Translating neural signals to text using a Brain-Computer Interface

Anonymous Author(s)
Affiliation
Address
email

Abstract

Brain-Computer Interfaces (BCI) help patients with faltering communication abilities due to neurodegenerative diseases produce text or speech output by direct neural processing. However, their practical realization has proven difficult due to limitations in speed, accuracy, and generalizability of existing interfaces. To this end, we aim to create a BCI that decodes text directly from neural signals. We implement a framework that initially isolates frequency bands in the input signal encapsulating differential information regarding production of various phonemic classes. These bands form a feature set that feeds into an LSTM which discerns at each time point probability distributions across all phonemes uttered by a subject. Finally, a particle filtering algorithm temporally smooths these probabilities incorporating prior knowledge of the English language to output text corresponding to the decoded word. Further, in producing an output, we abstain from constraining the reconstructed word to be from a given bag-of-words, unlike previous studies. The empirical success of our proposed approach, offers promise for the employment of such an interface by patients in unfettered, naturalistic environments.

1 Introduction

Neurodegenerative diseases such as amyotrophic lateral sclerosis (ALS) restrict an individual's potential to fully engage with their surroundings by hindering communication abilities. Brain-Computer Interfaces (BCI) have long been envisioned to assist such patients as they bypass affected pathways and directly translating neural recordings into text or speech output. However, practical implementation of this technology has been hindered by limitations in speed and accuracy of existing systems [5]. Many patients rely on devices that use motor imagery [10], or on interfaces that require them to individually identify and spell out text characters such as the "point and click" cursor method [11, 12]. Despite significant work in system optimization, inherent limitations in their designs render them significantly slower than spoken communication.

To address these shortcomings, several studies are using electrocorticography (ECoG) and local field potential (LFP) signals [3]. These invasive approaches provide superior signal quality with high temporal and spatial accuracy. Previous work attempted translation to continuous phoneme sequences using invasive neural data [9]; however, despite their reported higher translation speed, their applications are limited to a reduced dictionary (10-100 words). Other design choices meant to enhance phoneme classification capitalize on prior knowledge of the target words, hindering their generalization to unmodified scenarios. Additionally, a recent study synthesized speech using recordings from the speech cortex. Though the authors demonstrate partial transferrability of their decoder amongst patients, their accuracy is again limited to selection of the reconstructed word by a listener from a pool of 25 words and worsens as the pool size increases [4].

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

Thus, establishing the capability of these approaches to generalize to unconstrained vocabularies is not obvious and has to our knowledge not yet been studied. Here, we present the performance of a two-part decoder network comprising of an LSTM and a particle filtering algorithm on data gathered from six patients. We provide empirical evidence that our interface achieves an average accuracy of 32% calculated against a full corpus i.e. one encompassing all feasible English words that can be formulated using the entire set of phonemes uttered by a patient, thus marking an important, non-incremental step in the direction of viability of such an interface.

2 Methods

2.1 Experimental Design

Data was obtained from neurosurgical patients implanted with intracranial depth electrodes to identify seizure foci for potential surgical treatment of epilepsy [14]. Implantation of electrodes in the relevant cortical areas was based on clinical need. Six subjects performed between 55 and 603 trials, repeating either words (”yes”, ”no”) or phoneme strings (monopthongal vowels with or without preceding consonants). Number of phonemes per subject consequently varied from 8 (3 consonants, 5 vowels) to 16 (11 consonants, 5 vowels). The sampling rate of these recordings was 30 kHz. Before further processing, electrodes determined to have low signal-to-noise ratio (SNR) were removed. Criterion for electrode removal was either that its time-series signal was uniformly zero, or that it contained artifacts that were atleast one order of magnitude larger than the mean absolute signal amplitude.

