Linear and Nonlinear Reconstruction of Speech Envelope from EEG

Dat Quoc Ngo [1], Garret Oliver [2], Gleb Tcheslavski [2], Fei Chen [3], Chin-Tuan Tan [1]

[1] Erik Jonsson School Engineering and Computer Science, The University of Texas at Dallas, Richardson, USA

[2] College of Engineering, Lamar University, Beaumont, USA

[3] Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China

dqn1700000@utdallas.edu, gt.lamar@gmail.com, fchen@sustech.edu.cn, chin-tuan.tan@utdallas.edu

Abstract

Electroencephalography (EEG) is a non-invasive method of measuring cortical activities in association to speech communication. Recent studies have shown that EEG at multiple channels across head scalps were colinearly entrained to speech envelope. They were able to reconstruct speech envelope from EEG across scalp by using ridge regression. However, the predicted speech envelopes reconstructed by this linear model approach did not yield a high correlation when compared with the original speech envelopes. The outcome of those studies inspired us to explore a non-linear alternative with Deep Learning in reconstructing speech envelope from EEG. We proposed and developed an Encoder-Decoder model based on Convolutional (CONV) and Long-Short-Term-Memory (LSTM) layers to non-linearly reconstruct speech envelope from EEG. Our finding showed that correlation between the original speech envelope and the predicted speech envelope reconstructed with our model yielded a much higher value than the equivalence reconstructed with a linear model and a single-layer LSTM model. Our Encoder-Decoder model outperformed the regularized linear regression and the single-layer LSTM model with 134% and 21% improvement in correlation.

Index terms: Electroencephalography, Encoder-Decoder, Convolution, Long-Short-Term-Memory, Speech Envelope Reconstruction

1. Introduction

In our early study [9], we implemented a multivariate Temporal Response Function (mTRF) [4] to show the linear relationship between speech envelope and the EEG recorded at scalp electrode locations around the vertex of the head. We utilized the ridge regression in mTRF to reconstruct speech envelope of each single passage from the corresponding EEG recorded when subjects listen to the passage. The original mTRF model [5] was able to map the linear relationship between each EEG-speech envelope pair, but it did not generalize the relationship and assimilate into the model itself. To improve the generalization process in the model, we modified the mTRF model to update its randomly initialized weights at each EEG-Speech sample pair, instead of initializing separate weights for each sample pair. Despite the effort to optimize the model, the correlation with mTRF remained low, as shown in Table 1.

In the current study, we attempted to address the above issue using Deep Learning method, which is one promising technology to solve generation and classification problems. To date, Deep Learning methods [1, 4, 12, 13] have been deployed in the reconstruction of speech from brainwaves, including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Fully Connected Layers, and their combinations. Sakthi et al. [13] showed that the single-layer LSTM model with 100 modes in LSTM layer outperformed mTRF model [4] in reconstructing speech envelope from EEG. They found that the correlation (RHO) between the raw and predicted envelopes was 30% higher than that with the predicted envelope reconstructed by mTRF [5]. When our early work applied the single-LSTM model to our dataset, we did observe the improvement in RHO as compared to that with mTRF model. Despite improvement, the simplicity of the single-layer LSTM model may not be able to efficiently generalize EEG to speech envelope relationship. The benefit of the Encoder-Decoder architecture is to compress high-dimension representations of inputs into low-dimension and high-quality representations that the compression process reduces noise present in the input layer [3]. Raskov et al. [11] successfully implemented Autoencoder, an Encoder-Decoder variant, to establish the relationship between EEG and images.

The above-mentioned outcome has inspired us to combine the roles of Convolutional layer [14] and Long-Short-Term-Memory layer [13] to map EEG as input to speech envelope as output. We developed an Encoder-Decoder model composed of 1-Dimensional Convolutional layers as Encoder and LSTM layers as Decoder. The CONV layers were used in extracting feature representations of 2-dimensional data, which was the channel-by-time EEG data, and used in converting the feature representations into latent-space weights which would be learned by LSTM layers. During training, Gradient Back-Propagation technique [8] was used to minimize the error between the predicted and original speech envelopes to optimize latent-space weights [6]. As designed following the Encoder-Decoder architecture [3], our model attempted to yield latent-space weights which finally formed the mapping functions between EEG and speech envelopes.

