Speech Synthesis from Intracranial Neural Activity

Károly Gusztáv Borkó
kborko@edu.bme.hu
Milán Fodor
milanfodor@edu.bme.hu
Máté Szécsi
mate.szecsi@edu.bme.hu

Faculty of Electrical Engineering and Informatics
Budapest University of Technology and Economics

Abstract

The cognitive aspect of speech production is an intricate process involving many areas of the brain, and the underlying neural processes are not completely understood. Brain-Computer Interfaces that directly decode speech from neural activity has recently gained large attention as they could provide an intuitive means of communication. In this paper, the authors attempt to synthesize words in the forms of audio waveforms directly from intracranial neural activity measured with stereotactic electroencephalography. Proven deep-learning methods have been tested in an attempt to build models for personalized and generalized speech synthesis. Results mentioned in this paper demonstrate the viability of such approaches and conclude a clear next step toward achieving end-to-end speech synthesis.

1 Introduction and Background

We human beings have the unique ability to communicate with each other by producing intelligible speech. However, nearly 70 million people have speech disabilities around the world. In recent years there was a lot of research interest in developing assistive technologies to help with speech restoration for people with speaking disabilities. (Panachakel et al., 2019) Brain-Computer Interfaces (BCIs) that directly decode speech from neural activity has recently gained large attention, as they could provide an intuitive means of communication. (Mansoor et al., 2020)

The difficulty of the task mainly lies in the complexity of processing EEG signals. Their voltage range (μV), low signal-to-noise ratio (SNR), non-linearity, non-temporality, and low spatial resolution given by the EEG electrodes make it so that conventional ML methods are limited for the recognition of this type of signal. This poses an important challenge for designing new solutions to identify the characteristics of the EEG signal and, to select or design the proper classifiers. In conclusion, an ideal method should be able to automatically recognize the inherent characteristics of the EEG signal with its nonlinear and nonstationary properties. In consequence, Deep Learning (DL) methods have been proposed as it seems to be the practicable route. (Rashid, M. et al.,2020, Chinta and M, 2022)

Despite the fact that a full understanding of speech production is currently lacking, significant advances have been made in the field of speech neuroprostheses recently. The decoding of a textual representation by decoding phonemes, phonetic or articulatory features, words, full sentences, or spotting of speech keywords is possible from neural recordings during actual speech production. Most of this research is focused on speech recognition, classification of words, and producing a "brain-typing" solution. The complete synthesis of speech is a more laborious task. (Verwoert et al., 2022)

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Some studies attempted to directly synthesize an audio waveform of speech from neural data captured during speech production to enable more natural communication. According to preliminary findings, it is possible to decode speech processes from imagined speech using offline data and in real time. (Zhang and Yao, 2021) Electrocorticography (ECoG), is an invasive recording modality of neural activity that provides high temporal and spatial resolution and a high signal-to-noise ratio. (Hill et al., 2012) Several methods from the field of acoustic speech processing have been applied to ECoG signals, either for decoding into a sequence of words based on automatic speech recognition techniques (Moses, D.A. et al., 2016)or for conversion into speech based on speech synthesis strategies. (Krishna, G at al., 2016, Herff, C. et al., 2019),Herff et al., 2020)

In comparison to ECoG, stereotactic electroencephalography (sEEG) implants a series of penetrating electrode shafts, each consisting of multiple electrode contacts, into the brain. Despite their growing clinical usage potential for BCI applications in general (Krusienski and Shih, 2011), sEEG recordings have thus far received surprisingly limited attention for speech-related BCIs. sEEG recordings have been investigated in a perceived speech task, where recent advances in deep neural networks were applied to decode comprehensible speech from the auditory cortex. (Akbari et al., 2019).

In this paper, the authors attempt to synthesize full words in the form of audio waveforms directly from intracranial neural activity measured with EEG using proven deep learning methods.

2 Dataset And Methods

Authors used the SingleWordProductionDutch-iBIDS sEEG dataset (Herff, C. at al 2022) consisting of 10 participants speaking prompted words aloud while audio and intracranial EEG data are recorded simultaneously. The SingleWordProductionDutch-iBIDS dataset67 is available at `https://doi.org/10.17605/OSF.IO/NRGX6`. The iBIDS dataset passed a validation check using the BIDS Validator (https://bids-standard.github.io/bids-validator/) and manual inspection of each data file.

