Predicting mel-frequency cepstral coefficients from electrocorticographic signals during continuous speech production

Shreya Chakrabarti, Dean J. Krusienski, Gerwin Schalk, Jonathan S. Brumberg

Recent electrocorticography (ECoG) studies have investigated the neural bases for sensory and motor processes underlying speech and language [1-3]. Additional studies have successfully decoded words [6], phonemes [7,8], and speech acoustics [4,5] directly from intracranial recordings. In the present study, we employ techniques used in automatic speech recognition (ASR) as an alternative framework for predicting speech from neurological activity. Specifically, we decode Mel-frequency cepstral coefficients (MFCCs) of speech [9] from ECoG high-gamma band power using a spatial linear regression model. Our main goal is to investigate the differential contributions of neurological activity preceding or following self-productions and their usefulness in a speech prosthesis.

Methods: ECoG activity (58-120 electrodes per patient) was recorded from eight epileptic patients as they spoke aloud text that scrolled across computer screen. The ECoG and speech signals were simultaneously recorded at a sampling rate of 9600 Hz using g.USBamps (g.tec, Graz, AT). Twelve MFCCs were estimated using standard techniques [9] in Matlab. ECoG signals were highpass filtered at 0.01 Hz, re-referenced using a common average reference and decimated to 400 Hz. A zero-phase FIR filter was applied to extract the gamma band (70-170 Hz) and the band-power envelope was computed using the Hilbert transform. Decoded MFCCs were computed using a spatial linear regression model with 25% of the ECoG electrodes having the highest gamma-band correlation to the true speech MFCCs. The regression models were computed to predict MFCCs (N=12) from the gamma band power at latencies -200ms, -100ms and 0ms (lead condition) and at latencies 0ms, 100ms and 200ms (lag condition) relative to resultant speech output. The models were 10-fold cross-validated and p-values for the correlations were computed using a randomization test.

Results: The results of this analysis show ECoG signals in both the lead and lag conditions are useful in the prediction of the MFCCs of continuous speech. Leads and lags represent motor preparation and auditory feedback processing, respectively. Statistically significant correlations of predicted MFCCs to the true MFCCs in both conditions ranged from 0.17 to 0.30 with an average correlation of 0.24 in the lead condition and 0.23 in the lag condition. These correlation values are similar to those found in [4], which validates our approach. The major implication of this work is a demonstration of our method for predicting speech, via MFCCs, using ECoG activity preceding observation of the speech acoustic output. Our method for predicting MFCCs from ECoG activity may subsequently be used as an input to either speech-to-text recognition systems or speech synthesizers as a neural prosthesis for transcription or speech, respectively for individuals with severe paralysis.

References

[1] E. Edwards et al., "Spatiotemporal imaging of cortical activation during verb generation and picture naming.," *NeuroImage*, vol. 50, no. 1, pp. 291–301, Mar. 2010.

[2] X. Pei et al., "Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition.," *NeuroImage*, vol. 54, no. 4, pp. 2960–72, Feb. 2011.

[3] V. L. Towle., "ECoG gamma activity during a language task: differentiating expressive and receptive speech areas.," *Brain*, vol. 131, no. Pt 8, pp. 2013–27, Aug. 2008.

[4] B. N. Pasley et al., "Reconstructing speech from human auditory cortex.," *PLoS Biology*, vol. 10, no. 1, p. e1001251, Jan. 2012.

[5] F. H. Guenther et al., "A Wireless Brain-Machine Interface for Real-Time Speech Synthesis," *PLoS ONE*, vol. 4, no. 12, p. e8218, 2009.

[6] S. Kellis et al., "Decoding spoken words using local field potentials recorded from the cortical surface.," *Journal of Neural Engineering*, vol. 7, no. 5, p. 056007, Oct. 2010.

[7] X. Pei et al., "Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans.," *Journal of Neural Engineering*, vol. 8, no. 4, p. 046028, Aug. 2011.

[8] J. S. Brumberg et al., "Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex," *Frontiers in Neuroscience*, vol. 5, no. 65, 2011.

[9] R. Vergin, D. O'Shaughnessy, and A. Farhat, "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 525–532, 1999.

This work was supported in part by the NIH (R03-DC011304, R01-EB00856, R01-EB006356), the US Army Research Office (W911NF-07-1-0415, W911NF-08-1-0216) and NSF (1064912).

S. Chakrabarti and D. J. Krusienski are with Old Dominion University, Norfolk, VA 23529, USA (dkrusien@odu.edu; schak001@odu.edu)
G. Schalk is with the Wadsworth Center, Albany, NY (email: schalk@wadsworth.org)
J. S. Brumberg is with the University of Kansas, Lawrence, KS 66045, USA (email: brumberg@ku.edu)