Decoding Lip Movements During Continuous Speech using Electrocorticography

9th International IEEE EMBS Conference on Neural Engineering
San Francisco, CA, USA, March 20 - 23 , 2019

Decoding Lip Movements During Continuous Speech using Electrocorticography

Srdjan Lesaja, Christian Herff, Garett D. Johnson, Jerry J. Shih, Tanja Schultz,
Dean J. Krusienski, *Senior Member, IEEE*

***Abstract*— Recent work has shown that it is possible to decode aspects of continuously-spoken speech from electrocorticographic (ECoG) signals recorded on the cortical surface. The ultimate objective is to develop a speech neuroprosthetic that can provide seamless, real-time synthesis of continuous speech directly from brain activity. Instead of decoding acoustic properties or classes of speech, such a neuroprosthetic might be realized by decoding articulator movements associated with speech production, as recent work highlights a representation of articulator movement in ECoG signals. The aim of this work is to investigate the neural correlates of speech-related lip movements from video recordings. We present how characteristics of lip movement can be decoded and lip-landmark positions can be predicted.**

I. Introduction

Neuroprosthesis based on the decoding of speech processes would present a fast and intuitive way of communication for severely paralyzed patients [1]. The electrocorticogram (ECoG) is an invasive measurement of the electrical potentials generated from the neocortex of the brain [2] that is particularly well suited for the decoding of such speech processes due to its high spatiatl and temporal resolution [3]. The long-term viability of ECoG recordings have been established in humans [4], making it a suitable platform for the development of neuroprosthetic devices. Recent studies have shown that it is possible to decode certain aspects of continuously-spoken speech from ECoG signals including phonemes [5], words [6] and sentences [7] or reconstruct accoustic properties of perceived [8] and produced speech [9], [10]. Alternatively, motor representations of speech might be a better target for reconstruction due to the localized representation in speech motor cortex [11]. Conant et al., used ultrasound and video monitoring of the supralarungeal actuators in conjunction with ECoG to investigate the relationship between the articulator movement kinematics and the neural activity in the ventral sensory-motor cortex [12], [13]. Other studies used inverse mapping of acoustics to articulatory features to investigate the representation of speech production [14], [15].

In the present study, we use simultaneous ECoG and video recordings to investigate the neural correlates of lip movements and develop predictive models of lip-movement features to predict lip landmark positions.

This work was supported 01GQ1602 (BMBF) and 1608140 (NSF).

S Lesaja and GD Johnson are with the Biomedical Engineering Program, Old Dominion University, Norfolk, VA, USA slesa001@odu.edu; gjohn037@odu.edu

C Herff is with the School for Mental Health and Neuroscience, Maastricht University, The Netherlands c.herff@maastrichtuniversity.nl

JJ Shih is with the Neurology Department, UCSD Health, San Diego, CA, USA jerryshih@ucsd.edu

T Schultz is with the Cognitive Systems Laboratory, University of Bremen, Germany tanja.schultz@uni-bremen.de

DJ Krusienski is with the Department of Biomedical Engineering, Virginia Commonwealth University, Richmond, VA, USA djkrusienski@vcu.edu

II. Methodology

A. Data Acquisition

A subject with intractable epilepsy was implanted with a 64-contact ECoG array (Figure 1) for clinical monitoring at Mayo Clinic Florida.

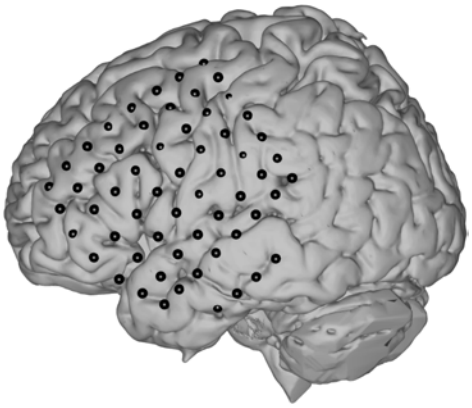

Fig. 1. Electrode locations.

The participant was asked to repeat sentences from the Harvard sentence corpus [16] that were presented both visually on a screen and aurally. The participant consented to participate in the study as approved by the IRB of both Mayo Clinic and ODU. During speech production, ECoG data, acoustic waveform and facial video were recorded and synchronized using BCI2000 [17]. ECoG data were digitized and sampled at 1200 Hz using g.tec gUSBamp biosignal amplifiers. Audio data were recorded using Snowball iCE microphone (Blue Microphones, California) and sampled at 48 kHz. *Sampling rates video* A diagram of the experimental setup is provided in Figure 2.

