2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Speech Synthesis from Stereotactic EEG using an Electrode Shaft Dependent Multi-Input Convolutional Neural Network Approach

Miguel Angrick[1], Maarten Ottenhoff[2], Sophocles Goulis[2], Albert J. Colon[3], Louis Wagner[3], Dean J. Krusienski[4], Pieter L. Kubben[2], Tanja Schultz[1], Christian Herff[2]

***Abstract*— Neurological disorders can lead to significant impairments in speech communication and, in severe cases, cause the complete loss of the ability to speak. Brain-Computer Interfaces have shown promise as an alternative communication modality by directly transforming neural activity of speech processes into a textual or audible representations. Previous studies investigating such speech neuroprostheses relied on electrocorticography (ECoG) or microelectrode arrays that acquire neural signals from superficial areas on the cortex. While both measurement methods have demonstrated successful speech decoding, they do not capture activity from deeper brain structures and this activity has therefore not been harnessed for speech-related BCIs. In this study, we bridge this gap by adapting a previously presented decoding pipeline for speech synthesis based on ECoG signals to implanted depth electrodes (sEEG). For this purpose, we propose a multi-input convolutional neural network that extracts speech-related activity separately for each electrode shaft and estimates spectral coefficients to reconstruct an audible waveform. We evaluate our approach on open-loop data from 5 patients who conducted a recitation task of Dutch utterances. We achieve correlations of up to 0.80 between original and reconstructed speech spectrograms, which are significantly above chance level for all patients ($p < 0.001$). Our results indicate that sEEG can yield similar speech decoding performance to prior ECoG studies and is a promising modality for speech BCIs.**

I. INTRODUCTION

Spoken communication and Brain-Computer Interfaces (BCIs) are starting to blend into each other, with the increasing interest to decode speech processes directly from electrophysiological recordings of neural signals. Millions worldwide [1] suffer from speech impairments caused by neurological disorders, such as stroke or amyotrophic lateral sclerosis, which in severe cases can even result in the complete loss of the ability to speak. Providing these affected people with an alternative modality for spoken communication would have a major impact on their quality of life. Recent advances in BCIs raise great hope for systems that directly transform these neural signals into intelligible representations [2], [3], such as written text [4] or audible speech [5].

Various measurement methods have been investigated [6] to model speech-related dynamics in the brain with the purpose of enabling an intuitive and natural way of communication. While non-invasive measurements (e.g. electroencephalography, EEG) have been successfully used in typing interfaces [7], their application in speech synthesis and speech recognition is limited due to motion artifacts and and filtering effects from scalp and skin [6]. Therefore, a common approach to capture the fast and complex processes of speech production are invasive methods, such as electrocorticography (ECoG) or implanted microelectrode arrays, whose recorded electrical potentials provide suitable characteristics for modeling and decoding.

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 [4], [8], [9] or for conversion into speech based on speech synthesis strategies [10], [5], [11].

In contrast to ECoG, stereotactic electroencephalography (sEEG) implants a series of penetrating electrode shafts, each containing multiple electrode contacts, into the brain. Despite their increasing clinical usage potential for BCI applications in general [12], [13], sEEG recordings have so far received very limited attention for speech-related BCIs. sEEG recordings have been investigated in a perceived speech task, where recent advances in deep neural networks were used to decode intelligible speech from the auditory cortex [14].

In this study, we investigate the use of sEEG to gain insight into the potential of speech decoding from deeper brain structures. We are building upon our prior work using convolutional neural networks (CNNs) to reconstruct audible speech from ECoG [10]. Since ECoG electrodes are generally arranged in 2D rectangular grids, the resulting neural activity can be spatially exploited by standard 2D convolution operations. For sEEG depth electrodes, the spatial dimensions span the linear directions of the individual electrode shafts. For this reason, we base our investigations on a decoding approach similar to [15], i.e. we propose a CNN architecture that uses an electrode shaft-dependent multi-input layer to extract features from coherent electrodes to estimate a spectral representation of spoken speech. Here, this multi-input approach enables the decoupling of convolution kernels across different electrode shafts, preventing influences from unrelated brain areas. To quantify the decoding performance, we use open-loop recordings from 5 patients performing a recitation task of Dutch utterances.