Platinum-iridium sEEG electrode shafts were in use (Microdeep intracerebral electrodes; Dixi Medical, Becanson, France) with a diameter of 0.8mm, a contact length of 2mm, and an inter-contact distance of 1.5mm. Each electrode shaft contained between 5 and 18 electrode contacts. Neural data were recorded using two or more Micromed SD LTM amplifier(s) (Micromed S.p.A., Treviso, Italy) with 64 channels each. Electrode contacts were referenced to a common white matter contact. Data were recorded at either 1024Hz or 2048Hz and subsequently downsampled to 1024Hz. We used the onboard microphone of the recording notebook (HP Probook) to record audio at 48kHz. Audio data was subsequently pitch-shifted to ensure our participants' anonymity using LibRosa63. We used LabStreamingLayer64 to synchronize the neural, audio, and stimulus data. Electrode locations were anatomically labeled. By far most electrodes are located in white matter (40.3 %) and unknown areas (12.6 %).

After preparing the data using scripts made available by Verwoert et al. (2022), two approaches were defined: one for "personalized speech synthesis" and a "general speech synthesis" approach. In the first scenario, only a single participant's data was used in the neural networks, in the second scenario: all 10 participant's data were used for training and testing. Two deep-learning models were implemented. For both approaches, a 1d convolutional neural network (1D-CNN), and a fully connected deep neural network(FC-DNN) have been tested. Results were evaluated by the network metrics and then further tested by the subjective assessment of the synthesized audio files.

Models were trained and then tested in Jupyter Notebook. The final system qualifies as low-cost using an NVIDIA RTX 3070 GPU as training hardware. All code described in this report can be found on Github:`https://github.com/MateSzecsi/BRAIN2SPEECH`. The code relies on Python 3.10.5, Numpy 1.22.4, Scipy 1.8.1, Scikit-learn 1.1.1, PyNWB 2.2.0, Keras 2.9.0, and Tensorflow 2.9.1 packages.

3 Implementation

3.1 Preparing data

In preparation for data, the Hilbert envelope of the broadband high-frequency activity (70–170Hz) was extracted for each contact using an IIR bandpass filter (filter order 4). To attenuate the first two harmonics of the 50Hz line noise, we used two IIR bandstop filters (filter order 4). All filters were

applied forward and backward so that no phase shift is introduced. The envelope was averaged over 50ms windows with a frameshift of 10ms. To include temporal information in the decoding process, non-overlapping neighboring windows up to 200ms into the past and future were stacked. Features are normalized to zero mean and unit variance using the mean and standard deviation of the training data. The same transform is then applied to the evaluation data.

The audio data is first downsampled to 16kHz. To extract features for the audio data, we subsequently calculated the Short-Term-Fourier-Transform in windows of 50ms with a frameshift of 10ms. As the frameshift between neural and audio data is the same, there is a correspondence between audio and neural feature vectors. The resulting spectrogram is then compressed into a log-mel representation71 using 23 triangular filter banks.

Compressing the feature space
To reduce the dimensionality of our decoding problem, we compress the feature space to the first 50 principal components. Principal components are estimated for each fold individually on the training data. The first 50 principal components explain between 29% and 76% of the variance depending on the participant.

3.2 Training

The single patient fully connected DNN: In the final model, 4 hidden layers are used. The activation function of the output layer is Relu. The dense hidden layers consist of 300, 200, 100, and 50 neurons in order. The activation function of the output layer is linear. The batch size is 256, with the Adam optimizer and a learning rate of 0.001. The maximum epoch number was set to 1000 but with an early stopping in place with a patience value of 8. The loss function is Mean Squared Error (MSE).

Single patient CNN: The final model contains six convolutional layers having 40, 40, 30, 30, 20, 20, and filters in order. The output is a linear layer, after flattening the output of the previous one. Convolution window was set to 8. This model proved to be the most successful, as can be seen in the figure.

Multi-patient FCDNN/CNN: We considered parameterizing the patients' brain structure, and the placement of the electrodes. The parameter space describing the human brain is very complicated, however, and with only ten patients, the generalization capability of the neural net would be minimal.

3.3 Evaluation

To evaluate the results, we calculated the mean absolute errors as the metric of the predictions and plotted spectrograms predicted by the neural networks. A comparison between the error values, and the spectrograms can be viewed below.

In figure 1 the FC DNN network predictions can be seen, which are similar to the linear regression prediction capabilities, with significant noise during the audio file.

Figure 1 shows the results from the 1D convolutional network. These are significantly better, with much less noise during silent parts of the audio file, and more recognizable frequency components during speeches.

We attempted filtering out the silence from the audio, this meant, however, that we ended up with many, shorter timeseries' (one for every word pronounced) which limited the effectiveness of the convolutional neural net. We used the fully connected net, but as mentioned before, it does not perform well on the data.

