SPEECH SYNTHESIS USING EEG

Gautam Krishna   *Co Tran*   *Yan Han*⋆   *Mason Carnahan*⋆   *Ahmed H Tewfik*

Brain Machine Interface Lab, The University of Texas at Austin

ABSTRACT

In this paper we demonstrate speech synthesis using different electroencephalography (EEG) feature sets recently introduced in [1]. We make use of a recurrent neural network (RNN) regression model to predict acoustic features directly from EEG features. We demonstrate our results using EEG features recorded in parallel with spoken speech as well as using EEG recorded in parallel with listening utterances. We provide EEG based speech synthesis results for four subjects in this paper and our results demonstrate the feasibility of synthesizing speech directly from EEG features.

Index Terms— Speech synthesis, EEG, Deep Learning

1. INTRODUCTION

Speech production is one of the most important abilities of human beings which helps humans to communicate with each other. In recent years there is lot of research interest in developing assistive technologies to help with speech restoration for people with speaking disabilities. In [2] authors demonstrated synthesizing speech directly from invasive electrocorticography (ECoG) neural recordings using a recurrent neural network (RNN) based speech decoder. Recently in [1, 3, 4] authors have demonstrated speech recognition using electroencephalography (EEG) features where EEG signals recorded in parallel with spoken speech [4, 3] as well as EEG signals recorded in parallel with listening utterances [1] are translated directly into text. EEG is a non invasive way of measuring electrical activity of human brain. EEG has high temporal resolution like the invasive ECoG signals. The subjects need not undergo a brain surgery like in the case of ECoG for recording EEG since EEG is completely a non invasive technique. EEG sensors are placed on the scalp of the subjects to obtain the recordings.

In [1] authors introduced three types of EEG feature sets which are useful for speech recognition and speech synthesis. Though in [1] authors demonstrated preliminary results for synthesizing speech directly from EEG signals using a long short term memory (LSTM) [5] based regression model and generative adversarial network (GAN) [6] based model, they didn't provide speech synthesis results per each subject for their experiments and they provided only EEG feature

---

⋆ Equal author contribution

set 1 based result for synthesizing speech from EEG signals recorded in parallel with spoken speech. Where as in this paper we provide speech synthesis results per each subject using all the three different EEG feature sets for EEG signals recorded in parallel with spoken speech as well as from EEG signals recorded in parallel with listening utterances. More over in this paper we introduce a new EEG data set consisting of EEG, speech signals recorded for commonly used voice commands and our overall results mentioned in this paper demonstrate a significant improvement compared to the results demonstrated for speech synthesis by authors in [1]. We demonstrate in this paper that using a gated recurrent unit (GRU) [7] based RNN network with dropout [8] instead of a LSTM network results in speech synthesis performance improvement as well as in this paper we perform speech synthesis experiments using more number of speech, EEG recording examples per sentence compared to the ones used by authors in [1].

For synthesizing speech from EEG signals recorded in parallel with spoken speech, with our approach we were able to achieve a mel cepstral distortion (MCD) [9] value as low as 0.433 compared to 5.737 demonstrated by authors in [1] and for synthesizing speech from EEG signals recorded in parallel with listening utterances, with our approach we were able to achieve a MCD value as low as 0.471 compared to 1.34 demonstrated by authors in [1]. Given the challenges outlined by authors in [1] for training a GAN model for speech synthesis using EEG, we didn't perform experiments with GAN model.

Results mentioned in this paper demonstrates a significant first step towards synthesizing speech directly from EEG features. Synthesizing speech directly from EEG signals might help amyotrophic lateral sclerosis (ALS) patients who lost ability to speak with speech restoration.

2. SPEECH SYNTHESIS MODEL

Our speech synthesis model consists of two layers of gated recurrent unit (GRU) [7] with 256 hidden units in first layer and 128 hidden units in second layer as shown in Figure 1. The final GRU layer is connected to a time distributed dense layer of 13 hidden units to predict acoustic features of dimension 13 at every time step. Between each GRU layers and between the final GRU layer and time distributed dense layer,

a dropout regularization [8] with dropout rate 0.2 is applied. The model takes EEG features as input at every time step and outputs acoustic features at every time step. The model was trained for 250 epochs with adam optimizer with learning rate 0.01 [10] to observe loss convergence. The batch size was set to 100 and mean squared error (MSE) was used as the regression loss function for the model. For each subject's data we used 80 % data to train the model and remaining 10 % for testing and rest 10 % for validation set. The validation set was used to identify the right values for hyper parameters for the model. All the scripts for the model were written using Keras deep learning python framework.
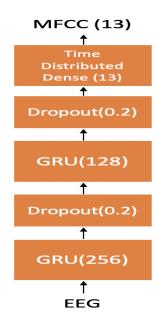

