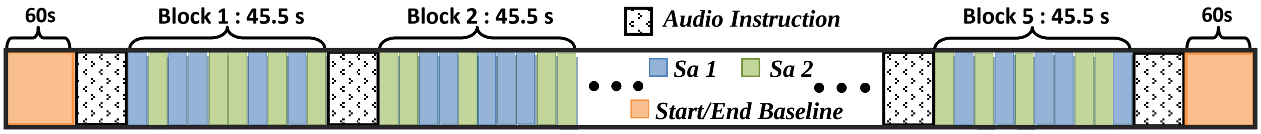
2. MOTIVATION AND RELATED WORK

Previous studies of speech envelope reconstruction from EEG indicate the existence of speech signatures in EEG [12–14]. Of the many frequency bands present in EEG, the delta band, constituting frequencies in the range 0.3Hz-3Hz, is seen to capture the syllabic rhythm of spoken speech.

Most of the previous works in the Speech-EEG domain have focused on phrase-level, vowel-level and imagined speech classification tasks. In [15], two syllables /ba/ and /ku/ are classified as a part of imagined speech experiments with an accuracy of 61%. [16] deals with the imagination of five vowels and their subsequent pairwise classification whereas [17] achieves a single trial classification of vowels /a/ and /u/. [18] aims to perform a two-way classification of imagined speech phonemes. Further [19] proposes a model to recognize two imagined words- ”yes” and ”no”. A majority of these experiments regardless of phoneme, vowel, syllable or word formats focus on a binary classification problem. Unlike these approaches, in this work we attempt to perform classification across 25 classes of syllables from co-speech EEG signals. Here, co-speech refers to the EEG signals recorded while the subject is listening to speech audio.

A closely related work in [20] intends to classify 50 phrases of speech EEG signals with a classification accuracy of 5% . However, the proposed method differs from the above in the band-based feature extraction and classification module and is shown to outperform the same with respect to accuracy and robustness.

The algorithm proposed in this paper is validated on EEG data obtained from 13 subjects performing a speech audio listening task over multiple sessions. The average classification accuracy of the proposed methodology in all cases is significantly above chance level - 37.12%, which suggests that EEG signals carry important information about auditory speech signals.



sa1: /ns_b1/ /she/ /had/ /your/ /dark/ /suit/ /in/ /greas/ /y/ /wash/ /wa/ /ter/ /all/ /year/ /ns_e1/
sa2: /ns_b2/ /don't/ /ask/ /me/ /to/ /car/ /ry/ /an/ /oil/ /y/ /rag/ /like/ /that/ /ns_e2/

Fig. 1: Data collection timeline description

3. EEG DATA ACQUISITION

Experiments were performed to collect EEG data in an acoustically isolated an-echoic chamber. The healthy volunteer subjects were seated in a comfortable position and directed to keep their eyes closed and minimize other voluntary movements throughout the experiment. Although this is unnatural, the objective was to set up controls with minimal interference due to artifacts in the EEG signals. A 128 channel net was used and the sampling rate for obtaining EEG data was set at 250 Hz.

3.1. Database collection and description

Data for this experiment were collected from 13 subjects with 5 subjects offering 2 sessions each amounting to a total of 18 sessions. The experiment begins with a baseline resting state of 1 minute followed by an instruction cue which requests the subject to pay careful attention to the spoken audio. Then a block of 5 sa1 and 5 sa2 sentences from the standard TIMIT database are played in a random order. These sentences were recorded by Indian English speaking volunteers (1 female and 4 males) to ensure that the subjects do not have difficulty in speech cognition due to dialectal effects. The syllables that make up these utterances are listed in Figure 1. Further, the instruction cue and the block are repeated 4 more times before the experiment concludes with an end baseline resting state of 1 minute. The instruction cues and the speech audio were communicated to the subject via speakers placed at a distance of approximately 4 feet facing the subject. The timeline of the experiment is represented pictorially in Figure 1.

