Continuous Speech Recognition from ECoG

Dominic Heger[1*], *Christian Herff*[1*], *Adriana de Pesters*[2,4], *Dominic Telaar*[1], *Peter Brunner*[2,3], *Gerwin Schalk*[2,3,4], *Tanja Schultz*[1]

[1] Cognitive Systems Lab, Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2] National Center for Adaptive Neurotechnologies, Wadsworth Center, New York State Department of Health, Albany, NY, USA
[3] Department of Neurology, Albany Medical College, Albany, USA
[4] Department of Biomedical Sciences, State University of New York at Albany, Albany, NY, USA

* These authors contributed equally to this work
christian.herff@kit.edu, dominic.heger@kit.edu

Abstract

Continuous speech production is a highly complex process involving many parts of the human brain. To date, no fundamental representation that allows for decoding of continuous speech from neural signals has been presented. Here we show that techniques from automatic speech recognition can be applied to decode a textual representation of spoken words from neural signals. We model phones as the fundamental unit of the speech process in invasively measured brain activity (intracranial electrocorticographic (ECoG)) recordings. These phone models give insights into timings and locations of neural processes associated with the continuous production of speech and can be used in a speech recognizer to decode the neural data into their textual representations. When restricting the dictionary to small subsets, Word Error Rates as low as 25% can be achieved. As the brain activity data sets are fairly small, alternative approaches to Gaussian models are investigated by relying on robust, regularized discriminative models.

Index Terms: electrocorticography, ECoG, speech recognition, brain-computer interface

1. Introduction

Numerous members of the scientific community, including linguists, speech processing technologists, and computational neuroscientists have studied the basic principles of speech and analyzed its fundamental building blocks. However, the high complexity and agile dynamics in the brain make it challenging to investigate speech production with traditional neuroimaging techniques. Therefore, previous work has mostly focused on isolated aspects of speech in the brain, but so far not on the analysis and fully automatic decoding of brain activity during continuously produced natural speech.

Studies provided evidence for a neural representation of phones and phonetic features during speech perception [1, 2], but did not investigate continuous speech production. Other research studies investigated the dynamics of the general speech production process [3, 4], which we extend by illustrating the differences between phones in continuous production of speech. Neural activity during the production of isolated phones [5, 6, 7, 8, 9] or words [10] has been classified in different brain imaging techniques. Extending this idea, the imagined production of isolated phones was classified in [11]. [12] recently demonstrated the classification of a full set of phones within manually segmented boundaries during isolated word production. In [13], we have shown that techniques from Automatic Speech Recognition (ASR) can be applied to neural data to decode a textual representation from intracranial electrocorticographic (ECoG) recordings. In addition, we show in this paper that phones can be modeled with traditional Gaussian models and compare them with a new modeling technique based on sparse robust discriminative optimization by our *DCR Framework* [14].

2. Material and methods

2.1. Participants and electrode placement

Seven epileptic patients (4 female) who underwent neurosurgical procedures for epilepsy at Albany Medical Center (Albany, New York, USA) participated in this study. All participants gave informed consent to participate in the study. The study was approved by the Institutional Review Board of Albany Medical College and the Human Research Protections Office of the US Army Medical Research and Materiel Command. Participants' ages varied between 18 and 56 (mean age of 31.0).
Electrodes were placed depending only on clinical needs of the patients. Implanted electrodes were on the left hemisphere and covered parts of the frontal and temporal lobes for all subjects. Electrode grids (Ad-Tech Medical Corp., Racine, WI; PMT Corporation, Chanhassen, MN) consisting of platinum-iridium electrodes (4 mm in diameter, 2.3 mm exposed) with distances of 0.6-1 cm, which were embedded in silicone were used. In a post-operative CT scan, electrode positions were registered and co-registered with a pre-operative MRI scan. To be able to compare activations across subjects, electrode positions of all subjects were co-registered in a common Talairach space [15]. Activation maps were rendered using the NeuralAct software package [16]. See Figure 1 for electrode placement of all subjects passing the data pre-selection process (see Section 2.2).

