Neural decoding of continuous speech EEG signals

Rini A Sharon[1]

[1]Indian Institute of Technology, Madras

ee15d210@smail.iitm.ac.in

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 our work 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(BCI). Fundamental syllabic units of speech in these phases are decoded accurately using temporal modelling based machine learning approaches generalizing over multiple sentences, trials, sessions, and subjects. [1]

1. Introduction and Contributions

As opposed to invasive(neurosurgical implants) and semi-invasive(electrocorticogram) BCI modules, EEG-based BCIs posses comforting prospects because of their noninvasive nature, convenience of recording and effortless deployability[1, 2]. In our work, we aim to investigate the reliability of speech-induced EEG signals in discriminating between distinct speech-like units in EEG. Datasets involving three common phases of communication, namely, speaking, listening, and imagining speech are considered for the same. Performance accuracies aside, the proposed framework offers three-fold design level advantages to potential BCI users as compared to popular speech-EEG decoding protocols as outlined below.

Large-set Decoding: Majority of works classify a closed-set vocabulary of units such as words[3, 4] and phrasal blocks[5]. 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.

Syllable recognition in continuous conversational speech: Existing syllable and vowel based classifiers disregard contextual dependencies by training and testing models on isolated units rather than continuous speech [6, 7, 8]. The proposed method performs context-independent decoding of units in continuous speech-EEG signals across mismatched sentences.

Model Generalization: Most neural decoding approaches perform binary classification [9, 10, 8]. Although there have been few successful multi-class attempts, they do not consider subject and session independence [11, 12]. Addressing the concerns of variability due to these factors [13, 14], the proposed approach provides generalization across multiple subjects and sessions while performing multi-class decoding.

Summarising, the proposed approach is a novel attempt to perform multi-class fundamental unit classification in continuous speech EEG generalizing over multiple subjects and sessions. The methodology involved and its reported results are published in [15].

[1]I thank my advisor, Hema A Murthy from IIT Madras and collaborative advisor Mriganka Sur from Massachusetts Institute of Technology for their valuable guidance in this work.

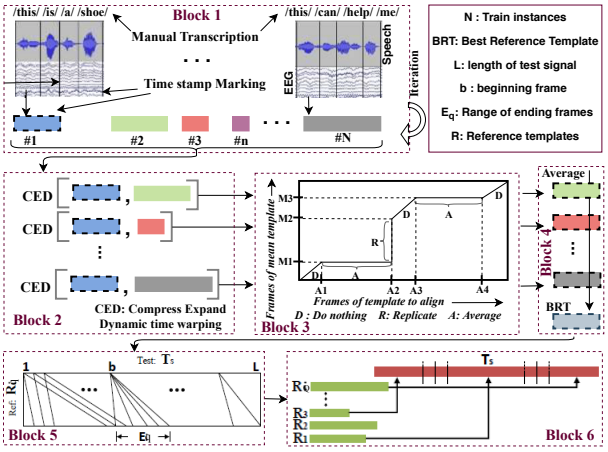
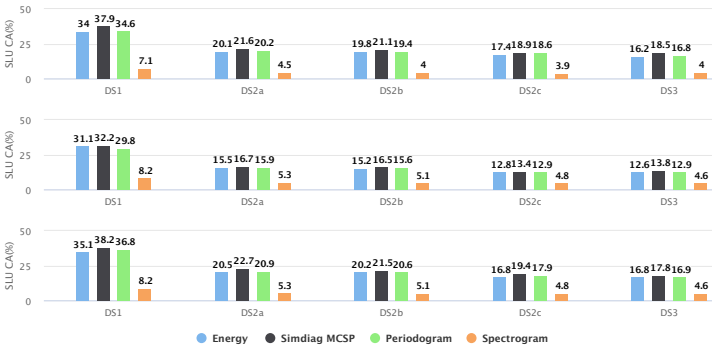


Figure 1: ***CWRT based 2LDP***: ***Block 1:*** *EEG data initial segmentation using manual markers obtained from the input/output speech signals, followed by iterative boundary correction.* ***Block 2,3:*** *CED algorithm to make all the templates equilength.* ***Block 4:*** *Average across the equilength templates to obtain the class-wise best reference template(BRT).* ***Block 5:*** *First level 2LDP distance score calculation* ***Block 6:*** *Second level path tracking and allocation of class labels.*