3.2. Pre-processing

After obtaining the EEG data, we band-pass the signal between 0.3 Hz and 60 Hz to retain the frequencies that contain relevant information and also apply a notch at 50 Hz to discard AC interference. Analysis is carried out using different frequency bands separately by band pass filtering each band, namely, delta(0.3-3Hz), theta(3-8Hz), alpha(8-13Hz), beta(13-30Hz) and gamma(30-50Hz). Data is then segmented with the help of markers set to discriminate sa1-sa2 sentences. These segments are considered as independent trials for training/classification.

4. PROPOSED METHOD

4.1. Feature Extraction

As the literature suggests, the parietal [21] and the temporal [22] lobes are responsible for language understanding, interpreting sounds and speech perception. Hence, out of the 128 channels, 36 channels corresponding to the temporal and parietal regions were considered. Two types of features were considered for analysis.

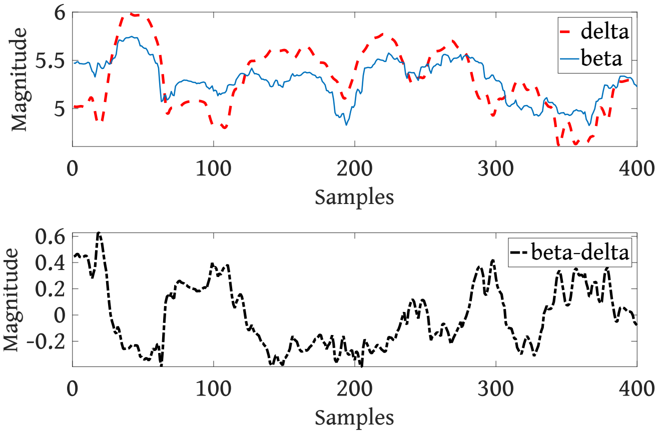


Fig. 2: Short Term Energy band analysis

4.1.1. *Short Term Energy (STE)*

Analogous to the speech stimulus, EEG signals can be considered as a non-stationary time-varying excitation. Akin to multiple speech processing approaches [23,24], we revert to short term processing of the EEG signal assuming it is stationary in a finite temporal block. The STE is calculated as given in Equation 1, where "w" is the hamming window function of length 125 samples and "x" is the input EEG signal.

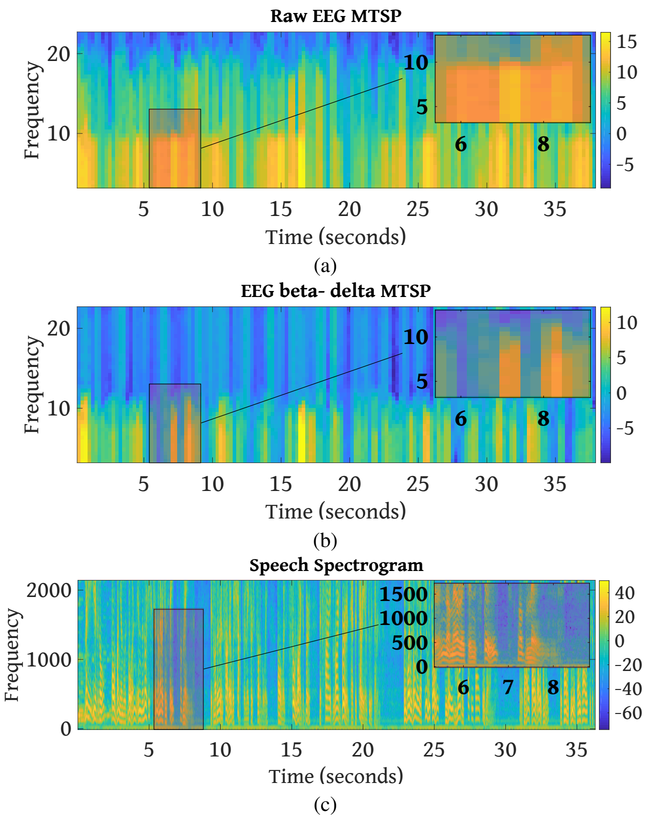
$$E_m = \sum_n [x(n)w(m-n)]^2 \tag{1}$$

A graphical analysis of the delta band STE, reveals signatures analogous to syllable rhythm/rate. Also noticeable is the delta band dominance in other frequency bands, especially beta. Hence when we subtract the delta band from the beta band, the spectral structure is evident. The characteristic of delta band and its influence on other bands is shown in Figure 2.