2.2 Feature Selection

In order to include as input to our network classifier differential information stored in the neural signals about production of various phonemes, an experiment was designed that mapped power in spectral bands of these recordings to the underlying phoneme pronunciation. While previous studies have used bands upto high gamma (70-150 Hz) for all speech uniformly, results tabulated in the prior work of Author et al. [2] show that there lies tuning to production of varied phonemes in a greater range of frequencies. Across multiple subjects vowels were captured using high frequency bands (> 600 Hz), meanwhile those lying between 150-400 Hz encapsulated consonant information.

2.3 LSTM Model Description

First part of our decoder is a stacked two-layer bLSTM. We use a bLSTM due to its ability to retain temporally distant dependencies when decoding a sequence [8]. Further, our analysis reveals that while a single-layer network can differentiate between phonemic classes such as nasals, semivowels, fricatives etc.; a two-layer model can distinguish between individual phonemes. There are 256 hidden units for each LSTM cell. The model is trained using ADAM optimizer to minimize weighted cross-entropy error, with weights inversely proportional to phoneme frequencies. The optimizer is initialized with $\beta 1$ = 0.9, $\beta 2$ = 0.999 and $\epsilon = 1e^{-8}$. Training occurs over 40 epochs with a learning rate of $1e^{-3}$. Using leave-one-trial-out, the recurrent network outputs a time sequence of probability distributions across all phonemes in the subject's dataset. Implementation was using Pytorch [1].

2.4 Language Model and temporal smoothing

A language model is used to apply prior knowledge about the expected output given the target domain of natural language. In this study, word frequencies were determined using the Brown corpus [6] translated into phonemic sequences using the CMU Pronouncing Dictionary [15]. A graphical representation of the English language is used that consists of states for every subsequence that starts a word in the corpus [13]. Each state links to every state that represents a superstring that is one character longer (Fig. 1).The language model accounts for the possibility of homophones by keeping a list of such words associated with each node in the model along with their relative frequency in the text corpus.

Laplacian smoothing is applied to the LSTM model output so that phonemes that were not seen during training are assigned a non-zero probability. A temporal model is then used to apply the language model to the resulting distributions. In this study, we applied a particle filtering (PF) method [13]. PF estimates the probability distribution of sequential outputs by creating a set of realities

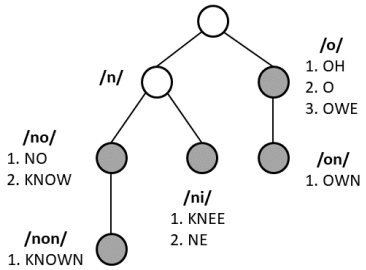

Figure 1: Example probabilistic automaton based on phonemes: /n/, /o/, and /i/. Shaded nadoes represent possible terminations for a word. Lists of words for a given node correspond to homophones.

(called particles) and projecting them through the model based on observed data and the language model [7]. Each particle contains a reference to a state in the model, a history of previous states, and an amount of time that it is going to remain in the current state. Distribution of states occupied by these particles represents an estimation of the true probability distribution.

3 Results

In our evaluation, output words were only considered correct if the phoneme sequence matched the labels and each phoneme overlaped at least partially with its respective audio label. Word accuracies varied between subjects, ranging from 54.55% (subject 1) to 13.46% (subject 2) (Table 1). On average, 32.16% of trials were classified completely correctly and an additional 23.06% had at least one phoneme match. Of the incorrect classifications 32.28% produced incorrect words either because none of the output phonemes were correct or because the sequences did not align temporally with the audio signal. In the remaining 12.49% of trials, the system did not detect speech signals, and produced an empty string as output.

Table 1: Word level performance of each subject.

Subject	ACC_W (%)	Partial (%)	Incorrect (%)	Omission (%)	Baseline (Random chance)
1	54.55	12.73	20.00	12.73	2.36
2	13.46	21.15	46.15	19.23	0.52
3	23.08	35.75	36.65	4.52	0.28
4	29.73	18.92	32.43	18.92	1.27
5	31.43	24.57	33.71	10.29	0.76
6	40.72	25.26	24.74	9.28	0.74
Average	32.16	23.06	32.28	12.49	0.99

4 Discussion

Each of the subjects in this study were able to communicate with significantly higher accuracy than chance. Nevertheless, the average word error rate seen in this study (67.8% on average) was higher than the 53% reported in [4]. There were several important differences in these studies, however. The primary difference is that their system produced an audio output that required a human listener to transcribe into a word selection. Despite advances in machine learning and natural language processing, humans have superior ability to use contextual information to find meaning in a signal. Furthermore, that study limited classifications to an output domain set of 50 words, which is generally not sufficient for a realistic communication system.

