Developing a real-time translator from neural signals to text: An articulatory phonetics approach

Lindy Comstock
Dept. of Applied Linguistics
University of California,
Los Angeles, CA, USA

Ariel Tankus
Dept. of Neurology and Neurosurgery
Tel Aviv University
Tel Aviv, ISR

Michelle Tran
Dept. of Neurosurgery
University of California,
Los Angeles, CA, USA

Nader Pouratian
Dept. of Neurosurgery
University of California,
Los Angeles, CA, USA

Itzhak Fried
Dept. of Neurosurgery
University of California,
Los Angeles, CA, USA

William Speier
Dept. of Radiological Sciences
University of California,
Los Angeles, CA, USA

1 Introduction

Practical implementation of brain-computer interface (BCI) technology has been hindered by limitations in the speed and accuracy of existing systems. Such systems primarily utilize the P300 evoked response potential (ERP) to identify a target character during repetitive serial presentations of possible characters (Farwell and Donchin, 1988). The use of a visual array means "P300 spellers" are available primarily to patients with gaze control. Moreover, accurate results come at the price of typing speed: no more than 15-19 characters per minute in healthy subjects (Townsend and Platsko, 2016; Speier et al., 2018), which has been deemed insufficient in patient surveys (Huggins et al., 2011). Furthermore, the device requires considerable time to set up and calibrate for each session (> 20 min.).

Electrocorticography (ECoG) and local field potential (LFP) signals provide superior data in that neural recordings are taken directly on top of the cortex or within the cortical layer. This allows single cell events to be recorded with great temporal and spatial accuracy, and the signals can be analyzed without external stimulus presentation. To date, only two studies have attempted translation from neural signals to phoneme sequences by means of continuous classification of invasive neural data (Herff et al., 2015; Moses et al., 2016).

While invasive systems promise better translation speed, the accuracy of these systems remains high only with use of a reduced dictionary (10-100 words). Other design features intended to facilitate phoneme classification limit applicability. Moses et al. (2016) made use of phoneme onset time relative utterance onset, and Herff et al., (2015) labeled phonemes by means of speech-recognition software prior to feature extraction; both approaches are capitalize on prior linguistic knowledge, making the schemes insufficient for classification of unmodified, naturalistic speech.

Thus, an important challenge to decoding naturalistic speech lies in identifying the appropriate inputs for a classificatory scheme. Data manipulation must not aid performance, and the number of inputs the classifier must learn should be few enough to be learned rapidly, yet extensible to the range of words used in naturalistic speech.

2 Current study

This study investigates whether isolating the neural signal of motor movements in articulatory phonetics provides more reliable inputs for classification than phoneme classes. Individual phonemes are defined by feature sets: while the set is unique, individual features overlap. Orienting to articulatory features can account for the similarities and differences that arise in the neural signals of each phoneme, resulting in better detection quality with a limited number of inputs. A classification scheme based on motor movements avoids distortion or ambiguity in recorded speech (Ohala, 1981) and interference from acoustic feedback during self-produced speech (Houde et al., 2002).

We utilize data from LFP signals, which can be processed more quickly and acquired with greater accurately than ERP data, and machine methods adapted from previous brain-computer interface research to improve classification performance (Speier et al., 2011; Speier et al., 2013).

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 322-325.
New York City, New York, January 3-6, 2019

3 Method

Phonemes were assigned a series of numbers representing their place and manner of articulation (Table 1). A subset of phonemes in different contexts was tested to determine if context-dependent phonetic models such as triphones (Jurafsky, 2000) are relevant for neural representations.

3.1 Speech stimuli

Three subjects performed between 56 and 603 trials, repeating words ("yes", "no") or phoneme strings (single vowels with and without preceding consonants). The inventory of phonemes pronounced varied from eight (three consonants, five vowels) to 16 (11 consonants, five vowels) depending on the number of trials.

