

DECODING SPEECH FROM HUMAN CORTEX AND
INTERPRETING SPEECH CORTICAL NETWORKS

DISSERTATION

Submitted in Partial Fulfillment

of the Requirements for the

Degree of

DOCTOR OF PHILOSOPHY (Electrical and Computer Engineering)

at the

NEW YORK UNIVERSITY

TANDON SCHOOL OF ENGINEERING

by

Ran Wang

Dec 2021

Approved:



Department Chair Signature

University ID: N16901684

Net ID: rw1691

Dec 21 2021

Date

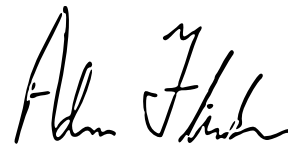
Approved by the Guidance Committee :

Major : Electrical and Computer Engineering



Yao Wang (Advisor)

Professor
Electrical & Computer Engineering,
Tandon School of Engineering



Adeen Flinker (Co-Advisor)

Assistant Professor
Department of Neurology,
Grossman School of Medicine



Anna Choromanska

Assistant Professor
Electrical & Computer Engineering,
Tandon School of Engineering

Microfilm or copies of this dissertation may be obtained from

UMI Dissertation Publishing
ProQuest CSA
789 E. Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Vita

Ran Wang was born in Anhui, China in 1990. He received his B.S. degree in Electronic Engineering from Tsinghua University, China in 2012. He left his home and came to the states for further education in order to fulfill his curiosities. He started his doctoral training in Tandon School of Engineering of New York University in Fall 2015. After brief explorations on general computer vision and machine learning projects for the first year, he focused his research on neural engineering and related neuroscience topics as his dissertation. During his PhD years, He also shortly worked as research intern at Cisco and Alibaba on machine learning research.

To the past, the present, and the future.

Acknowledgements

This thesis and the related work would not be possible without the help of a large number of people who provided both academic and moral support over the years.

First and foremost, I would like to thank my advisor Prof. Yao Wang for her vision and support whenever I needed throughout my doctoral studies. I also would like to thank my co-advisor Prof. Adeen Flinker for both generous career guidance as well as insightful advises on understanding the neuroscience theories which I had no experience before my PhD. Moreover, I owe thanks to Prof. Anna Choromanska for her kindness and valuable academic suggestions.

Furthermore, I wish to thank to my colleague PhD students Xupeng Chen, Amirhossein Khalilian-Gourtani, and Leyao, as well as graduate assistant Zhaoxi Chen and Yicheng Ma, who have assisted in many aspects of the research.

In addition to the people involved in the research for this thesis, I also would like present my special thanks to Prof. Jonathan Viventi, Prof. Yong Liu, Prof. Edward Wong, and Prof. Eero Simoncelli and people that have provided collaboration, assistance and support in various research topics during my years in NYU.

Last but not least, I would like to thank my family and my friends both in United States and China for being there for me over the long years of my PhD. They have made my wonderful already doctoral journey even more pleasant.

Ran Wang, New York University, Tandon School of Engineering
December 21, 2021

ABSTRACT

**DECODING SPEECH FROM HUMAN CORTEX AND
INTERPRETING SPEECH CORTICAL NETWORKS**

by

Ran Wang

Advisor: Yao Wang

Co-advisor: Adeen Flinker

Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy (Electronic and Computer Engineering)

December 2021

Directly decoding speech from human cortical activity enables neuroprosthetic applications. An accurate auditory decoding model also facilitates neuroscientific research for dynamic brain mechanisms through model interpretation. This work aims to decode speech from human brain activities recorded by electrocorticography (ECoG) devices implanted in the human cortex. The main challenge in developing such decoding models lies in the scarcity of training data. We demonstrate that models designed with domain knowledge of speech perception and production can decode high-quality speech with limited training data. The interpretation of corresponding models corroborates established neuroscience theories and provides evidence for novel neuroscientific phenomena. **Our earlier effort** focuses on decoding perceived speech from neural activity in the superior temporal gyrus (STG) cortex. We show that acceptable decoding results are achievable even with a small amount of training data, and the trained decoder reflects the phoneme selectivity of STG. **In the second approach**, we further improve the stimuli decoding performance by introducing a transfer learning method with a pre-trained GAN using a large corpus of speech data to produce spectrograms from an abstract representation space. We then transfer it to a bigger network with an encoder attached before, which maps the neural signal to the representation space. By visualizing the attention mask embedded in the encoder, we observe brain dynamics that are consistent with findings from neuroscience studies investigating dynamics in the superior temporal gyrus (STG), precentral gyrus (motor), and inferior frontal gyrus (IFG). **Our recent approach** focuses on speech production, and we develop a novel deep learning architecture that translates neural signals to an interpretable representational space that can directly synthesize speech. By using either causal or anti-causal temporal convolution, we can disentangle feedforward vs. feedback processing. Unlike prevailing neuroscientific models for speech production, we find a mixed cortical architecture wherein the frontal cortex processes both feedforward and feedback information in tandem across precentral and inferior frontal sites.

Contents

Vita	iv
Acknowledgements	vi
Abstract	vii
1 Introduction	1
1.1 Overview and motivation	1
1.2 History and previous research	1
1.3 Contributions	2
1.4 Organization of the thesis	4
2 SpecWaveNet: Wavenet Model for Stimuli Decoding	6
2.1 SpecWaveNet Structure	6
2.2 Experimental Performance of SpecWaveNet	8
2.2.1 Data Acquisition	8
2.2.2 Spectrogram Reconstruction	9
3 Stimuli Decoding with Generative Transfer Learning	12
3.1 Transfer-GAN: a generative network transfer learning framework for spectrogram decoding	12
3.2 Generator Network Structure	13
3.3 ECoG decoder Network Structure	15
3.4 Fine-tuning the ECoG decoder and the Generator Together	16
3.5 Reconstruction of Audible Waveform	17
3.6 Experimental Performance of Transfer-GAN	17
3.6.1 Dataset Acquisition and Preprocessing	17
3.6.2 Stimuli Reconstruction	18
3.6.3 Production Reconstruction	20
4 Stimuli Decoding from Non-grid Electrodes Inputs	22
4.1 Spatial-temporal Separable Conv-Attention Layer	23
4.2 MNI Positional Encoding	25
4.3 Attention Based Encoder	27
4.4 Experimental Performance of attention based encoder	28

5	Generate Produced Speech with Differentiable Speech Synthesizer	30
5.1	Speech decoding framework	30
5.1.1	Differentiable speech synthesizer	34
5.1.2	ECoG decoder and speech encoder	37
5.2	Loss and training hyper-parameters	38
5.3	Produced speech decoding performance	39
5.4	Perceived speech decoding performance	42
5.5	Discussion	43
6	Feedforward-Feedback Contribution Analysis for Speech Decoding Models	46
6.1	Background	46
6.2	Method	47
6.2.1	Revealing delay-dependent contribution of different cortical regions from the trained ECoG to speech model	47
6.3	Result	50
6.3.1	Feedforward and feedback cortical contributions to speech production	50
6.3.2	Temporal dynamics and receptive fields of speech production	51
6.3.3	Contribution analysis for speech parameters	55
6.4	Comparing feedback cortical contribution between speech perception and production	55
6.5	Discussion	58
7	Attention Mask Reflects Speech Cortical Network Dynamics during Perception-Production Task	62
7.1	Attention Mask Visualization	62
7.2	Attention Mask Reflect Perception-Production Cortical Dynamic	63
8	Impulse Response of Deep Speech Stimuli Decoding Network Reflects Phoneme Selectivity of STG	64
8.1	Impulse Response of a Deep Neural Network	64
8.2	Phonetic Feature of Model Impulse Response	65
9	Summary and Future Work	67
9.1	Summary	67
9.2	Future Work	68
	Publication List	75