2.2. Experiment and data pre-selection

In this study, brain activity during overt speech production of seven participants was recorded using electrocorticographic

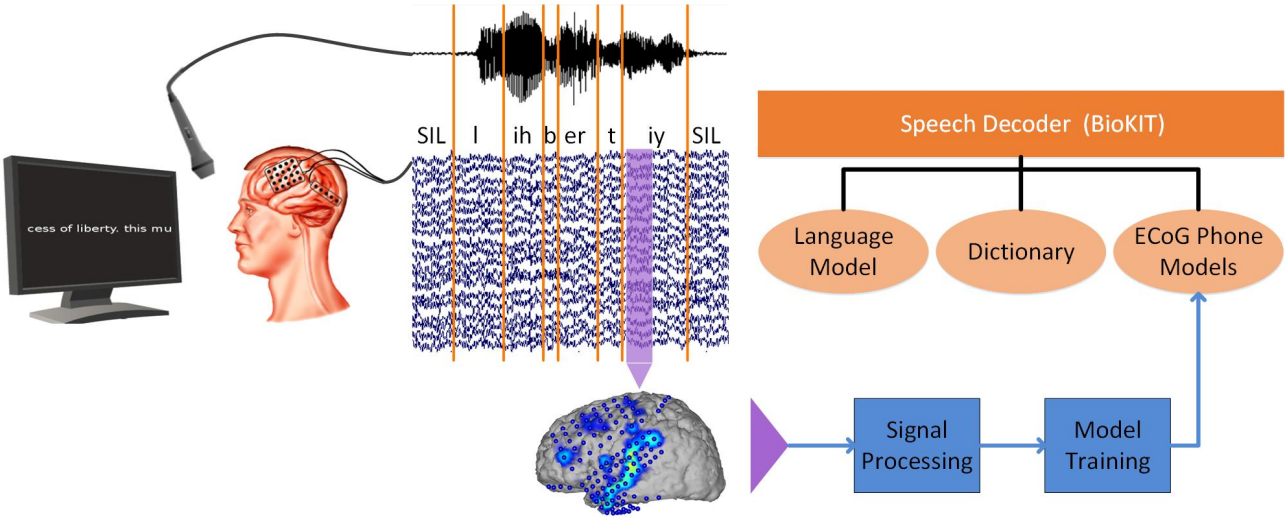

Figure 2: *Data recording and model training. Acoustic data and ECoG data are recorded synchronously. Acoustic data is then labeled on phone-level using BioKIT. Labels from the acoustic data are then imposed on the neural data. After pre-processing, phone models are trained from the neural data. These models are combined with a dictionary and a language model for automatic speech recognition based on neural data.*

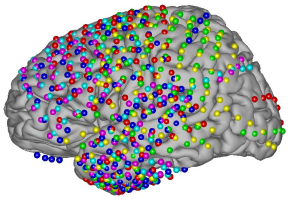

Figure 1: *Combined electrode montage of all participants after pre-selection. Participant 1 (yellow), participant 2 (magenta), participant 3 (cyan), participant 5 (red), participant 6 (green) and participant 7 (blue). Participant 4 did not yield sufficient activations related to speech activity and thus was excluded from combined analysis.*

(ECoG) grids that had been implanted as part of presurgical procedures preparatory to epilepsy surgery. ECoG measures electrical potentials directly on the brain surface with high temporal and spatial resolution. Due to the location directly on the brain surface, signals are unfiltered by skull and scalp. We used BCI2000 [17] and eight 16-channel g.USBamp biosignal amplifiers (g.tec, Graz, Austria) to record ECoG signals in this study. Additionally, we recorded the acoustic waveform of the participants' speech in synchronization with the ECoG signals. Both ECoG and acoustic signals were digitized at 9600 Hz.