II. MATERIAL AND METHODS

A. Experiment Design and Recording Setup

We conducted an experiment on 5 native speakers of Dutch suffering from intractable epilepsy, who were implanted with sEEG electrodes to identify the epileptogenic zone. Locations

[1]Cognitive Systems Lab, University of Bremen, Bremen, Germany
[2]Department of Neurosurgery, School of Mental Health and Neurosciences, Maastricht University, Maastricht, Netherlands
[3]Epilepsy Center Kempenhaeghe, Kempenhaeghe, Netherlands
[4]ASPEN Lab, Biomedical Engineering Department, Virginia Commonwealth University, Richmond, VA, United States

of electrode shafts were purely determined based on clinical needs. Patients performed a speech production task for which they were asked to read aloud 100 short utterances randomly drawn from the Mozilla Common Voice Dutch corpus [16], resulting in 8:20 min to 20:00 min of speech data. For each trial, the target utterance was presented for 4-10 seconds on a monitor in front of the patient (depending on the patient's reading speed), followed by a pause of 1 second.

sEEG and acoustic data were recorded in parallel using LabStreaming Layer [17] . Neural data was digitized using a Micromed SD LTM amplifier (Micromed S.p.A., Treviso, Italy, sampling rate: 1024 Hz). Audio data was recorded at 48 kHz using the recording notebook's on-board microphone.

The experiment design was approved by the IRB of Maastricht University and Epilepsy Center Kempenhaeghe and was conducted in a clinical environment under the supervision of experienced healthcare staff.

B. Data processing

sEEG recordings were processed to extract features in the broadband gamma band (70-170 Hz), using a bandpass filter (4th order Butterworth filter). The broadband gamma band is known to contain highly localized information about speech production [18], [19], and we have successfully employed it in previous studies [10], [15] to decode speech processes. After the extraction of the high-gamma band, two bandstop filters (98-102 and 148-152, respectively, both 4th order Butterworth filters) were applied to attenuate the first and second harmonic of the line noise at 50 Hz. The resulting signals were segmented into 50 ms windows with a 10 ms frameshift and the signal envelope was determined using the Hilbert transform. To capture the temporal context of speech processes, we augmented each feature vector with 8 consecutive intervals, which span a time range from -200 ms to 200 ms. After processing, the data had a shape of $|\text{ frames }| \times |\text{ electrodes }| \times 9$. In order to handle the electrode shaft dependent multi-input layer in the network architecture, we parsed the data in the electrode dimension to only contain electrodes within the same shaft.

Recordings of acoustic speech were downsampled to 16 kHz and spectral features were calculated in windows of 50 ms with a frameshift of 10 ms. Triangular mel-scale filter banks were applied to reduce the complexity of the speech spectrogram to 40 coefficients and apply the logarithm as the final step in the preprocessing to extract logarithmic mel-scaled spectral coefficients (logMels).

C. Network Architecture

The network architecture is designed as a feed-forward multilayer network, which considers the individual electrode shafts in the input layer separately. This approach enables the identification of shaft-dependent neural activity from speech production processes by first employing a dedicated stack of layers to extract spatio-temporal dynamics from within a related group of electrodes, then combining the activations across all electrode shafts for a final estimation of spectral coefficients. An overview of the network architecture

