Towards direct speech synthesis from ECoG: A pilot study

Christian Herff[1], Garett Johnson[2], Lorenz Diener[1], Jerry Shih[3], Dean Krusienski[2] and Tanja Schultz[1]

***Abstract*— Most current Brain-Computer Interfaces (BCIs) achieve high information transfer rates using spelling paradigms based on stimulus-evoked potentials. Despite the success of this interfaces, this mode of communication can be cumbersome and unnatural. Direct synthesis of speech from neural activity represents a more natural mode of communication that would enable users to convey verbal messages in real-time.**

In this pilot study with one participant, we demonstrate that electrocoticography (ECoG) intracranial activity from temporal areas can be used to resynthesize speech in real-time. This is accomplished by reconstructing the audio magnitude spectrogram from neural activity and subsequently creating the audio waveform from these reconstructed spectrograms. We show that significant correlations between the original and reconstructed spectrograms and temporal waveforms can be achieved. While this pilot study uses audibly spoken speech for the models, it represents a first step towards speech synthesis from speech imagery.

I. Introduction

Brain-Computer Interface (BCI) research has made tremendous advances in the last several decades. The most prominent BCIs for communication rely on stimulus-evoked potentials in conjunction with spelling paradigms to type a single letter at a time [1], [2]. Because these systems operate in a stimulus-locked fashion, users can only communicate in predefined intervals. While speech can be synthesized via text-to-speech methods, these systems cannot operate in real-time. Direct synthesis of speech from neural activity represents a more natural mode of communication that would enable users to convey verbal messages in real-time. Additionally, such systems could convey other important aspects of speech communication such as accentuation and prosody.

Neuroscientific investigations show detailed insights into the production of speech [3], [4] and show that speech production and perception are processed very differently in motor areas [5]. Studies have shown that a complete set of English phonemes can be classified from electrocorticography (ECoG) [6], [7]. Others showed that speech recognition technology can be used to reconstruct a textual representation of spoken phrases using ECoG [8], [9]. Despite their innovative direction, these approaches suffer from the same limitations as typing approaches, as additional information of the spoken phrases is lost.

Pasely et al. were able to reconstruct perceived speech from neural activity [10] and Martin et al [11] showed reconstruction of low dimensional spectral representations from audible and imagined speech. We extend on these ideas by reconstructing a complete spectrogram from neural activity. We then use these reconstructed spectrograms to synthesize a waveform of the speech signal. This approach enables users to not only convey a message, but also add extra information such as accentuation, prosody and accent.

In this pilot study, we recorded audible speech and ECoG activity simultaneously from one participant and showed that the speech spectrogram can be reconstructed with promising correlations in an offline analysis. Furthermore, we show that this scheme is fast enough for real-time, online synthesis of speech from the neural signals.

II. Materials and Methods

A. Data Acquisition

Data were collected from a 42 year-old female patient with medically intractable epilepsy who underwent clinical evaluation to localize the epileptogenic zone prior to surgical resection. The patient consented to participate in the study as approved by the IRB of both Mayo Clinic and Old Dominion University. The patient had temporary placement of bilateral temporal depth electrodes (8 contacts apiece, 5 mm spacing), as well as three additional subdural strips placed on the cortex of the left temporal lobe (6 contacts apiece, 1 cm spacing). Electrode (Ad-Tech Medical Instrument Corporation, Wisconsin) placement and duration of intracranial monitoring were solely based on clinical evaluation, with no consideration given to this study. Electrode placements were verified using a postoperative CT. Figure 1 illustrates locations of subdural electrodes.

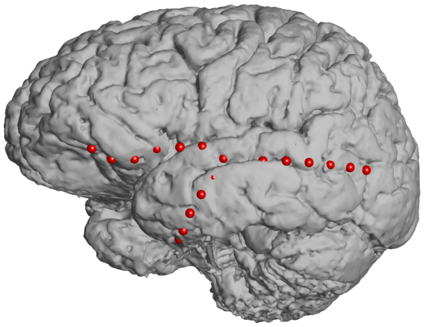

Fig. 1. Electrode positions for the pilot study participant.