In the experiment, participants had to read out text excerpts that were displayed on a screen about one meter in front of the participant. Text consisted of historical political speeches (i.e., Gettysburg Address [18]), JFK's Inaugural Address [19], a childrens' story (Humpty Dumpty [20]) or *Charmed* fan-fiction [21]. The texts scrolled through the screen from right to left at a constant rate, which was adjusted to the participants comfort (rate of scrolling text: 42-76 words/min). Subjects had to read the displayed text aloud as it appeared. Sessions were repeated 2-3 times. Table 1 summarizes data recording details for every session.

The recorded texts were cut along pauses into 21 to 49 phrases, depending on the session length. We used our in-house

Table 1: *Details for every recording session. Texts are abbreviated as follows: GA is Gettysburgh address, JFK is John F. Kennedy's inaugural speech, HD is Humpty Dumpty and Charmed are Charmed fan-fiction texts.*

Participant	Session	Text	|Phrases|	Recording length (s)
1	1	GA	36	279.87
	2	JFK	38	326.90
2	1	HD	21	129.87
	2	HD	21	129.07
	3	HD	21	126.37
3	1	Charmed	42	310.27
	2	Charmed	40	310.93
	3	Charmed	41	307.50
4	1	GA	38	299.67
	2	GA	38	311.97
5	1	JFK	49	341.77
	2	GA	39	222.57
6	1	GA	38	302.83
7	1	JFK	48	590.10
	2	GA	38	391.43

speech recognition toolkit BioKIT [22] to phone-label the audio recordings (see Section 2.3). Figure 2 shows experimental setup and model training (see Section 2.6).

As a first step, we evaluated whether speech activity segments could be distinguished from segments with no speech activity. We fitted a Gaussian model to all feature vectors containing speech activity and one to feature vectors when the participant was not speaking. Timings of speech and non-speech segments were extracted from the audio recordings. In a leave-one-phrase-out validation, we then evaluated whether these models could be used to identify speech activity above chance level. Both sessions of participant 4 and session 2 of participant 5 did not show classification rates significantly above chance level (paired t-test, $p > 0.05$) and were excluded based on this analysis. The comparison against random activations was performed as described in Section 3.1.

2.3. Cross-modality phone labeling

Acoustic recordings were phone-labeled using an English ASR system, which was trained on broadcast news. The sequence of phones was calculated by Viterbi forced alignment given the transcribed texts and acoustic models of the ASR system. We then adapted the Gaussian mixture model (GMM)-based acoustic models using maximum likelihood linear regression (MLLR). Finally, we repeated the Viterbi forced alignment using the adapted models of each session. These final phone alignments were then imposed on the ECoG data.
As the training data sets are rather small, we reduced the amount of distinct phones by grouping similar phones together into 20 ECoG models.

2.4. Feature extraction

The neural signal data was continuously segmented into 50 ms intervals with 25 ms overlap. This enabled the capturing of the fast cortical process underlying phones and was still long enough to robustly extract broadband gamma (70 – 170 Hz) activity. Each segment was labeled with the corresponding phone from the audio labeling. To calculate features, we first removed linear trends in the raw signals from each channel. The signals were then down-sampled to 600 Hz. Noisy channels were identified and excluded. We used common average re-referencing on the remaining channels and used elliptic IIR low-pass and high-pass filters to represent broadband gamma activity. To attenuate the first harmonic of 60 Hz line noise, we applied an elliptic IIR notch filter. For each channel c and interval i, we calculated the signal energy $E_{i,c}$ and applied the logarithm. We concatenated the logarithmic broadband gamma power of all channels into one feature vector $E_i = [E_{i,1}, \ldots, E_{i,d}]$. Temporal dynamics and context information were integrated by including neighboring intervals up to 200 ms prior to and after the current interval. Context of similar sizes have been found relevant in other speech perception studies [23]. Resulting feature vectors were thus stacked with four feature vectors in the past and four in the future, i.e. $F_i = [E_{i-4}, \ldots, E_i, \ldots, E_{i+4}]^\top$.