	FCDNN	CNN
Validation MAE	0.48	0.52
Test MAE	0.61	0.62

3.4 Comperative Testing

For the subjective audio testing, we first use a method by Griffin and Lim utilized for waveform reconstruction, in which the phase is initialized with noise and then iteratively modified. After comparing the newly synthesized audio files with the ones reconstructed from the original spectrograms

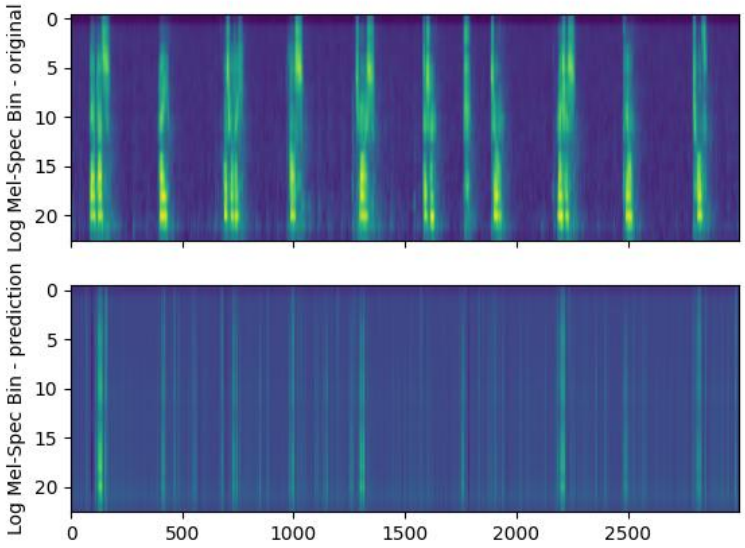

Figure 1: Predicted spectrogram using an FC DNN network structure

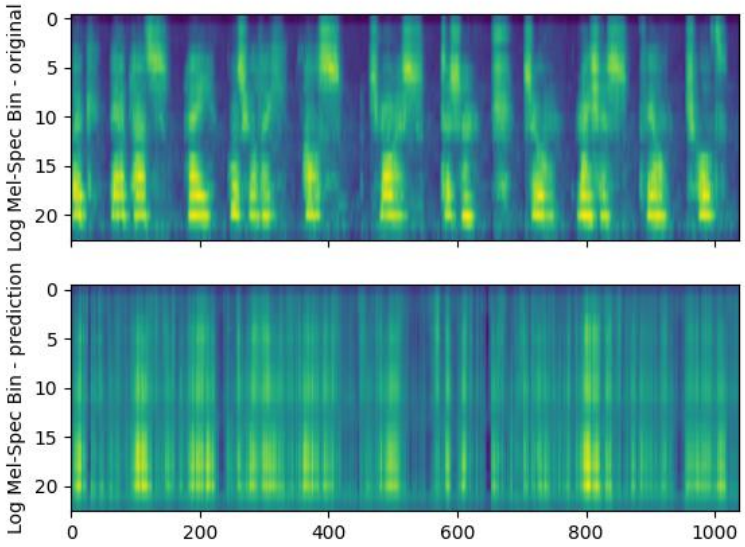

Figure 2: Predicted spectrogram using an FC DNN network structure, after filtering out silence from the data

similarities can be observed for the single-subject DNN results, in other cases, the best-case scenario is the separation of silence and speech.

4 Discussion

The 1D CNN appeared to be the more suitable approach. An attempt with a 2D convolutional network might be worth a try, as the relation between the presence of different frequency components over