While this study makes a significant addition to existing BCI literature in terms of its avoidance of the traditional bag-of-words approach, our accuracies are lower than those reported in ERP-based BCI studies [12]. Moreover, in order for a BCI system based on translating neural signals to become a practical solution, improvements need to be made either in signal acquisition, machine learning translation, or user strategy. One approach could be to sacrifice some of the speed advantages by having users repeat words multiple times. While this would reduce communication speeds below

natural speeking rates, it would still greatly exceed ERP-based methods, while increasing the signals available for classification which could improve system accuracy.

4.1 Limitations and Future Work

The language model used in this study was designed to be general enough for application in a realistic BCI system. This generality may have been detrimental to its performance in the current study, however, as language models based on natural language will bias towards words that are common in everyday speech. The current study design produced many words that are infrequent in the training corpus. As a result, the language model actually biased away from this output, making it almost impossible to correctly classify. The results presented in this study are promising, but they represent offline performance which does not include several factors such as user feedback that occur in an online implementation.

5 Conclusion

The proposed system serves as a step in the direction of a generalized BCI system that can directly translate neural signals into written text in naturalistic scenarios. However, communication accuracies are currently insufficient for a practical BCI device, so future work must focus on improving these and developing an interface to present feedback to users.

References

[1] Code available on author's github page; data can be acquired upon personal request to authors.

[2] Identifying input features for development of real-time translation of neural signals to text. *Accepted to appear at Interspeech 2019*.

[3] H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani. Towards reconstructing intelligible speech from the human auditory cortex. *Scientific Reports*, 10:874, 2019.

[4] G. K. Anumanchipalli, J. Chartier, and E. F. Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019.

[5] L. A. Farwell and E. Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalogr. Clin. Neurophysiol*, 79(6), 1988.

[6] W. Francis and H. Kucera. Brown corpus manual dept of linguistics. Technical report, Brown University, 1979.

[7] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEEE Proceedings F (Radar and Signal Processing)*, pages 107–113, 1993.

[8] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Netw.*, 18:602–601, 2005.

[9] C. Herff, D. Heger, A. Pesters, D. Telaar, P. Brunner, G. Schalk, and T. Schultz. Brain-to-text: decoding spoken phrases from phone representations in the brain. *Front. Neurosci.*, 9:217, 2015.

[10] D. J. McFarland, L. A. Miner, T. M. Vaughan, and J. R. Wolpaw. Mu and beta rhythm topographies during motor imagery and actual movements. *Brain Topogr.*, 12(3):177–186, 2000.

[11] C. Pandarinath, P. Nuyujukian, C. Blabe, B. Sorice, J. Saab, F. Willett, L. Hochberg, K. Shenoy, and J. Henderson. High performance communication by people with paralysis using an intracortical brain-computer interface. *eLife*, 6:e18554, 2017.

[12] W. Speier, C. Arnold, J. Lu, R. K. Taira, and N. Pouratian. Natural language processing with dynamic classification improves p300 speller accuracy and bit rate. *J. Neural Eng.*, page 016004, 2011.

[13] W. Speier, C. Arnold, A. Deshpande, J. Knall, and N. Pouratian. Incorporating advanced language models into the p300 speller using particle filtering. *J. Neural Eng.*, 12:046018, 2015.

[14] A. Tankus, I. Fried, and S. Shoham. Structured neuronal encoding and decoding of human speech features. *Nat. Commun.*, 3:1015, 2012.

[15] R. L. Weide. The cmu pronouncing dictionary. Technical report, Carnegie Melon University, 1998.