2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)
April 3-7, 2020, Iowa City, Iowa, USA

STIMULUS SPEECH DECODING FROM HUMAN CORTEX WITH GENERATIVE ADVERSARIAL NETWORK TRANSFER LEARNING

Ran Wang* , *Xupeng Chen**, *Amirhossein Khalilian-Gourtani**, *Zhaoxi Chen**, *Leyao Yu*†, *Adeen Flinker*†, *Yao Wang**

* Electrical and Computer Engineering Department, New York University, Brooklyn, NY, USA
†Langone School of Medicine, New York University, New York, NY, USA

ABSTRACT

Decoding auditory stimulus from neural activity can enable neuroprosthetics and direct communication with the brain. Some recent studies have shown successful speech decoding from intracranial recording using deep learning models. However, scarcity of training data leads to low quality speech reconstruction which prevents a complete brain-computer-interface (BCI) application. In this work, we propose a transfer learning approach with a pre-trained GAN to disentangle representation and generation layers for decoding. We first pre-train a generator to produce spectrograms from a representation space using a large corpus of natural speech data. With a small amount of paired data containing the stimulus speech and corresponding ECoG signals, we then transfer it to a bigger network with an encoder attached before, which maps the neural signal to the representation space. To further improve the network generalization ability, we introduce a Gaussian prior distribution regularizer on the latent representation during the transfer phase. With at most 150 training samples for each tested subject, we achieve a state-of-the-art decoding performance. By visualizing the attention mask embedded in the encoder, we observe brain dynamics that are consistent with findings from previous studies investigating dynamics in the superior temporal gyrus (STG), pre-central gyrus (motor) and inferior frontal gyrus (IFG). Our findings demonstrate a high reconstruction accuracy using deep learning networks together with the potential to elucidate interactions across different brain regions during a cognitive task.

Index Terms— speech decoding, generative adversarial networks (GAN), transfer learning, electrocorticographic (ECoG), superior temporal gyrus (STG)

1. INTRODUCTION

Our understanding of speech processing in human cortex has come a long way in the past century. Specifically, the superior temporal gyrus (STG) cortex has been shown to play an important role in speech recognition on a phonetic level [1–3]. One approach to study the activity of different cortical regions during hearing is to reconstruct the speech stimuli from intracranial Electrocorticographic (ECoG) recordings. In addition to finer scope study of the STG, better understanding of the speech processing can help development of better brain computer interface (BCI) systems to help patients with neurological conditions that lead to loss of communication.

Towards this goal, linear models have been utilized to quantitatively demonstrate STG cortical representations [4]. Although the intelligibility of the recovered speech is limited, this approach provides a means to study how the STG area reacts to speech stimulus. More recently, a WaveNet like deep network structure was adopted to decode the stimulus speech from the ECoG recordings in the STG area and have obtained significant improvement over the linear models [5].

In addition to stimulus speech decoding, deep learning models have been applied to ECoG signals for other applications. For example, neural activity prediction in an animal model have been attempted by using multi-scale deep neural networks [6, 7]. Long short-term memory (LSTM) networks are also shown to be effective in predicting human arm movement from brain activity [8]. In addition, recurrent neural networks (RNN) are used in decoding cortical activity into articulatory movement during speech production [9].

Our goal in this work is to leverage deep learning models to decode intelligible audio stimuli from ECoG recordings of the cortical regions including the STG area. One major challenge that limits the success of deep learning methods is the scarcity of the training data. In this study, we tackle this challenging problem with a network structure containing an encoder followed by a generator (Sec. 2.1 and Fig. 1). The encoder performs feature extraction and maps the ECoG signal to a representation space (Sec. 2.3). The generator predicts realistic spectrograms from the representation space (Sec. 2.2). We choose to encourage an independent and identically distributed (i.i.d.) standard Gaussian distribution for the representation vector to be able to pre-train the generator without prior knowledge of the distribution of the encoder output. Additionally, this maximizes the capacity of the representation space in terms of entropy.