2.5. Identification of relevant regions and times

ECoG recordings have high temporal and spatial resolution which allows us to trace the temporal dynamics of speech production in the brain. We investigate the cortical regions with high relevance by calculating the mean symmetrized Kullback-Leibler divergence (KL-div) among phone models for each recording position at every time interval. The Kullback-Leibler divergence (KL-div) can be interpreted as the amount of discriminability between the neural activity models in bits and can be calculated in closed form for normal distributions. We estimate the discriminability of a feature $E_{i,c}$ (log broadband gamma at a recording position for a specific time interval) by calculating the mean KL-div between all phone-pairs for this feature. The mean of all these divergences estimates the discriminability of the feature $E_{i,c}$ in bits. Figure 3 shows interpolated KL-divs on the combined electrode montage of all participants for 200 ms prior to the onset of phone production, at phone production and 200 ms after the phone production onset.

2.6. ECoG feature selection and phone model training

To limit model complexity for our Gaussian models, we selected features with the largest average distance between phone models based on the KL-div in the training data. The number of features was selected automatically based on the distributions of KL-div values. Feature selection was purely based on KL-divs in the training data and did not include any prior knowledge about suitable brain regions or time offsets. The feature space was further reduced by applying a linear discriminant analysis (LDA) using the phone labels to 20 dimensions.
We modeled each phone by a normal distribution. Thus, each phone is characterized by the mean broadband gamma activity and variance of the neural activity measurements at each electrode and time offset. Due to the extremely limited amount of data, we only trained one context-independent Gaussian model for each phone, despite the fact that context effects during speech production have been shown in neural data [24]. It is important to keep in mind that a modeling of phones does not contradict the representation of articulatory features during speech perception ([2, 25]) and production ([26, 27]) in neural recording, as various representations of acoustic phenomenons are likely. The general idea of the model training is depicted in Figure 2.

2.7. Decoding

To decode continuous speech from neural data, we used our in-house speech recognition toolkit BioKIT [22]. We substituted the acoustic model for our ECoG phone models, used a bi-gram language model estimated on the texts read by the participants and a standard English dictionary.

2.8. Vowel classification using discriminative models

In recent years, alternatives to Gaussian Mixture Models (GMM) have become popular for acoustic modeling. Specifically, deep neural networks (DNN, [28]) have become a de facto standard. As our data sets are very limited in size and DNNs are known to require large amounts of data, we investigate an alternative phone modeling approach that (1) uses discriminative models instead of the generative Gaussian models, (2) combines multiple sessions of the participants, and (3) models invariance against non-stationarities in the training process.
Using this modeling approach we evaluated the frame-wise classification of the five vowels (/a/, /e/, /i/, /o/, /u/) from neural data. This classification task is particularly relevant for speech decoding and challenging as vowels share multiple articulatory properties and their production involves similar motor actions. To the best of our knowledge, vowel classification from neural data has not been investigated for continuous speech before.
We applied the *DCR framework* [14], our new Brain-Computer Interface recognition framework using joint convex optimization that is particularly designed to handle small amounts of data by learning sparse discriminative models. In the *DCR framework*, multiple so-called robustness directions in the feature space can be defined, whose influence is reduced during the optimization. This enables to learn models that are robust against signal variabilities, such as signal changes between multiple recording sessions and changes of the feature distributions over time (non-stationarities). To incorporate invariance against non-stationarities, we chose the robustness directions as follows (similar idea as in [29]): For each of the five vowels, we split the training features of a vowel c into 10 blocks of equal length per session, and calculated the mean feature vectors B_k^c for each block. We set the robustness directions d_k^c as the difference between the average feature vector for this vowel in the training data μ^c and the average feature vector of each of the blocks in the training data set B_k^c, i.e. $d_k^c = \mu^c - B_k^c$. In a joint optimization, sparse discriminative models are trained that are regularized to be invariant in the robustness directions, i.e. are