B. Signal Processing

A common-average reference (CAR) was applied to each ECoG channel and the resulting signals were band-pass filtered between 70 and 115 Hz using a zero-phase FIR filter in order to extract the gamma band. An FFT was performed

978-1-5386-7921-0/19/$31.00 ©2019 IEEE

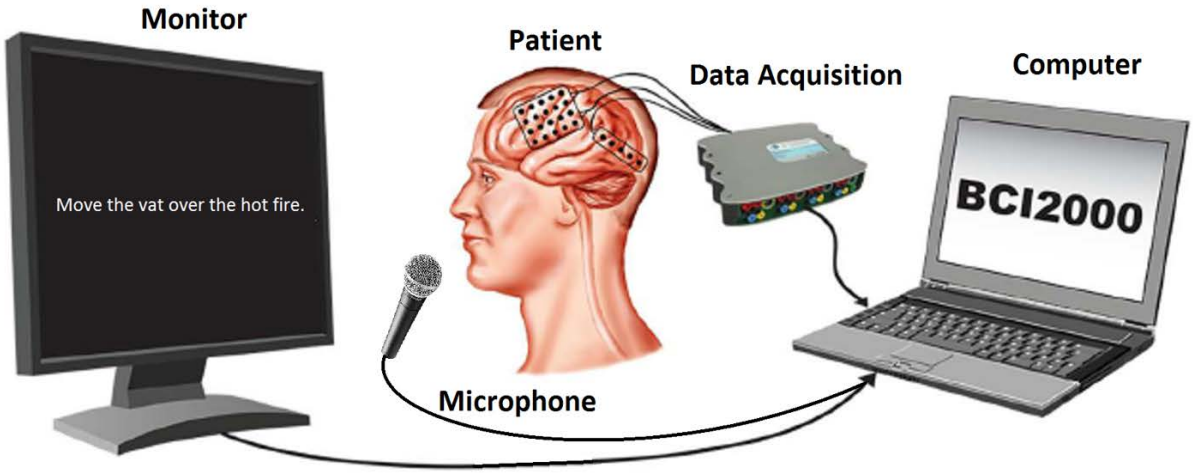

Fig. 2. Diagram of the experimental setup. *Figure does not contain video camera*

using 60 ms segments with 30 ms overlap, and the average of the average power of the entire gamma band was used as the feature for each channel.

C. Video Features

Dlib, an open source facial recognition library [18], was used to detect the subjects face in each video frame, and draw facial landmarks. Of 68 facial landmark points, 20 were used to outline the inner and outer lines of the lips, as shown in Figure 3. These 20 points were used to calculate four lip features: area of outer lip perimeter, area of inner lip perimeter, distance between outer top and bottom lips, distance between inner top and bottom lips.

Each frame of the video taken during the patient trial was input into the facial recognition library. If the library could find a face, then it would also assign the facial landmarks for the face, and the lip features could be extracted for that frame. Due to the inconsistent nature, and the angle of the patient's head position during trials made it so a face could not be detected by the library on all frames. We excluded trials in which the face could not be detected. This was the case for large portions of several trials. The analysis presented here is based on the two trials that had the greatest number of frames with detected faces, and thus, available facial landmarks.