In order to address the shortage of ECoG and speech stimuli pairs, we propose a training scheme which trains the generator using a large corpus of natural speech data. Additionally, we introduce a regularization term for the fine-tuning loss function that not only encourages the output of the encoder to follow the desired distribution, but also helps the network increased generalization. By introducing an attention mechanism in the encoder, we can reveal the activation of different cortical regions during speech perception. Our results (Sec. 3) show state-of-the-art decoding of English word stimuli from cortical area including STG. Additionally, the visualization of the attention mechanism (Sec. 3.3) reveals the dynamics of brain regions that is consistent with prior neuroscientific findings. Although the focus of this study is on stimulus speech decoding from recorded ECoG signals, the proposed methods are general and can be of interest in other applications with limited training data.

2. METHOD

2.1. Transfer-GAN: a generative network transfer learning framework for spectrogram decoding

One of the main challenges in reconstructing the speech stimulus from brain activity with deep neural networks is the limitation of

This material is based upon work supported by the National Science Foundation IIS-1912286 and the National Institute of Health NINDS 1R01NS109367 grants.

978-1-5386-9330-8/20/$31.00 ©2020 IEEE

pairs of ECoG and speech data. In our preliminary experiments, we discovered that it is nearly impossible to directly learn the relation between the ECoG signal and the very complex speech structure from limited number of data pairs. On the other hand, there are plenty of natural speech data that allows one to discriminate "fake" from "real" speech. One naive solution is to use a GAN loss while training a network to map the ECoG signal to the Speech signal. However, this is still limited by the number of paired ECOG and speech data. To circumvent this challenge, we propose the transfer-GAN framework, an innovative transfer learning approach for generative network.

The transfer-GAN framework (Fig. 1) contains an encoder that maps an ECoG signal to a representation space with a prescribed distribution, followed by a generator that generates a spectrogram from the representation vector (output of the encoder). Finally, the spectrogram is converted to the sound waveform using another network (vocoder). Both the generator and the vocoder can be pre-trained using any large corpus of speech data. To encourage realist spectrograms generation, a GAN loss is applied during generator pre-training. Then, the encoder and the generator can be refined together using the paired data. This approach not only allows us to efficiently exploit the prior information about real speech spectrograms and waveforms, but also prevents the mode collapse problem often associated with small training data [10].

In the following, we will introduce the structure of the generator and the encoder, the transfer learning approach to fine-tune the encoder and generator together, as well as the vocoder used to reconstruct waveforms.

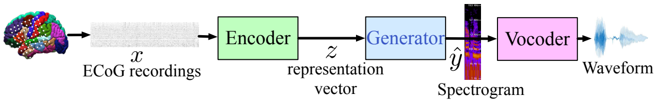

Fig. 1. Overview of the transfer-GAN framework.

2.2. Generator Network Structure

The limitation of the ECoG and speech training pairs is tackled by using a generator network which is pre-trained on a larger corpus dataset. The generator takes a representation vector z and generates a spectrogram of a spoken word. The structure of the generator network is shown in Fig. 2. Here, we introduce a structure inspired by WaveNet [11] which has been shown to successfully generate waveforms. The efficiency of WaveNet is that it encodes the input with multiple temporal scales. Convolutions with different dilation rates allow filters to span small to large temporal duration without increasing the number of parameters. We show that a similar structure is also suitable for generating speech spectrograms.

First, the generator projects the input vector into temporal domain with a fully connected layer and reshaping. Then, several WaveNet residual blocks follow the initial convolution layer. In each block, signal is processed with a gated unit with certain temporal filtering scale controlled by a dilation rate. The output of each gated unit flows into two paths of 1×1 convolution with temporal filter width of 1. One path is further fed into the next residual block with another temporal scale and deeper feature extraction, while the other path (skip convolution) contributes to the final spectrogram generation by adding the features to the sub-network.

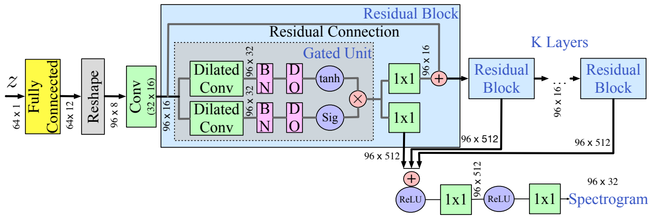

Fig. 2. Overview of the generator network. Total $K = 5$ residual blocks are used. The BN, DO, and 1×1 in the figure denote batch normalization, dropout, and temporal convolution with filter width 1, respectively.

