Spee[ch reconstruction from human auditory cortex with deep neural networks

Minda Yang$^1$, Sameer A. Seth$^2$, Catherine A. Schevon$^3$, Guy M. McKhann II$^2$, Nima Mesgarani$^1$

$^1$Department of Electrical Engineering, Columbia University, New York, NY, USA.
$^2$Department of Neurosurgery, Columbia University Medical Center, New York, NY, USA.
$^3$Department of Neurology, Columbia University Medical Center, New York, NY, USA.

my2407,cas2044@columbia.edu, ss4451, gm317@cumc.columbia.edu, nima@ee.columbia.edu

Abstract

We examined the accuracy of the reconstructed speech spectrograms from neural responses recorded invasively in human auditory cortex. Electrodes were implanted in the cortex of epilepsy patients for the localization of seizures, and the neural responses were recorded as the subjects passively listened to continuous speech. We compared the reconstructed spectrograms estimated with two different models: a linear regression model and a deep neural network. Compared with linear regression model, the reconstructed spectrograms from the deep neural network achieved a higher average correlation with the original spectrograms. In addition, the reconstructed spectrograms from the neural network better preserved the average acoustic features of phonemes. We further investigated how changing the number of hidden layers in the network affects the reconstruction accuracy and found a better performance with deeper networks, particularly in the reconstruction of spectrotemporal modulation content of speech. These findings reveal the efficacy of deep neural network models in decoding speech signals from neural responses and provide a method for improving the performance of brain computer interfaces with prosthetic applications.

Index Terms: speech reconstruction, deep neural networks, brain computer interfaces

1. Introduction

Stimulus reconstruction [1], [2] is an inverse mapping technique in which the neural responses are used to approximate the acoustic representation of the sound that was heard by a subject. This method can be used to investigate what stimulus features are encoded in the neural responses [3] by examining the neural information in the stimulus space, where it is better understood. In addition, stimulus reconstruction is used in brain-computer interfaces [4], [5] and neural signal processing.

In this study we used recorded multi-electrode neural responses to speech from the human auditory cortex to reconstruct the spectrogram of the speech stimulus presented passively to the subjects. Previous studies have performed stimulus reconstruction using linear regression (LR) on spectrogram [2], or spectrotemporal modulation representation of speech [4], here we use an alternative technique based on deep neural network (DNN) models, motivated by their recent success in various machine learning applications [6]–[8]. Deep learning can improve the reconstruction accuracy both by imposing a better prior on the reconstructed spectrogram, due to their superior ability to learn its structure, and by formulating a better mapping for the nonlinearly encoded features of speech in the neural data, which is ubiquitous through out the auditory pathway [9]–[12].

To quantify the difference between these models, we compared the correlation values between actual and reconstructed speech samples to measure the reconstruction accuracy, along with the correlation between average phoneme spectrograms estimated from the original and reconstructed signals. We also tested the performance of DNNs with different number of hidden layers on the reconstruction problem to see if a deeper network can achieve a better performance.

2. Experiments

2.1. Speech perception task

Two subjects with normal hearing underwent the placement of subdural electrodes as part of their clinical treatment for epilepsy. The population neural activity recorded from 13 electrodes in total was included for this research. The speech presented to the subjects included words and sentences recorded from two voice actors (one male, one female). Speech sounds were presented aurally to the subjects from a loud speaker and the duration of speech presented from each speaker was approximately 10 minutes. The speech corpus contains sentences that are commonly used in daily communication. An example sentence being "Can I have a cup of green tea please". Speech stimuli were phonetically transcribed using Penn Phonetics Lab Forced Aligner [13]. With the phonetically transcribed speech stimuli, we can estimate the average phoneme spectrograms by aligning all the
instances of a certain phoneme by the onset time and taking the mean over a fixed duration.

2.2. Signal processing

Cortical local field potentials (LFP) were recorded with multi-electrode arrays and connected to a digital signal processor (Natus Medical Incorporated, Pleasanton, CA, USA). The neural signals were acquired at 500 Hz.

