Reconstructing Speech Stimuli from Human Auditory Cortex Activity
Using GAN based Approach

THESIS

Submitted in Partial Fulfillment of
the Requirements for
the Degree of

MASTER OF SCIENCE (Computer Engineering)

at the

NEW YORK UNIVERSITY
TANDON SCHOOL OF ENGINEERING

by

Zhaoxi Chen

May 2019

Approved:

________________________________________
Advisor Signature

________________________________________
Date

________________________________________
Department Chair Signature

________________________________________
Date

University ID:  N18963553
Net ID:  zc1134
Vita

Zhaoxi Chen was born in Sichuan, China on July 5, 1995. She started her undergraduate education at University of Science and Technology of China in September 2013 and in July 2017 she graduated with a Bachelor of Eng. In Electronic Information Engineering. In September 2017, she enrolled in New York University, Tandon School of Engineering. Her purpose is to graduate with a master's degree in Computer Engineering. She studied the thesis in NYU Video Lab from September 2018 to May 2019.
Acknowledgment

Thanks to Ran Wang for sharing results for WaveNet and pre-trained approach on the low-density dataset. Thanks to professor Adeen Flinker for providing the data and interpretation of data and results.
ABSTRACT

Reconstructing Speech Stimuli from Human Auditory Cortex Activity Using GAN based Approach

by

Zhaoxi Chen

Advisor: Prof. Yao Wang

Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science (Computer Engineering)

May 2019

The superior temporal gyrus (STG) is the cortical region responsible for the sensation of sound. In this work, we show that two GAN based approaches, with limited available data, are able to reconstruct speech stimuli from STG intracranial recordings. Using a mapping-generator structure, the mapping part is unique for each subject, mapping his/her STG recording to a common latent space, from which a shared generator reconstructs speech stimuli. In the first approach, we pre-train the generator part with a large number of speech clips to improve its reconstructing ability. The other CycleGAN based method is used to combine supervised learning with unsupervised, which also take advantage of the large extra amount of speech clips to increase regression accuracy.
Table of Contents

1. Introduction......................................................................................................................................... 1
2. Related Works....................................................................................................................................... 3
3. Method.................................................................................................................................................. 8
   3.1 Mapping-generator ....................................................................................................................... 8
   3.2 Pretrained Network ..................................................................................................................... 10
   3.3 CycleGAN based Network .......................................................................................................... 11
4. Experiments ......................................................................................................................................... 13
   4.1 Data Acquisition ....................................................................................................................... 13
   4.2 Spectrogram reconstruction ...................................................................................................... 14
       4.2.1 Pretrained Network ........................................................................................................... 14
       4.2.2 CycleGAN ......................................................................................................................... 15
   4.3 Result ........................................................................................................................................... 15
5. Conclusion .......................................................................................................................................... 18
6. References .......................................................................................................................................... 19
List of Figures

Fig. 1. The lateral surface of the left cerebral hemisphere…………………………………1
Fig. 2. ECoG coverage viewed from the side…………………………………………………1
Fig. 3. Modified WaveNet structure…………………………………………………………3
Fig. 4. (a) CycleGAN structure, (b) forward cycle-consistency loss, and (c) backward cycle-consistency loss…………………………………………………………………..7
Fig. 5. Mapping-generator Network Structure………………………………………………9
Fig. 6. Pre-training approach………………………………………………………………….10
Fig. 7. CycleGAN approach. (a) structure, (b) forward cycle-consistency loss and (c) backward cycle-consistency loss…………………………………………………………….12
Fig. 8. Results on the low-density dataset, on each subfigure the upper row is predicted spectrogram and bottom one is ground truth. (a) WaveNet trained from scratch, (b) WaveNet with pre-trained GAN and finetune…………………………………………………………….16
Fig. 9. Results on the high-density dataset, on each subfigure the upper row is predicted spectrogram and bottom one is ground truth. (a) WaveNet trained from scratch, (b) WaveNet with pre-trained GAN and finetune, (c) WaveNet with CycleGAN………………….17
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>DILATION RATES FOR EACH RESIDUAL BLOCK IN THE GENERATOR</td>
<td>8</td>
</tr>
<tr>
<td>II</td>
<td>HYPERPARAMETERS FOR MAPPING-GENERATOR STRUCTURE</td>
<td>10</td>
</tr>
<tr>
<td>III</td>
<td>QUANTITATIVE EVALUATION</td>
<td>16</td>
</tr>
</tbody>
</table>
1. Introduction