Generator Pre-training. For pre-training the generator network, we use random vectors $\hat{z}$ with an i.i.d. standard Gaussian distribution as the input and try to predict real spectrograms. In addition to common practice in training GAN networks, the sampling from i.i.d. Gaussian distribution is used here since

- the posterior distribution of the encoder output is not known ahead of time and a common output/input distribution should be agreed for the encoder/generator to follow.
- Gaussian distribution has the largest entropy among all sources with the same variance and hence maximizes the representation capacity of the representation space [12].

We use a Wasserstein Generative Adversarial Network (wGAN) [13] scheme to help with generator pre-training. The wGAN has been proved to be a stable variation in GAN family and has shown success for image and audio generation [13–17].

Wasserstein Generative Adversarial Network. wGAN [13] uses the Wasserstein distance (w-distance) to measure the distance between real and generated data distributions. Compared with the Jensen–Shannon (JS) distance used in previous GAN structure, w-distance is continuous and differentiable with respect to the distributions. Although the computation of such distances is not trivial, a critic network is used in wGAN to estimate the w-distance.

During training, generator learns to produce "fake" spectrograms that look like the "real" spectrograms, $\hat{y}$, from input vector $\hat{z}$ by minimizing the estimated w-distance provided by the critic network. The critic network $C(y, \hat{y})$ takes "real" spectrograms randomly sampled from the training data, y, and the "fake" spectrograms $\hat{y}$ as inputs and learns to update the measured w-distance between y and $\hat{y}$. The generator and critic network are trained alternatively. In our experiments, we use ResNet [18] as our critic network and set the input vector dimension to 64.

2.3. Encoder Network Structure

The encoder in our framework serves as a feature extractor that maps the ECoG signals to a representation vector z. We use a ResNet [18] as the backbone of the encoder. Fig. 3 demonstrates the structure of the encoder. In the beginning layers, only temporal filtering along the signal from each electrode is used because the ECoG signal has less correlations between electrodes than across time. Along with the last temporal residual block, an attention gated unit is applied that allows the network to focus only on more significant electrodes at each time step. This not only helps to improve the network accuracy by ignoring the uninformative electrodes with low signal-to-noise ratio (SNR), but also provides a way to analyze the dynamics of each cortical area. We discovered that the evolution of the visualized attention mask over time follows prior neuroscientific findings of the brain dynamics (see Sec. 3.3).

After extracting temporal features for several layers, the later parts of the encoder further extract spatiotemporal features with residual blocks using 3D convolution. At the output layer, instead of embedding to latent vector z directly, we use the “reparameterization trick” [19] to encourage the generated vector z to follow the desired i.i.d. Gaussian distribution as the input to the pre-trained generator. This also regularizes the encoder during fine-tuning, which in turn results in better generalization with small amount of training data.

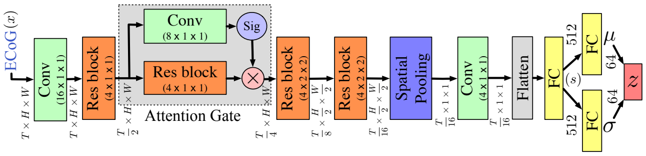

Fig. 3. Overview of the encoder network. Initial convolution layers only performs temporal filtering within each channel (corresponding to signal from one electrode). Attention mechanism helps with the feature extraction and interpretability of the results.