4.1.2. *Multi-taper spectrogram*

The multi-taper (MT) spectrogram is used for visualization purpose to validate the approach of removing delta band influence from beta band in EEG signals. The multi-taper method as introduced in [25], provides multiple independent estimates from the same sample by multiplying the signal with pairwise orthogonal data taper windows. The final spectrum is obtained by averaging over all the statistically independent tapered spectra. In our experiment we set the parameters of the tapers as follows: time-bandwidth product is 5, the number of tapers used is 9, the moving window length was set to 150 samples with 1 sample shift.

In agreement with the STE scenario, the MT spectrogram of the beta$-$delta band signal brings to light syllable like spectral structures



(a)

(b)

(c)

Fig. 3: (a) Raw EEG signal's MT spectrogram is plotted with an expanded inset. (b) After subtracting the delta band from the beta band MT spectrogram is plotted. (c) The corresponding speech spectrogram is plotted

as shown in Figure 3(b). The raw-EEG signal's MT spectrogram can be visualized as a noisy version of the beta−delta band EEG signal's MT spectrogram. Silence gaps similar to that of speech are seen while comparing Figures 3(b) and 3(c), albeit with some delays.

4.2. Proposed Protocol for Syllable Classification

A dynamic time warping (DTW) based cross-word reference template (CWRT) matching approach was adopted to perform an iterative segmentation algorithm to obtain syllable level templates and to classify the test EEG signals into the 25 syllable classes.

4.2.1. Segment boundary initialization from speech

Let G_t be a single feature-extracted train EEG signal with $1 \leq t \leq T$, where T total training examples are available. Post feature-extraction, the initial level of segmentation is performed on G_t using the syllabic boundary information that is obtained by the manual alignment of the speech waveform. Unlike speech, where we can clearly identify silence regions, EEG does not guarantee one-to-one correspondence between speech silence and brain signal silence. Hence, the beginning and ending silence part of the speech waveform is considered to be some distinct non-speech segments for EEG. These classes are named with the prefix "ns" referring to non-speech in Figure 1. The initial boundaries so obtained are then adjusted iteratively as described in section 4.2.2.

Algorithm 1 Iterative CWRT matching for segmentation

```
Input: $D_{init}$ , Train EEG signals $G_t, t \in T$
Output: Best Reference Templates (BRT).
procedure OBTAIN $N$ BRTS ( $N$ syl classes)
    Initialize Use CWRT on $D_{init}$ to obtain $N$ MBTs
    ref-temp-concat : Concatenate MBT based on ground truth transcription
    for each i in #iter do
        for each t in $T$ do
            warp-path = dtw(ref-temp-concat, $G_t$)
            for each n in $N$ do
                Segment $G_t$ based on warp path
                $D_{new}$(t,n) = new segment
            end for
        end for
        Use CWRT on $D_{new}$ to obtain new MBT
        ref-temp-concat : Concatenate new MBT
    end for
    BRT = mean along columns of $D_{new}$
end procedure
```

4.2.2. Iterative Template-level Segmentation

The initial boundary segmentation as described in section 4.2.1 yields D_{init} as described below. Assume T training instances are available with N syllable classes each for $sa1$. Then $D_{init}(t,n) = \left\{ k_n^t \mid n = 1 \; to \; N \; \& \; t = 1 \; toT \right\}$, where k is a variable length template. Hence, every syllable is represented by T varying length templates. Due to computational constraints, efficient ways of choosing the best templates for each class proves to be advantageous. For this purpose, Crossword-reference template (CWRT) algorithm as described in [26] is implemented. Out of the T extracted templates from the training set, a base-reference template (BRT) per class is chosen such that it's length is closest to the average length of all the extracted templates of that class. Now, the other templates are time aligned by DTW such that their lengths match the base-template length. These time-aligned templates are further averaged to obtain the mean-base-template (MBT) for that particular class as depicted in Figure 4. The CWRT algorithm is implemented iteratively as described in Algorithm 1 for both $sa1$ and $sa2$ sentences.