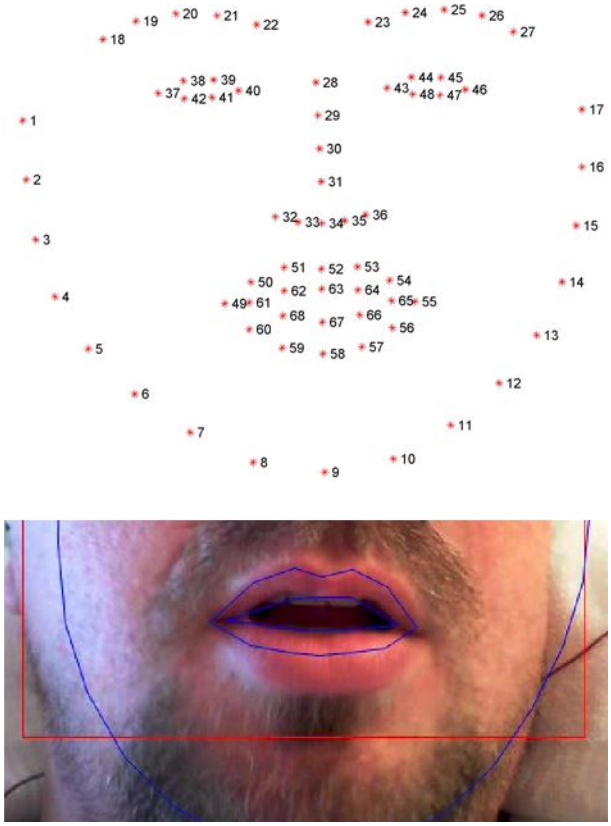

Fig. 3. The 68 facial landmarks defined in Dlib (top). Facial landmarks in blue superimposed on a video frame (bottom).

D. Correlations

The first step in the analysis was to assess the correlation between each lip feature and the average power features. To analyze whether and where the lip features were correlated to brain activity we calculated correlations by shifting the two time series signals. The correlation between the two unshifted series represents the correlation between a lip feature and the brain signals happening at that moment. In order to see the correlation between brain signals that occurred, for example 100 ms before the lip features, the time series were shifted by the appropriate time steps. Correlations for every electrode location were calculated in this manner at 25 ms time lags, from 1000 ms before to 1500 ms after the current reference frame.

To ensure the correlations were significant, a randomization test was performed. The normal maximum likelihood parameters were estimated for each electrode signal. One thousand simulated random electrode signals were generated and the correlations were computed. These simulated correlations were used to construct an empirical distribution. Each

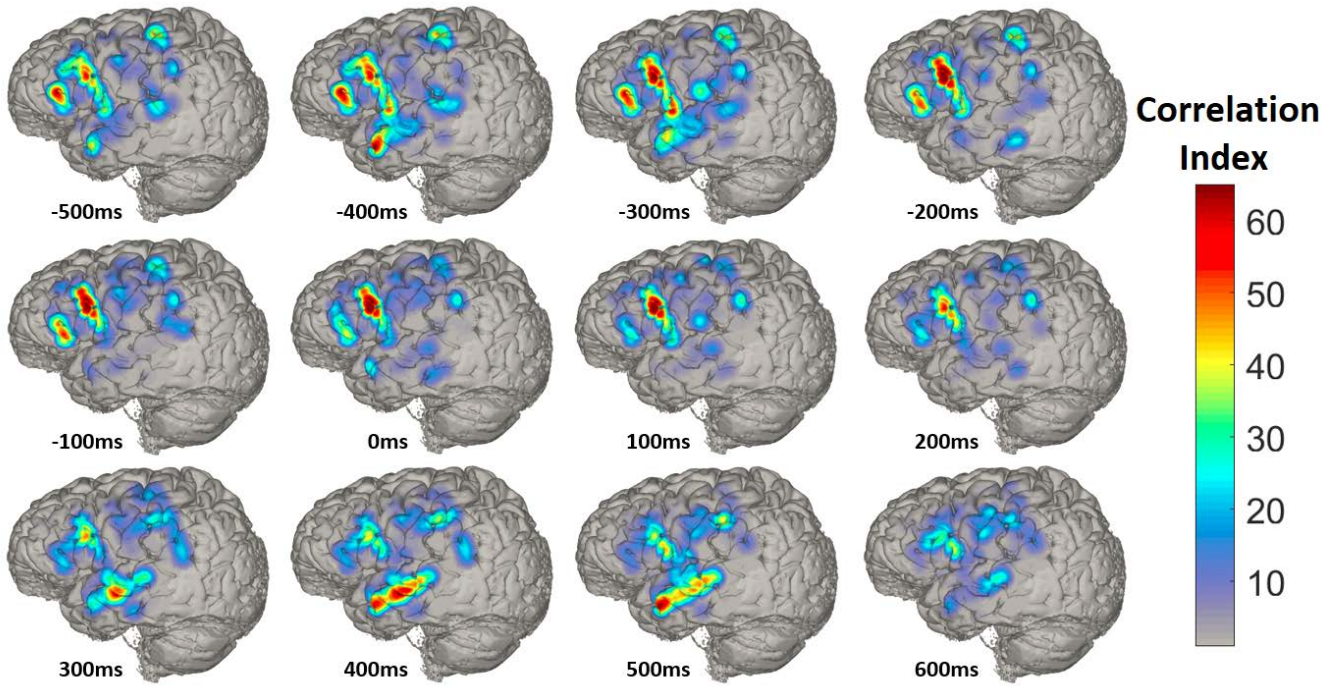

Fig. 4. Topographies of the correlation index at different time lags.

actual measurement was compared to this distribution and its p-value calculated to ensure the correlations were greater than random chance.