The ECoG data were bandpass filtered between 0.5-500 Hz, digitized, and recorded using two 16-channel g.USB amplifiers (Guger Technologies, Austria) at a 1200 Hz

[1] C.H., L.D. and T.S. are with the Cognitive Systems Lab, University of Bremen, Bremen, Germany
[2] G.J. and D.K. are with the Advanced Signal Processing in Engineering and Neuroscience (ASPEN) Lab, Old Dominion University, VA, USA
[3] J.S. is with Mayo Clinic, Jacksonville, FL, USA
christian.herff@uni-bremen.de

sampling rate. Simultaneously, a Snowball iCE microphone (Blue Microphones, California) sampled the voice data at 48 kHz. The data recordings were synchronized using the general-purpose BCI system BCI2000 [12].

B. Experiment

For this study, a sentence was presented to the participant visually and aurally for 4 seconds. Subsequently, the participant had 4 seconds to recite the phrase aloud from memory. Sentences from the Harvard Sentence corpus [13] were used. A total of 50 sentences were recited, which resulted in a total of 200 seconds of data.

C. Feature Extraction

The recorded ECoG data were segmented into 50 ms intervals with 25 ms overlap. This duration is short enough to capture the cortical processes associated with speech production and are long enough to extract broadband gamma (70-170 Hz) activity, which is known to be highly task-related [14], [15].

To extract broadband-gamma, linear trends were first removed and data were subsequently downsampled to 600 Hz. The first harmonic of 60 Hz line noise was attenuated using an elliptic IIR notch filter. Elliptic IIR low-pass and high-pass filters were used to isolate the gamma band. Signal energy was then calculated on the filtered signal. A logarithm was applied to the energy estimates to give the power features a more Gaussian distribution.

Context information was included by concatenating 4 neighboring feature vectors up to 200 ms before and after the current interval. This resulted in a total of $18 \cdot 9 = 162$ features in each feature vector x_n for a time interval n.

The audio data was downsampled to 12 kHz to reduce the total spectrogram size. The audio spectrogram is calculated by taking the Short-Time Fourier Transform (STFT) in 50 ms intervals with 25 ms overlap, windowed using Hanning windows. This results in 301 frequency bins per interval. Only the magnitude of the STFT was utilized, as phase information can not be reconstructed from neural signals. The spectral information of a time interval n is denoted as f_n. As ECoG data and audio data are recorded simultaneously, each ECoG feature vector x_n can be assigned a corresponding audio spectrum f_n.

With the phase information missing, the audio signal can not be trivially reconstructed anymore and an approximation method as described in Section II-E is needed.

D. Spectrogram Reconstruction

A linear mapping between ECoG features and log power is estimated in a specific frequency bin. This mapping is obtained using a Lasso regression [16]. The optimal regularization weight α was determined using a nested 10-fold cross-validation. This results in a weight-vector v_i for each spectral bin i and a scalar intercept b_i. Once the models are trained for all spectral bins, all weight-vectors and intercepts can be combined to form a mapping matrix v and an intercept vector b. Using this combined representation, a new frame x_n of ECoG activity can be transfered to the spectral power representation f_n of the audio by simply calculating

$$f_n = v * x_n + b \tag{1}$$

Using a simple linear model for ECoG to speech mapping might not be optimal. Spectral reconstruction methods using deep learning methods have achieved great results in the past [17], but are usually orders of magnitude slower in training and require more time for reconstruction of each spectrum than the simple matrix multiplication needed in our approach. Since this is a pilot study, a linear model was used knowing that more complex methods should be investigated in the future.