List of Figures

2.1	SpecWavenet structure.	7
2.2	Reconstruction samples of (a) linear convolution, (b) Resnet, and (c) WaveNet.	9
3.1	Overview of the transfer-GAN framework.	13
3.2	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.	14
3.3	Samples from real and generated spectrogram on big corpus English word dataset.	15
3.4	Overview of the ECoG decoder 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.	16
3.5	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.	18
3.6	Decoding results on each one of HD and HB testing set. GT and SW denote ground truth and SpecWaveNet, respectively.	20
3.7	Samples of decoded produced speech spectrograms with Transfer-GAN framework.	21
4.1	Structure of Conv-Att block. The input and output of the block are in shape of $T \times N \times F$ (<i>time</i> \times <i>nodes</i> \times <i>feature</i>). The block operates temporal and spatial/node dimensions with convolution and attention separately. The attention coefficient computation is augmented with MNI positional encoding to achieve more expressive power.	25
4.2	Overview structure of the attention based ECoG decoder network.	27
4.3	Reconstructed samples from the decoding network with attention based encoder. Upper row: ground truth spectrogram, lower row: reconstructed spectrogram	28

- 5.1 Structure of the decoding framework. (a) the overview of the overall network architecture. An auto-encoder structure (b) is used to pretrain the speech encoder by training a speech encoder to generate proper speech parameters that can reconstruct input spectrograms through the speech synthesizer. (c) The ECoG decoder is a modified three-dimensional residual network. After an initial temporal convolutional layer and eight residual blocks (constructed by three-dimensional convolution layers), multiple convolutional layers (each has temporal kernel size of 1) generate speech parameters separately. (d) The speech encoder in (a) is constructed by a convolutional network, with three convolutional layers backbone and the same multihead output structure as in (c). (e) Illustrates the processes within the speech synthesizer. The harmonics (in voice pathway) and white noise (unvoice pathway) are generated and filtered (multiplication in spectrogram domain) by voice and unvoice filters, respectively. The filtered results are then weighted averaged according to a mixing parameter and then amplified by the loudness parameter. 33
- 5.2 Comparison of original and decoded speech produced by the model. (a) Spectrograms of decoded (left) and original (right) speech exemplar words. (b) Decoded loudness parameter with the voiced (mostly vowel) or unvoiced (mostly consonant) mixing parameter color coded over the loudness curves. (c) Frequencies of the first two formants (F1, F2) and the pitch (F0). The matching between decoded (dashed) curves and reference (solid) curves in both averaged frequencies during each phoneme as well as the overall temporal dynamic leads to intelligible and naturalistic decoding of voiced sounds. (d) Evaluation of the decoded speech quality in objective metrics. The correlation coefficient of spectrogram (CC, left), short-time objective intelligibility (STOI, middle), and Mel cepstral distortion (MCD, right) are used for the evaluation. Note that lower MCD values represent better performance. Both the reconstructed speech from the speech auto-encoder (yellow boxes) and the speech decoded by the ECoG decoder (green boxes) are reported. Besides, the performance of a model trained on a shuffled dataset (trained by matching the decoded spectrogram from a neural signal in a given duration to randomly selected segments of spectrograms during the entire recording session) is also reported as a control. (e) Comparison of CC metric among noncausal (green), causal (blue), and anticausal (red) models. Compared to the shuffled model (the same shuffled model as in Fig. 5.2d), the close performance among noncausal, causal, and anticausal models demonstrates adequate information for decoding speech in both feedforward and feedback signals during speech production. 41

- 5.3 Comparison of original and decoded speech produced by the model for stimulus decoding. (a) Spectrograms of decoded (left) and original (right) speech exemplar words. (b) Decoded loudness parameter with the voiced (mostly vowel) or unvoiced (mostly consonant) mixing parameter color coded over the loudness curves. (c) Frequencies of the first two formants (F1, F2) and the pitch (F0). The matching between decoded (dashed) curves and reference (solid) curves in both averaged frequencies during each phoneme as well as the overall temporal dynamic leads to intelligible and naturalistic decoding of voiced sounds. (d) Evaluation of the decoded speech quality in objective metrics. The correlation coefficient of spectrogram (CC). The speech decoded by the ECoG decoder (purple boxes) are reported. Besides, the performance of a model trained on a shuffled dataset (magenta boxes, trained by matching the decoded spectrogram from a neural signal in a given duration to randomly selected segments of spectrograms during the entire perception period) is also reported as a control. 43
- 6.1 (a) averaged signal of input ECoG projected on the standardized MNI anatomical map. The colors reflect the percentage change of high gamma compared to the baseline level during the pre-stimulus baseline period. (b) shows the anticausal contribution of different cortical locations (red indicates higher contribution), while (c) illustrates the causal contribution. (d) noise level of the contribution analysis evaluated by the contributions from the shuffled model. Contributions below noise level are not shown in (b) and (c). (e) the contrast obtained by taking the difference of the anticausal and causal contribution maps (red means higher anticausal contribution, while blue means higher causal contribution). The boxplots (f) show the average difference in each cortical region (*: P-value<0.05, **: P-value<0.01,***: P-value<0.001,****: P-value<0.0001). 52
- 6.2 Spatial-temporal receptive fields based on decoding contribution. The contribution to decoding the current speech from cortical neural responses with certain temporal delays. (a) and (b) are the feedforward spatial-temporal receptive fields derived from the causal model by evaluating the contribution of past (negative delays) neural signals during a period before production onset (a) and after onset (b). (c) represents the feedback spatial-temporal receptive fields derived from the anticausal models that evaluate the contribution of future (positive delays) neural signals during feedback after articulation. Contributions below significance ($p < 0.05$) representing the noise level are clipped and not shown in the plots. 54

6.3	The temporal receptive field across anatomical regions. The contribution to decoding the current speech from cortical neural responses with certain temporal delays. (a) and (b) are the feedforward temporal receptive fields derived from the causal model by evaluating the contribution of past (negative delays) neural signals during a period before production onset (a) and after onset (b). (c) represents the feedback temporal receptive fields derived from the anticausal models that evaluate the contribution of future (positive delays) neural signals during feedback after articulation. The temporal propagation of the shuffled model estimates the noise level dynamics (grey curves in plots). Only regions significantly above noise level (Wilcoxon sign rank test on across-time averaged data, $P < 0.05$) are reported.	56
6.4	Contribution maps for speech components of (a) anticausal model and (b) causal model.	57
6.5	Comparison of feedback cortical contribution between perception and production periods. contribution during (a) preception and (b) production feedback process. (c) contribution contrast between (a) and (b).	57
6.6	Comparison of cortical contribution between different speech tasks during perception period. Contribution map for (a) passive listening task, (b) imaginary speaking task, and (c) active speaking task. The contrast contribution map (d) between imaginary speaking and passive listening tasks, (e) between active speaking and passive listening task, and (f) between active speaking and imaginary listening tasks.	58
7.1	The averaged evolution of the attention mask during active listening task. 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. . . .	63
8.1	Impulse response for subject S1 (a) and S2 (b). Each subfigure corresponds to one ECoG electrode site.. . . .	66