2.4. Fine-tuning the Encoder and the Generator Together

In our transfer-GAN framework, once the generator is pre-trained, we then transfer it to the overall network by fine-tuning it together with the encoder. To encourage the output distribution of the encoder to follow an i.i.d. standard Gaussian distribution, we want to minimize the Kullback–Leibler (KL) divergence between the two distributions. Let x denote a segment of ECoG signal input to the encoder. Following the “reparameterization trick” introduced in the variational auto-encoder (VAE) [19], the output of encoder uses two separate fully connected layers to estimate mean, $\mu(x) \in \mathbb{R}^d$, and variance, $\sigma(x) \in \mathbb{R}^d$, vectors. The representation vector z is then constructed to follow a Gaussian distribution $N(\mu(x), \Sigma(x))$ (where $\Sigma(x) = \text{diag}\left(\sigma^2(x)\right)$) by rescaling an i.i.d. standard Gaussian random vector (i.e. $z = \mu(x) + \sigma(x) \circ n$, where $n \sim N(0, I)$).

KL divergence between encoder output distribution $p(z)$ and generator input distribution $p(\hat{z})$ during pre-training is

$$\begin{aligned}\text{KL}\left(p\left(z\right)\|p\left(\hat{z}\right)\right) &= \text{KL}\left(N\left(\mu\left(x\right), \Sigma\left(x\right)\right)\|N\left(0, I\right)\right) \\ &= \frac{1}{2}\sum_{i=1}^{d}(\mu_i^2(x) + \sigma_i^2(x) - \log \sigma_i^2(x) - 1)\end{aligned} \tag{1}$$

To jointly train the encoder and refine the generator, we minimize a loss function that is a weighted sum of the mean squared error (MSE) between reconstructed $\hat{y}(x)$ and the ground truth spectrogram y and the KL divergence term:

$$\text{loss} = \text{MSE}\left(y, \hat{y}(x)\right) + \lambda\, \text{KL}\left(p\left(z\right)\|p\left(\hat{z}\right)\right) \tag{2}$$

where λ is a hyper-parameter and is set to 0.1 in our experiment.

2.5. Reconstruction of Audible Waveform

To transform the predicted spectrogram back to an acoustic waveform, we adapt a WaveNet vocoder model [11] to reconstruct waveforms of good quality. The WaveNet vocoder takes spectrogram as its input and learns the mapping between the spectrogram and the corresponding speech waveform. We pre-train the vocoder on large corpus datasets. Once the encoder and generator are fine-tuned, we then fine-tune the vocoder separately with pairs of the stimuli waveforms and the corresponding predicted spectrograms in the training set.

3. EXPERIMENTAL RESULTS

3.1. Dataset Acquisition and Preprocessing

We collected datasets of three different electrode density and coverage: A higher density dataset (HD) with 2 subjects and 4 mm inner-electrode spacing; A lower density dataset (LD) with 12 subjects and 10 mm spacing; A hybrid density dataset (HB) with 2 subjects and overall 10 mm spacing with particular regions inserted by 5 mm spacing sub-grid electrodes. The STG region is sampled for all subjects, and other cortical regions (including Broca’s area and motor cortex) are also sampled in LD and HB data. Fig. 4 illustrates grid placement examples for the three datasets. HD data was based on previously published data [20], other HB and LD data was acquired at NYU Langone working with patients undergoing treatment for refractory Epilepsy who gave written consent to participate in the study. The study protocol was approved by the NYU Langone Medical Center Committee on Human Research. During the task, all subjects were instructed to listen to speech audio of 50 different English words/pseudo-words recorded by a native English female speaker. The 50 words are repeated 2-4 times depending on the dataset. One subject in the HD dataset passively listened to each word while all other subjects are required to reproduce each word after listening.

The ground truth speech spectrograms were generated from waveforms by applying a 32 band-pass filter bank. Filters with bandwidth of of $1/12$ octave and center frequencies spaced logarithmically from 180-7000 Hz were used. The spectrograms are then down-sampled 125 Hz in time [21]. 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 125 Hz to match the sampling rate of the spectrogram. Silent period of 250ms before each stimuli is used as reference . We normalize the signal from each electrode by the mean and standard deviation of the reference period.