3.2 Neural recordings

Data was obtained from neurosurgical patients implanted with intracranial depth electrodes to identify seizure foci for potential surgical treatment of epilepsy (Tankus et al., 2012). LFPs were recorded from microwires in temporal and frontal lobe sites. Preprocessing of the data followed the procedure outlined by Moses et al. (2016).

3.3 Feature selection

Neural signals converted into feature vectors were used to train a convolutional neural network architecture for the classification process (Figure 1). The statistical probability of a phoneme was calculated based on speech corpora (Weide, 2005; Francis and Kucera, 1979), and classifiers providing the optimal performance for probability distributions over phonemes at each time point were then generated by means of a particle filter and model of natural language. Our model takes into account frequency components, signal latency, and the context of phonemic representations.

Phoneme	Reference	Position	Height	Rounding	Tenseness
/i/	beet	1	1	1	1
/ɪ/	bit	2	2	1	2
/eɪ/	bait	1	3	1	1
/ɛ/	bet	1	5	1	2
/æ/	bat	1	6	1	2
/ə/	about	3	4	1	2
/ʌ/	but	5	5	1	2
/ɑ/	cot	5	7	1	1
/u/	boot	5	1	2	1
/ʊ/	book	4	2	2	2
/oʊ/	boat	5	3	2	1
/ɔ/	bought	5	5	2	2
/aʊ/	cow	5+4	7+2	2	1
/aɪ/	hide	5+2	7+2	1	1
/ɔɪ/	toy	5+2	5+2	2	1

Table 1: Articulatory features for vowel phonemes in the CMU pronouncing dictionary. Tongue position varies from front (1) to back (5); tongue height varies from close (1) to open (7); lip rounding is rounded (1) or unrounded (2); and tenseness can be tense (1) or lax (2).

3.4 Evaluation metrics

Trial accuracy is the number of trials classified correctly in entirety by the model, divided by the total number of trials. Phoneme sequences must match their word set label and each phoneme must overlap with its label. Classification can be considered either a true positive (i.e., correct phoneme overlapping the label), false positive (i.e., classified phoneme that either doesn't match the corresponding label or occurs during silence), or false negative (i.e., no detected phoneme during a label). Phoneme precision is the proportion of classified phonemes that match known labels, and phoneme recall is the proportion of phonemes that were correctly detected.

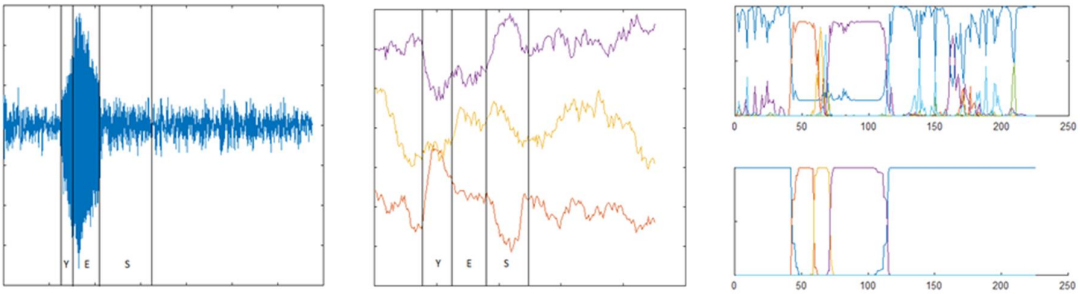

Figure 1: Machine learning process for decoding speech. Ground truth values are determined by manually labeling phonemes in the audio signal (a). Feature vectors are created at each time point based on the associated neural signal (b) and a probability distribution over phonemes is created using linear discriminant analysis and a Gaussian mixture model (c). Probabilities are smoothed using a temporal process model and prior knowledge of probabilities in natural language (d).

4 Results

Inter-subject variation in the number of phonemes obtained precluded an exact comparison across subjects. Nonetheless, we observed that for each subject classification accuracy was higher for vowels within words, suggesting phonetic context does play a role in neural signals (Table 2).