E. Classification

A classifier was developed to predict the state of the lips from the ECoG signals. The outer lip feature was dichotomized into 'open' if the area was above a threshold, or 'closed' otherwise. Because there was no clear distinction in the 'open' and 'closed' states based on the lip features distribution, the threshold was chosen for a number of image pixels that split the lip features into approximately equal sets for the two categories.

The average power was computed from 0.5 s before to 0.5 s after a video frame for 1, 5, 10 and 20 uniformly-spaced time intervals. This was done to explore the effect of time resolution on classification accuracy. These measures for all electrodes constituted the features for classification. The classification was performed using a Support Vector Machine (SVM) with a linear kernel.

F. Location Predictive Model

In addition to the classification model, the ECoG signals were used to predict the location of the mouth landmarks. The relative lack of available data make Bayesian framework or artificial neural network models impractical. Instead, a multivariate normal regression model was implemented. For each frame, the pixel positions of each landmark were normalized in reference to the left mouth corner landmark (location 49 in Figure 3), which was set to (0,0). The independent variables in the multivariate regression were the horizontal and vertical coordinates of each mouth landmark,

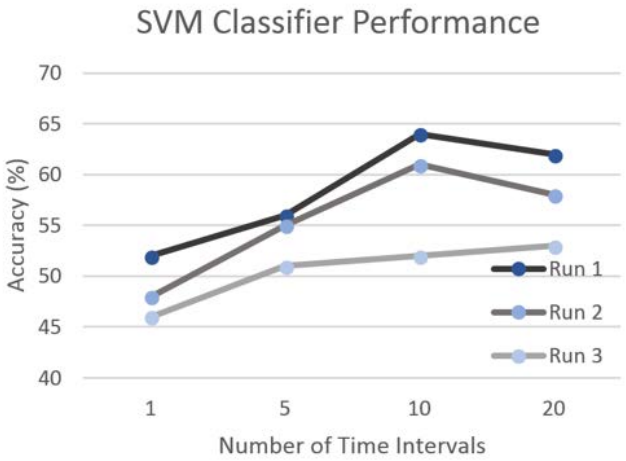

Fig. 5. Results of SVM classifier prediction of mouth 'open' versus 'closed'.

except for the reference landmark. The dependent variables were the average gamma power for all electrodes. Different time spans for average power were evaluated, ranging from 500 ms to 41.67 ms (the frame rate of the video).

III. Results

Figure 4 shows the correlation for different time lags, with predominant activation of Brocas area from -500 ms to 200 ms, and the superior temporal gyrus from 300 ms to 500 ms. The maximum correlation across channels was 0.26, with no simulated electrode reaching above 0.03. This result suggests that the correlations between lip features and

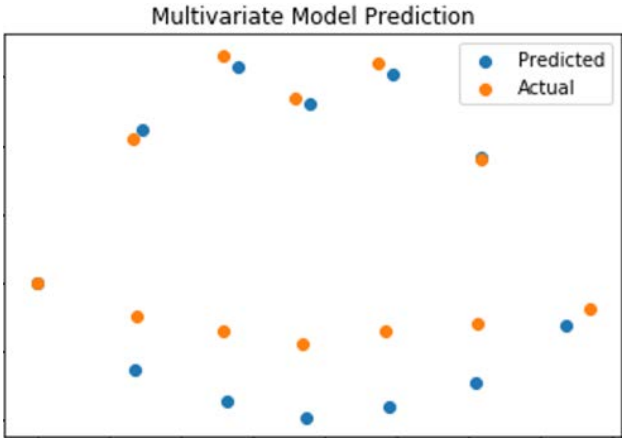

Fig. 6. Exemplary lip-landmark frame predicted from multivariate model.

the ECoG signal are significantly better than random chance. An analysis of the facial feature performance indicated that outer lip area was the highest correlated feature.

