TOWARDS CLOSED-LOOP SPEECH SYNTHESIS FROM STEREOTACTIC EEG: A UNIT SELECTION APPROACH

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

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

ABSTRACT

Neurological disorders can severely impact speech communication. Recently, neural speech prostheses have been proposed that reconstruct intelligible speech from neural signals recorded superficially on the cortex. Thus far, it has been unclear whether similar reconstruction is feasible from deeper brain structures, and whether audible speech can be directly synthesized from these reconstructions with low-latency, as required for a practical speech neuroprosthetic. The present study aims to address both challenges. First, we implement a low-latency unit selection based synthesizer that converts neural signals into audible speech. Second, we evaluate our approach on open-loop recordings from 5 patients implanted with stereotactic depth electrodes who conducted a read-aloud task of Dutch utterances. We achieve correlation coefficients significantly higher than chance level of up to 0.6 and an average computational cost of 6.6 ms for each 10 ms frames. While the current reconstructed utterances are not intelligible, our results indicate promising decoding and run-time capabilities that are suitable for investigations of speech processes in closed-loop experiments.

Index Terms— neuroprosthesis, speech synthesis, stereotactic EEG, low-latency processing of neural signals

1. INTRODUCTION

Research related to neuroprosthesis, which use brain activity data to synthesize audible speech, has gained an increasing interest in recent years [1, 2, 3, 4]. The potential to regain spoken communication for people who have lost their ability to speak due to neurological disorders, such as amyotrophic lateral sclerosis (ALS) or severe paralysis, would significantly impact the quality of life of those affected. Joint research between the Brain-Computer Interface (BCIs) [5] and speech processing communities raises hope to serve these needs by identifying and decoding speech processes from brain activity and directly transforming them into audible speech to provide a spoken communication modality beyond acoustics [6].

Recent studies have demonstrated initial success with the reconstruction of intelligible speech from open-loop recordings of produced and perceived speech. Anumanchipalli et al. [7] employed a recurrent neural network to decode articulatory kinematic trajectories from electrocorticographic (ECoG) recordings, which were then synthesized into acoustic speech in a subsequent step. Herff et al. [8] relied on a concatenative approach by selecting waveform segments from a database of neural and acoustic pairs. This approach has been successfully replicated using intracortical electrode arrays with very similar results [9]. Akbari et al. [10] used penetrating depth electrodes to reconstruct perceived speech through a deep neural network model.

While these findings greatly advanced the state-of-the art in the field of speech related neuroprosthetics, many challenges remain and practical application for target users has not yet been achieved. One of these open challenges refers to the decoding capabilities from deeper brain structures. To date, the majority of studies acquire neural signals either from the scalp, such as electroencephalography (EEG), or directly from the cortex (ECoG) [11]. Despite the large potential of signals from deeper structures [12], the general adoption for BCI remains to be seen. A second open challenge refers to the requirement of a low-latency synthesis output to enable a natural conversation. While the related work has achieved high-quality reconstructions using open-loop recordings, it is still unclear whether these systems are applicable for real-time experiments.

In this study, we contribute to both of these open challenges. We developed a speech synthesizer that uses a unit

Correspondence to Miguel Angrick, miguel.angrick@uni-bremen.de.
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.

selection technique [13], a technique known to provide high-quality output even when little training data is available, to generate an acoustic waveform in real-time. For this, we build upon our prior findings for reconstructing intelligible speech based on concatenative approaches [8] and an initial attempt to decode imagined speech processes from closed-loop recordings [14]. Real-time capabilities are necessary to enable continuous low-latency feedback in closed-loop experiments. Our evaluation relies on open-loop data acquired from stereotactic depth electrodes (sEEG) to capture neural dynamics in cortical and deeper brain structures. This study aims to provide preliminary results and to explore the decoding capabilities and limitations of the proposed approach before deployment in closed-loop experiments.