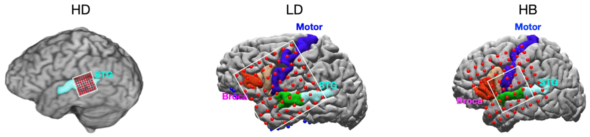

Fig. 4. Examples of electrode grid of HD, LD and HB datasets. Electrodes within the cropped regions are included in the input to the network. For HD dataset, this covers mostly the STG area. For LD and HB datasets, the motor and Broca’s areas are also covered. In all cases, the input covers a 8×8 grid. Some subjects have missing or bad electrodes in the area chosen. For simplicity, we assumed the signals are all zero in those locations.

3.2. Stimuli Reconstruction

The generator has 5 residual blocks with dilation rate of $\{1, 2, 4, 8, 16\}$ and filter size 2 to cover 62 ms temporal perceptive field. The encoder network covers 51 ms temporal field. Adam optimizer [22] is used to fit models with hyper-parameters as following: generator pre-training ($lr = 10^{-4}$, $\beta_1 = 0$, $\beta_2 = 0.9$), encoder-generator fine-tuning ($lr = 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$), vocoder pre-training ($lr = 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$), and vocoder fine-tuning ($lr = 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$). We use a corpus of spoken English words from the Shtooka project [23] to pre-train the generator. It contains 4876 individual words pronounced by a female speaker. For pre-training the vocoder, a combination of dataset is used. It includes Shtooka project [23] and LJ speech dataset [24], which contains 24 hours speech waveforms of English sentences in female voice.

During fine-tuning, we separate the dataset into testing and training sets. For each subject, 50 unique words are used for testing, and the rest of the samples (each word appeared 1 to 3 times depending on the subject) are used for training. An individual model is trained

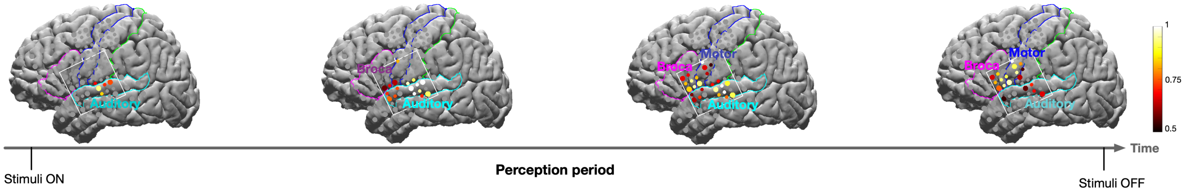

Fig. 5. The averaged evolution of the attention mask for a subject with the hybrid grid. The color in each electrode indicates the value of the attention mask, following the color bar. The white square shows the 8×8 grid used in the experiment. Similar dynamic is also observed in the other HB subject.

and tested for each subject. For reconstructing the stimuli, only the perception period ECoG signal is used, which begins at the stimulus onset and runs up to 419 ms -791 ms. During fine-tuning, for data augmentation purpose, we randomly pick a sliding window of 768ms from a combination of the perception period and the corresponding silent period for each sampled word. As shown in Fig. 4, signals from a 8×8 grid are used as the input.

We compare our transfer-GAN framework with a spectrogram based WaveNet (SpecWaveNet) [5], which has shown good reconstruction and outperformed linear models and residual network based approaches. To allow for a fair comparison, the number of parameters for the encoder and generator is the same as that of SpecWaveNet. We also compared with the linear regression model decoding accuracy reported by Pasley et al [4]. Table 1 compares the averaged mean squared error (MSE) and correlation coefficient (CC) on our dataset. The transfer-GAN has better performance by a large margin on all three datasets in terms of both MSE and CC metrics. Fig. 6 compares samples of reconstructed spectrograms on the testing set of HD and HB datasets for transfer-GAN and SpecWaveNet methods. The results show our proposed approach captures spectrogram dynamics more accurately. The consonants (short segment of high frequency components in the figures) are important for intelligible reconstruction but they can be easily overlooked by decoding models due to their low energy. Our method provides better consonant reconstruction compared to the SpecWaveNet. Some audio examples of the decoded waveforms are provided *here*[1].