We visually and quantitatively inspected the time series recorded from each channel for artifacts and excluded channels with excessive noise. To reduce the effect of the reference, the average of channels was subtracted from all signals (Common Reference Average). The time-varying analytic amplitude was extracted from eight band-pass filters (Gaussian filters, centered logarithmically between 70 and 150 Hz (high-gamma band) and semi-logarithmically increasing band-widths) with the Hilbert transform. The high-gamma power was taken as the averaged amplitude across these eight bands. The neural signals were then down-sampled to 100 Hz and normalized to have zero mean and unit standard deviation for each channel. We identified speech responsive channels by comparing responses during speech and silence (t-test), and channels with t-value over 30 were included for subsequent analysis.

2.3. Methods

We performed stimulus reconstruction using both LR [2] and DNNs to map the population neural activity to the spectrogram of the speech stimulus. We used optimal prior reconstruction as a linear model to reconstruct the original stimulus from population neural responses [14], [15]. The response at electrode n at time \( t = 1 \ldots T \) is denoted as \( R(t, n) \). The spectrogram \( S(t, f) \) is a function of the time \( t \) and frequency \( f \). As neural responses in human auditory cortex are not phase-locked to the modulations in the original sound pressure waveform, the inverse filter \( g(t, f, n) \) is defined as a linear function that maps \( R(t, n) \) to the time-frequency representation of speech \( \hat{S}(t, f) \) as follow

\[
\hat{S}(t, f) = \sum_n \sum_\tau g(f, \tau, n) R(t - \tau, n) \quad (1)
\]

The inverse filter \( g \) is estimated by minimizing the mean squared error between actual and reconstructed stimulus

\[
\min_e \sum_n \sum_f [S(t, f) - \hat{S}(t, f)]^2 \quad (2)
\]

Resulting in normalized reverse correlation solution [14], [15]:

\[
g = C_{RR}^{-1}C_{RS} \quad (3)
\]

Where \( C_{RR} \) and \( C_{RS} \) are the auto-correlation of neural responses and cross-correlation of stimulus and neural responses at different lags, respectively. We also applied regularization when calculating the inverse filter \( g \) by truncating the eigen vectors of \( C_{RR} \) to optimize the reconstruction accuracy for the validation set [16], [17], which also improved the average correlation for the test data.

The DNN on the other hand was used as a nonlinear regression model to reconstruct the speech spectrogram from the time-shifted neural responses. Unlike the neural networks used in classification problems, here we replaced the softmax output layer with sigmoid. The activation \( a_k^L \) of \( j^{th} \) unit in the \( L^{th} \) layer (the output layer) is related to the activations in the \( (L-1)^{th} \) layer by the equation

\[
a_k^L = \sigma \left( \sum_j w_{kj}^L a_j^{L-1} + b_k^L \right) \quad (4)
\]

Where \( w_{kj}^L \) is the weight connecting \( k^{th} \) unit in the \( (L-1)^{th} \) layer and \( j^{th} \) unit in the \( L^{th} \) layer and \( b_k^L \) is the bias for \( j^{th} \) unit in the \( L^{th} \) layer. The sigmoid function is denoted by \( \sigma \). The training of the DNN model consists of unsupervised pre-training with Restricted Boltzmann Machines (RBM) [6] followed by supervised fine-tuning using backpropagation algorithm [18]. The parameters of RBM are trained on consecutive layers with the approximate contrastive
approaches aim as speech has a characteristic spectrotemporal modulation pattern, the early 5 for the set. For pre-stopping technique between the actual and reconstructed parameters are divergence algorithm [6]. For the supervised fine-tuning, the parameters are adjusted to minimize the mean squared error between the actual and reconstructed stimulus

$$\min e_z = \sum \sum [S(t, f) - \hat{S}_z(R, W, b)]^2 \quad (5)$$

Where $\hat{S}_2 = (a_1, a_2, \ldots, a_f)$ is the reconstructed spectrogram and $\hat{S}_2$ is a function of the input $R$, weight $W$ and bias $b$ of the trained network. To prevent over-fitting, early-stopping technique was used to terminate the training process when the performance no longer improves on the validation set. For pre-training, the number of epochs was 20, and the learning rate was 0.001. The learning rate for fine-tuning was 5 for the first 30 epochs and decreases by 5% afterward until the early-stopping criteria was met.