E. Speech Synthesis

Given the spectrogram reconstructed from the measured ECoG activity f, one can reconstruct an audio waveform by iteratively modifying the spectral coefficients of a signal initialized with noise. Griffin and Lim [18] proposed Algorithm 1 to reconstruct the waveform from the spectrogram. With STFT & ISTFT being the Short-Term Fourier Transform and the Inverse Short-Term Fourier Transform, respectively. This allows the reconstruction of a complete audio waveform from the reconstructed spectrograms. Generally, only few iterations l of this procedure are necessary to yield sufficient audio quality. A value of $l = 8$ was chosen as no improvements could be seen with more iterations and processing was still very fast for 8 iterations. This algorithm can be used either on the complete reconstructed spectrogram in offline-analyses, or on each individual spectrum for online-synthesis. In this study, waveform reconstruction was performed on the entire reconstructed spectrogram.

Algorithm 1: Waveform reconstruction

```
Data: Spectrogram $f$
Result: Waveform $w$
$w \leftarrow$ noise;
for $i \leftarrow 1$ to $l$ do
    $X \leftarrow \mathrm{STFT}(w)$;
    $Z \leftarrow f \exp(i\angle X)$;
    $w \leftarrow \mathrm{ISTFT}(Z)$;
```

III. Results

A. Spectrogram and Waveform Reconstructions

Figure 2 illustrates an original and reconstructed (log) spectrogram. Figure 3 shows an example of original and reconstructed speech waveforms.

B. Computation Time

To assess the feasibility of our approach for online synthesis of speech from neural signals, all involved components were evaluated in terms of computational time and the thus induced time lag. As hardware offsets induced by data recording and audio output are not within the scope of this analysis, they have not been included. All calculations are

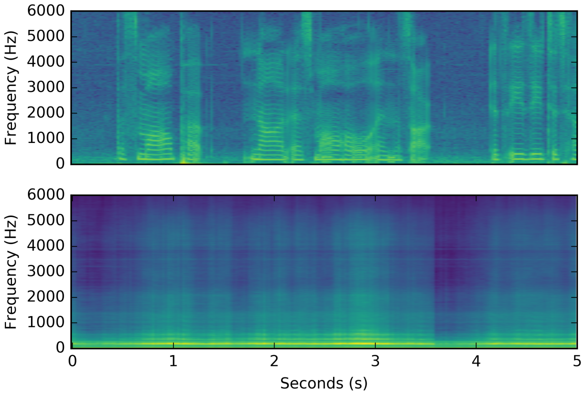

Fig. 2. Original (top) and reconstructed (bottom) spectrograms.

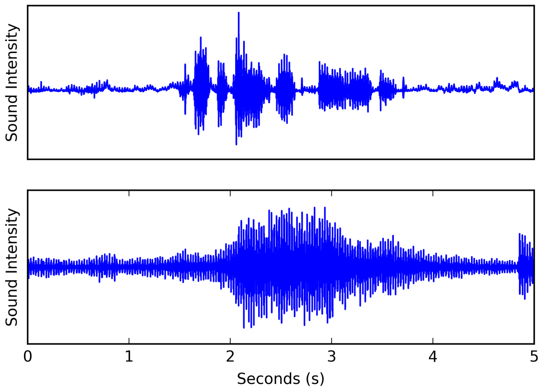

Fig. 3. 5 seconds of original (top) and reconstructed (bottom) waveform. Only very broad characteristics of the waveform can be seen in the reconstructed waveform.

performed on an Intel Core i7 processor running at 3.6 GHz . The time needed for data filtering, feature calculation, spectrogram reconstruction using the linear filter described in Section II-D and the waveform reconstruction described in Section II-E of one frame x_n of ECoG features resulting in 50 ms of audio were measured.

As can be seen in Table I, all operations can be performed in under 1 ms resulting in a total offset far smaller than the 50 ms interval length. Speech synthesis from neural signals can thus be performed in real time.

TABLE I
TIME NEEDED FOR COMPONENTS.