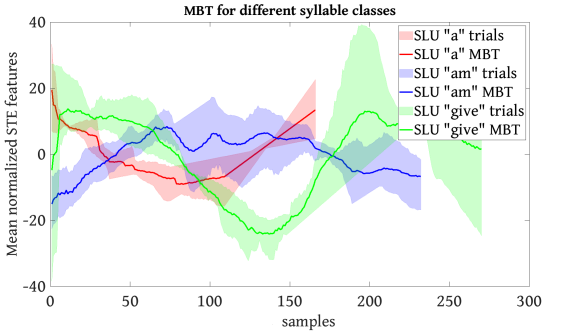
2. Results

2.1. Current Status of work

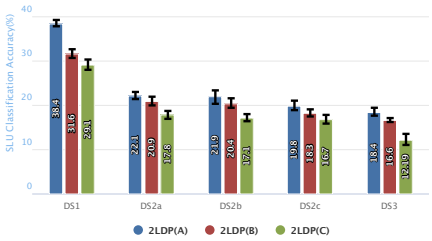
Neural decoding of speech using non-invasive techniques necessitates optimal choice of signal analysis and translation protocols. By employing selection-by-exclusion based temporal modelling strategies, an optimum feature-model pair was chosen for the task of speech-EEG unit classification[16]. Features derived from short term energy(STE), periodogram, spectrogram and multi-class common spatial patterns(MCSP) were considered in conjunction with classifiers built using dynamic programming(2LDP), Gaussian mixture-hidden Markov models(GMM-HMM) and convolutional neural networks(CNN)(Figure 2a). Since our datasets involve multiple sessions and subjects, the variability induced by them is also addressed by formulating three cases, namely, intra-subject+intra-session(Case-A), intra-subject+inter-session(Case-B) and inter-subject(Case-C). Real-time decoding devices require accurate operation in Case-C scenarios. The novel method proposed in this work using common word reference template(CWRT) matching algorithm coupled with 2LDP classifier as illustrated in Figure 1, performs best for Case-C. Input EEG data is segmented using markers from the corresponding continuous speech stimuli and CWRT is implemented to obtain one reference template from the many training templates across trials. Given the test signal and the references, 2LDP determines the sequence of patterns and their boundaries. At the first level, the algorithm matches each pattern with an arbitrary portion of T_s



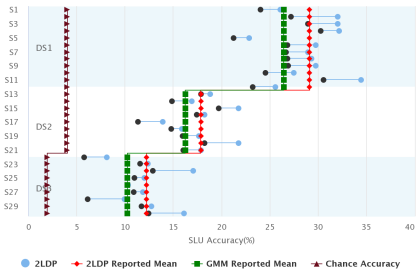
(a) *2LDP(top pane), GMM-HMM(middle pane), CNN(bottom pane)*



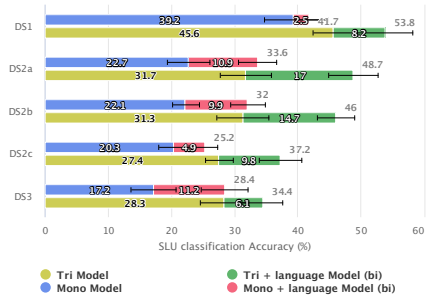
(b) *Visualization of temporal structures of syllable like units(SLU)*



(c) *2LDP- Case A,B,C generalization*



(d) *LOS accuracy for 30 subjects*



(e) *Acoustic and transcription context modelling*

Figure 2: *Visualization and classification of Syllable like units averaged over multiple subjects and sessions across five Datasets(DS1,DS2a,DS2b,DS2c,DS3).* ***a,****Classification accuracy comparison for feature-model pair variants,* ***b,*** *Mean Best Template(MBT) plotting for three syllable classes -"a", "am" and "give",* ***c,*** *Protocol Generalization abilities for 3 cases of testing,* ***d,*** *Leave one subject out testing performance across datasets using 2LDP and GMM-HMM classifiers,* ***e,*** *Accuracy boost provided by incorporating context modelling approaches.*

and generates a matrix of scores. The second level then pieces together the individual scores to minimize the overall accumulated distance and backtracks the optimal path and sequence of patterns matching T_s[17].

2.2. Statistical Results

DS1, DS2 and DS3 comprise of input sentence cues formed using a set of 25, 25 and 54 syllables respectively occurring in varying contexts. To differentiate between the different phases of speech-based cognition in DS2, we define DS2a as the hearing phase, DS2b as the imagining phase and DS2c as the speaking phase. EEG signals are classified in syllable, word, phrase and sentence levels to discern the fundamental unit that best captures distinct speech signatures. Classification of fundamental syllable-like units(SLU) yield best results and also posses unique temporal patterns when visualized(Figure 2b).