Fig. 1. Speech Synthesis Model

3. DESIGN OF EXPERIMENTS FOR BUILDING THE DATABASE

Four subjects took part in the EEG recording experiments. All were UT Austin undergraduate students in their early twenties. Three were females and one subject was male. Each subject was first asked to listen to four different natural utterances and then speak out loud the utterances that they listened to. The EEG was recorded in parallel while they were listening to the utterances as well as EEG was recorded in parallel while they were speaking out the utterances that they listened to. Throughout this paper we will refer to the EEG recorded in parallel with listening as listen EEG and EEG recorded in parallel with speech as spoken EEG. The listening utterances were simultaneously recorded with listen EEG and the subject's speech was recorded simultaneously with spoken EEG. The four natural utterances that the subjects listened were

"Hi Bixby", "Call Mom", "Open Camera" and " What's the weather". We collected 70 Speech - EEG recordings per each subject per each sentence.

We used Brain Vision EEG recording hardware. Our EEG cap had 32 wet EEG electrodes including one electrode as **ground** as shown in Figure 2. We used EEGLab [11] to obtain the EEG sensor location mapping. It is based on standard 10-20 EEG sensor placement method for 32 electrodes [12].
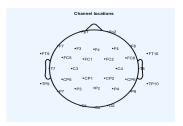

Fig. 2. EEG channel locations for the cap used in our experiments

4. EEG AND SPEECH FEATURE EXTRACTION DETAILS

EEG signals were sampled at 1000Hz and a fourth order IIR band pass filter with cut off frequencies 0.1Hz and 70Hz was applied. A notch filter with cut off frequency 60 Hz was used to remove the power line noise. EEGlab's [11] Independent component analysis (ICA) toolbox was used to remove other biological signal artifacts like electrocardiography (ECG), electromyography (EMG), electrooculography (EOG) etc from the EEG signals. We then extracted the three EEG feature sets explained by authors in [1]. The details of each EEG feature set are covered in [1]. Each EEG feature set was extracted at a sampling frequency of 100 Hz for each EEG channel [3].

The recorded speech signal was sampled at 16KHz frequency. We extracted mel-frequency cepstral coefficients (MFCC) as features for speech signal. We extracted MFCC features of dimension 13. The MFCC features were also sampled at 100Hz same as the sampling frequency of EEG features.

5. EEG FEATURE DIMENSION REDUCTION ALGORITHM DETAILS

We used kernel principal component analysis (KPCA) [13] to de-noise the EEG feature space by performing dimension reduction for each EEG feature set. By following the dimension reduction methods explained by authors in [1] we reduced EEG feature set 1 to a dimension of 30, EEG feature set 2 was reduced to a dimension of 50 and EEG feature set 3 was kept at original dimension of 93. More details of explained

variance plots used to identify the right feature dimensions are covered in [1].

6. RESULTS

Like mentioned earlier, for each subject we used 10 % of data as test set. During test time, EEG features from test set are fed as input to the trained speech synthesis model to output MFCC or acoustic features.

We computed three types of performance metrics namely the mel cepstral distortion (MCD) [9], root mean squared error (RMSE) and normalized RMSE between the predicted MFCC during test time and ground truth MFCC from test set to evaluate the performance of the model on test set for each subject. The RMSE values were normalized by dividing the RMSE values with the absolute difference between the maximum and minimum value in the test set observation vector.

Tables 1,2,3 and 4 shows results obtained during test time for predicting listen MFCC features from listen EEG features for each of the four subjects for various EEG feature set inputs. Tables 5,6,7 and 8 shows results obtained during test time for predicting spoken MFCC features from spoken EEG features for each of the four subjects for various EEG feature set inputs.

We used Griffin Lim reconstruction [14] algorithm to convert the predicted test time MFCC or acoustic features to audio or speech waveforms. We observed comparable MCD values for each subject for different EEG feature sets for each of the speech synthesis experiment as seen from the results shown in tables. Our results indicate if sufficient amount of data is available for training the speech synthesis model, then the effect on choice of EEG feature sets is negligible.

However we observed some what significant variation in MCD values across different subjects which indicates each brain generates unique set of signals during speech perception and production even though the different subjects were listening to the same natural utterances and speaking out loud the same sentences.