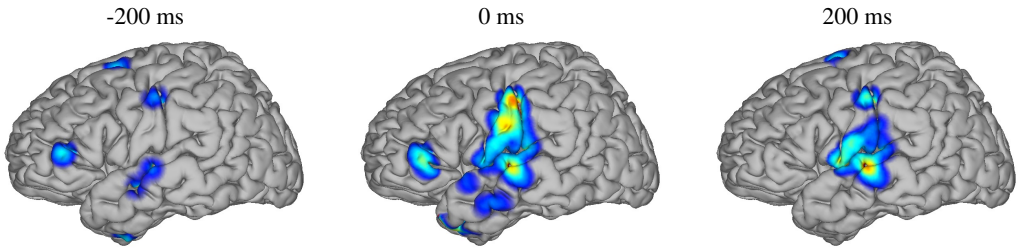

Figure 3: *Temporal course of regions with high discriminability between phone models. Heat maps show regions of high discriminability (red) according to mean Kullback-Leibler Divergences between models on a combined electrode montage of all participants (Talairach space) that exceed chance level (larger than 99% of randomized discriminabilities). Starting 200 ms before the actual phone production, early differences are present in diverse areas. Concurrent with production, high discriminability in sensorimotor areas can be observed. 200 ms after production, regions of highest discriminabilities correspond to auditory regions of the superior temporal gyrus.*

invariant to within-class fluctuations in the data.

3. Results

3.1. Word decoding results

We evaluated our models in a leave-one-phrase out cross-validation. To obtain a robust baseline for chance levels, we randomized the labels by shifting the ECoG features by half of the recording length. This way, data still showed typical ECoG behavior and label priors remained unchanged, but no correspondence between data and labels should be found. Figure 4 shows Word Error Rates (WER) for participant 7 session 1 (other participants are omitted due to the limited space available) for different dictionary sizes when decoding continuous speech from neural data using Gaussian models. All WER are significantly lower than the random models (paired t-test, $p < 0.001$). Figure 4 also shows phone true positive rates extracted from the most likely phone path, compared to the acoustic labeling. Again, our models are significantly better than chance level (paired t-test, $p < 0.001$). Average true positive rates remain stable over all dictionary sizes.

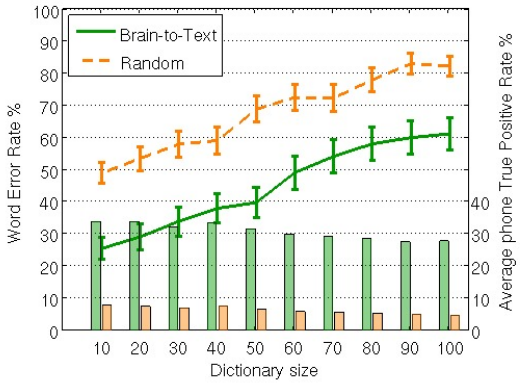

Figure 4: *Word Error Rates depending on dictionary size. ECoG speech models (green line) outperform random models (red line) for all dictionary sizes. Bars (ECoG models green, random models red) depict average true-positive rates across phones depending on dictionary size.*

3.2. Vowel Classification

For vowel classification, features were calculated as described in section 2.4 for the Gaussian models. For the discriminative models, no KLdiv feature selection and LDA compression has been performed as the *DCR framework* implicitly performs a feature selection by ℓ_1-norm regularization. The feature vectors were restricted to vowel frames and classified frame-wise using a 10-fold cross-validation with splits between phrases. Multi-class classification was performed using the 1-vs-rest strategy. Figure 5 shows the recognition results for the six participants in terms of f-scores weighted by the prior distribution of the phones. Whiskers indicate standard deviations across the different vowels. Randomization tests showed that all recognition results were significantly above chance level, except for participant 6 (one-sided, paried Wilcoxon signed rank tests, $p < 0.05$). The classification using the *DCR framework* shows improvements in recognition rates over Gaussian models for all participants except participant 1. It achieved significant improvements in weighted f-score over Gaussian models (participants 2, 3 and 5, paired Wilcoxon signed rank tests, $p < 0.05$) by up to 6.8% (absolute) and 2.8% (absolute) on average.