List of Tables

2.1	Quantitive evaluation of SpecWavenet for S1 dataset	10
2.2	Quantitive evaluation of SpecWavenet for S2 dataset	10
3.1	Quantitive comparison between transfer-GAN (proposed), SpecWaveNet [79], and linear model [61] in MSE (lower is better) and CC (higher is better) on test data. “-” refers to number not reported.	20
4.1	Quantitive comparison between convolutional and attention based decoder for spectrogram reconstruction.	29
8.1	Discovered phonetic features for ECoG electrodes on subject S1	65
8.2	Discovered phonetic features for ECoG electrodes on subject S2	65

Chapter 1

Introduction

1.1 Overview and motivation

Our understanding of speech processing in the human cortex has come a long way in the past century. Cardinal among human high order functions is the ability to perceive and understand, as well as to plan and execute complex speech sequences which carry semantic and linguistic meaning [11,43]. One approach to studying different cortical regions' activity during speech processing is to reconstruct the perceived or produced speech from intracranial Electrocorticographic (ECoG) recordings. In addition to a finer scope study of the speech cortical networks, a better understanding of speech processing can help the development of better brain-computer-interface (BCI) systems to help patients with neurological conditions that lead to loss of verbal communication.

1.2 History and previous research

Towards decoding speech perception, linear models have been utilized to quantitatively demonstrate STG cortical representations [61]. Although the intelligibility of the recovered speech is limited, this approach provides a means to study how the STG area reacts to speech stimulus. Compared to speech stimulus decoding, speech production decoding for neural speech prosthesis is generally a more challenging problem. The advances in modern deep neural networks have brought new trust towards higher quality neural decoding for produced

speech. Both convolutional neural networks (CNN) and recurrent neural networks (RNN) have been adapted towards this goal. Akbari et al. [2] trained a feedforward neural network to map the neural responses into vocoder parameter space for decoding the speech with a vocoder. Densely connected 3D convolutional network has been adapted to decode speech spectrogram directly [4, 38]. RNN has shown promising performance towards intelligible produced speech decoding from motor cortex by introducing articulatory trajectory as an intermediate representation [5]. In addition, natural language processing (NLP) models have also played an essential role in decoding text from brain activity. Inspired by natural language translation, Makin et al. [55] employed LSTM to translate text directly from the neural signal. NLP has also been used to model the transition probability of decoded words when reconstructing text sentences from brain [58].

Many existing studies have shown promising decoding performance towards speech neural prosthetics. However, the invasive nature of ECoG impedes convenient large-scale neural data collection. Thus, the obtained training data is generally limited compared to the common practice of deep learning applications (such as classification on ImageNet with 14 million images). Also, despite the good quality of the decoded results, few existing works systematically study the speech decoding network as a tool for interpreting speech cortical networks.

We focus our research on novel speech neural decoding frameworks with higher data efficiency and better interpretability to tackle the challenges mentioned above.

1.3 Contributions

Our goal in this work is to leverage deep learning models to decode intelligible audio stimuli from ECoG recordings of the cortical regions, including superior temporal, inferior frontal, precentral, and postcentral gyri. One major challenge that limits the success of deep learning methods is the scarcity of training data. In this study, we tackle this challenging problem from two different perspectives. We propose a network structure containing an ECoG decoder followed by a generator (Sec. 3.1 and Fig. 3.1). The ECoG decoder performs feature extraction and maps the ECoG signal to a representation space (Sec. 3.3). The

generator predicts realistic spectrograms from the representation space (Sec. 3.2).

For the first approach, we show that a proposed deep network inspired by WaveNet, trained with limited available data, can reconstruct speech stimuli from STG intracranial recordings. We further investigate the impulse response of the fitted model for each recording electrode and observe phoneme level temporospectral tuning properties in some recorded areas. This discovery is consistent with previous studies implicating the posterior STG (pSTG) in speech phonetic speech representation. It provides detailed acoustic features that specific electrode sites possibly extract during speech recognition.

For the second approach, the generator is a full capacity convolutional neural network. We choose to encourage an independent and identically distributed (i.i.d.) standard Gaussian distribution for the representation vector to pre-train the generator without prior knowledge of the distribution of the ECoG decoder output. To address the shortage of ECoG and speech stimuli pairs, we propose a training scheme that trains the generator using a large corpus of natural speech data. Additionally, we introduce a regularization term for the fine-tuning loss function that encourages the ECoG decoder’s output to follow the desired distribution and helps the network increase generalization. By introducing an attention mechanism in the ECoG decoder, we can reveal the activation of different cortical regions during speech perception. Our results show state-of-the-art decoding of English word stimuli from the cortical areas, including STG. Additionally, the visualization of the attention mechanism reveals brain regions’ dynamics that are consistent with prior neuroscientific findings.

For the third approach, we propose a novel speech synthesizer as the generator to reconstruct a speech spectrogram from a set of interpretable speech parameters. The speech synthesizer is constructed by several parameter-free signal processing equations so that it dramatically reduces the network capacity; thus entire framework is less prone to over-fit. Also, the speech parameter space we chose is biologically plausible so that the mapping from the brain signal to the intermediate representation is efficient. This further reduces the number of trainable parameters needed and increases the training data efficiency. We demonstrate state-of-the-art performance for decoding produced speech. By learning neural network architectures that apply either casual, anticausal (or both) spatial-

temporal convolutions, we can analyze the overall feedforward and feedback contributions, respectively, and elucidate the temporal receptive fields of recruited cortical regions. Our analyses reveal a surprisingly mixed architecture of causal and anticausal processing across the cortex while achieving speech decoding performance on-par or better than previously reported. We also leverage the proposed framework to decode perceived speech and yield accurate performance.

Although the focus of this study is on speech decoding from recorded ECoG signals, the philosophy of intermediate representation design is general. It can be of interest in other applications with limited training data.

1.4 Organization of the thesis

This dissertation is organized as follows: Part I (Chapter 2 - Chapter 4) will introduce details of model design for leveraging limited training data. Part II (Chapter 6 - Chapter 8) will introduce how we interpret the decoding models that either confirm discoveries of previous neuroscience studies or reveal new cortical phenomena during speech processing.

Finally, Chapter 9 will summarize our contributions and the plan for future works.

PART I

Decoding Speech from Human Cortex

Chapter 2

SpecWaveNet: Wavenet Model for Stimuli Decoding

This chapter explores the performance of a very straightforward speech neural decoding solution: applying generic (convolutional) neural network designs and training the network end-to-end. We show that a deep network inspired by WaveNet, trained with limited available data, can reconstruct speech stimuli from STG intracranial recordings. We further investigate the impulse response of the fitted model for each recording electrode and observe phoneme level temporospectral tuning properties in some recorded areas. This discovery is consistent with previous studies implicating the posterior STG (pSTG) in speech phonetic speech representation. It provides detailed acoustic features that specific electrode sites possibly extract during speech recognition.