To study the temporal and spectral structure in the reconstructed speech from the neural signals, we estimated the spectrotemporal components of the reconstructed speech using a model of sound processing in the auditory cortex [19]. Since speech has a characteristic spectrotemporal modulation pattern, this analysis can indicate how these components are decoded from the brain signals, and how phonological information such as syllabic rates in temporal and formant and pitch information in the spectral domain are preserved [20], [21].

Figure 1 shows the schematic of the framework. The input data to both the LR and the DNN models was time shifted high-gamma neural response over 250 ms window. We used cross-validation technique to explore each sample in the speech data set with the LR and the DNN models. Both approaches aim to minimize the mean square error between the original and reconstructed spectrogram. To explore the effect of number of layers in the neural networks on reconstruction accuracy, we varied the number of hidden layers from one to four and kept the training procedure unchanged. For all models, each hidden layer has 64 units. All the results reported in this paper are from the neural network with three hidden layers unless otherwise specified.

3. Results

The resulting accuracy of reconstruction from the LR and DNN models were compared as well as their corresponding average accuracy across acoustic features of phonemes and different manners of articulation. We also tested the effects of DNNs with different number of hidden layers on reconstruction accuracy and the spectrotemporal modulation content as described below.

3.1. Reconstruction accuracy for the LR and the DNN model

We used the correlation between original and reconstructed spectrograms to evaluate the accuracy for the two models. Figure 2A shows the comparison of correlation values with original speech spectrogram for the LR (horizontal axis) and the DNN (vertical axis). Each point corresponds to one sentence. The majority of speech samples have higher correlation value for the DNN compared with the linear reconstruction. The average and standard deviation of the correlation between reconstructed and original spectrograms over 272 sentences were 0.69 ± 0.14 for the DNN model, compared with 0.58 ± 0.16 for the linear model. Three representative examples are shown in Figure 2B demonstrating the range of correlation accuracies. To examine the reconstruction accuracy for time and frequency dimensions separately, we averaged the spectrograms along these dimensions and compared to the original signals. These averages are shown in Figure 2C for the first two examples, where the DNN (red plots) show a better match to the target (black plots). We also quantified this effect by measuring the correlation between time averages of spectrograms (0.74 ± 0.15 for the DNN model, compared with 0.67 ± 0.17 for the...
Correlation on phoneme spectrograms

We report reconstruction accuracy for all the phonemes is significantly improved from the linear model. The correlation accuracy for all the phonemes is significantly larger in the DNN model than in the linear model. The reconstruction accuracy for all the phonemes is significantly larger in the DNN model compared to the linear model (0.67 ± 0.04, compared to 0.56 ± 0.05 for the linear model).

Since phonemes with different manners of articulation exhibit distinct spectral and temporal acoustic features that could impact how they are reconstructed with LR and DNN models, we examined the improved average reconstruction accuracy of phonemes separately for different manners of articulations. These include plosives (oral occlusive in which the vocal tract is blocked before the airflow is released) such as /p/ /d/, vowels (sound produced with no constrictions in the vocal tract) such as /a/ /e/ /i/ /o/ /u/, fricatives (sound produced by forcing air through a narrow channel made by placing two articulators close) such as /s/ /f/ and nasals (voice resonating in the nose) such as /m/ /n/ /ng/ /n/ [22]. We observed that the improvement in the reconstruction accuracy was highest for the plosives (Figure 3B) that are featured by sudden energy burst in their spectrograms due to the release of the airflow.

The average phoneme spectrogram for one example phoneme (/er/) is shown in Figure 3C. The DNN reconstruction in this case maintains the spectral and temporal features of the phoneme such as the formants, consistent with the higher correlation results shown in Figure 3A.

3.2. Reconstructing average phoneme spectrograms

Phonemes are the smallest contrastive units in a given language [22] and their robust perception is critical for speech communication. We therefore studied how the acoustic features of various phonemes are preserved in reconstructed spectrograms from LR and DNN models.