4.2.3. 2-Level Dynamic Programming (2LDP)

Given the test signal, T_s and the final MBT references R_n of N syllable classes where $1 \leq n \leq N$, a 2LDP as discusssed in [27] is implemented to determine the sequence of reference templates

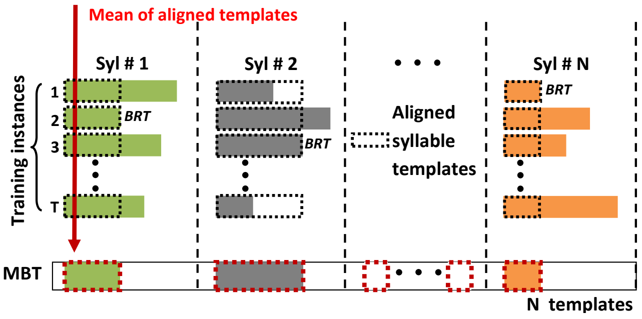


Fig. 4: Computation of MBT based on CWRT

Table 1: Protocol sanity check results (%avg accuracy)

Reference	sa1	sa1	sa2	sa2	sa1	sa2
Test	sa1	sa2	sa2	sa1	rest	rest
Avg Acc %	42.51	2.01	48.15	1.27	1.24	1.37

and their boundaries. A range of end frames E_n for each reference template R_n is determined in accordance with its length L_n as $b + \frac{L_n}{2} \leq e \leq b + 2L_n$, where b is the beginning frame and $e \in E_n$ is the end frame in T_s. The following steps are followed to procure an $\bar{M}$ matrix of scores.

$$\widehat{M}(R_n, b, e) = \text{dtw-distance-measure}(R_n, T_s(b : e))$$

$$\widetilde{M}(b, e) = \min_{1 \leq n \leq N}[\widehat{M}(R_n, b, e)]\text{: retain best template match}$$

$$\widetilde{P}(b, e) = \underset{1 \leq n \leq N}{\text{argmin}}[\widehat{M}(R_n, b, e)]\text{: retain best path index}$$

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

Keeping $\bar{M}$ as evidence of template match, backtracking of path using $\widetilde{P}$ is done to obtain the output classification labels. Once the classification labels are obtained, we perform median filtering based smoothing to remove ambiguous classification errors. A grid size of 3×1 was used for filtering.

5. RESULT ANALYSIS

5.1. Protocol sanity check

Since the classification accuracy is significantly above chance-level, we perform further analysis to investigate the integrity of the results. The following experiments are carried out to verify if the results are coherent and the resulting accuracies are recorded in Table 1.

- **Cross-sentence verification**: Test sa2 sentence is classified using templates from sa1 sentence.
- **Cross-session templates**: Test instances from Subject A's session a are compared against the references templates from Subject A's session b. This yielded an average accuracy of **31.45%**.
- **Cross-subject templates**: Test instances from Subject A are compared against the references templates from Subject B. This yielded an average accuracy of **29.12%**.
- **Rest state verification**: Test is taken from the baseline rest state and compared with sa1/sa2 sentence templates.

When $sa1$ and $sa2$ sentences are compared in the cross-sentence verification case, their confusion matrix reveals that the syllable $/y/$ from the word "oily" gets classified as the syllable $/y/$ from the word "greasy" frequently. This contributes to the scant accuracy achieved.

5.2. Performance Observations

The following are a few notable observations derived from the above set of experiments. The initial segmentation taken from speech alignments outperformed the conventional flatstart segmentation. CWRT method of choosing the best templates proved to be useful and reduced the computation cost substantially. While analyzing frequency band structures, the delta band signals and the beta minus delta signals show syllabic structure and hence provide better classification as outlined in Table 2. The MT approach used for visualization shows the beta minus delta band MTSP closely resembles the speech spectrogram. Choosing the channels in the temporal and parietal regions gave the best performance. The results of the sanity check experiments show that cross-session templates perform better classification than cross-subject templates. This is in accordance with the fact that some subject information is embedded in the EEG signals [28]. Also, using templates across subjects reduces the accuracy as in the cases of DTW based isolated word recognition problems in speech domain. The initial and final non-speech segments are often cross-classified, and median filtering helps alleviate these spikes in the output.