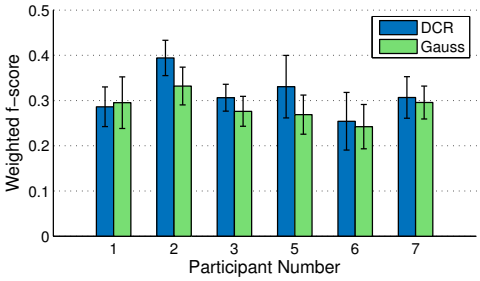

Figure 5: *Weighted f-scores of frame-wise vowel classification. Figure shows results of the DCR framework models (blue) in comparison to Gaussian models (green). Whiskers indicate standard deviations across the different vowels.*

4. Conclusion

In this paper we showed that techniques from ASR can be used to continuously decode speech from neural data. While good WER can be achieved for small dictionary sizes, they increase drastically with increasing size. We therefore investigated an alternative approach for ECoG phone models. The evaluation of frame-wise vowel classification using the *DCR framework* showed promising results.

5. Acknowledgements

We would like to thank Dr. Anthony Ritaccio for patient interaction, Dr. Aysegul Gunduz for support in data recording and Dr. Cuntai Guan for valuable discussions.

6. References

[1] E. F. Chang, J. W. Rieger, K. Johnson, M. S. Berger, N. M. Barbaro, and R. T. Knight, "Categorical speech representation in human superior temporal gyrus," *Nature neuroscience*, vol. 13, no. 11, pp. 1428–1432, 2010.

[2] N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang, "Phonetic feature encoding in human superior temporal gyrus," *Science*, p. 1245994, 2014.

[3] N. Crone, L. Hao, J. Hart, D. Boatman, R. Lesser, R. Irizarry, and B. Gordon, "Electrocorticographic gamma activity during word production in spoken and sign language," *Neurology*, vol. 57, no. 11, pp. 2045–2053, 2001.

[4] N. E. Crone, D. Boatman, B. Gordon, and L. Hao, "Induced electrocorticographic gamma activity during auditory perception," *Clinical Neurophysiology*, vol. 112, no. 4, pp. 565–582, 2001.

[5] E. C. Leuthardt, C. Gaona, M. Sharma, N. Szrama, J. Roland, Z. Freudenberg, J. Solis, J. Breshears, and G. Schalk, "Using the electrocorticographic speech network to control a brain–computer interface in humans," *Journal of neural engineering*, vol. 8, no. 3, p. 036004, 2011.

[6] F. H. Guenther, J. S. Brumberg, E. J. Wright, A. Nieto-Castanon, J. A. Tourville, M. Panko, R. Law, S. A. Siebert, J. L. Bartels, D. S. Andreasen *et al.*, "A wireless brain-machine interface for real-time speech synthesis," *PloS one*, vol. 4, no. 12, p. e8218, 2009.

[7] E. Formisano, F. De Martino, M. Bonte, and R. Goebel, ""who" is saying "what"? brain-based decoding of human voice and speech," *Science*, vol. 322, no. 5903, pp. 970–973, 2008.

[8] T. Blakely, K. J. Miller, R. P. Rao, M. D. Holmes, and J. G. Ojemann, "Localization and classification of phonemes using high spatial resolution electrocorticography (ECoG) grids," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*. IEEE, 2008, pp. 4964–4967.

[9] X. Pei, D. L. Barbour, E. C. Leuthardt, and G. Schalk, "Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans," *Journal of neural engineering*, vol. 8, no. 4, p. 046028, 2011.

[10] S. Kellis, K. Miller, K. Thomson, R. Brown, P. House, and B. Greger, "Decoding spoken words using local field potentials recorded from the cortical surface," *Journal of neural engineering*, vol. 7, no. 5, p. 056007, 2010.

[11] J. S. Brumberg, E. J. Wright, D. S. Andreasen, F. H. Guenther, and P. R. Kennedy, "Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex," *Frontiers in neuroscience*, vol. 5, 2011.