EEG Feature Set	Average MCD	Average RMSE	Average Normalized RMSE
Set 1	0.4713	5.557	0.0114
Set 2	0.4723	5.540	0.0114
Set 3	0.4744	5.536	0.0113

Table 1. Results for predicting **listen MFCC from listen EEG for subject 1**

EEG Feature Set	Average MCD	Average RMSE	Average Normalized RMSE
Set 1	0.829	7.636	0.0225
Set 2	0.80	7.419	0.0227
Set 3	0.814	7.521	0.0221

Table 2. Results for predicting **listen MFCC from listen EEG for subject 2**

EEG Feature Set	Average MCD	Average RMSE	Average Normalized RMSE
Set 1	0.64	6.69	0.016
Set 2	0.638	6.70	0.016
Set 3	0.63	6.69	0.016

Table 3. Results for predicting **listen MFCC from listen EEG for subject 3**

EEG Feature Set	Average MCD	Average RMSE	Average Normalized RMSE
Set 1	1.759	13.11	0.05
Set 2	1.765	13.14	0.05
Set 3	1.758	13.12	0.05

Table 4. Results for predicting **listen MFCC from listen EEG for subject 4**

EEG Feature Set	Average MCD	Average RMSE	Average Normalized RMSE
Set 1	0.433	4.867	0.0105
Set 2	0.435	4.881	0.0105
Set 3	0.435	4.911	0.0106

Table 5. Results for predicting **spoken MFCC from spoken EEG for subject 1**

EEG Feature Set	Average MCD	Average RMSE	Average Normalized RMSE
Set 1	0.856	8.04	0.0237
Set 2	0.847	8.03	0.0236
Set 3	0.841	7.96	0.0232

Table 6. Results for predicting **spoken MFCC from spoken EEG for subject 2**

7. CONCLUSION

In this paper we demonstrated synthesizing speech directly from non invasive EEG neural recording using a simple GRU based regression model for four subjects using different types of EEG feature sets. For synthesizing speech from EEG signals recorded in parallel with spoken speech, with our approach we were able to achieve a mel cepstral distortion

EEG Feature Set	Average MCD	Average RMSE	Average Normalized RMSE
Set 1	0.647	6.442	0.0155
Set 2	0.650	6.43	0.0156
Set 3	0.645	6.437	0.0156

Table 7. Results for predicting **spoken MFCC from spoken EEG for subject 3**

EEG Feature Set	Average MCD	Average RMSE	Average Normalized RMSE
Set 1	1.733	13.19	0.053
Set 2	1.736	13.19	0.054
Set 3	1.741	13.23	0.053

Table 8. Results for predicting **spoken MFCC from spoken EEG for subject 4**

(MCD) value as low as 0.433 compared to 5.737 demonstrated by authors in [1] and for synthesizing speech from EEG signals recorded in parallel with listening utterances, with our approach we were able to achieve a MCD value as low as 0.471 compared to 1.34 demonstrated by authors in [1]. Our results demonstrate adding regularization factors like dropouts to the speech synthesis model and training with more number of examples per sentence helps in reducing the test time MCD values. Our results demonstrates a significant first step towards synthesizing speech directly from EEG features. We further plan to publish the data sets used in this work to help advancement of research.

8. ACKNOWLEDGEMENT

9. REFERENCES

[1] Gautam Krishna, Yan Han, Co Tran, Mason Carnahan, and Ahmed H Tewfik, "State-of-the-art speech recognition using eeg and towards decoding of speech spectrum from eeg," *arXiv preprint arXiv:1908.05743*, 2019.

[2] Gopala K Anumanchipalli, Josh Chartier, and Edward F Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493, 2019.

[3] Gautam Krishna, Co Tran, Jianguo Yu, and Ahmed Tewfik, "Speech recognition with no speech or with noisy speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2019 IEEE International Conference on*. IEEE, 2019.

[4] Gautam Krishna, Co Tran, Mason Carnahan, and Ahmed Tewfik, "Advancing speech recognition with no speech or with noisy speech," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019.

[5] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[8] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[9] John Kominek, Tanja Schultz, and Alan W Black, "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion," in *Spoken Languages Technologies for Under-Resourced Languages*, 2008.

[10] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[11] Arnaud Delorme and Scott Makeig, "Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis," *Journal of neuroscience methods*, vol. 134, no. 1, pp. 9–21, 2004.

[12] Frank Sharbrough, "American electroencephalographic society guidelines for standard electrode position nomenclature," *J clin Neurophysiol*, vol. 8, pp. 200–202, 1991.

[13] Sebastian Mika, Bernhard Schölkopf, Alex J Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch, "Kernel pca and de-noising in feature spaces," in *Advances in neural information processing systems*, 1999, pp. 536–542.

[14] Daniel Griffin and Jae 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.