2. MATERIAL AND METHODS

2.1. Experiment Design and Recording Setup

We conducted an experiment with 5 native speakers of Dutch, who were being monitored for intractable epilepsy via implanted sEEG electrodes to identify the epileptogenic zone. Placement of electrode shafts was purely determined based on clinical needs , ranging from 107 to 127 electrodes across patients. The patients performed a speech production task for which they read aloud 100 short utterances randomly drawn from the Mozilla Common Voice Dutch corpus [15], 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 patients' reading speed), followed by a pause of 1 second (except for patient 5, who conducted the experiment with a 2 second pause).

Neural data was digitized using a Micromed SD LTM amplifier (Micromed S.p.A., Treviso, Italy). Audio data was recorded at 48 kHz using the recording notebook's on-board microphone. LabStreaming Layer [16] was used to record sEEG and acoustic data in parallel.

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.

2.2. Data Processing

In order to compute meaningful features from the sEEG signals, we focused on the high-gamma band, which is known to contain highly localized information about speech production [17, 18] and has been successfully employed in previous speech-related decoding studies [19, 20]. We used a bandpass (70 - 170, 4th order Butterworth filter) filter to extract the high-gamma band and two bandstop filters (98 - 102 and 148 - 152, respectively, both 4th order Butterworth filters) 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. For each window, we calculated the signal power and applied a natural logarithm to normalize the distribution. We appended 4 non-redundant frames of preceding context to each window to model temporal dependencies of up to -200 ms in the past. Resulting features are of dimension $|\text{ frames }| \times |\text{ electrodes }| \cdot 4$.

Acoustic speech was resampled to 16 kHz and segmented into 150 ms windows with 10 ms frameshift to match the temporal alignment and number of windows of the high-gamma features. A length of 150 ms for the acoustic data was used to have sufficient acoustic samples to enable smooth transitions in the concatenation procedure. Finally, we employed feature selection to maintain a manageable number of high-gamma features by selecting the top 150 features yielding the highest correlation with the signal energy of the acoustic speech [14] for each patient. The number of 150 features was determined empirically in order to preserve as much information as possible and at the same time to be able to perform the computation steps in real time.

2.3. Unit Selection based on Brain Activity Data

In our decoding step, we synthesize acoustic speech using a unit selection paradigm – a technique that originates from the text-to-speech domain [13]. In addition, unit selection has been successfully deployed for voice conversion [21] and speech synthesis from facial EMG [22], an approach that is also applicable for similar silent-speech-interfaces [6]. Unit selection utilizes a database, also known as codebook, containing time-aligned pairs of high-gamma activity and waveform segments. This database enables the mapping from brain activity data to acoustic speech by selecting the pair with the highest similarity. Figure 1 illustrates the general concept.

In the training phase, we populate the database with corresponding pairs extracted from the training set. Each pair is responsible for the mapping of a single window (50 ms) of brain activity data to a single window (150 ms) of acoustic speech, where the brain activity data is aligned with the left third of the acoustic window. In the decoding phase, high-gamma windows are extracted in real-time from the sEEG stream, resulting in units of the same size as in the codebook. The extracted windows are then compared to all pairs in the database. Here, we follow the approaches from Herff et al. [8] and Wilson et al. [9] by relying on the cosine similarity distance metric:

$$\text{similarity}(A, B) = \frac{A \cdot B}{||A|| \cdot ||B||} = \frac{\sum_{i}^{n} A_i B_i}{\sqrt{\sum_{i}^{n} A_i^2} \sqrt{\sum_{i}^{n} B_i^2}}$$

By picking the pair with the closest cosine distance, resulting waveform segments are concatenated to obtain the final speech signal. We applied a Hamming window to reweight the acoustic amplitudes for smooth transitions.

