ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

Interpretation of Convolutional Neural Networks for Speech Regression from Electrocorticography

Miguel Angrick[1*], Christian Herff[1*], Garett Johnson[2], Jerry Shih[3],
Dean Krusienski[2] and Tanja Schultz[1]

1- University of Bremen - Cognitive Systems Lab
Enrique-Schmidt-Straße 5, 28359 Bremen - Germany

2- Old Dominion University - ASPEN Lab
5115 Hampton Blvd, Norfolk, VA 23529 - USA

3- UC San Diego Health - Epilepsy Center
200 West Arbor Drive San Diego, CA 92103 - USA

Abstract.
The direct synthesis of continuously spoken speech from neural activity is envisioned to enable fast and intuitive Brain-Computer Interfaces. Earlier results indicate that intracranial recordings reveal very suitable signal characteristics for direct synthesis. To map the complex dynamics of neural activity to spectral representations of speech, Convolutional Neural Networks (CNNs) can be trained. However, the resulting networks are hard to interpret and thus provide little opportunity to gain insights on neural processes underlying speech. Here, we show that CNNs are useful to reconstruct speech from intracranial recordings of brain activity and propose an approach to interpret the trained CNNs.

1 Introduction

Brain-Computer Interfaces (BCIs) that continuously decode neural activity into audible speech could provide a communication means for otherwise mute users [1]. State-of-the-art BCIs enable users to input words letter-by-letter or to move a cursor on the screen - at slow speed. A close to real-time decoding of continuously spoken speech would allow for much faster and more intuitive interfaces. Earlier work indicates that Electrocorticography (ECoG) provides signal properties that are suitable for the decoding of speech processes from neural data [2]. Few studies demonstrate reliable decoding of phonemes [3, 4] and continuous speech [5] from ECoG. Recent results give hope that even imagined speech could be successfully decoded [6, 7, 8]. While the decoding of speech from related neural signals into the corresponding textual representation enables fast device control, several speech characteristics that are crucial in spoken communication are not captured, such as emphasis, rhythm, and prosody. The direct mapping of neural signals into audible speech would enable users to regain the full expressive power of speech.

Both continuous speech and ECoG data have complex spatio-temporal dynamics indicating that a mapping from one to the other might not be simply linear. CNNs have recently produced promising results on neural data even with

[*] These authors contributed equally.

limited amounts of training data [9]. Here, we demonstrate that CNNs can be applied to solve a regression problem, namely to successfully reconstruct spectral speech features from ECoG data. However, CNNs models are notoriously difficult to interpret and thus hinder to gaining knowledge. For classification tasks, activation maximization [10, 11] may provide insights into the trained models but it cannot be directly applied to the given regression problem. In this paper, we modified the application of activation maximization to fit it to our regression problem and thus to investigate and verify the trained network.

2 Material and Methods

2.1 ECoG dataset

We simultaneously recorded ECoG activity and acoustic speech data from 3 participants (1 female) suffering from intractable epilepsy. We asked them to speak aloud phrases from the Harvard sentences [12] which were aurally and visually presented to them for 4 seconds. Participants repeated between 50 and 150 phrases. Participants were implanted with different numbers of ECoG electrodes (participant 1: 18 electrodes, participant 2: 16, participant 3: 68 electrodes) on the left hemisphere covering at least some areas relevant for speech production. Electrode locations and duration of intracranial monitoring were purely based on clinical needs. All participants gave written informed consent and the experiment was approved by the IRB of both Mayo Clinic and Old Dominion University.

2.2 Signal Preprocessing and Feature Extraction

ECoG data was preprocessed by linear detrending. Additionally an elliptic IIR notch filter was used to attenuate the first harmonic of the line noise at 120 Hz. As a meaningful feature, logarithmic broadband gamma (70-170 Hz) power was extracted and normalized to zero mean and unit variance per channel.

To capture the complex dynamics of neural activity associated with speech production, we use 9 consecutive intervals of 50 ms (downsampled to 20 Hz) of broadband gamma activity, which can be interpreted as a two dimensional spatio-temporal pattern of brain activity. Resulting patterns are of dimension $|electrodes| \times 9$ and form the input of the CNN architecture to reconstruct the spectral features of speech.

Acoustic speech data was downsampled to 16 kHz and the spectral features were calculated in windows of 50 ms with a frameshift of 10 ms. To capture the speech-relevant spectral dynamics and to reduce the feature dimensionality, we applied traditional mel-scaling using triangular filter banks to finally extract 40 logarithmic mel-scaled spectral coefficients per frame [13].

2.3 Deep Convolutional Neural Network

As acoustic and neural data were recorded in parallel, the acoustic feature vectors in terms of mel-scaled spectral coefficients can be aligned to ECoG features.