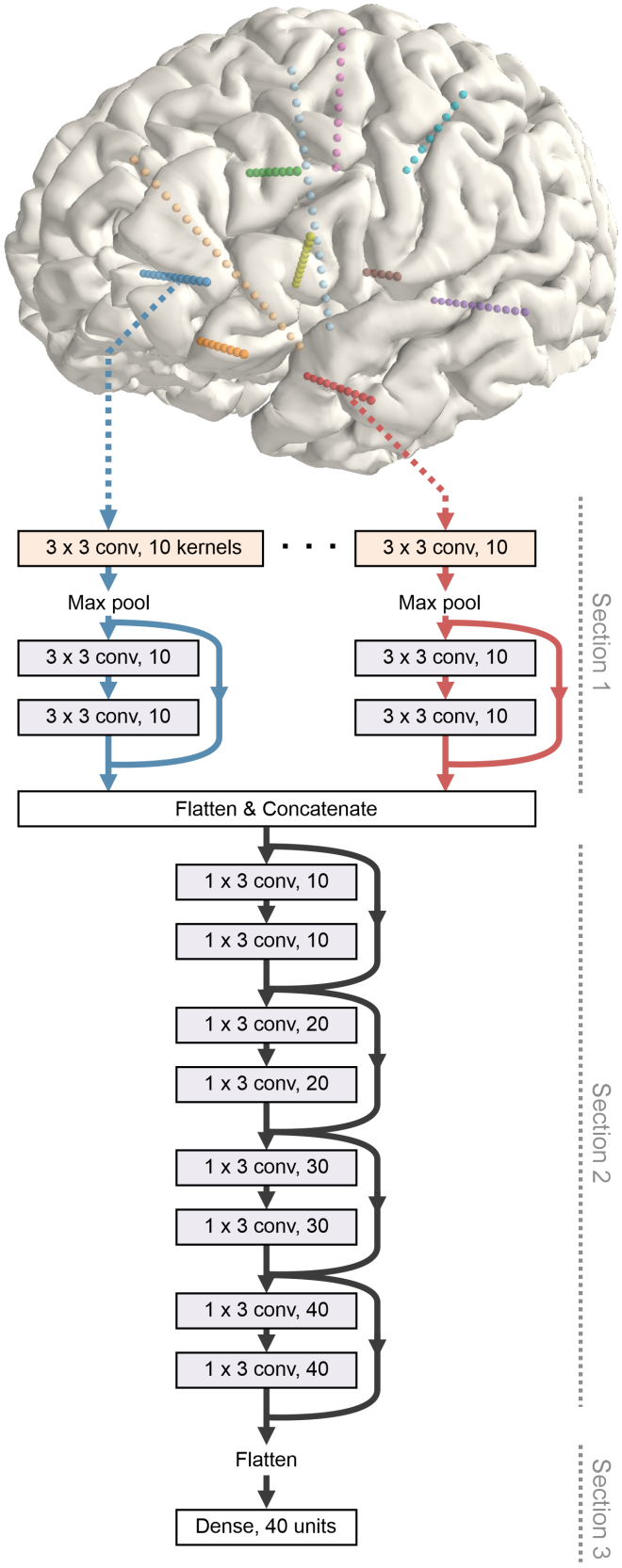

Fig. 1. Overview of the network architecture. In the first section, each electrode shaft is processed by a stack of convolutions. Section 2 extracts features across all shaft-dependent feature maps, before estimating spectral coefficients (Section 3). The network employs shortcut-connections for improved gradient flow in a residual learning framework.