Table 2: Performance of the proposed method for different bands of EEG signal with tuned parameters (avg accuracy%)

	Delta	Beta-delta	Beta	Gamma
2LDP	36.21	**37.12**	34.77	34.15

5.3. Speech-syllable analysis

For comparison purpose, the input speech stimuli to the subject was considered for analysis. An approach identical to the proposed protocol was employed to perform syllable-level classification in the above mentioned speech signals. Three notable differences in achieving segment classification between speech and EEG in this case are as follows:

1. Since the speech signals were used just as stimuli for the experiment, only one recording per sentence- per speaker was obtained. Hence the train and test instances for syllable classification were taken from different speakers (inter-speaker case).
2. Since accurate ground truth segment boundaries were available for the speech signals, the iterative segmentation was not performed.
3. Mel-frequency cepstral coefficients (MFCC) are known to best represent speech features. Hence along with STE, MFCC features were also considered for evaluation.

The average syllable classification accuracy for the speech stimuli is seen to be 29.18%. This is comparable to the cross-subject template results (29.12%) reported for EEG in section 5.1. This clearly confirms that the EEG signal does contain syllable signatures.

6. CONCLUSION

This paper proposes a novel method to analyze temporal syllable structure in co-speech EEG signals and further attempts to perform syllable level classification. Careful analysis of the frequency bands of the EEG signals suggest that the delta band signatures are present in other frequency bands. Removing the presence of these signatures from beta band signals result in magnifying the syllabic content of EEG signals. The proposed protocol for classification employs a common word reference template based iterative segmentation coupled with two level dynamic programming algorithm and achieves a superior accuracy (37.12% for intra-subject and 29.12% for inter-subject) than chance level (4%) across 13 subjects. The sanity check experiments verify the robustness of the results. The speech signals were also considered for syllable classification under a similar protocol proceeding and the results are closely similar to the EEG case. In conclusion, this paper provides proof of existence of speech signatures in co-speech EEG signals.

7. REFERENCES

[1] Swati Vaid, Preeti Singh, and Chamandeep Kaur, "EEG signal analysis for BCI interface: A review," in *Advanced Computing & Communication Technologies (ACCT), 2015 Fifth International Conference on*. IEEE, 2015, pp. 143–147.

[2] Martin Spüler, "A high-speed brain-computer interface (BCI) using dry EEG electrodes," *PloS one*, vol. 12, no. 2, pp. e0172400, 2017.

[3] Kang Wang, Xueqian Wang, and Gang Li, "Simulation experiment of BCI based on imagined speech EEG decoding," *arXiv preprint arXiv:1705.07771*, 2017.

[4] Sergio Machado, Fernanda Araújo, Flávia Paes, Bruna Velasques, Mario Cunha, Henning Budde, Luis F Basile, Renato Anghinah, Oscar Arias-Carrión, Mauricio Cagy, et al., "EEG-based brain-computer interfaces: an overview of basic concepts and clinical applications in neurorehabilitation," *Reviews in the Neurosciences*, vol. 21, no. 6, pp. 451–468, 2010.

[5] Sarah N Abdulkader, Ayman Atia, and Mostafa-Sami M Mostafa, "Brain computer interfacing: Applications and challenges," *Egyptian Informatics Journal*, vol. 16, no. 2, pp. 213–230, 2015.

[6] Walter Glannon, "Ethical issues with brain-computer interfaces," *Frontiers in systems neuroscience*, vol. 8, pp. 136, 2014.

[7] Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F Chang, "Phonetic feature encoding in human superior temporal gyrus," *Science*, vol. 343, no. 6174, pp. 1006–1010, 2014.

[8] Stephanie Martin, José del R Millán, Robert T Knight, and Brian N Pasley, "The use of intracranial recordings to decode human language: challenges and opportunities," *Brain and language*, 2016.

[9] Ivan Kopeček, "Speech recognition and syllable segments," in *International Workshop on Text, Speech and Dialogue*. Springer, 1999, pp. 203–208.