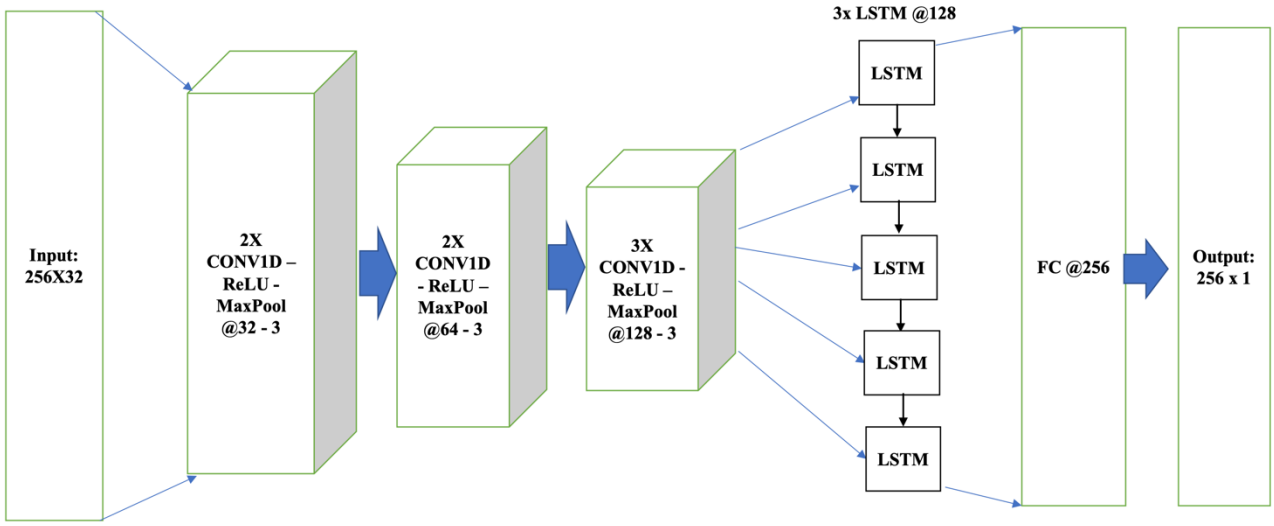

Figure 1: *Encoder-Decoder model with CNN for Encoder and LSTM for Encoder. CONV1D: 1-Dimension Convolutional layer; ReLU: Rectified Linear Unit; MaxPool: 1-Dimension MaxPooling layer; @#-#: filter size – kernel size; LSTM – Long-Short-Term-Memory; FC: Fully Connected layer.*

Figure 1 shows the structure of our proposed Encoder-Decoder model . The Encoder consists of 3 Convolution blocks of 1-Dimensional Convolutional layers followed ReLU activation functions. Each Convolution block is followed by a 1-Dimensional MaxPooling layer to downsize captured features. The Decoder consists of 3 LSTM layers which are commonly used in Machine Translation [10]. The final layer is a Fully Connected (Dense) layer to reconstruct speech envelopes from corresponding EEGs.

2. Experiment

In this experiment, we implemented a linear model and two non-linear models, i.e., mTRF [5], a single-layer LSTM model [13], and our proposed model, respectively, and compared their performance on the same EEG-Speech dataset in the task to reconstruct speech envelopes.

2.1 Data and Materials

2.1.1 Participants

This study was approved by the Lamar University IRB committee. Eight native English speakers (5 males and 3 females) with normal hearing (verified at a local hearing clinic) of age from 20 to 24 participated in this study. All participants reported no history of neurophysiological disorders and were not tired or sleepy at the time of data acquisition.