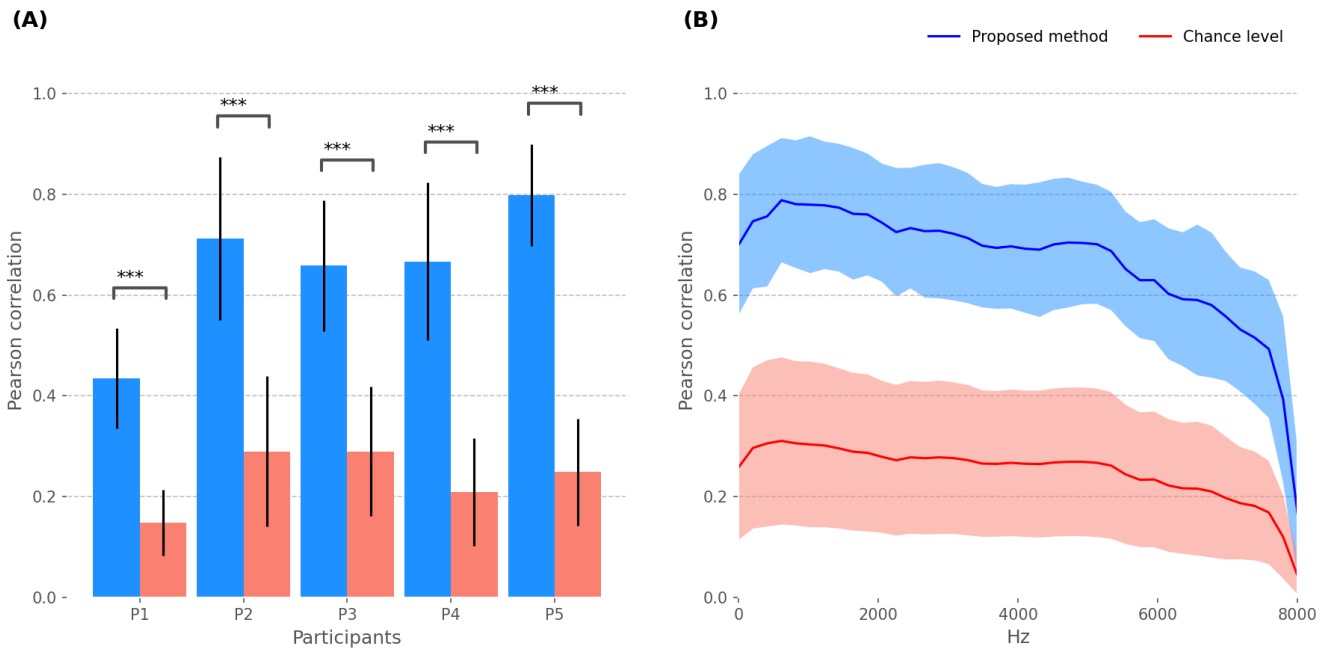

Fig. 2. Correlation results of the speech synthesis approach using an electrode-shaft dependent multi-input CNN. (A) Mean Pearson correlations between original and reconstructed spectrograms. Blue bars represent decoding scores of the proposed method and red bars denote chance level (whiskers indicate standard deviation). Significance brackets represent the Mann-Whitney-U test statistics, which show that $p < 0.001$ (***) consistently across all participants. (B) Mean correlation scores along the frequency range of 40 logarithmic mel-scaled spectral coefficients, considering all 5 participants. Shaded areas show standard deviations.

is depicted in Figure 1 and consists of three sections: 1) the multi-input section to extract shaft dependent local features, 2) a feature extraction section considering all feature maps and 3), a linear output layer which maps the neural activations onto spectral coefficients of a logarithmic mel-scaled speech spectrogram.

The first two sections employ a stack of convolutional layers that use shortcut-connections between former and subsequent layers to enable a residual learning framework for an improved gradient and information flow [20]. Although the network architecture is comprised of a moderate number of layers, we observed that it is more effective to optimize when the convolutional layers are assembled in a residual block. The separated feature extractions in the first section enable the network to prioritize the single shafts based on their contribution to the regression problem – and likewise assign a lower weight to shafts that do not represent speech-relevant neural correlates. The third section sets up the network to predict 40 mel-scaled spectral coefficients in a continuous space.

The network configuration is composed of the following layers and properties: In the first section, we use a 2D convolutional layer with 10 kernels (size: 3×3) and a max pooling operation (stride: 2×2) to downsample the feature space. Subsequently, we use 2 convolutional layers that are organized in a residual block. In case an electrode shaft contains fewer than 4 electrodes, we omitted the residual learning since the input dimension would end up being too compressed. All feature maps of each shaft are concatenated and reshaped into a 1D tensor as the input for the second section, for which we use a stack of 4 consecutive residual blocks, each followed by a max pooling operation. Starting with 10 kernels for the first residual block, the number of kernels are gradually increased by 10 for each block. The third section is composed of a flattening operation followed by a fully connected layer with 40 units and a linear activation function.

A residual block consists of two convolutional layers which are connected through batch normalization and a ReLU non-linearity. The previous layer input is added through a shortcut-connection to the feature maps of the second convolution, before the second activation function. Both convolutions employ a kernel size of 3 and a stride of 1.

We used Adam [21] as the optimizer and trained for a fixed number of 50 epochs. In total, network architectures are comprised of 43.400-50.900 trainable weights.

D. Waveform Reconstruction

Similar to our previous studies [10], we use the Griffin-Lim [22] algorithm to recover lost phase information and convert the reconstructed speech spectrogram into audible audio due to its simplicity and time efficiency. In this iterative procedure, we limited ourselves to a number of 8 iterations to approximate the phase spectrogram since further improvement was not observed for additional iterations.

III. RESULTS

Network architectures were trained for each participant since the number of electrode shafts and their locations

differed across participants. We used a 5-fold cross validation to reconstruct the complete speech spectrogram from each participant's experiment run. In order to approximate a random chance level, a randomization test was performed. The acoustic data were split into two partitions at a random time point and the resulting partitions were swapped to break the temporal alignment with the sEEG signals. The identical reconstruction pipeline was performed using the swapped data. This randomization process was repeated 60 times per participant to obtain an approximation of the chance level.

In accordance with previous research studies, we measured the speech synthesis performance by computing the Pearson correlation for each spectral bin between a reconstructed spectrogram and its original counterpart – both for runs with proper and broken alignment. Figure 2.(A) shows the decoding results for our proposed method. We achieve correlation scores (averaged across all 5 folds, whiskers indicate standard deviation) of $r_1 = 0.43$ (± 0.10), $r_2 = 0.71$ (± 0.16), $r_3 = 0.66$ (± 0.13), $r_4 = 0.67$ (± 0.16), $r_5 = 0.80$ (± 0.10), respectively, that consistently outperform the chance level across all patients (Mann-Whitney-U test, $p < 0.001$).

In Figure 2.(B), we examine the reconstruction results averaged across all 5 participants in more detail by considering each bin of the 40 logarithmic mel-scaled spectral coefficients individually. It is observed that all frequency coefficients containing human voice information can be reconstructed with correlations above 0.5.