1.1 Gyrus and ECoG

Gyrus is a ridge on the cerebral cortex. Many gyri contain important structures of the brain; for example, the superior temporal gyrus (STG) is the cortical region responsible for the sensation of sound, and the postcentral gyrus is the location of the primary somatosensory cortex, the main sensory receptive area for the sense of touch.

Electrocorticography (ECoG) is a type of electrophysiological monitoring devices that uses electrodes placed directly on the exposed surface of the brain to record electrical activity from the cerebral cortex.

In the experiment, ECoG was located on one side of the cerebral hemisphere, mainly on STG area. It recorded while the subject was listing to five-minutes of English words presented auditorily.

1.2 Motivation and Challenges

Decoding recognizable speech stimuli requires not only recovering complex acoustic features for the speech waveform but also uncovering the non-linear relationship between the ECoG signal and the acoustic features.
As mentioned in [2], since the orientation and offset of the ECoG implanted grid varied across patients, obtained signals were different for different subjects, making combining data from multiple subjects very challenging. To use data from all subjects, we apply the mapping-generator structure. The mapping layers transfer ECoG signals to a latent space, in this space, data from all subjects have a common representation. From which, the generator reconstructs speech stimuli.

The other challenge is the dataset is much smaller than typical deep learning applications. To prevent the model from overfitting while keeping as many fitting parameters as possible to enhance the model's expressive power, we utilized big corpus dataset and trained with two GAN based approach. Since the generator mainly focuses on speech reconstruction, one approach pre trains the generator to contain prior knowledge of audio representation for speech decoding task, then fine-turn it together with mapping layers. The other trains like CycleGAN to combine supervised training with unsupervised one. In this approach, our goal is to learn the mapping between an input ECoG signal and an output stimuli speech, and couple it with an inverse mapping: from stimuli speech to ECoG.

We compared the decoding results of the pre-trained model and cycleGAN based model with a previously developed adapted WaveNet model [2]. Both quantitative and qualitative comparisons show that the pre-trained model generates reconstructions closer to the original stimuli and is a novel promising approach to generate intelligible speech from brain signals.
2. Related Works

2.1 WaveNet

WaveNet [4] is a state-of-art model for waveform generation. The original WaveNet takes noise samples as input and sequentially generates random-meaning speech audio. The model is able to produce real sounding speech while training on natural speech dataset. WaveNet also is a novel promising approach to generate intelligible speech from brain signals, as showing in [2]. The modified network is a regression model that takes time series ECoG signal input and outputs the spectrogram of speech.

Dilated convolution in the gated units contributes the most to capturing spectro-temporal patterns. This convolution has a perspective field that is larger than its filter length by skipping input values with steps of a certain dilation rate. By stacking residual blocks with dilated convolution layers of exponentially increasing dilation rate, the WaveNet is able to cover a large range of temporal samples with the number of parameters approximately equal to the logarithm of the temporal range. This significantly improves the parameter efficiency in terms of temporal coverage. Meanwhile, exponentially increasing dilation rate extracts a multi-scale representation of the ECoG signal. The summation of skip-connections from the residual blocks allows each block to only process residual
information on a certain scale. Skip-connections, along with residual connections, further improves parameter efficiency by decomposing the model to process multi-scale information in an incremental manner.