2.1.2 Stimuli

Four speech fragments spoken by a male speaker and four speech fragments spoken by a female speaker were extracted from two separate audio books and adjusted to three sound pressure levels (SPL's): (75, 65 and 55 dB). Two types of noise, i.e., White and Babble noise, were generated at three SPL's (75, 65 and 55 dB) and mixed with the speech to produce stimuli at different signal-to-noise ratios (SNRs) and speech levels. Note that speech at 75 dB SPL and noise at 65 dB SPL was considered a different listening condition than a speech at 65 dB SPL and noise at 55 dB SPL, even though both stimuli had same SNRs of 10 dB. Therefore, the stimuli pool consisted of 24 original and 96 noisy speech fragments.

2.1.3 EEG Recording

Audible stimulation was delivered diotically to the participants via Etymotic insert earphones (Etymotic Research ER3A, 10 Ω impedance). Stimuli presentation was synchronized with the EEG recording via evoked EEG was recorded via the ASA-Lab40 acquisition system by ANT Neuro, Netherlands. Continuous EEG was pre-filtered in 0.3-50 Hz range, notch-filtered at 60 Hz, sampled at 1,024 Hz, and recorded from 32 electrodes positioned according to the extended International 10/20 placement map. EEG was processed offline by first fragmenting it into epochs synchronized with stimulation. Each epoch was baseline-corrected and filtered with a CAR spatial filter to reduce surface currents.

2.1.4 Data Augmentation

Both EEG and speech recordings were down-sampled to 256 Hz. Before training, normalization was performed to scale EEG signals to the range of -1 and 1, and to scale speech envelopes to the range of 0 and 1. Data normalization is a necessary for time-series forecasting [2]. EEG-Speech recordings in our dataset also varied in length, which also posed difficulties on fixed-length computations. Each EEG-Speech recording was processed in a 1-s short-time window moving at the step of 125 ms. This

method allowed us to generate more data to better generalize our model.

2.2 Methods

For the linear model, we implemented an mTRF model, and trained the model as according to the original study [5], which was based on Ridge Regression, on our dataset for baseline performance verification. For comparison with the other nonlinear methods under the similar optimization criterion, we modified the training process and updated the weights of mTRF model by minimizing the Mean Squared Error (MSE) between the predicted and original speech envelopes for each EEG-Speech sample pair. For the nonlinear models, we first implemented a single-layer LSTM model with 100 nodes followed by a Dense layer of 256 nodes and trained the model for 500 epochs as according to [13] as one of the comparing models. The loss function used for training was the Mean Absolute Error (MAE) [16] and was optimized by Adam protocol [7]. Finally, in our proposed model, all Convolutional layers in Encoder were configured with a kernel size of 3 and with same padding. Three Convolution blocks had two 32-node CONV layers, two 64-node CONV layers, and three 128-node CONV layers respectively. In Decoder, 3 LSTM layers were set to 128 nodes. The output layer was a Dense layer with 256 nodes to reconstruct speech envelopes at an interval of 1-s. Our model's loss function was calculated by MSE; and the model was trained and optimized by Adam protocol for 200 epochs.

3 Results

Table 1: *Mean Pearson R (RHO), MSE, and MAE of predictions by mTRF, the single-layer LSTM model, and our model on validation dataset.*

	RHO	MSE	MAE
mTRF	0.261	0.056	N/A
Single-LSTM	0.506	N/A	0.104
Encoder-Decoder (ours)	0.613	0.019	N/A

Our dataset was randomly split: 80% for training and 20% for validation. We trained all 3 models on the training dataset and validated them on the validation dataset. Table 1 shows the performance of 3 models on the validation dataset.

In Table 1, we can observe that our model outperforms the single-LSTM model and mTRF, i.e., with RHO 0.613 vs. 0.506 and 0.261 of the single-LSTM model and mTRF, respectively. This results into 134% and 21% improvement in RHO.