[10] Lija V Bondarko, "The syllable structure of speech and distinctive features of phonemes," *Phonetica*, vol. 20, no. 1, pp. 1–40, 1969.

[11] Jacques Mehler, "The role of syllables in speech processing: Infant and adult data," *Phil. Trans. R. Soc. Lond. B*, vol. 295, no. 1077, pp. 333–352, 1981.

[12] Brian N Pasley, Stephen V David, Nima Mesgarani, Adeen Flinker, Shihab A Shamma, Nathan E Crone, Robert T Knight, and Edward F Chang, "Reconstructing speech from human auditory cortex," *PLoS biology*, vol. 10, no. 1, pp. e1001251, 2012.

[13] James A O'sullivan, Alan J Power, Nima Mesgarani, Siddharth Rajaram, John J Foxe, Barbara G Shinn-Cunningham, Malcolm Slaney, Shihab A Shamma, and Edmund 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, 2014.

[14] Minda Yang, Sameer A Sheth, Catherine A Schevon, Guy M McKhann II, and Nima Mesgarani, "Speech reconstruction from human auditory cortex with deep neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

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

[16] Beomjun Min, Jongin Kim, Hyeong-jun Park, and Boreom Lee, "Vowel imagery decoding toward silent speech BCI using extreme learning machine with electroencephalogram," *BioMed research international*, vol. 2016, 2016.

[17] Charles S DaSalla, Hiroyuki Kambara, Makoto Sato, and Yasuharu Koike, "Single-trial classification of vowel speech imagery using common spatial patterns," *Neural networks*, vol. 22, no. 9, pp. 1334–1339, 2009.

[18] Xuemin Chia, John B Hagedorna, Daniel Schoonovera, and Michael D'Zmuraa, "EEG-based discrimination of imagined speech phonemes," *International Journal of Bioelectromagnetism*, vol. 13, no. 4, pp. 201–206, 2011.

[19] Noramiza Hashim, Aziah Ali, and Wan-Noorshahida Mohd-Isa, "Word-based classification of imagined speech using EEG," in *International Conference on Computational Science and Technology*. Springer, 2017, pp. 195–204.

[20] Marianna Rosinová, Martin Lojka, Ján Staš, and Jozef Juhár, "Voice command recognition using EEG signals," in *ELMAR, 2017 International Symposium*. IEEE, 2017, pp. 153–156.

[21] Sonia LE Brownsett and Richard JS Wise, "The contribution of the parietal lobes to speaking and writing," *Cerebral Cortex*, vol. 20, no. 3, pp. 517–523, 2010.

[22] Bernard J Baars and Nicole M Gage, *Cognition, brain, and consciousness: Introduction to cognitive neuroscience - Chapter 7,11*, Academic Press, 2010.

[23] Sergio Suárez Guerra, José Luis Oropeza Rodríguez, Edgardo M Felipe Riveron, and Jesús Figueroa Nazuno, "Speech recognition using energy parameters to classify syllables in the spanish language," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2005, pp. 161–170.

[24] Tan Tian Swee, Sheikh Hussain Shaikh Salleh, and Mohd Redzuan Jamaludin, "Speech pitch detection using short-time energy," in *Computer and Communication Engineering (ICCCE), 2010 International Conference on*. IEEE, 2010, pp. 1–6.

[25] David J Thomson, "Spectrum estimation and harmonic analysis," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 1055–1096, 1982.

[26] Waleed H Abdulla, David Chow, and Gary Sin, "Cross-words reference template for dtw-based speech recognition systems," in *TENCON 2003. Conference on Convergent Technologies for the Asia-Pacific Region*. IEEE, 2003, vol. 4, pp. 1576–1579.

[27] Hiroaki Sakoe, "Two-level dp-matching–a dynamic programming-based pattern matching algorithm for connected word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 6, pp. 588–595, 1979.

[28] Katharine Brigham and BVK Vijaya Kumar, "Subject identification from electroencephalogram (EEG) signals during imagined speech," in *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*. IEEE, 2010, pp. 1–8.