	MSE (±sd) / CC (±sd)		
	transfer-GAN	SpecWaveNet	linear model
HD	**0.58**(0.09) / **0.69**(0.05)	0.68(0.08) / 0.61(0.05)	- / 0.41(0.03)
HB	**0.53**(0.03) / **0.72**(0.01)	0.64(0.01) / 0.66(0.02)	- / -
LD	**0.73**(0.14) / **0.60**(0.04)	0.79(0.15) / 0.54(0.05)	- / 0.3(0.05)

Table 1. Quantitative comparison of transfer-GAN (proposed), SpecWaveNet [5], and linear model [4] in MSE (lower is better) and CC (higher is better) on test data. "-" refers to number not reported.

3.3. Attention Mask Visualization

The attention mask generated in the encoder for each input ECoG signal helps the network to focus on electrodes with useful information at each time step. Additionally, it provides a way to visualize the brain signal dynamics of different cortical areas at different times. As an example, we visualize the dynamics of the mask for one subject of HB. The hybrid grid provides both satisfying resolution on STG area and large span over other perisylvian cortical areas related to language. Fig. 5 shows the averaged evolution of the attention mask for all the test samples.

By observing the attention mask as in Fig. 5, we notice that STG, Broca's area and motor cortices are attended sequentially. Accessory auditory cortex in the STG is consistently attended during the perception period. Another observation is that the attention in Broca's area increases shortly after auditory cortex is initially activated. This suggests that Broca's area is active during speech perception and is likely involved in sequencing articulatory information prior to speech articulation [25]. Moreover, motor cortex is active during the stimuli perception and prior to speech production. Similar phenomenon has been observed in the literature [26]. In order to confirm that the observation on motor area activation is related to perception rather than early articulatory preparation, further study is required. The fact that the attention mask generated by the learnt network matches with recent findings in the neuroscientific literature [25, 26] of cortical dynamics is reassuring. It also suggests that such a deep learning architecture with an embedded attention mask can potentially help elucidate the functions of cortical regions during different cognitive tasks.

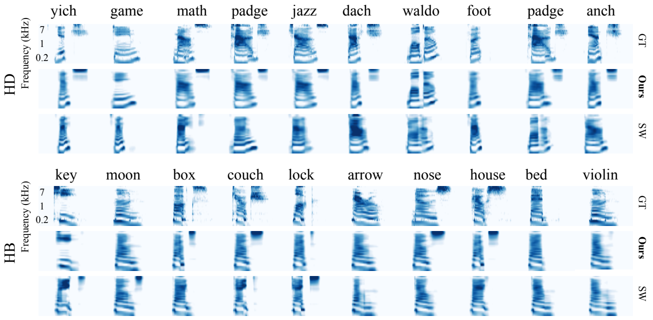

Fig. 6. Decoding results on each one of HD and HB testing set. GT and SW denote ground truth and SpecWaveNet, respectively.

4. CONCLUSION

In this study, we developed a framework for stimulus speech decoding from human cortex ECoG recordings. We proposed a new framework containing an encoder to extract features from and map the ECoG signals to a representation space. The generator in our framework is pre-trained to predict realistic spectrograms from the representation space. This approach allows us to tackle the challenge of limited training data and achieve state-of-the-art reconstruction from cortical areas including STG. Additionally, the attention mechanism introduced in the encoder allows for better generalization of the network and interpretation of the results for neuroscientific discoveries.

In future work we aim to use the developed techniques to study speech processing in human cortex in finer details. For instance, dynamics of speech perception, understanding and production can be studied by developing not only stimulus speech decoder, but also semantic word decoder and response speech decoder. Our developed framework might also be useful for other medical applications where limitations in training data hinder the use of deep learning.

[1] https://wp.nyu.edu/videolab/ecog_demo/

5. REFERENCES

[1] Gregory Hickok and David Poeppel, “The cortical organization of speech processing,” *Nature Reviews Neuroscience*, vol. 8, no. 5, pp. 393, 2007.

[2] Josef P Rauschecker and Sophie K Scott, “Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing,” *Nature Neuroscience*, vol. 12, no. 6, pp. 718, 2009.

[3] Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F Chang, “Phonetic feature encoding in human superior temporal gyrus,” *Science*, vol. 343, no. 6174, pp. 1006–1010, 2014.