Isolated vowels	Word vowels
71.8%	87.5%
31.9%	34.8%
41.1%	50.0%

Table 2: Classification accuracy of vowels in isolation versus vowels occurring in words.

These results indicate the addition of a consonant improved vowel accuracy, even when fewer instances of a type were available for classifier training. Consonants reflect better classification performance than vowels. This is indicated by precision and recall metrics, averaged across subjects (Table 3).

Place of articulation was most informative for phoneme precision and exhibited greater overall consistency. Articulation that required tongue movements (e.g., alveolars) produced the greatest precision. An interaction between place and manner of articulation likely played a role in the classification of stops and fricatives. Manner of articulation was most informative for phoneme recall. This metric showed wide variation, with strong results for "vowel-like" nasals and liquids. Voiceless consonants were detected least often.

5 Discussion

The results of our study suggest that phenomena relevant to the overt pronunciation of phonemes, such as co-articulation in context and the articulatory feature set of phonemes, may be encoded in their neural representation. These findings allow for speculation as to the extent to which such neural signals may also be incorporated into the human brain's system of speech discrimination, perhaps lending renewed support for a revised version of the motor theory of perception; that is, the idea that the identification of vocal tract gestures contributes to speech perception (e.g., Galantucci et al., 2006; Liberman et al., 1967).

Similarly, it is notable that even in the neural representation of articulatory features, vowels are easier to detect, but more difficult to reliably classify. These finding parallels those in speech perception research, which finds the isolation of vowels to be substantially more difficult, as vowels lack clear temporal markers of onset (Johnson, 1988; Hermes, 1990). Further research may reveal whether parallels exist between the relative stability of a phoneme's articulatory feature set over time and the robustness of neurological encoding.

6 Future directions

Future directions include full integration of this post-hoc analysis into the machine learning model. Learning articulatory features rather than individual phonemes can significantly reduce the dimensionality of the model, allowing for more accurate estimation of probabilities and more efficient use of training data

Two innovations in data collection are planned. Firstly, individual consonant or consonant/vowel combinations will be contrasted in minimal pairs so as to isolate the neural signal of each feature (i.e., /z/ vs. /s/: *voiced/voiceless* alveolar fricative, /ð/ vs. /θ/: *voiced/voiceless* dental fricative, /z/ vs. /ð/: voiced *alveolar/dental* fricative, /s/ vs. /θ/: voiceless *alveolar/dental* fricative, etc.).

Phoneme	Articulation	Precision	Phoneme	Articulation	Recall
/d/	voiced alveolar stop	0.46	/m/	voiced bilabial nasal	0.63
/l/	voiced alveolar liquid	0.42	/l/	voiced alveolar liquid	0.51
/h/	voiceless glottal fricative	0.35	/r/	voiced palatal liquid	0.48
/m/	voiced bilabial nasal	0.30	/d/	voiced alveolar stop	0.27
/r/	voiced palatal liquid	0.30	/v/	voiced labiodental fricative	0.15
/v/	voiced labiodental fricative	0.29	/g/	voiced velar stop	0.15
/p/	voiceless bilabial stop	0.25	/h/	voiceless glottal fricative	0.13
/g/	voiced velar stop	0.20	/p/	voiceless bilabial stop	0.03

Table 3: Precision and recall for consonants and their associated articulatory features.

Secondly, a subset of silent trials have been collected and may be used to compare the classification scheme in covert versus overt language production to ascertain if motor signals remain relevant for covert speech. Subjects will be asked to silently mouth minimal pairs in addition to their covert and overt production, allowing articulation, pre-planning, and associated production noise to be teased apart in the neural signal.

A final consideration concerns verification of the results of the study in a larger subject pool and with patients who possess an alternative electrode placement. Electrodes were implanted based on clinical need, rather than according to any expectation of their optimal placement for the differentiation of articulatory feature sets. Alternative electrode placement may reveal that the robust results exhibited by certain features are merely indicative of more optimal electrode coverage of the area associated with control of that feature.