2.1 SpecWaveNet Structure

With the constraint of a small dataset on which trained models are easily overfitted, a trade-off between expressive power and generalization performance must be considered. To achieve the same representation power, the model with higher parameter efficiency should have fewer trainable parameters and, thus, less likely to overfit.

Based on the above consideration, we modified the WaveNet model, the state-of-art model for waveform generation, to recover speech spectrograms from recorded ECoG re-

sponses. The original WaveNet takes noise samples as input and sequentially generates random-meaning audio. If trained on the natural speech dataset, the model can produce real sounding speech. It has shown success in many audio generation applications [20, 59, 63, 68, 74]. Here, we modified Wavenet to a regression model that takes time series ECoG signal input and outputs the spectrogram of speech.

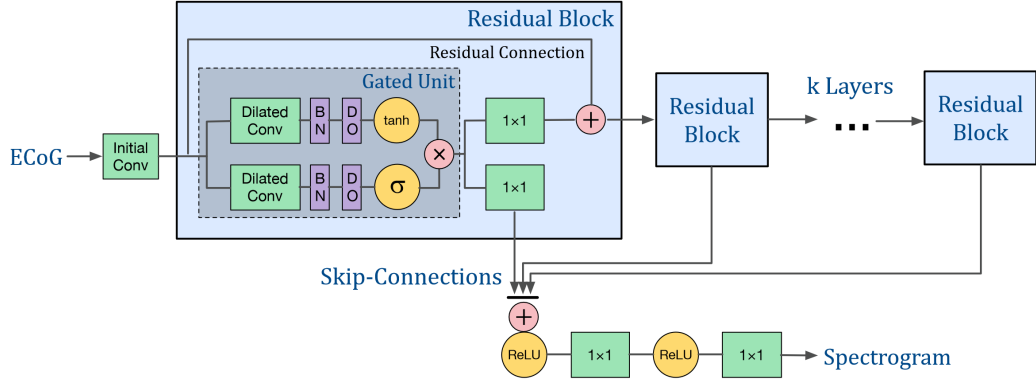


Figure 2.1: SpecWavenet structure.

The structure of WaveNet is illustrated in Figure 2.1. The model is based on a 1d convolution that filters along the time axis. The multiple electrodes of ECoG and frequency bands of the spectrogram are treated as channels. After the initial convolutional layer, the model is constructed with several residual blocks. In each block, a gated unit is used as nonlinear activation to model the possible modulation effect of certain input temporal patterns 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 are skipped and summed at the final post-processing layers that convert the extracted ECoG representation to the final spectrogram output.

Dilated convolution in the gated units contributes the most to capturing spectro-temporal patterns. This convolution has a larger perspective field than its filter length by skipping input values with steps of a certain dilation rate. Convolution with a larger dilation rate has a larger perspective field with the same filter length. By stacking residual blocks with dilated convolution layers of exponentially increasing dilation rate, the WaveNet can 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 of a certain scale. Skip connections and residual connections improve parameter efficiency by incrementally decomposing the model to process multi-scale information. Batch normalization and dropout are included between each convolutional layer to prevent the model from overfitting.

2.2 Experimental Performance of SpecWaveNet

2.2.1 Data Acquisition

The brain activities were obtained from two patients with epilepsy and undergoing neurosurgery with an ECoG recording device [23]. The ECoG arrays have 8 by 8 electrodes with inter-electrode spacing of 4mm and sampling rate of 3051 Hz. Electrode arrays were implanted to cover the posterior lateral surface of the superior temporal gyrus. ECoG signals were recorded simultaneously when subjects were participating in short tasks within five minutes. During the task, both subjects were instructed to listen to speech audio (24 kHz sampling rate) of 50 different English words recorded by a native English female speaker. The first subject (S1) passively listened to each word. The same 50 words were repeated three times in different pseudo-random order. The second subject (S2) participated in two tasks with the same stimuli as S1. The subject was required to repeat the words they heard during the first task and translate the perceived words to Spanish during the second task. Each word was played twice for both tasks thus four ECoG signal examples of hearing the same word were recorded for S2. For the current work, only the response during the “listening” period is used to reconstruct the stimuli speech.

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.

2.2.2 Spectrogram Reconstruction

For spectrogram reconstruction, we used a WaveNet with 10 residual blocks. Convolution with a filter width of 2 is used for dilated convolutions of each residual block. With the exponentially increasing dilation rate, the network covers 1240 ms temporal perspective field.

We also implemented linear convolution (equivalent to linear regression) and ResNet [37] as benchmark models. One-layer linear convolution has previously been used for correlating pSTG cortex responses and speech stimuli and for decoding speech from ECoG signals [61]. ResNet is widely used in various machine learning based audio signal processing tasks [39, 48, 62, 75]. Linear convolution and ResNet are designed to have the same temporal perspective field width (1240ms) as the WaveNet. The linear convolution model uses one-layer of 1d convolution without activation function to transform ECoG signal with 64 channels to speech spectrogram with 32 channels. The ResNet model contains 8 residual blocks with filter length 4 and feature number 32. These settings are also optimized by hyper-parameter search.

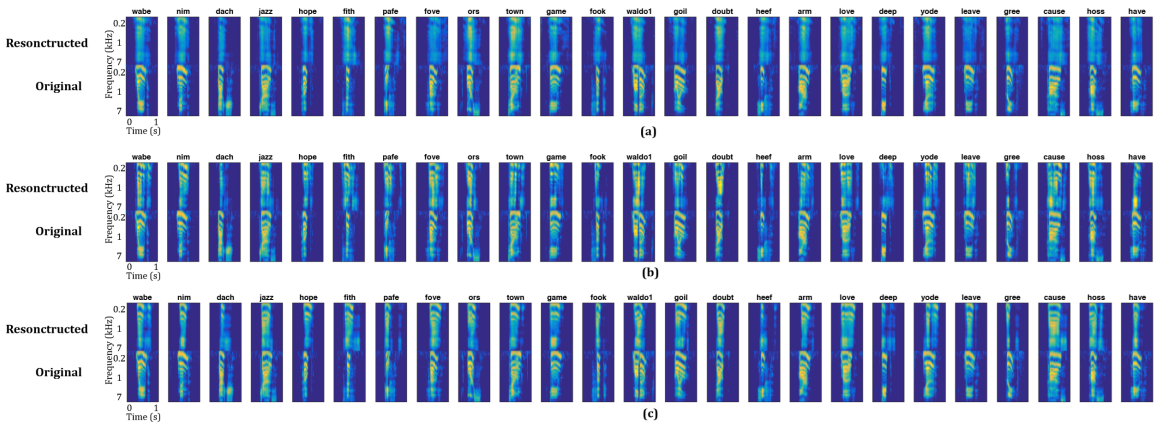


Figure 2.2: Reconstruction samples of (a) linear convolution, (b) Resnet, and (c) WaveNet.