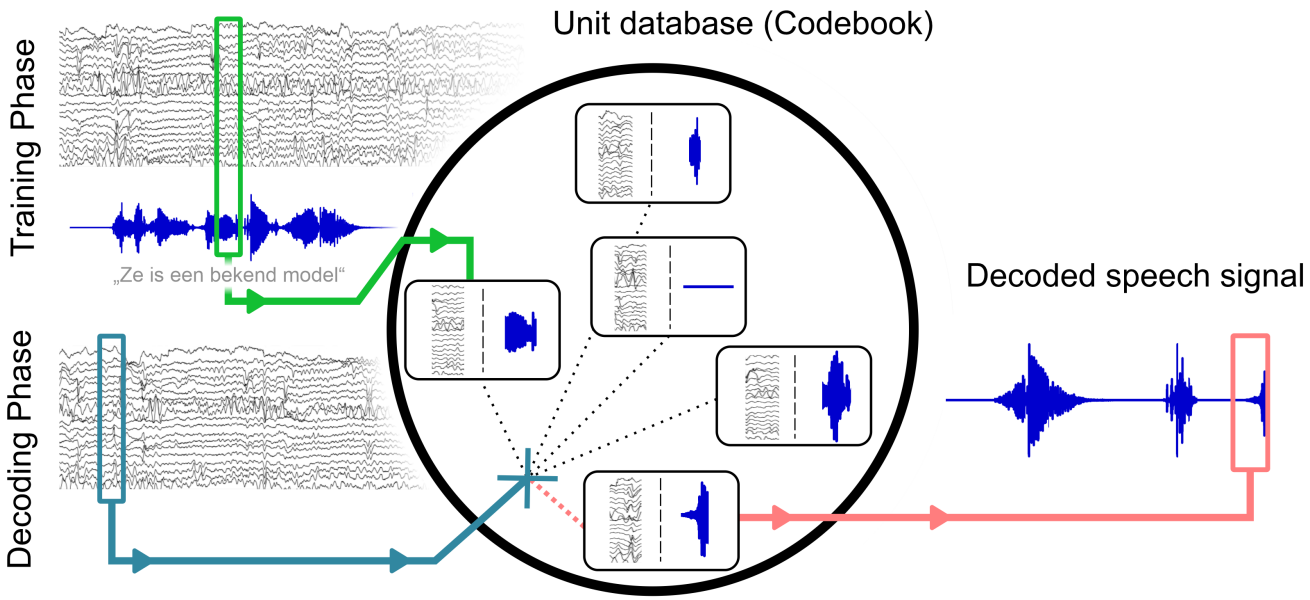

Fig. 1. Illustration of the unit selection concept for speech synthesis from invasive brain signals. In the training phase, pairs of brain activity and acoustic speech are stored in a unit database. In the decoding phase, windows of brain activity are compared to all entries in the database using cosine similarity. The approach selects the pair with the highest similarity score and concatenates its acoustic data to the output waveform. To ensure smooth transitions between previously selected and current waveform segments, we used a re-weighting based on a Hamming window function.

2.4. Closed-Loop Decoder Design

The implementation of our closed-loop decoder is build upon a node-based framework, which is specifically tailored for conducting experiments that require real-time stream processing of various biosignals to provide a continuous low-latency feedback for its participants. In brief summary, the framework enables to abstract certain functionality in self-contained nodes. These nodes can be linked together to provide a processing chain, where each node can receive multiple inputs, performs its calculation and propagates the results to the next nodes. Here, each node can be configured to initiate a new process to enable simultaneous computations in a multi-processor environment. For further details, we refer to our previous study about real-time decoding of imagined speech processes [14].

The unit selection approach is implemented across three subsequent nodes: In the first node, the decoding model computes the similarity with respect to each entry in the codebook by using the cosine similarity, and picks the waveform corresponding to the unit with the smallest distance. This segment gets inserted into a ringbuffer for storing a fixed amount of previous selections that are used in the concatenation step to enable smooth acoustic transitions. In the end, a reconstruction node is responsible for the final audio generation by using a hamming window to re-weight individual waveform samples from the second ringbuffer and placing them together.

We incorporated our unit selection approach to the system architecture of our previous study [14] by exchanging the linear models with our proposed method. Hence, the final decoding pipeline consists of one processing chain of nodes for the conversion from raw neural signals into an acoustic waveform. We used concurrent nodes to store intermediate and final results to quantify performances in the evaluation.