2.2 GANs

Generative adversarial networks (GANs) [3], are generative models that learn to map samples \( z \) from a prior distribution \( \mathcal{Z} \) to samples \( x \) from another distribution \( \mathcal{X} \). The component within the GAN structure that performs the mapping is called the generator (G), of which the main task is to learn an effective mapping that can imitate the real data distribution to generate novel samples related to those of the training set. G learns by adversarial training, where the other component, discriminator (D) takes part in. D is typically a binary classifier, and its inputs are either real samples, coming from the dataset that G is imitating, or fake samples, made up by G. The adversarial characteristic comes from the fact that D has to classify the samples coming from \( \mathcal{X} \) as real, whereas the samples coming from G, have to be classified as fake. This leads to G trying to fool D, and the way to do so is that G adapts its parameters such that D classifies G’s output as real. During back-propagation, D gets better at finding realistic features in its input and, while G corrects its parameters to move towards the real data manifold described by the training data. This adversarial learning process is formulated as a minimax game between G and D, with the objective function

\[
\min_G \max_D V(D, G) = \mathbb{E}_{x \sim P_{data}(x)} \left[ \log D(x) \right] + \mathbb{E}_{z \sim P_{data}(z)} \left[ \log (1 - D(G(z))) \right] \tag{1}
\]

2.3 WGAN

According to the loss of the discriminator defined by GAN, we can get the form of the optimal discriminator

\[
\mathbb{E}_{x \sim P_g} \left[ \log (1 - D(x)) \right] \tag{2}
\]
under the optimal discriminator, we can transform the generator loss equivalent of the original GAN definition to minimize Jensen-Shannon (JS) divergence between the real distribution $P_r$ and generated distribution $P_\theta$.

$$JS(P_r, P_g) = KL(P_r || P_m) + KL(P_g || P_m)$$ (3)

$$KL(P_r || P_g) = \int \log \left( \frac{P_r(x)}{P_g(x)} \right) P_r(x) d\mu(x)$$ (4)

Where both $P_r$ and $P_g$ are assumed to be absolutely continuous, with respect to the same measure $\mu$ defined on $\mathcal{X}^2$. $P_m$ is the mixture $(P_r + P_g)/2$. The more we train the discriminator, the closer it is to optimality, and the closer the loss of the generator is to the minimized JS divergence between $P_r$ and $P_\theta$. Since $P_r$ and $P_\theta$ are almost impossible to have non-negligible overlap, no matter how far apart they are, the JS divergence is constant log2, which eventually causes the generator's gradient to be near zero and disappear. So JS distance is very easy to cause gradient unstable and model collapse.


$$W(P_r, P_g) = \inf_{\gamma \in \Pi(P_r, P_g)} \mathbb{E}_{(x,y) \sim \gamma} [||x - y||]$$ (5)

where $\Pi(P_r, P_g)$ denotes the set of all joint distributions $\gamma(x, y)$ whose marginals are respectively $P_r$ and $P_g$. Intuitively, $\gamma(x, y)$ indicates how much “mass” must be transported from $x$ to $y$ in order to transform the distributions $P_r$ into the distribution $P_g$. The Wasserstein-1 distance then is the “cost” of the optimal transport plan.

In Eq. (7), $-\mathbb{E}_{x \sim P_r} [D(x)] + \mathbb{E}_{x \sim P_g} [D(x)]$ forms the Wasserstein-1 distance. $D(x)$ in this equation is no longer the probability that the input is true, but a representation of distance. To enhance training stability, we add gradient penalty [6] on Wasserstein-1 distance to limit its gradient to around one, the discriminator loss function changes to
\[ L(D) = -\mathbb{E}_{x \sim p_r} [D(x)] + \mathbb{E}_{x \sim p_g} [D(x)] + \lambda \mathbb{E}_{x \sim p_{\hat{x}}} [||\nabla_x D(x)||_2 - 1]^2 \]  \hspace{1cm} (7)

Here \( \hat{x} \) is a random interpolation sample on the line connecting \( x_r \) and \( x_g \).

\[ \hat{x} = \epsilon x_r + (1 - \epsilon) x_g, \epsilon \sim \text{Uniform}[0,1] \]  \hspace{1cm} (8)