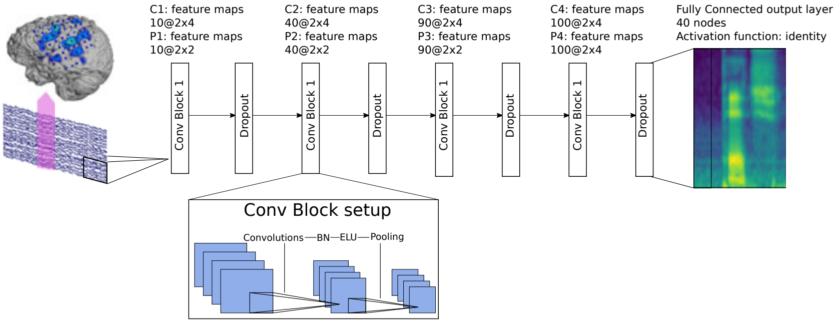

Fig. 1: ECoG Gamma activity is arranged to form a 2-dimensional, spatio-temporal pattern of *electrodes* × *time*. This pattern is fed into 4 subsequent convolutional blocks and a linear output layer to produce the log mel speech features.

Thus, a CNN can be trained to predict spectral features (spectrogram) from the spatio-temporal patterns of ECoG activity.

For this regression task we designed a deep CNN inspired by the architecture of Schirrmeister et al [9]. The convolutional layers are intended to exploit the coherence between electrode location and temporal progression. Figure 1 illustrates our CNN architecture that is applied to reconstruct the spectral speech features from the neural data. The network consists of four convolutional blocks followed by a fully-connected layer with a linear activation function as the output layer to map the ECoG input features to the 40 logarithmic mel-scaled spectral coefficients. Each convolutional block starts with a convolutional layer followed by a batch normalization layer. Figure 1 shows the number of feature maps, the size of the receptive field, and the pooling size for the subsequent subsampling layer. For the batch normalization (BN) layer we used a constant momentum of 0.9. Non-linearity is introduced by exponential linear units [14]. A convolutional block ends with a max pooling layer to reduce the dimensionality. Dropout is applied with a probability of 0.5 after each block.

Network training has been applied in a 5-fold cross-validation to synthesize a spectrogram of the entire session. We used Adam as the optimizer and trained for a fixed number of 80 epochs.

2.4 Activation Maximization for Regression Problems

The activation maximization [10, 11] technique identifies the input pattern of a trained neural network that maximizes the activation of a specific (hidden) unit. This optimization problem that can be solved by performing gradient ascent in the input space and by modifying the input sample according to the gradient. Common applications of activation maximization are input visualizations for image categories in object recognition, where the class unit activation in the output layer is maximized [15].

For the speech regression task, we adapted this idea to enable interpretation of our CNN. Instead of maximizing the output of a single unit, we minimized the mean-squared error between the network output for a given input sample and the mean spectrogram target. Two targets were used for the activation maximization, speech and silence, i.e. we calculated one mean log mel-scaled spectrum of segments when the participant was speaking and one mean when s/he was silent.

Gradient descent is used to move the input sample to a local minimum in the input space by subtracting the partial error of each input feature in an iterative manner. To generate input samples for the mean speech and mean silence spectrogram targets, we used a constant learning rate of 0.01 and a fixed number of 600 update steps in the gradient descent optimization. Initial input samples for both target classes comprise the maximum value from each feature over all training samples.

3 Results

3.1 Speech Reconstruction Results

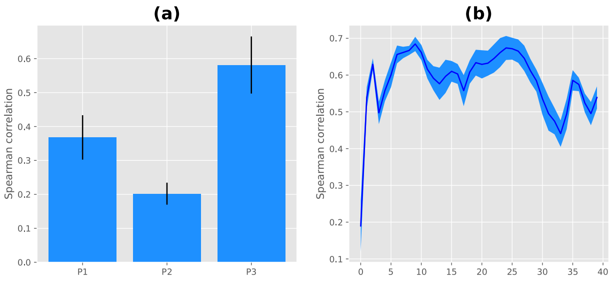

Fig. 2: Correlation results of the speech synthesis approach using CNN. (a) Mean Spearman's rank correlations between original and reconstructed spectrograms over all spectral bins for all participants. (b) The correlations over all 40 logarithmic mel-scaled spectral coefficients for participant 3. Whiskers/shaded areas show standard deviations.