Figure 3A shows the average correlation values between original and reconstructed phoneme spectrograms, for LR (horizontal axis) and the DNN model (vertical axis). The reconstruction accuracy for all the phonemes is significantly larger in the DNN model (0.67 ± 0.04, compared to 0.56 ± 0.05 for the linear model).

Since phonemes with different manners of articulation exhibit distinct spectral and temporal acoustic features that could impact how they are reconstructed with LR and DNN models, we examined the improved average reconstruction accuracy of phonemes separately for different manners of articulations. These include plosives (oral occlusive in which the vocal tract is blocked before the airflow is released) such as /p/ /d/, vowels (sound produced with no constrictions in the vocal tract) such as /a/ /e/ /i/ /o/ /u/, fricatives (sound produced by forcing air through a narrow channel made by placing two articulators close) such as /s/ /f/ and nasals (voice resonating in the nose) such as /m/ /n/ /ng/ /n/ [22]. We observed that the improvement in the reconstruction accuracy was highest for the plosives (Figure 3B) that are featured by sudden energy burst in their spectrograms due to the release of the airflow.

The average phoneme spectrogram for one example phoneme (/er/) is shown in Figure 3C. The DNN reconstruction in this case maintains the spectral and temporal features of the phoneme such as the formants, consistent with the higher correlation results shown in Figure 3A.

3.3. Effect of number of hidden layers on reconstruction accuracy

We varied the number of hidden layers in the DNN model and evaluated the reconstruction accuracy using average correlation on phoneme spectrograms and spectrotemporal modulations. The correlation value between reconstructed and original spectrograms for each manner group using neural networks with different number of layers are shown in Figure 4. The reconstruction accuracy improved significantly from the LR to the DNN models, with further improvement up to 3 hidden layers. Reconstruction accuracy was not improved with adding hidden layers beyond 3 (Figure 4), which could be due to overfitting resulted from limited training data.

It has been shown that spectral and temporal modulations of speech play a critical role in speech perception and have been used reliably to predict its intelligibility [23]. The temporal modulation rate for speech is relatively low (2-16 Hz), reflecting the speed of the phonetic and syllabic rates of speech [20], [24]. Here, we examined how the spectrotemporal modulation content of the reconstructed speech from the DNN models varied as the number of hidden layers increased from 1 to 3. Figure 5 shows the spectral and temporal modulations for the original spectrogram and reconstructed spectrograms from the DNNs with different number of hidden layers. The high spectral modulation (scale) and high temporal modulation (rate) components outside the normal range for speech may indicate the noise in the reconstructed spectrogram. These components were reduced when the DNNs with more hidden layers were used, suggesting a regularization role for deeper layers. The correlation values between the spectral and temporal modulations of original and reconstructed speech for DNNs with one to three hidden layers were 0.404, 0.499 and 0.613, respectively.

4. Conclusion

In this study, we proposed a method of reconstructing the speech stimulus using subdural multi-electrode recordings with the DNN as a regression model. We demonstrated that the DNN outperformed the LR model in correlation accuracy, the average acoustic features of phonemes, as well as temporal and spectral profiles. Through multiple layers of nonlinear transformations, the DNN can model the nonlinearity in the stimulus-response transformation, and learn the structure in the speech signal more effectively then the LR model, resulting in a better overall performance.

We further investigated the effects of number of hidden layers in neural networks and found that DNNs with more hidden layers achieved improved performance. This was partly due to a better regularization of the stimulus-response mapping, resulting in reduced spectrotemporal distortion in the reconstructed signals.

By adopting state-of-the-art machine learning algorithms such as deep neural networks we improved speech reconstruction from invasive neural recordings in human auditory cortex. This finding has significant implications for brain computer interfacing technologies and neural prostheses, in addition to providing a powerful technique for studying the representational properties of speech in human auditory cortex.

5. Acknowledgement

All experiments were approved by the Columbia University Institutional Review Board. The authors would like to thank Elliot H. Smith, Taylor B. Nelp, Chuck Mikell and Tasha A. Nagamine for technical help. Minda Yang was partly funded by the National Institute of Health, NIDCD, DC014279.
6. Reference