### 2.4 CycleGAN

CycleGAN [7] learns a mapping function between two domains \( \mathcal{X} \) and \( \mathcal{Y} \). This model includes two mappings \( G: \mathcal{X} \rightarrow \mathcal{Y} \) and \( F: \mathcal{Y} \rightarrow \mathcal{X} \). In addition, it has two adversarial discriminators \( D_X \) and \( D_Y \), where \( D_X \) aims to distinguish between signals \( \{x\} \) and translated signals \( \{F(y)\} \); in the same way, \( D_Y \) aims to discriminate between \( \{y\} \) and \( \{G(x)\} \). CycleGAN’s objective contains two types of terms: adversarial losses for matching the distribution of generated signals to the data distribution in the target domain:

\[ \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{\text{data}}(y)} [\log D_Y(y)] + \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log (1 - D_Y(G(x)))] \]  \hspace{1cm} (9)

and cycle consistency losses to prevent the learned mapping \( G \) and \( F \) from contradicting each other:

\[ \mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{x \sim p_{\text{data}}(x)} [||F(G(x)) - x||_1] + \mathbb{E}_{y \sim p_{\text{data}}(y)} [||G(F(y)) - y||_1] \]  \hspace{1cm} (10)

The full objective is:

\[ \mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) + \mathcal{L}_{\text{GAN}}(F, D_X, Y, X) + \lambda \mathcal{L}_{\text{cyc}}(G, F) \]  \hspace{1cm} (11)

Where \( \lambda \) controls the relative importance of the two objectives. The network aims to solve:
\[ G^*, F^* = \arg \min_{G,F} \max_{D_X,D_Y} \mathcal{L}(G,F,D_X,D_Y) \] (12)

Fig. 4. (a) CycleGAN structure, (b) forward cycle-consistency loss, and (c) backward cycle-consistency loss
3. Method

3.1 Mapping-generator

Based on the above consideration, we break the whole network into two parts—the mapping layers which transfer ECoG signals to hidden features, and a generator which takes in them and produces corresponding stimuli result. This structure combines data from different subjects by assigning each subject a separate mapping layer, but the generator shares weights.

The structure of the whole approach is illustrated in Fig. 5. The mapping part is constructed with a sequential convolutional layer and 4 residual blocks. It compressed the 64 by 160 (channels by time axis) ECoG array to a 1 by 64 vector. The Generator is modified WaveNet as in [2], constructed with several residual blocks after initial layers. The initial part contains a fully connected layer, which enlarges the 1 by 64 vector to 1024 and reshapes it to a 64 by 16 matrix, and a temporal convolutional layer. In each residual block, a gated unit is used as nonlinear activation to model the possible modulation effect of certain input temporal pattern on the output of the speech spectrogram. The signal then passes two branches of 1×1 convolution. The first branch is added to the input of the block as the residual. The second branch of all blocks is skipped and summed at the final post-processing layers that convert the extracted ECoG representation to the final spectrogram output.

Convolution layers with a filter width of 2 are used for dilated convolutions of each residual block, the network covers 1240ms temporal perspective field. As a result of the grid search, Table. I and II show the optimized parameters of the network structure.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>DILATION RATES FOR EACH RESIDUAL BLOCK IN THE GENERATOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residual block No.</td>
<td>1</td>
</tr>
<tr>
<td>Dilation rate</td>
<td>1</td>
</tr>
</tbody>
</table>
Fig. 5. Mapping-generator Network Structure
TABLE II
HYPERPARAMETERS FOR MAPPING-GENERATOR STRUCTURE

<table>
<thead>
<tr>
<th>Conv layer</th>
<th>Filter length</th>
<th>Number of features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mapping-initial conv</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Mapping-residual conv</td>
<td>8</td>
<td>32</td>
</tr>
<tr>
<td>Generator-initial conv</td>
<td>32</td>
<td>16</td>
</tr>
<tr>
<td>Generator-initial res conv</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Generator-dilated conv</td>
<td>2</td>
<td>32</td>
</tr>
</tbody>
</table>

3.2 Pretrained Network

Since a small dataset on which trained models are easily overfitted, a pre-trained generator with prior knowledge is considered. In this strategy, we train the generator separately with large corpus dataset to learn prior knowledge of audio expression.