[12] E. M. Mugler, J. L. Patton, R. D. Flint, Z. A. Wright, S. U. Schuele, J. Rosenow, J. J. Shih, D. J. Krusienski, and M. W. Slutzky, "Direct classification of all american english phonemes using signals from functional speech motor cortex," *Journal of Neural Engineering*, vol. 11, no. 3, p. 035015, 2014. [Online]. Available: http://stacks.iop.org/1741-2552/11/i=3/a=035015

[13] C. Herff, D. Heger, A. de Pesters, D. Telaar, P. Brunner, G. Schalk, and T. Schultz, "Brain-to-text: Decoding spoken phrases from phone representations in the brain," *Frontiers in Neuroscience*, vol. 9, no. 217, 2015.

[14] D. Heger, C. Herff, F. Putze, and T. Schultz, "Joint optimization for discriminative, compact and robust brain-computer interfacing," in *Neural Engineering (NER), 2015 7th International IEEE/EMBS Conference on*, April 2015.

[15] J. Talairach and P. Tournoux, *Co-planar stereotaxic atlas of the human brain. 3-Dimensional proportional system: an approach to cerebral imaging*. Thieme, 1988.

[16] J. Kubanek and G. Schalk, "Neuralact: A tool to visualize electrocortical (ecog) activity on a three-dimensional model of the cortex," *Neuroinformatics*, pp. 1–8, 2014.

[17] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "Bci2000: a general-purpose brain-computer interface (bci) system," *Biomedical Engineering, IEEE Transactions on*, vol. 51, no. 6, pp. 1034–1043, 2004.

[18] E. Roy and P. Basler, *The gettysburg address, in The Collected Works of Abraham Lincoln*. New Brunswick, NJ: Rutgers University Press, 1955.

[19] J. F. Kennedy, *Inaugural Addresses of the Presidents of the United States. (Washington, DC)*. Available online at: www.bartleby.com/124/, 1989.

[20] W. Crane, S. Gilbert, John, W. McConnell, S. Tenniel, John, H. Weir, and J. B. Zwecker, *Mother Gooses Nursery Rhymes. A Collection of Alphabets, Rhymes, Tales and Jingles*. London: George Routledge and Sons, 1867.

[21] unknown, *"Traitor among us" and "Split Feelings"*. available on https://www.fanfiction.net/, 2009.

[22] D. Telaar, M. Wand, D. Gehrig, F. Putze, C. Amma, D. Heger, N. T. Vu, M. Erhardt, T. Schlippe, M. Janke, C. Herff, and T. Schultz, "BioKit - Real-time decoder for biosignal processing." in *Interspeech*, 2014.

[23] N. T. Sahin, S. Pinker, S. S. Cash, D. Schomer, and E. Halgren, "Sequential processing of lexical, grammatical, and phonological information within brocas area," *Science*, vol. 326, no. 5951, pp. 445–449, 2009.

[24] E. Mugler, M. Goldrick, and M. Slutzky, "Cortical encoding of phonemic context during word production," in *Engineering in Medicine and Biology Society, 2014. EMBS 2014. 36th Annual International Conference of the IEEE*. IEEE, 2014.

[25] F. Pulvermüller, M. Huss, F. Kherif, F. M. del Prado Martin, O. Hauk, and Y. Shtyrov, "Motor cortex maps articulatory features of speech sounds," *Proceedings of the National Academy of Sciences*, vol. 103, no. 20, pp. 7865–7870, 2006.

[26] K. E. Bouchard, N. Mesgarani, K. Johnson, and E. F. Chang, "Functional organization of human sensorimotor cortex for speech articulation," *Nature*, vol. 495, no. 7441, pp. 327–332, 2013.

[27] F. Lotte, J. S. Brumberg, P. Brunner, A. Gunduz, A. L. Ritaccio, C. Guan, and G. Schalk, "Electrocorticographic representations of segmental features in continuous speech," *Frontiers in Human Neuroscience*, vol. 9, no. 97, 2015.

[28] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[29] P. von Bünau, F. C. Meinecke, F. C. Király, and K.-R. Müller, "Finding stationary subspaces in multivariate time series," *Physical Review Letters*, vol. 103, no. 21, p. 214101, 2009.