The accuracy of the SVM classifier for each of the time interval sets is shown in Figure 5. There was a distinct increase in the accuracy as the number of time intervals increased form 1 to 10, with the performance peaking at 65% (50% chance) for 10 intervals. This implies that a temporal resolution near 100 ms may be optimal for capturing the disrciminable features. The right panel of Figure 6 shows the actual facial landmarks and landmarks predicted by the location predictive model with 100 ms resolution for a representative image frame.

IV. Discussion

The results indicate clear neural correlates of lip movements obtained from video recordings, and that there is potential for these neural signatures to provide predictive information that may benefit the development of future neuroprothetics. The neural correlates appear in the anticipated brain regions (i.e., Broca's area, motor cortex) and temporal onsets. However, there is later activity in the superior temporal gyrus that warrants further examination.

There are obvious limitations to this preliminary analysis. Video recordings were only available from a single subject. Furthermore, limited data were available due to the patient's inconsistent head position, which caused the facial landmarking to fail for a large number of video frames. This issue can be easily mitigated in the future with a simple changes to the test protocol to ensure consistent video and face angle. While the lip position classifier approached an accuracy of 65% for discriminating 'open' versus 'closed', this result is expected to significantly improve with additional training data.

Future research will work to create a more robust video acquisition protocol to facilitate better usage and tuning of facial recognition libraries. The ultimate objective is to use this information in conjunction with recent findings from the speech signal and other biosignals to improve understanding of the brain processes and the predictive power of speech decoding models.

References

[1] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-based spoken communication: A survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2257–2271, 2017.

[2] G. Schalk and E. C. Leuthardt, "Brain-computer interfaces using electrocorticographic signals," *IEEE reviews in biomedical engineering*, vol. 4, pp. 140–154, 2011.

[3] C. Herff and T. Schultz, "Automatic speech recognition from neural signals: a focused review," *Frontiers in neuroscience*, vol. 10, p. 429, 2016.

[4] E. S. Nurse, S. E. John, D. R. Freestone, T. J. Oxley, H. Ung, S. F. Berkovic, T. J. O'Brien, M. J. Cook, and D. B. Grayden, "Consistency of long-term subdural electrocorticography in humans," *IEEE Trans Biomed Eng*, 2017.

[5] 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.

[6] 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.

[7] 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, p. 217, 2015.

[8] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang, "Reconstructing speech from human auditory cortex," *PLoS biology*, vol. 10, no. 1, p. e1001251, 2012.

[9] S. Martin, P. Brunner, C. Holdgraf, H.-J. Heinze, N. E. Crone, J. Rieger, G. Schalk, R. T. Knight, and B. N. Pasley, "Decoding spectrotemporal features of overt and covert speech from the human cortex," *Frontiers in neuroengineering*, vol. 7, p. 14, 2014.

[10] C. Herff, G. Johnson, L. Diener, J. Shih, D. Krusienski, and T. Schultz, "Towards direct speech synthesis from ecog: A pilot study," in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE, 2016, pp. 1540–1543.

[11] 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, p. 327, 2013.

[12] D. F. Conant, K. E. Bouchard, M. K. Leonard, and E. F. Chang, "Human sensorimotor cortex control of directly-measured vocal tract movements during vowel production," *Journal of Neuroscience*, pp. 2382–17, 2018.

[13] K. E. Bouchard, D. F. Conant, G. K. Anumanchipalli, B. Dichter, K. S. Chaisanguanthum, K. Johnson, and E. F. Chang, "High-resolution, non-invasive imaging of upper vocal tract articulators compatible with human brain recordings," *PLoS One*, vol. 11, no. 3, p. e0151327, 2016.

[14] E. M. Mugler, M. C. Tate, K. Livescu, J. W. Templer, M. A. Goldrick, and M. W. Slutzky, "Differential representation of articulatory gestures and phonemes in precentral and inferior frontal gyri," *Journal of Neuroscience*, 2018. [Online]. Available: http://www.jneurosci.org/content/early/2018/09/26/JNEUROSCI.1206-18.2018

[15] J. Chartier, G. K. Anumanchipalli, K. Johnson, and E. F. Chang, "Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex," *Neuron*, vol. 98, no. 5, pp. 1042–1054, 2018.

[16] E. Rothauser, "Ieee recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

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

[18] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1755–1758, 2009.