At the pre-training stage, the generator takes in random inputs, each of them is a 1 by 64 vector from gaussian noise field, then sequentially generates random-meaning audio. To reduce the instability and module collapse problem in pre-training GAN process, the Wasserstein distance is used in loss function and a gradient penalty is added to enforce a Lipschitz constraint. The model is able to produce real sounding speech after adversarial training on natural speech dataset.

Then we fine-tune it with the mapping layer transferring recorded ECoG responses to the input of the generator by simple L2 loss.
3.3 CycleGAN based Network

To enlarge dataset with unpaired data, the other way is to combine supervised learning with unsupervised. We use the training structure called CycleGAN, in which the mapping-generator network is used for both G and F.

The CycleGAN based network is trained from scratch with both paired ECoG-Spectrogram dataset and big corpus dataset. For big corpus dataset, it optimizes with unsupervised loss in Eq. (10). For paired data, the network optimizes with supervised loss. It adds mean square error for both G and F in loss function:

\[ \mathcal{L}_{\text{sup}}(G, F, \mathcal{D}_X, \mathcal{D}_Y) = \mathcal{L}(G, F, \mathcal{D}_X, \mathcal{D}_Y) + \lambda_1 \text{MSE}(G, X, Y) + \lambda_2 \text{MSE}(F, Y, X) \]  

(13)

\[ \text{MSE}(G, X, Y) = \text{sum}(G(X) - Y)^2 \]  

(14)

Where \( \lambda_1 \) and \( \lambda_2 \) control the relative importance of objectives.
Fig. 7. CycleGAN approach. (a) structure, (b) forward cycle-consistency loss, and (c) backward cycle-consistency loss
4. Experiments

4.1 Data Acquisition

The brain activities were obtained from patients with epilepsy and undergoing neurosurgery with an ECoG recording device [3], which is an 8 by 8 electrodes array. There are two types of ECoG sensors, one has 4mm spacing between electrodes, from which the dataset mentioned in [2] as S1 and S2 is collected, called as ‘high-density dataset’. The other sensor has 10mm spacing between electrodes, from which 6 more subjects are collected, as ‘low-density dataset’. ECoG signals were implanted to cover the STG area and recorded simultaneously when subjects were participating in short tasks within five minutes. During the task, the subject was instructed to listen to speech audio (24 kHz sampling rate) of 50 different English words recorded by a native English female speaker. Data from S1 and S2 is collected as in [2]. For low-density dataset, the same 50 words were repeated two times in different pseudo-random order, each subject was required to repeat the words they heard. For current work, only the response during ‘listening’ period is used to reconstruct the stimuli speech.

We adopt the same preprocessing method as [2]. After synchronizing the speech waveform with ECoG signal by lagging speech with 168 ms behind, the speech spectrogram was generated by applying a 128 band-pass filter bank on the waveform. Center frequencies of filters are logarithmically spaced from 180-7000 Hz and have a bandwidth of 1/12 octave. The spectrogram is then subsampled to 32 bands in frequency and 100 Hz in time. ECoG signals were preprocessed with high gamma band-pass filter (70-150 Hz). The envelope of the filtered signal was then extracted by a Hilbert-Huang transform and downsampled to 100 Hz to match the sampling rate of the spectrogram.

The extra speech dataset we used is the free audio collection of English words (eng-wimsmary) dataset [8] which contains 4876 single English word speech. Each speech clips is down-sampled to the same frequency as the high and low-density dataset and transformed to spectrogram with the same strategy as showing above.
4.2 Spectrogram reconstruction

For the paired dataset, with the same strategy in [1], it’s separated into training and testing set for k-fold (k=3 for S1, k=4 for S2 and k=2 for all subjects from low-density dataset) cross-validation. Each partition contains 50 individual words in the testing set. The quantitative evaluation is calculate as the averaged testing result for each cross-validation. The datasets were segmented into short sequences of 1000 ms with overlapping 10 ms. All of the short sequences contain a single word. 17k training sequences are included in each cross-validation partition. Despite a large number of segments, the actual words contained in the training set are limited (50 words, each repeated twice). The two approaches are trained on both low and high density dataset separately. Mean squared error (MSE) is used to evaluate each method.