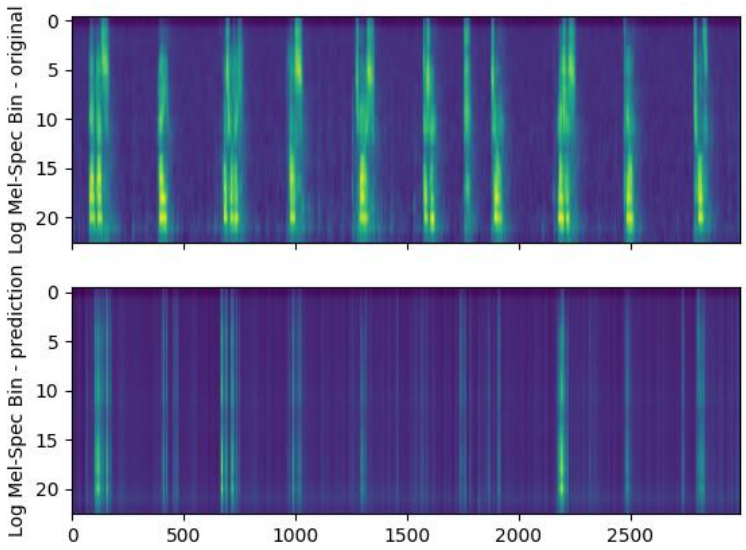

Figure 3: Predicted spectrogram using an FC DNN network structure

time possibly contains relevant information too. It is concluded, that the data set might not be robust enough, and the amount of data might not be enough, to create a general speech synthesis model. Recent advances highlight that deeper brain structures, such as the hippocampus (Piai et al., 2016) and thalamus (Hebb Ojemann, 2013), are also heavily involved in speech production, which are underrepresented in our dataset. A clear way forward is to extend the dataset with more test subjects, giving our model more training data as per different brain areas and also making it easier to account for individual differences in the cognitive process underlying the given speech production task.

5 References

Panachakel, J.T., Ramakrishnan, A.G. and Ananthapadmanabha, T.V. (2019) "Decoding imagined speech using wavelet features and deep neural networks," 2019 IEEE 16th India Council International Conference (INDICON) [Preprint]. Available at: https://doi.org/10.1109/indicon47234.2019.9028925.

A. Mansoor, M. W. Usman, N. Jamil, and M. A. Naeem, "Deep Learning Algorithm for Brain-Computer Interface," Scientific Programming, vol. 2020, p. e5762149, Aug. 2020, doi: 10.1155/2020/5762149.

Rashid, M. et al. (2020) "Current status, challenges, and possible solutions of EEG-based brain-computer interface: A comprehensive review," Frontiers in Neurorobotics, 14. Available at: https://doi.org/10.3389/fnbot.2020.00025.

Chinta, M.B. and M, D.D.M. (2022) "Efficient automatic speech recognition from EEG signals using optimal deep learning approach," SSRN Electronic Journal [Preprint]. Available at: https://doi.org/10.2139/ssrn.4211331.

Verwoert, M., Ottenhoff, M.C., Goulis, S. et al. Dataset of Speech Production in intracranial Electroencephalography. Sci Data 9, 434 (2022). https://doi.org/10.1038/s41597-022-01542-9

Zhang, X. and Yao, L. (2021) "Deep learning for EEG-based brain–computer interfaces." Available at: https://doi.org/10.1142/q0282.

Hill, N.J. et al. (2012) "Recording human Electrocorticographic (ECoG) signals for neuroscientific research and real-time functional cortical mapping," Journal of Visualized Experiments [Preprint], (64). Available at: https://doi.org/10.3791/3993.

Moses, D.A. et al. (2016) "Neural speech recognition: Continuous phoneme decoding using spatiotemporal representations of human cortical activity," Journal of Neural Engineering, 13(5), p. 056004. Available at: https://doi.org/10.1088/1741-2560/13/5/056004.

Krishna, G., Tran, C., Han, Y., Carnahan, M. (2020). Speech Synthesis using EEG (arXiv:2002.12756). arXiv. http://arxiv.org/abs/2002.12756

Herff, C. et al. (2019) "Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices," Frontiers in Neuroscience, 13. Available at: https://doi.org/10.3389/fnins.2019.01267.

Herff, C., Krusienski, D.J. and Kubben, P. (2020) "The potential of stereotactic-EEG for Brain-Computer interfaces: Current progress and Future Directions," Frontiers in Neuroscience, 14. Available at: https://doi.org/10.3389/fnins.2020.00123.

Krusienski, D.J. and Shih, J.J. (2011) "Control of a brain–computer interface using stereotactic depth electrodes in and adjacent to the hippocampus," Journal of Neural Engineering, 8(2), p. 025006. Available at: https://doi.org/10.1088/1741-2560/8/2/025006.

Akbari, H. et al. (2019) "Towards reconstructing intelligible speech from the human auditory cortex," Scientific Reports, 9(1). Available at: https://doi.org/10.1038/s41598-018-37359-z.

Piai, V. et al. (2016) "Direct brain recordings reveal hippocampal rhythm underpinnings of Language Processing," Proceedings of the National Academy of Sciences, 113(40), pp. 11366–11371. Available at: https://doi.org/10.1073/pnas.1603312113.

Hebb, A.O. and Ojemann, G.A. (2013) "The thalamus and language revisited," Brain and Language, 126(1), pp. 99–108. Available at: https://doi.org/10.1016/j.bandl.2012.06.010.