We separate the datasets into training and testing set for k-fold ($k=3$ for S1 and $k=4$ for S2) cross-validation. Each partition contains 50 individual words in the testing set. The quantitative evaluation is based on the averaged testing result for each cross-validation.

	linear conv	ResNet	SpecWaveNet
MSE	0.86	0.79	0.68
CC	0.57	0.61	0.65

Table 2.1: Quantitive evaluation of SpecWavenet for S1 dataset

	linear conv	ResNet	SpecWaveNet
MSE	0.71	0.70	0.69
CC	0.69	0.70	0.71

Table 2.2: Quantitive evaluation of SpecWavenet for S2 dataset

The datasets were segmented into short sequences of 1000 ms with overlapping 10 ms. Note that each word lasts around 400 ms and the spacing between words is about 1000 ms. Some of the short sequences contain only a single word, others have a partial word, and some contain only the silent period. 17k training sequences for S1 dataset and 26k for S2 dataset 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 for S1 and 50 words, each repeated 3 times for S2).

Both mean squared error (MSE) and correlation coefficient (CC) are used as metrics to evaluate each method. CC is calculated based on the averaged correlation coefficient of all 32 spectrogram frequency channels. Table 2.1 and 2.2 report the testing result for subjects S1 and S2. Overall, WaveNet achieved the highest CC and lowest MSE for both datasets and deep models (WaveNet and ResNet) performed better than the linear convolution network. The number of trainable parameters for WaveNet, ResNet, and linear convolution was 509K, 132K, and 204K respectively. The fact that WaveNet obtained better results with more parameters suggests that the proposed WaveNet structure has a richer expressive power while having a better generalization capability with limited training data. Figure 2.2 illustrates reconstructed samples of each model. Linear convolution leads to over-smoothed spectrograms and fails to recover detailed spectral features. ResNet, on the contrary, has over-sharpened results and suffers from artificial spectro-temporal aliasing that causes discontinuity across time.

Even though the spectrograms of recovered and original words can be visually similar, to decode the word waveform intelligibly and correctly requires more precise reconstruction

of the complex acoustic features, such as Formant frequencies and fast spectro-temporal fluctuation. We inverted the decoded spectrogram to a waveform using an iterative convex projection approach [7] for qualitative evaluation. We have found that, although the words in the testing set have already appeared in the training set, the ECoG recordings for the repeated words are quite different. There is no significant higher correlation between ECoG recordings of the same repeated words than of different words. Without sufficient data to uncover the complex non-linear relationship between ECoG signal and speech spectrogram, intelligible reconstruction of all speech stimuli is not expected. Nevertheless, several decoded words from WaveNet (“waldo”, “yich” and “pave”) are definitely intelligible. Even though some of the reconstructed spectrograms by ResNet looked quite similar to those by WaveNet, the reconstructed speech by ResNet is much less intelligible. The intelligible reconstruction further validates better expressive power and generalization ability of WaveNet than benchmark models.

Chapter 3

Stimuli Decoding with Generative Transfer Learning

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

In our preliminary experiments (Chapter 2), we discovered that it is nearly impossible to directly reconstruct intelligible word with complex speech structure from ECoG signal with 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. 3.1) contains an ECoG decoder 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 ECoG decoder). 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 realistic spectrograms generation, a GAN loss is applied during generator pre-training. Then, the ECoG decoder 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 [65].

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

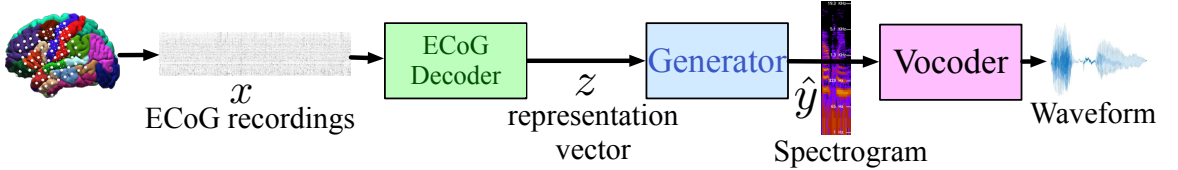


Figure 3.1: Overview of the transfer-GAN framework.

3.2 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. 3.2. Here, we introduce a structure inspired by WaveNet [74] 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 durations 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.

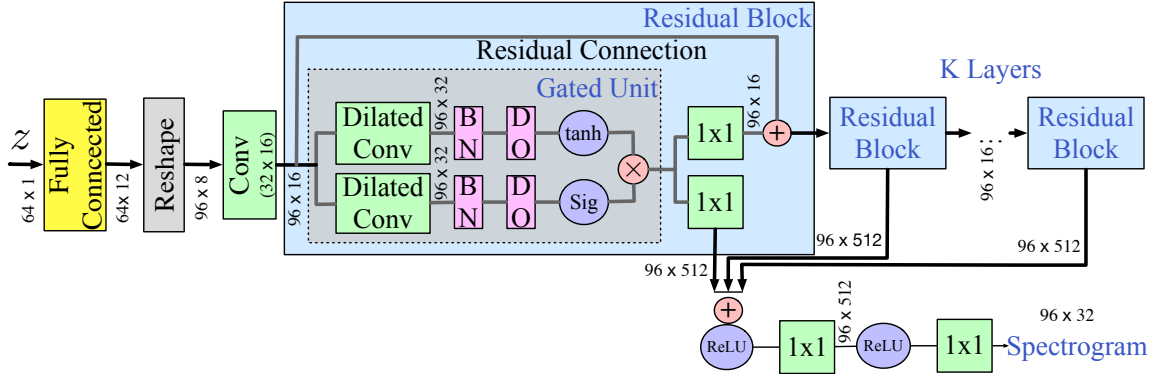


Figure 3.2: Overview of the generator network. Total $K = 5$ residual blocks are used. The BN, DO, and 1×1 in the figure denote batch normalization, dropout, and temporal convolution with filter width 1, respectively.

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 ECoG decoder output is not known ahead of time and a common output/input distribution should be agreed for the ECoG decoder/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 [14].

We use a Wasserstein Generative Adversarial Network (wGAN) [6] 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 [1, 6, 17, 64, 81].

Wasserstein Generative Adversarial Network. wGAN [6] 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.

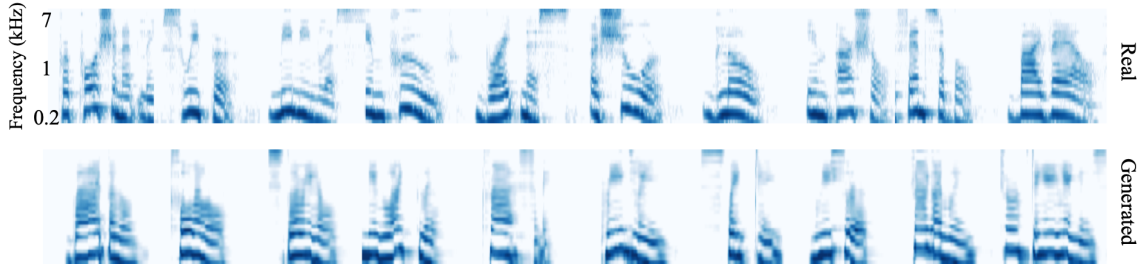


Figure 3.3: Samples from real and generated spectrogram on big corpus English word dataset.