While merging data from different subject, testing set would contain 1 partition of a subject, and training set contains all partitions of other subjects and the rest partitions in the subject. The testing set varies between subjects, and the final result is the average MSE of all testing results.

4.2.1 Pretrained Network

During the pre-training phase, all unpaired speech clips are used in the WGAN with WaveNet as the generator. After running 350k epochs, the network can generate speech-like results from random inputs. The pre-trained WaveNet is then fine-tuned with mapping layer with paired data. During fine-tuning phase, we use the Adam Optimizer. The learning rate is fixed to 0.001 for mapping layers, and 0.0001 for WaveNet.
4.2.2 CycleGAN

In each batch, both paired and unpaired data are fed in, with an approximate proportion of \(\frac{\text{number}_{\text{paired data}}}{\text{number}_{\text{unpaired data}}} = 1:20\). For paired data, each subject has a separate mapping layer. During training, the unpaired data is randomly distributed to each mapping layer in equal proportions.

Since the amount of unpaired speech data is much larger than the paired signal, we set the weight of MSE, \(\lambda_1\) and \(\lambda_2\) in the loss function in Eq.(13), to 40 to make supervised learning comparable to unsupervised.

4.3 Result

Table.III reports testing result. The row for "S1" and "S2" refer to results when we train and test the adapted WaveNet and the pre-trained WaveNet with mapping layers on two individual datasets, S1 and S2. The two approaches are trained on two repetitions of the words from one subject and then test on the remaining repetition of the same subject.

For the "high density data set" and "low density dataset " results, we select one repetition of one subject as the test subset and use all repetitions of other subjects and remaining repetitions of the test subject for training. The results are the average results when different subjects/repetitions are chosen as the test data.

Overall, the pre-trained model achieved a lower MSE for all the training/testing scenarios. It suggests that the fine-tuned generator has more prior knowledge to express audio information which leads to a better generalization capability with limited training data in original ECoG-Stimuli dataset. Furthermore, the pretrained approach achieved more accurate reconstruction for high density data when using both subjects’ data for training (result indicated for "high-density dataset") than using the single subject data alone (result indicated for S1 and S2). The CycleGAN approach performed much worse. More explorations on how to train the network using the CycleGAN approach is needed.
Figure 8 and Figure 9 illustrate reconstructed samples of each model. As we can see, pre-trained approach generated spectrograms more similar to the ground truth.

TABLE III
QUANTITATIVE EVALUATION

<table>
<thead>
<tr>
<th>MSE</th>
<th>WaveNet</th>
<th>WaveNet (with mapping layers)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Pre-trained WaveNet</td>
</tr>
<tr>
<td>S1</td>
<td>0.57</td>
<td>0.55</td>
</tr>
<tr>
<td>S2</td>
<td>0.53</td>
<td>0.43</td>
</tr>
<tr>
<td>High-density dataset</td>
<td>0.6</td>
<td>0.44</td>
</tr>
<tr>
<td>Low-density dataset</td>
<td>0.4</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Fig. 8. Results on the low-density dataset, on each subfigure the upper row is predicted spectrogram and bottom one is ground truth. (a) WaveNet trained from scratch. (b) WaveNet with pre-trained GAN and finetune
Fig. 9. Results on the high-density dataset, on each subfigure the upper row is predicted spectrogram and bottom one is ground truth. (a) WaveNet trained from scratch, (b) WaveNet with pre-trained GAN and finetune, (c) WaveNet with CycleGAN.
5. Conclusion

In this work, we adopt the mapping-generator network to merge ECoG signals from different subjects for speech decoding task. Despite a relatively small dataset, the pre-trained network with large extra speech clips for prior knowledge overcomes the overfitting problem and generate reconstructed spectrograms with intelligible quality. The CycleGAN based network combines supervised learning with unsupervised, also has the potential to utilize big corpus dataset. However, our training and testing results so far with this approach is not satisfactory.
6. References