Operation	Computation time
Data filtering	<1 ms
Feature calculation	<1 ms
Spectrogram reconstruction	<1 ms
Waveform synthesis	<1 ms

C. Reconstruction Quality

All evaluations were performed using a 10-fold cross-validation: The Lasso regression models were trained on 90% of the data and were used to reconstruct spectrograms for the remaining 10%. This procedure was repeated 10 times, so that all data were used for testing once. The Lasso regularization parameter α was optimized using a nested 10-fold cross-validation on the training data. The models need approximately 1.5 seconds to be trained for each frequency bin. This would result in a total training time of about 450 seconds for the complete model.

We calculated the Spearman correlation coefficient ρ between the original and reconstructed spectrogram for each frequency bin to assess which parts of the spectral information can be robustly reconstructed. Figure 4 illustrates correlation coefficients over frequency bin. The mean overall correlation over all frequency bins is $\rho = 0.36$. Correlations below 200 Hz are around chance level as no speech information is present in this frequency range. From 200 Hz onwards, rank correlation coefficients increase until reaching a level of approximately 0.4 at around 300 Hz. As the first formant of vowel production usually starts around 300 Hz, high correlation in these frequency ranges is especially important. Rank correlations remain stable up to approximately 5 kHz, after which only little speech information is left in the spectrogram and correlation coefficients deteriorate rapidly in our evaluations. Despite these very promising results, it is evident from the short excerpt in Figure 2 that only very broad aspects of the spectrogram are reconstructed and improvements are still necessary to capture all delicate processes in the speech spectrogram. The achieved correlation coefficients are similar to those achieved by average subjects in [11].

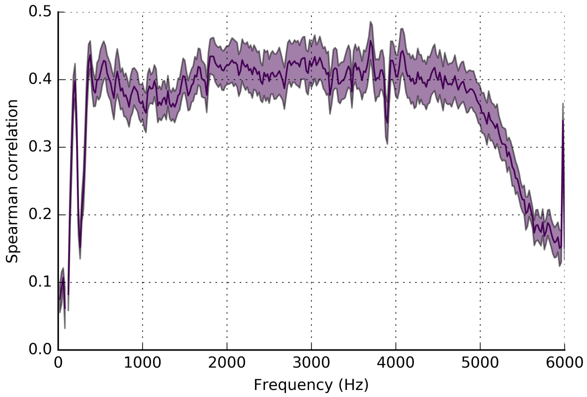

Fig. 4. Spearman correlation coefficients between original and reconstructed spectrograms for different frequency ranges. Purple shaded region denotes standard error of the mean over folds. Reconstruction remains relatively stable between 500 and 5000 Hertz.

To evaluate the synthesized waveform, Spearman correlations between the mean absolute Hilbert envelope in 50 ms intervals of the original and reconstructed waveforms were calculated. This yielded a Spearman correlation of $\rho = 0.41$, which is significantly better than chance (Randomization Tests, $p < 0.001$). As can be seen in Figure 3, the reconstructed waveform broadly captures the envelope of speech activity, but no detailed resemblance can be observed. Unsurprisingly, the reconstructed waveforms are not intelligible for our pilot study participant. We hypothesize that this might be due to the suboptimal electrode montage only covering areas in the temporal lobe with low density and thus not providing any information from motor areas which have been found to contain a lot of relevant information about speech production [5], [7], [8].

D. Interpretation of Regression Models

To visualize which neural activity is used to reconstruct the spectrogram, the corresponding forward models to the Lasso backward models v were estimated. This is done using the method described by Haufe et al. [19]. Figure 5 visualizes the mean forward model over all frequency bins for the pilot participant. Highest model weights are on regions in the auditory cortex. Activations are rendered using the NeuralAct Software package [20].