3. RESULTS

To test the proposed approach, we conducted a simulated online experiment. For each participant, we reconstructed the audio waveform of their experimental runs by using a 10-fold cross validation to split the brain activity and acoustic data into non-overlapping partitions for model training and testing. We then evaluated our results with respect to two criteria: (1) decoding performance of the generated output in comparison to its original speech and (2) the computational costs to enable such a conversion using the unit selection approach in a real-time capable synthesizer.

3.1. Decoding Performance

In accordance with previous investigations based on ECoG recordings [8, 20], we used the Pearson correlation as our performance metric. For each trial, we transformed both the original and reconstructed waveforms into the time-frequency

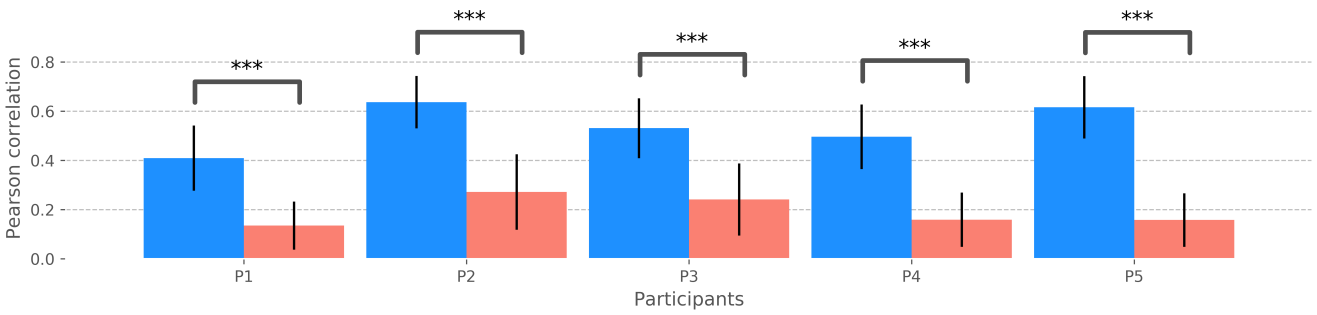

Fig. 2. Mean Pearson correlation scores across all trials of each participant. Blue bars indicate our proposed approach while red bars correspond to the chance level. Whiskers indicate standard deviation.

domain and averaged the correlation scores across frequency bins. The chance level was established as a baseline by splitting the neural data into two partitions at a random point in time and swapping them to break the alignment to the acoustic data. We then applied the same decoding pipeline to the rotated data and calculated the results. We repeated this step 100 times to get a closer approximation of the chance level. Figure 2 shows our decoding results. Our proposed method achieves scores of 0.41, 0.64, 0.53, 0.50 and 0.62, as indicated by the blue bars. Across all participants, the correlation results from the proposed method are significantly higher than the chance level (Man-Whitney-U test, $***P < 0.001$, red bars). While reconstructed audio sounds very natural, the generated speech is not intelligible at this point.

3.2. Computational Cost

We estimated the computational cost for each node in the network to ensure that the calculations can be executed in real-time. Based on a runtime of 1 min, we measured the respective processing times for our 50 ms windows with a 10 ms frameshift and calculated their average. The results are shown in Table 1. The data processing nodes need approximately 1.4 ms to execute their calculations, while the unit selection nodes require a runtime of 5.2 ms. The runtime of the unit-selection is dependent on the number of units n in the codebook. Assuming single comparison in constant time, the processing cost increases linearly as n grows, since in the decoding process, one feature frame is compared to all units in the codebook. This corresponds to a computational complexity of $\mathcal{O}(n)$. In total, the proposed approach needs 6.6 ms for the conversion of sEEG signals to audible speech, which is below the threshold of the 10 ms frameshift. We conducted our evaluation on a laptop with 16 GB of RAM and an Intel(R) Core(TM) i5-6600 CPU (3.3 GHz, 3301 MHz). To limit the interprocess communication, we used only one process for the proposed approach and dedicated nodes for storing intermediate results regarding the evaluation to circumvent delays from I/O operations.