We used Spearman's rank correlation to compare reconstructed spectrograms with their original counterpart. Figure 2 (a) shows that mean correlations over all spectral bins for all 3 participants are significantly better than chance level. Correlations for participant 3 clearly outperform the other two with a mean $\rho = 0.58$, this is most likely due to the larger number of electrodes (68 compared to 16 and 18, respectively) and the better coverage of brain areas involved in speech production. Figure 2 (b) examines results for participant 3 in more detail by looking at rank correlations for each of the 40 logarithmic mel-scaled spectral coefficients individually. It can be seen that except for the first coefficient, which contains signal energy, all bins can be reconstructed with high correlations.

Subsequently to reconstruction, a resulting spectrogram can be synthesized into an acoustic speech waveform, although phase information is lost [16].

3.2 Interpretation of Trained CNNs

Figure 3 highlights those brain areas for which expected activities strongly differ between speech segments and silence segments with regards to deviation from mean normalized broadband gamma-activity, as identified by our modified activation maximization approach. Figure 3 (a) shows differences for participant 3 in the lower motor cortex, an area which is responsible for facial muscle control that are involved in articulation. Figure 3 (b) highlights differences for participant 1 in the auditory cortex, an area that processes the perception of the participant's own voice. As our proposed approach pinpoints brain areas and associates them to time-aligned speech processes, it facilitates a sanity check to verify if CNN models learn patterns in accordance to known speech processes [17], and thus provides a form of validation.

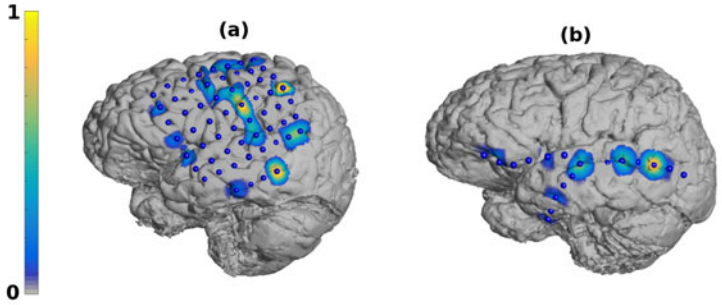

Fig. 3: Activation patterns in terms of one standard deviation of log-gamma activity identified by our approach. (a) Relevant gamma activity concurrent to speech production in lower parts of the motor cortex (Participant 3). (b) For participant 1, most differences in activation can be found in auditory areas for acoustic processing.

4 Conclusion

The direct synthesis of neural activity into audible speech could provide an intuitive means of communication and thus help people who do not have a voice. In this paper, we have shown that convolutional neural networks can be used to reconstruct spectral features of speech solely from invasively measured brain activity. The reconstructed speech features can be synthesized into an audible speech waveform. To interpret the dynamics learned by the CNN models, we present a modification of the activation maximization approach for regression tasks. The areas of brain activity indicated by our approach confirm common theories of speech production.

References

[1] J.R. Wolpaw, N. Birbaumer, D.J. McFarland, G. Pfurtscheller, and T.M. Vaughan. Brain–computer interfaces for communication and control. *Clinical neurophysiology*, 113(6):767–791, 2002.

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

[3] N.F. Ramsey, E. Salari, E.J. Aarnoutse, M.J. Vansteensel, M.G. Bleichner, and Z.V Freudenburg. Decoding spoken phonemes from sensorimotor cortex with high-density ECoG grids. *NeuroImage*, 2017.

[4] 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*, 11(3):035015, 2014.

[5] 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*, 9, 2015.

[6] 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*, 8(4):046028, 2011.

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

[8] S. Martin, P. Brunner, I. Iturrate, J.d.R. Millán, G. Schalk, R.T. Knight, and B.N. Pasley. Word pair classification during imagined speech using direct brain recordings. *Scientific reports*, 6:25803, 2016.

[9] R.T. Schirrmeister, J.T. Springenberg, L.D.J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human brain mapping*, 2017.

[10] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341:3, 2009.

[11] G. Montavon, W. Samek, and K.-R. Müller. Methods for interpreting and understanding deep neural networks. *arXiv preprint arXiv:1706.07979*, 2017.

[12] E.H. Rothauser. IEEE recommended practice for speech quality measurements. *IEEE Trans. on Audio and Electroacoustics*, 17:225–246, 1969.

[13] S. Imai. Cepstral analysis synthesis on the mel frequency scale. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'83.*, volume 8, pages 93–96. IEEE, 1983.

[14] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[15] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[16] 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*, pages 1540–1543. IEEE, 2016.

[17] J.S. Brumberg, D.J. Krusienski, S. Chakrabarti, A. Gunduz, P. Brunner, A.L. Ritaccio, and G. Schalk. Spatio-Temporal Progression of Cortical Activity Related to Continuous Overt and Covert Speech Production in a Reading Task. *PloS one*, 11(11):e0166872, 2016.