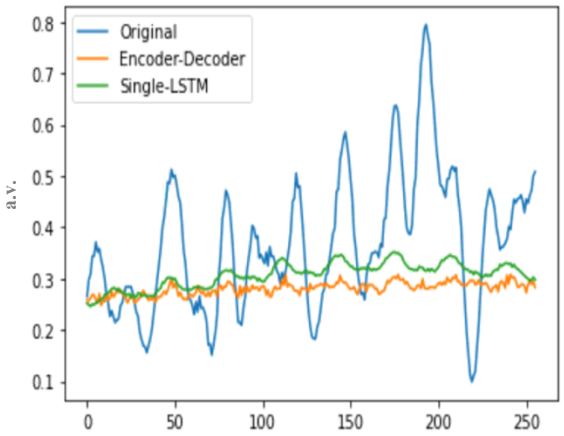

Figure 2*: Plot of a 1-s sample of original speech envelope and the corresponding speech envelope predicted by our model. Original: original speech envelopes; Encoder-Decoder: speech envelopes predicted by our model; Single-LSTM: speech envelopes predicted by the Single-LSTM model*

Similarly, our model outperforms mTRF with MSE at 0.019 compared with MSE at 0.056 of mTRF. These results showed that the non-linear LSTM-based models performed better than the linear-like models in generalizing EEG-Speech mappings.

In Fig. 2, we plot an original sample of speech envelopes and the corresponding speech envelopes predicted by the single-LSTM model and our model. We observed that the fluctuation of the predicted speech envelopes reconstructed with both our model and the single-LSTM model corresponded to the original speech envelopes, even the magnitude was not scaled to the same range. However, the predicted speech envelopes reconstructed with our model fluctuated less than the speech envelopes reconstructed with the single-LSTM model, as shown in Fig. 3. These results showed that the Encoder-Decoder model were able to reconstruct more highly correlated speech envelopes than the single-LSTM model.

In Fig. 3, during the training phase, we observed that our model encountered underfitting in correlation despite

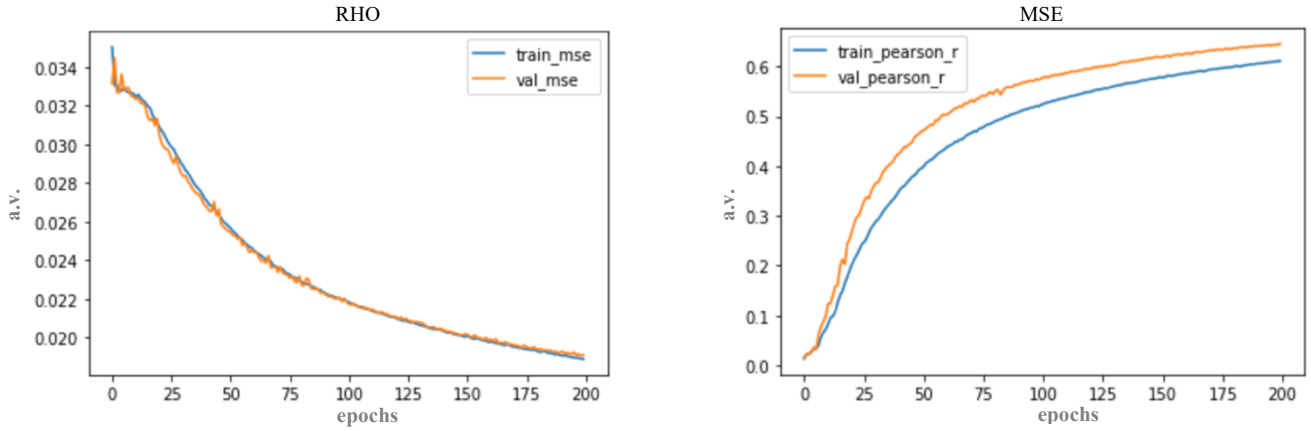

Figure 3: *Plot Mean training and validation MSE and Pearson R (RHO) for our model's predictions. Train_mse: training MSE; val_mse: validation MSE; train_pearson_r: training Pearson R; val_pearson_r: validation Pearson R.*

encountering neither overfitting nor underfitting in loss. The observed underfitting may result into the predicted speech envelopes which are less identical in shape to the original speech envelopes. The underfitting correlation could be resolved by increasing training time and complicating model's parameters. Also, we observed our model converged faster than the single-LSTM model. Our model achieved the validation correlation larger than 0.6 after 200 epochs, as shown in Fig. 3; and the single-LSTM achieved the validation correlation larger than 0.5 after 500 epochs, as shown in Table 1.