[4] Brian N Pasley, Stephen V David, Nima Mesgarani, Adeen Flinker, Shihab A Shamma, Nathan E Crone, Robert T Knight, and Edward F Chang, “Reconstructing speech from human auditory cortex,” *PLoS Biology*, vol. 10, no. 1, pp. e1001251, 2012.

[5] Ran Wang, Yao Wang, and Adeen Flinker, “Reconstructing speech stimuli from human auditory cortex activity using a WaveNet approach,” in *2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. IEEE, 2018, pp. 1–6.

[6] Ran Wang, Yilin Song, Yao Wang, and Jonathan Viventi, “Long-term prediction of μECOG signals with a spatiotemporal pyramid of adversarial convolutional networks,” in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 1313–1317.

[7] Yilin Song, Jonathan Viventi, and Yao Wang, “Multi resolution LSTM for long term prediction in neural activity video,” *arXiv preprint arXiv:1705.02893*, 2017.

[8] Nancy XR Wang, Ali Farhadi, Rajesh PN Rao, and Bingni W Brunton, “AJILE movement prediction: Multimodal deep learning for natural human neural recordings and video,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[9] Gopala K. Anumanchipalli, Josh Chartier, and Edward F. Chang, “Speech synthesis from neural decoding of spoken sentences,” *Nature*, vol. 568, no. 7753, pp. 493–498, 2019.

[10] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen, “Improved techniques for training gans,” in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[11] Akira Tamamori, Tomoki Hayashi, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda, “Speaker-dependent WaveNet vocoder.,” in *Interspeech*, 2017, pp. 1118–1122.

[12] Thomas M Cover and Joy A Thomas, *Elements of information theory*, John Wiley & Sons, 2012.

[13] Martin Arjovsky, Soumith Chintala, and Léon Bottou, “Wasserstein generative adversarial networks,” in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 214–223.

[14] Jonas Adler and Sebastian Lunz, “Banach Wasserstein GAN,” in *Advances in Neural Information Processing Systems*, 2018, pp. 6754–6763.

[15] Jinfu Ren, Yang Liu, and Jiming Liu, “EWGAN: Entropy-based Wasserstein GAN for imbalanced learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 10011–10012.

[16] Chris Donahue, Julian McAuley, and Miller Puckette, “Adversarial audio synthesis,” *arXiv preprint arXiv:1802.04208*, 2018.

[17] Yi Zhao, Shinji Takaki, Hieu-Thi Luong, Junichi Yamagishi, Daisuke Saito, and Nobuaki Minematsu, “Wasserstein GAN and waveform loss-based acoustic model training for multi-speaker text-to-speech synthesis systems using a WaveNet vocoder,” *IEEE Access*, vol. 6, pp. 60478–60488, 2018.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[19] Diederik P Kingma and Max Welling, “Auto-encoding variational Bayes,” *arXiv preprint arXiv:1312.6114*, 2013.

[20] A Flinker, EF Chang, NM Barbaro, MS Berger, and RT Knight, “Sub-centimeter language organization in the human temporal lobe,” *Brain and Language*, vol. 117, no. 3, pp. 103–109, 2011.

[21] Adeen Flinker, Werner K Doyle, Ashesh D Mehta, Orrin Devinsky, and David Poeppel, “Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries,” *Nature Human Behaviour*, vol. 3, no. 4, pp. 393, 2019.

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

[23] SHTOOKA, “The shtooka project,” `http://shtooka.net/`, 2019.

[24] Keith Ito, “The LJ speech dataset,” `https://keithito.com/LJ-Speech-Dataset/`, 2017.

[25] Adeen Flinker, Anna Korzeniewska, Avgusta Y Shestyuk, Piotr J Franaszczuk, Nina F Dronkers, Robert T Knight, and Nathan E Crone, “Redefining the role of Broca’s area in speech,” *Proceedings of the National Academy of Sciences*, vol. 112, no. 9, pp. 2871–2875, 2015.

[26] Connie Cheung, Liberty S Hamilton, Keith Johnson, and Edward F Chang, “The auditory representation of speech sounds in human motor cortex,” *Elife*, vol. 5, pp. e12577, 2016.