Processing step	Processing costs $\left[\frac{ms}{\text{per 10 ms}}\right]$
Data Processing	
$\rightarrow$ Channel Selection	0.056
$\rightarrow$ High-Gamma	0.154
$\rightarrow$ Noise Filtering	0.282
$\rightarrow$ log Power Features	0.311
$\rightarrow$ Temporal Context	0.414
$\rightarrow$ Feature Selection	0.189
Unit Selection Approach	
$\rightarrow$ Unit Selection	4.209
$\rightarrow$ Waveform Ringbuffer	0.353
$\rightarrow$ Waveform Concatenation	0.599
Total:	**6.565**

Table 1. Mean processing costs for the data processing steps and the unit selection approach.

4. CONCLUSION

Here, we address two open challenges in the field of speech-related neuroprostheses using a unit selection technique to generate acoustic speech from sEEG recordings in real-time. Our results indicate reliable decoding performances yielding significantly higher correlations than the random chance level. While the synthesized speech is not yet intelligible and not on par with recent findings from ECoG recordings using a similar decoding paradigm, our results show promise in utilizing deeper brain structures for speech-related BCIs. The real-time capabilities presented by our approach are a mandatory requirement for moving towards closed-loop experiments. Further work is needed to assess the decoding performance of speaking modes beyond produced speech,

5. REFERENCES

[1] 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.

[2] S. Chakrabarti, H. M. Sandberg, J. S. Brumberg, and D. J. Krusienski, "Progress in speech decoding from the electrocorticogram," *Biomedical Engineering Letters*, vol. 5, no. 1, pp. 10–21, 2015.

[3] J. S. Brumberg, A. Nieto-Castanon, P. R. Kennedy, and F. H. Guenther, "Brain–computer interfaces for speech communication," *Speech communication*, vol. 52, no. 4, pp. 367–379, 2010.

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

[5] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," *Clinical neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[6] 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.

[7] 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.

[8] 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, pp. 1267, 2019.

[9] G. H. Wilson, S. D. Stavisky, F. R. Willett, D. T. Avansino, J. N. Kelemen, L. R. Hochberg, J. M. Henderson, S. Druckmann, and K. V. Shenoy, "Decoding spoken english from intracortical electrode arrays in dorsal precentral gyrus," *Journal of Neural Engineering*, vol. 17, no. 6, pp. 066007, 2020.

[10] 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, pp. 1–12, 2019.

[11] C. Cooney, R. Folli, and D. Coyle, "Neurolinguistics research advancing development of a direct-speech brain-computer interface," *IScience*, vol. 8, pp. 103–125, 2018.

[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, pp. 123, 2020.

[13] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. IEEE, 1996, vol. 1, pp. 373–376.

[14] M. Angrick, M. C. Ottenhoff, L. Diener, D. Ivucic, G. Ivucic, S. Goulis, J. Saal, A. J. Colon, L. Wagner, D. J. Krusienski, et al., "Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity," *Communications Biology*, vol. 4, no. 1, pp. 1–10, 2021.

[15] 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.

[16] C. Kothe, "Lab streaming layer (LSL)," *https://github.com/sccn/labstreaminglayer. Accessed on October 2020*, vol. 26, pp. 2015, 2014.

[17] 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.

[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, pp. 99, 2012.

[19] 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, pp. 217, 2015.

[20] 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," *Journal of neural engineering*, vol. 16, no. 3, pp. 036019, 2019.

[21] Z. Wu, T. Virtanen, T. Kinnunen, E. Chng, and H. Li, "Exemplar-based unit selection for voice conversion utilizing temporal information.," in *INTERSPEECH*. Lyon, 2013, pp. 3057–3061.

[22] M. Zahner, M. Janke, M. Wand, and T. Schultz, "Conversion from facial myoelectric signals to speech: a unit selection approach," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.