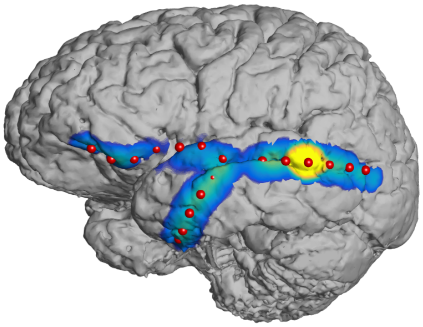

Fig. 5. Average activation pattern of regression models for spectrogram reconstruction.

IV. Conclusions

In a pilot study with one participant, we have shown that intracranial ECoG recordings can be used to synthesize speech. This is achieved by mapping the neural activity directly to magnitude spectrograms which allow for a reconstruction of a speech waveform. Our method yields reconstruction similar to previously reported spectral reconstructions despite a suboptimal electrode montage. Even though the reconstructed waveforms in this pilot study are not intelligible, performance is expected to improve with better coverage of more relevant brain areas. Most significantly, we verified that our approach is capable of achieving real-time online synthesis of speech from neural recordings, which is key in the development of future speech neuroprosthetics.

References

[1] E. Donchin, K. M. Spencer, and R. Wijesinghe, "The mental prosthesis: assessing the speed of a p300-based brain-computer interface," *Rehabilitation Engineering, IEEE Transactions on*, vol. 8, no. 2, pp. 174–179, 2000.

[2] G. R. Müller-Putz, R. Scherer, C. Brauneis, and G. Pfurtscheller, "Steady-state visual evoked potential (ssvep)-based communication: impact of harmonic frequency components." *Journal of neural engineering*, vol. 2, no. 4, pp. 123–130, 2005.

[3] 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, pp. 327–332, 2013.

[4] K. Bouchard and E. Chang, "Neural decoding of spoken vowels from human sensory-motor cortex with high-density electrocorticography," in *Engineering in Medicine and Biology Society, 2014. EMBS 2014. 36th Annual International Conference of the IEEE*. IEEE, 2014.

[5] C. Cheung, L. S. Hamiton, K. Johnson, and E. F. Chang, "The auditory representation of speech sounds in human motor cortex," *eLife*, vol. 5, p. e12577, mar 2016. [Online]. Available: https://dx.doi.org/10.7554/eLife.12577

[6] E. Mugler, M. Goldrick, and M. Slutzky, "Cortical encoding of phonemic context during word production," in *Engineering in Medicine and Biology Society, 2014. EMBS 2014. 36th Annual International Conference of the IEEE*. IEEE, 2014.

[7] 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.

[8] 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, no. 217, 2015.

[9] D. Heger, C. Herff, A. d. Pesters, D. Telaar, P. Brunner, G. Schalk, and T. Schultz, "Continuous speech recognition from ecog," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[10] 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.

[11] 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*, vol. 7, no. 14, 2014.

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

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

[14] E. C. Leuthardt, X.-M. Pei, J. Breshears, C. Gaona, M. Sharma, Z. Freudenberg, 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. 99–99, 2011.

[15] K. J. Miller, E. C. Leuthardt, G. Schalk, R. P. Rao, N. R. Anderson, D. W. Moran, J. W. Miller, and J. G. Ojemann, "Spectral changes in cortical surface potentials during motor movement," *The Journal of neuroscience*, vol. 27, no. 9, pp. 2424–2432, 2007.

[16] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[17] L. Diener, M. Janke, and T. Schultz, "Direct conversion from facial myoelectric signals to speech using deep neural networks," in *Neural Networks (IJCNN), 2015 International Joint Conference on*. IEEE, 2015, pp. 1–7.

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

[19] S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, and F. Bießmann, "On the interpretation of weight vectors of linear models in multivariate neuroimaging," *Neuroimage*, vol. 87, pp. 96–110, 2014.

[20] J. Kubanek and G. Schalk, "Neuralact: A tool to visualize electrocortical (ecog) activity on a three-dimensional model of the cortex," *Neuroinformatics*, pp. 1–8, 2014.