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 [37] as our critic network and set the input vector dimension to 64. Figure 3.3 show some generated spectrogram after generator pretraining. The generation looks very realistic.

3.3 ECoG decoder Network Structure

The ECoG decoder in our framework serves as a feature extractor that maps the ECoG signals to a representation vector z . We use a ResNet [37] as the backbone of the ECoG decoder. Fig. 3.4 demonstrates the structure of the ECoG decoder. 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. 7.2).

After extracting temporal features for several layers, the later parts of the ECoG de-

coder 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” [52] 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 ECoG decoder during fine-tuning, which in turn results in better generalization with small amount of training data.

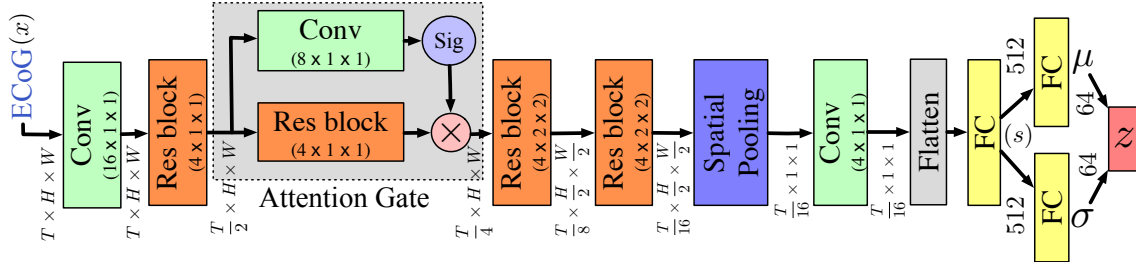


Figure 3.4: Overview of the ECoG decoder network. Initial convolution layers only performs temporal filtering within each channel (corresponding to signal from one electrode). Attention mechanism helps with the feature extraction and interpretability of the results.

3.4 Fine-tuning the ECoG decoder 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 ECoG decoder. To encourage the output distribution of the ECoG decoder 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 ECoG decoder. Following the “reparameterization trick” introduced in the variational auto-ECoG decoder (VAE) [52], the output of ECoG decoder 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 $\Sigma(x) = \text{diag}(\sigma^2(x))$) 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 ECoG decoder output distribution $p(z)$ and generator input

distribution $p(\hat{z})$ during pre-training is

$$\begin{aligned} KL(p(z) | p(\hat{z})) &= KL(N(\mu(x), \Sigma(x)) | N(0, I)) \\ &= \frac{1}{2} \sum_{i=1}^d (\mu_i^2(x) + \sigma_i^2(x) - \log \sigma_i^2(x) - 1) \end{aligned} \quad (3.1)$$

To jointly train the ECoG decoder 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:

$$\text{loss} = (y, \hat{y}(x)) + \lambda KL(p(z) | p(\hat{z})) \quad (3.2)$$

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

3.5 Reconstruction of Audible Waveform

To transform the predicted spectrogram back to an acoustic waveform, we adapt a WaveNet vocoder model [74] 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 ECoG decoder 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.

3.6 Experimental Performance of Transfer-GAN

3.6.1 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. 3.5 illustrates grid placement examples for the three datasets. 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 1/12 octave and center frequencies spaced logarithmically from 180-7000 Hz were used. The spectrograms are then down-sampled 125 Hz in time [25]. 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.

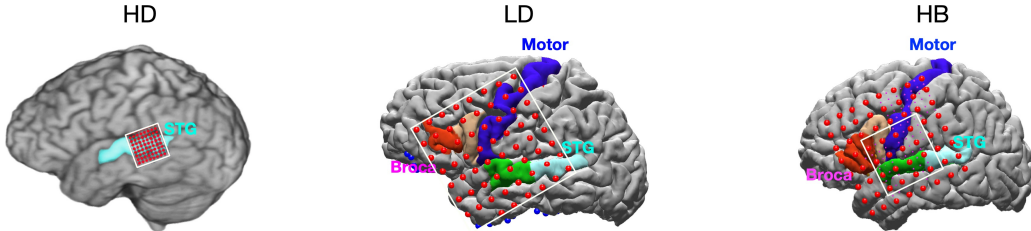


Figure 3.5: Examples of electrode grid of HD, LD and HB datasets. Electrodes within the cropped regions are included in the input to the network. For HD dataset, this covers mostly the STG area. For LD and HB datasets, the motor and Broca’s areas are also covered. In all cases, the input covers a 8×8 grid. Some subjects have missing or bad electrodes in the area chosen. For simplicity, we assumed the signals are all zero in those locations.

3.6.2 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 ECoG decoder network covers 51 ms temporal field. Adam optimizer [51] is used to fit models with hyper-parameters as following: generator pre-training ($lr = 10^{-4}$, $\beta_1 = 0$, $\beta_2 = 0.9$), ECoG decoder-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 [69] 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 [69] and LJ speech dataset [47], 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 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) [79], 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 ECoG decoder and generator is the same as that of SpecWaveNet. We also compared with the linear regression model decoding accuracy reported by Pasley et al [61]. Table 3.1 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. 3.6 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.

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

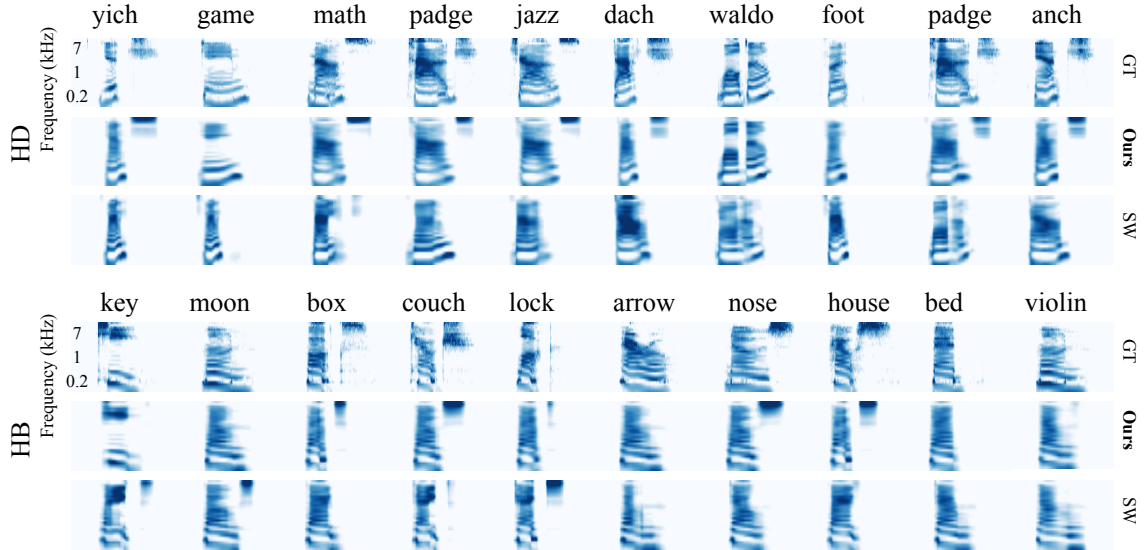


Figure 3.6: Decoding results on each one of HD and HB testing set. GT and SW denote ground truth and SpecWaveNet, respectively.