References

Farwell, Lawrence Ashley, and Emanuel Donchin. 1988. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical Neurophysiology*, *70*(6), 510-523.

Francis W. N. and Kucera H. 1979 *Brown Corpus Manual* (Providence, RI: Dept of Linguistics, Brown University)

Galantucci, Bruno, Carol A. Fowler, and Michael T. Turvey. 2006. The motor theory of speech perception reviewed. *Psychonomic bulletin & review*, *13*(3), 361-377. https://doi.org/10.3758/bf03193857.

Herff, Christian, Dominic Heger, Adriana De Pesters, Dominic Telaar, Peter Brunner, Gerwin Schalk, and Tanja Schultz. 2015. Brain-to-text: decoding spoken phrases from phone representations in the brain. *Frontiers in neuroscience*, *9*, 217. https://doi.org/10.3389/fnins.2015.00217.

Hermes, Dik. J. 1990. Vowel-onset detection. *The Journal of the Acoustical Society of America*, *87*(2), 866-873.

Houde, John F., Srikantan S. Nagarajan, Kensuke Sekihara, and Michael M. Merzenich. 2002. Modulation of the auditory cortex during speech: an MEG study. *Journal of cognitive neuroscience*, *14*(8), 1125-1138.

Huggins, Jane E., Patricia A. Wren, and Kirsten L. Gruis. 2011. What would brain-computer interface users want? Opinions and priorities of potential users with amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis*, *12*(5), 318-324. https://doi.org/10.3109/17482968.2011.572978.

Johnson, Keith A. 1988. *Processes of speaker normalization in vowel perception* (Doctoral dissertation, The Ohio State University).

Jurafsky, Dan. 2000. *Speech & language processing*. Pearson Education India.

Liberman, Alvin M., Franklin S. Cooper, Donald P. Shankweiler, and Michael Studdert-Kennedy. 1967. Perception of the speech code. *Psychological review*, *74*(6), 431. http://dx.doi.org/10.1037/h0020279.

Moses, David A., Nima Mesgarani, Matthew K. Leonard, and Edward F. Chang. 2016. Neural speech recognition: continuous phoneme decoding using spatiotemporal representations of human cortical activity. *Journal of neural engineering*, *13*(5), 056004. https://doi.org/10.1088/1741-2560/13/5/056004.

Ohala, John J. Articulatory constraints on the cognitive representation of speech. 1981. In *Advances in Psychology* (Vol. 7, pp. 111-122). North-Holland.

Speier, William, Corey Arnold, Nand Chandravadia, Dustin Roberts, Shrita Pendekanti, and Nader Pouratian. 2018. Improving P300 spelling rate using language models and predictive spelling. *Brain-Computer Interfaces*, *5*(1), 13-22. https://doi.org/10.1080/2326263X.2017.1410418.

Speier, William, Corey Arnold, Jessica Lu, Ricky K. Taira, and Nader Pouratian. 2011. Natural language processing with dynamic classification improves P300 speller accuracy and bit rate. *Journal of neural engineering*, *9*(1), 016004. https://doi.org/10.1088/1741-2560/9/1/016004.

Speier, William, Itzhak Fried, and Nader Pouratian. 2013. Improved P300 speller performance using electrocorticography, spectral features, and natural language processing. *Clinical Neurophysiology*, *124*(7), 1321-1328. https://doi.10.1016/j.clinph.2013.02.002.

Tankus, Ariel, Itzhak Fried, and Shy Shoham. 2012. Structured neuronal encoding and decoding of human speech features. *Nature Communications*, 3, 1015. https://dx.doi.org/10.1038%2Fncomms1995.

Townsend, G., and V. Platsko. 2016. Pushing the P300-based brain–computer interface beyond 100 bpm: Extending performance guided constraints into the temporal domain. *Journal of neural engineering*, *13*(2), 026024. https://doi:10.1088/1741-2560/13/2/026024.

Weide, Robert. 2005. The Carnegie Mellon pronouncing dictionary [cmudict. 0.]