4 Discussion

Instead of having a single-layer of LSTM, our proposed model takes advantages of the noise reduction of the Encoder-Decoder architecture [3] to reconstruct the predicted speech envelopes with higher correlation when compared with the original speech envelopes. The results in this work showed that the Encoder-Decoder architecture was beneficial in tasks of generating the generalized mapping function between EEGs and speech envelopes. On the other hand, the limit of this model is the large difference in magnitude between the original and predicted speech envelopes. This may be due to the small size of the original dataset that each distinct EEG-Speech record has only 144 samples. Despite effort in augmenting data to generate more data, the small size of the original dataset limits our model's performance. In future work, more data will be collected and/or augmented to generalize our proposed model; and the current model will be modified to address the large magnitude difference and the underfitting issues.

5 References

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

[2] S. Bhanja and D. Abhishek, "Impact of Data Normalization on Deep Neural Network for Time Series Forecasting," Jan. 2019. [Online]. Available: http://arxiv.org/abs/1812.05519.

[3] D, Bhowmick, D. K. Gupta, S. Maiti, U. Shankar, "Stacked Autoencoders based Machine Learning for Noise Reduction and Signal Reconstruction in geophysical data,", Jul. 2019, [Online]. Available: https://arxiv.org/pdf/1907.03278.pdf

[4] Botchkarev, "Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology," *Interdiscip. Jrnl. Inf.*, *Knowl., and Mgmt*, vol. 14, pp. 045–76, 2019.

[5] M. J. Crosse, G. M. Di Liberto, A. Bednar, E. C. Lalor, "The Multivariate Temporal Response Function (MTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli," *Frontiers in Human Neuroscience*, vol. 10, 2016.

[6] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, L. Wang, G. Wang, J. Cai, T. Chen, "Recent Advances in Convolutional Neural Networks," Oct. 2017, [Online]. Available: http://arxiv.org/abs/1512.07108.

[7] D. P. Kingma, J. Ba. "Adam: A Method for Stochastic Optimization," Jan. 2017, [Online]. Available: http://arxiv.org/abs/1412.6980.

[8] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, "Gradient-Based Learning Applied to Document Recognition," in *Proc. IEEE*, vol. 86, no. 11, pp. 2278–324, Nov. 1998.

[9] D. Q. Ngo, G. Oliver, G. Tcheslavski C. Tan, "Neural Entrainment to Speech Envelope in response to Perceived Sound Quality," *9th Internat. IEEE/EMBS Conf. Neural Engin. (NER), California, USA,* pp. 920-923, 2019.

[10] B. Nouhaila, A. Habib, A. Abdellah, I. E. F. Abdelhamid, "Arabic Machine Translation using Bidirectional LSTM Encoder-Decoder," 2016, [Online]. Available: https://indabaxmorocco.github.io/materials/posters/Bensalah.pdf.

[11] G. Rashkov, A. Bobe, D. Fastovets, M. Komarova, "Natural Image Reconstruction from Brain Waves: A Novel Visual BCI System with Native Feedback," *Oct. 2019.*

[12] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, J. Faubert, "Deep Learning-Based Electroencephalography Analysis: A Systematic Review," *Journal of Neural Engineering*, vol. 16, no. 5, p. 051001, Aug. 2019.

[13] M. Sakthi, A. Tewfik, B. Chandrasekaran, "Native Language and Stimuli Signal Prediction from EEG*," ICASSP-IEEE International Conference. Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom,* pp. 3902-3906, 2019.

[14] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Apr. 2015, [Online]. Available: http://arxiv.org/abs/1409.1556.

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Mache Learning Research*, vol. 15, no. 56, pp. 1929–58, 2014.

[16] J. Willmott, K. Matsuura, "Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance," *Climate Research*, vol. 30, pp. 79–82, 2005.