IV. DISCUSSION & CONCLUSION

In this study, we describe a CNN architecture specifically designed to find spatio-temporal patterns in neural activity from sEEG recordings that decode spoken speech processes. The method uses a multi-input layer, which is specifically tailored to first extract neural features from each electrode shaft separately before the final estimation considering all features. The resulting decoding performance is on par with previous results using ECoG recordings. Although the generated audio is not yet intelligible, these results show great potential for speech-related BCIs based on sEEG recordings. Further work is needed to understand how activity from the deeper structures relates to the cortical speech production networks, and how this information is distinct from that obtained from cortical recordings to determine which modality (or combination) will ultimately be most feasible for long-term implantation of a speech neuroprosthetic.

V. ACKNOWLEDGEMENT

C.H. acknowledges funding by the Dutch Research Council (NWO) through the research project 'Decoding Speech In SEEG (DESIS)' with project number VI.Veni.194.021. T.S., D.J.K and M.A. acknowledge funding by BMBF (01GQ2003) and NSF (2011595) as part of the NSF/NIH/BMBF Collaborative Research in Computational Neuroscience Program.

REFERENCES

[1] P. Coppens *et al.*, *Aphasia and related neurogenic communication disorders*. Jones & Bartlett Publishers, 2016.

[2] Q. Rabbani, G. Milsap, and N. E. Crone, "The potential for a speech brain–computer interface using chronic electrocorticography," *Neurotherapeutics*, vol. 16, no. 1, pp. 144–165, 2019.

[3] 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.

[4] 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.

[5] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, 2019.

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

[7] D. J. Krusienski, E. W. Sellers, F. Cabestaing, S. Bayoudh, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, "A comparison of classification techniques for the P300 Speller," *Journal of neural engineering*, vol. 3, no. 4, p. 299, 2006.

[8] D. A. Moses, N. Mesgarani, M. K. Leonard, and E. F. Chang, "Neural speech recognition: continuous phoneme decoding using spatiotemporal representations of human cortical activity," *Journal of neural engineering*, vol. 13, no. 5, p. 056004, 2016.

[9] J. G. Makin, D. A. Moses, and E. F. Chang, "Machine translation of cortical activity to text with an encoder–decoder framework," *Nature neuroscience*, vol. 23, no. 4, pp. 575–582, 2020.

[10] M. Angrick, C. Herff, E. Mugler, M. C. Tate, M. W. Slutzky, D. J. Krusienski, and T. Schultz, "Speech Synthesis from ECoG using Densely Connected 3D Convolutional Neural Networks," *bioRxiv*, p. 478644, 2018.

[11] C. Herff, L. Diener, M. Angrick, E. Mugler, M. C. Tate, M. A. Goldrick, D. J. Krusienski, M. W. Slutzky, and T. Schultz, "Generating Natural, Intelligible Speech From Brain Activity in Motor, Premotor, and Inferior Frontal Cortices," *Frontiers in Neuroscience*, vol. 13, p. 1267, 2019.

[12] C. Herff, D. J. Krusienski, and P. Kubben, "The Potential of Stereotactic-EEG for Brain-Computer Interfaces: Current Progress and Future Directions," *Frontiers in Neuroscience*, vol. 14, p. 123, 2020.

[13] D. J. Krusienski and J. J. Shih, "Control of a brain–computer interface using stereotactic depth electrodes in and adjacent to the hippocampus," *Journal of neural engineering*, vol. 8, no. 2, p. 025006, 2011.

[14] H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani, "Towards reconstructing intelligible speech from the human auditory cortex," *Scientific Reports*, vol. 9, no. 1, p. 874, 2019.

[15] M. Angrick, C. Herff, G. Johnson, J. Shih, D. Krusienski, and T. Schultz, "Interpretation of convolutional neural networks for speech spectrogram regression from intracranial recordings," *Neurocomputing*, vol. 342, pp. 145–151, 2019.

[16] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[17] C. Kothe, "Lab streaming layer (LSL)," *https://github.com/sccn/labstreaminglayer. Accessed on October*, vol. 26, p. 2015, 2014.

[18] E. Leuthardt, X.-M. Pei, J. Breshears, C. Gaona, M. Sharma, Z. Freudenburg, D. Barbour, and G. Schalk, "Temporal evolution of gamma activity in human cortex during an overt and covert word repetition task," *Frontiers in human neuroscience*, vol. 6, p. 99, 2012.

[19] N. Crone, L. Hao, J. Hart, D. Boatman, R. P. 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.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.