EEG signals from different electrode cap regions and bands are studied to understand their importance in speech-induced EEG. Results suggest that temporal and parietal region channels consistently perform better than channels extracted from other regions(average absolute accuracy gain of 1.85% and 1.68% over other regions respectively). Concerning the frequency bands, the gamma band gives the best classification performance(absolute gain of 2.62% over other bands). Given these observations, the two best performing regions and bands in combination were extracted and analysed across all datasets. Figure 2c compares the SLU classification accuracies using 2LDP for the three cases of generalization testing(inter/intra Sessions and Subjects). Leave-one-subject(LOS) out accuracy(Case C) is plotted for all 30 subjects in Figure 2d for the audition phase. The reported mean and chance accuracies for every dataset are also marked for comparison.

Transcription level modelling determines which linguistic paths are more probable than the others and helps improve the confidence and correctness of the decoded output. It is observed that incorporating a bi-gram context model built using the wall street journal(WSJ) text vocabulary greatly improved the performance(Figure 2e). 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 $\pm$ 1.7% across subjects as outlined by examples given in Table 1. The average GMM-HMM decode duration per test trial is 10sec$\pm\sigma$(3sec), making the protocol ideal for online decoding of speech EEG for the convenience of BCI users.

Significantly higher than chance accuracies are recorded for single trial multi-unit EEG classification using machine learning approaches over five datasets across 30 subjects. In addition to result-based experimentation, a variety of control checks are also 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 distinct speech EEG signatures.

2.3. Future Work

The transcribed outputs from the decoded EEG signals can be synthesised as speech using a trained text to speech synthesizer thus making it a functional communication interface. Further, experimental setup for online decoding of 4-channel MUSE imagined speech-EEG data is in progress. The data that support the findings of this study and details of their elicitation protocols are freely available in https://www.iitm.ac.in/donlab/cbr/cospeech_eeg_dataset/.

Table 1: *Single trial decoding of perception phase EEG signals.* ***o****: Original sentence,* ***d****: decoded sentence. Substitutions are highlighted, deletions are striked out and insertions are in red.*

Syllable error rate	Original sentences Vs decoded sentences
wd$\sim$50% od$\sim$30%	o: nice to meet and know you d: nice to meet ~~and know~~ you there
wd$\sim$60% od$\sim$40%	o: what is wrong with that d: there is wrong with him

3. References

[1] 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 research bulletin*, vol. 75, no. 6, pp. 796–803, 2008.

[2] J. Del R. Millán, P. W. Ferrez, F. Galán, E. Lew, and R. Chavarriaga, "Non-invasive brain-machine interaction," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 22, no. 05, pp. 959–972, 2008.

[3] 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 Transactions on Biomedical Engineering*, vol. 65, no. 10, pp. 2168–2177, 2017.

[4] C. Cooney, A. Korik, F. Raffaella, and D. Coyle, "Classification of imagined spoken word-pairs using convolutional neural networks," in *The 8th Graz BCI Conference, 2019*. Verlag der Technischen Universitat Graz, 2019, pp. 338–343.

[5] M. Rosinová, M. Lojka, J. Staš, and J. Juhár, "Voice command recognition using EEG signals," in *2017 International Symposium ELMAR*. IEEE, 2017, pp. 153–156.

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

[7] B. M. Idrees and O. Farooq, "EEG based vowel classification during speech imagery," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 2016, pp. 1130–1134.

[8] B. Min, J. Kim, H.-j. Park, and B. Lee, "Vowel imagery decoding toward silent speech BCI using extreme learning machine with electroencephalogram," *BioMed research international*, vol. 2016, 2016.

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

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

[11] 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.

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

[13] 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 in human neuroscience*, vol. 11, p. 150, 2017.

[14] R. A. Sharon, S. Aggarwal, P. Goel, R. Joshi, M. Sur, H. A. Murthy, and S. Ganapathy, "Level-wise subject adaptation to improve classification of motor and mental eeg tasks," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 6172–6175.

[15] R. A. Sharon, S. S. Narayanan, M. Sur, and A. H. Murthy, "Neural speech decoding during audition, imagination and production," *IEEE Access*, vol. 8, pp. 149 714–149 729, 2020.

[16] R. A. Sharon and H. A. Murthy, "Comparison of feature-model variants for cospeech-EEG classification," in *2020 National Conference on Communications (NCC)*. IEEE, 2020, pp. 1–6.

[17] 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 *ICASSP 2019-2019 IEEE*.