	MSE (\pm sd) / CC (\pm 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 3.1: Quantitive comparison between transfer-GAN (proposed), SpecWaveNet [79], and linear model [61] in MSE (lower is better) and CC (higher is better) on test data. “-” refers to number not reported.

3.6.3 Production Reconstruction

Besides decoding stimuli from neural data, we also investigated decoding produced speech with transfer-GAN framework following aforementioned network architecture and training strategy. Figure 3.7 reports the decoded samples. Compared with decoded spectrogram illustrated in figure 3.6, the decoded production spectrogram are more blurry and the resulted waveform are not as intelligible compared to the stimuli ones. In terms of objective metrics, the decoded production has averaged MSE 0.673 and CC 0.59 while decoding stimuli achieves 0.53 MSE and 0.72 CC. Both objective metrics and perceptual quality imply that decoding production is a more complicated task than decoding stimuli within Transfer-GAN framework.

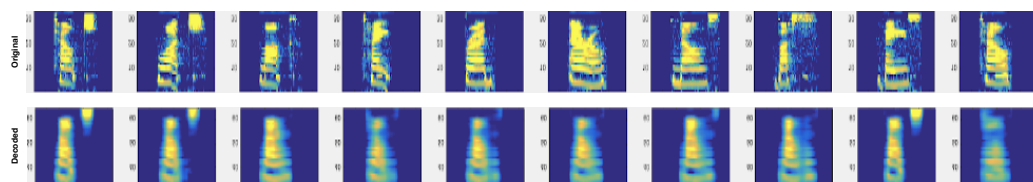


Figure 3.7: Samples of decoded produced speech spectrograms with Transfer-GAN framework.

Chapter 4

Stimuli Decoding from Non-grid Electrodes Inputs

In the previous chapter, we introduced a neural decoding framework that is able to project grid ECoG signals to spectrogram/waveform. In real applications, the electrode nodes are usually not in a grid shape. For example, there may be missing electrodes in the grid. One possible solution is to pad zeros or interpolate for the missing electrodes, but this may still lead to sub-optimal results. In many other cases, the electrodes array is constructed with non-grid placed stripes. In these cases, it is difficult to use convolution in the spatial dimension.

Motivated by the previous literature on attention based networks [76, 77, 80], we propose an attention based network to solve the non-grid input problem. In attention based networks, such as non-local networks and graph attention networks (GAT) input are aggregated according to the signal values instead of grid neighbors, thus naturally bypass the difficulty of convolution on non-grid data. In the following sections, we will first introduce the attentional building blocks and the positional encoding used in the encoder. Then the overall encoder structure will be described in section 4.3.

4.1 Spatial-temporal Separable Conv-Attention Layer

The input to the attention block is a set of feature sequences, $\mathbf{x} = \{\vec{x}_1(t), \vec{x}_2(t), \dots, \vec{x}_N(t)\}$, $\vec{x}_i(t) \in \mathbb{R}^F$, where $0 < t \leq T$, N is the number of nodes and F is the number of features in each node. The layer generates a new set of node feature sequences, $z = \{\vec{z}_1(t), \vec{z}_2(t), \dots, \vec{z}_N(t)\}$, $\vec{z}_i(t) \in \mathbb{R}^F$, as its output.

For such a spatial-temporal signal, we can apply the attention mechanism on both dimension. Or we can apply attention to spatial dimension only and perform convolution on the temporal axis. During experiments, we discover that the attention on both temporal and spatial is not very stable. So we propose a separable convolution-attention block described as following:

We first obtain the *query*, *key* and *value* for attention via convolution on the temporal axis of \mathbf{x} . For $\forall i \in [1, N]$,

$$\vec{q}_i(t) = \text{Conv1D}(\vec{x}_i(t), \Theta) \quad (4.1)$$

$$\vec{k}_i(t) = \text{Conv1D}(\vec{x}_i(t), \Phi) \quad (4.2)$$

$$\vec{v}_i(t) = \text{Conv1D}(\vec{x}_i(t), \mathbf{G}) \quad (4.3)$$

where Θ , Φ and \mathbf{G} are the temporal convolution kernel shared among nodes.

From *query* $\vec{q}_i(t)$ and *key* $\vec{k}_i(t)$, we want to obtain a static attention matrix for each time step for stability purpose. To this end, a max-pooling layer is applied on $\vec{q}_i(t)$ and $\vec{k}_i(t)$ along temporal dimension,

$$\vec{q}_i = \max(\vec{q}_i(t) | 0 < t \leq T) \quad (4.4)$$

$$\vec{k}_i = \max(\vec{k}_i(t) | 0 < t \leq T) \quad (4.5)$$

Then the attention coefficient between i th and j th node α_{ij} is calculated by:

$$e_{ij} = \text{LeakyReLU}(\mathbf{W}[\vec{q}_i \parallel \vec{k}_j]) \quad (4.6)$$

$$\alpha_{ij} = \text{softmax}(e_{ij}) \quad (4.7)$$

$$= \frac{\exp(e_{ij})}{\sum_{j \leq N} \exp(e_{ij})} \quad (4.8)$$

where \parallel represents concatenation operation, \mathbf{W} represents weights of a fully connected layer.

The obtained attention coefficient is used as summation weights to aggregate features of different node of value $\vec{v}(t)$:

$$\vec{v}'_i(t) = \text{ReLU} \left(\sum_{j \leq N} \alpha_{ij} \vec{v}_j(t) \right) \quad (4.9)$$

To allow more expressive power, we follow similar *multi-head attention* as introduced in [76, 77]. Namely, K sets of independent attention coefficients and *values* are generated to form K different sets of features follow the transformation of Equation 4.9. These features are then concatenated and followed by nonlinear transformation for sufficient feature combinations.

$$\vec{v}'_i(t) = \parallel_{k=1}^K \text{ReLU} \left(\sum_{j \leq N} \alpha_{ij}^k \vec{v}_j^k(t) \right) \quad (4.10)$$

$$= \parallel_{k=1}^K \text{ReLU} \left(\sum_{j \leq N} \alpha_{ij}^k \text{Conv1D}(\vec{x}_j(t)) \right) \quad (4.11)$$

$$\vec{y}'_i(t) = \text{LeakyReLU}(\mathbf{P} \vec{v}'_i) \quad (4.12)$$

where \vec{x}_j is input feature vector of j -th electrode of shape $T \times 1 \times F$. We also follow a similar residual design in the non-local network. The output of the convolution-attention block then becomes:

$$\vec{z}_i(t) = \vec{x}_i(t) + \vec{y}'_i(t) \quad (4.13)$$

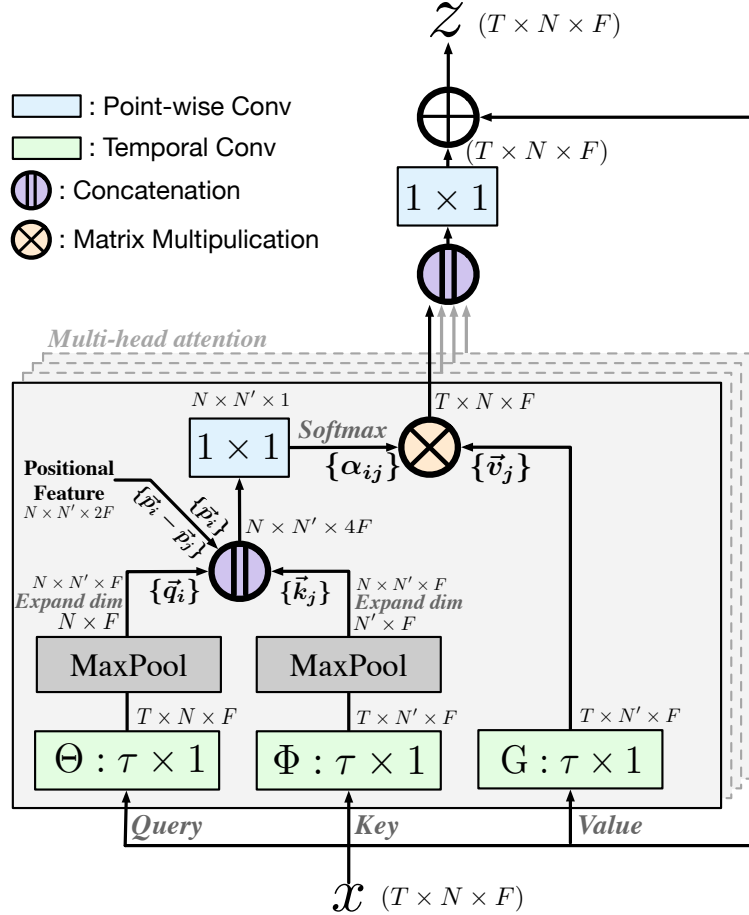


Figure 4.1: Structure of Conv-Att block. The input and output of the block are in shape of $T \times N \times F$ (*time* \times *nodes* \times *feature*). The block operates temporal and spatial/node dimensions with convolution and attention separately. The attention coefficient computation is augmented with MNI positional encoding to achieve more expressive power.

4.2 MNI Positional Encoding

Despite that the attention based encoder has the benefit of processing non-grid ECoG signals, we notice that the attention based encoder is out-performed by the convolution based encoder with grid ECoG input. The similar phenomenon has also been discovered in [8] where 2D convolution ResNet achieve higher accuracy than non-local network in image classification tasks. Experiments in [8, 67] show positional encoding on Euclidean space is able to improve attention based network performance. Inspired by this work, we propose a positional encoding approach on non-grid ECoG input.

Existing positional encoding method in attention based network usually requires the

encoded position on an Euclidean grid. To overcome this restriction, we introduce a positional encoding approach based on MNI coordinate of the electrodes. The MNI template is a 3D brain template proposed by Montreal Neurological Institute [18] that allows MRI scan of the individual brain to transform to. With a fixed origin on the template, The MNI coordinate of the electrodes can be defined from the corresponding 3D coordinate of the electrodes after the MNI mapping.

For the purpose of encoding position information, we first use a mapping layer that transfers the MNI coordinate to a representation feature of a higher dimension. The MNI feature $\vec{v}_i \in \mathbb{R}^F$ for the i th electrode is

$$\vec{p}_i = \text{LeakyReLU}(\mathbf{U}\vec{c}_i) \quad (4.14)$$

where the transformation matrix $\mathbf{U} \in \mathbb{R}^{F \times 3}$ is shared among all electrodes and \vec{c}_i is the MNI coordinate.

Equipped with the above MNI feature, the attention coefficient computation in section 4.1 can be augmented by concatenating MNI features with *query* and *key*. Specifically, Equation 4.6 is modified as:

$$e_{ij} = \text{LeakyReLU}(\mathbf{W}[\vec{q}_i \parallel \vec{k}_j \parallel \vec{p}_i \parallel (\vec{p}_i - \vec{p}_j)]) \quad (4.15)$$

the concatenation of \vec{p}_i and $\vec{p}_i - \vec{p}_j$ make use information of both the absolute position and relative position. \vec{p}_i allows the network to be aware of the actual brain area of electrode i to perform potentially different functions accordingly. And $\vec{p}_i - \vec{p}_j$ provides spatial relationship between electrode i j .

Figure 4.1 demonstrates how MNI positional encoding augmented Convolution-Attention block works with tensor inputs. To illustrate more clearly how tensor dimensions are manipulated, we refer spatial dimension of *query* and *key* with different notations, N and N' . In practice *query* and *key* has same amount of nodes, namely $N = N'$.

4.3 Attention Based Encoder

In this section, we will introduce the attention based encoder that maps non-grid ECoG neural signals to the representation vector z as is described in Chapter 3. Figure 4.2 illustrates the structure of attention based encoder. After several temporal convolution and residual blocks on temporal dimension with temporal down-sampling, two convolution-attention blocks follow. The conv-attention blocks combine features crossing in both temporal and spatial (electrode nodes specifically) dimensions. The MNI features fed to the blocks are obtained by another 1×1 convolution mapping layer with MNI coordinates as input.

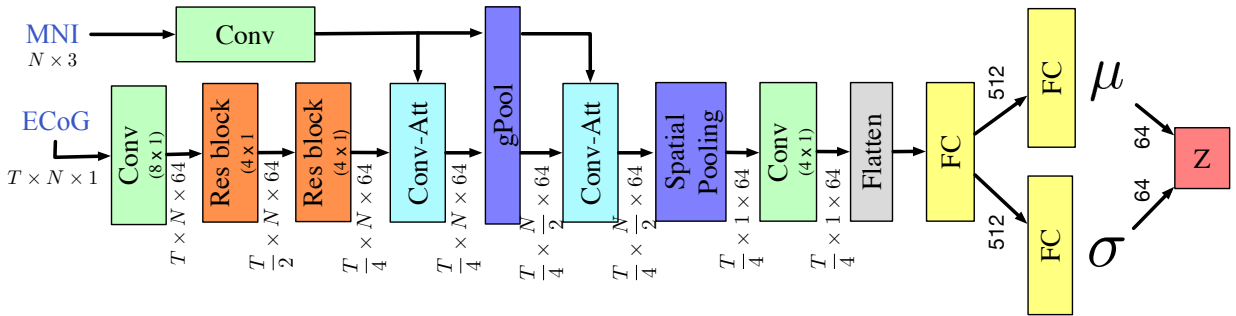


Figure 4.2: Overview structure of the attention based ECoG decoder network.

Similar to spatial convolution networks that usually perform spatial downsampling, we allow the network to pool among electrode nodes. This is achieved by inserting a gPool layer between the two conv-attention blocks. The gPool method was first introduced in [30] to solve the problem of nodes pooling in graph neural networks. We adapt the method to take input of non-grid sequential signals. By average pooling, a temporal collapsed version of the last layer output is fed to a gPool layer. Within the gPool layer, a node selection array is generated. We use such array to index the selected electrode nodes for both ECoG and MNI features after pooling.

The rest layers of the decoder follow the same design of the later parts of the encoder introduced in Chapter 3 to generate a representation vector z .

4.4 Experimental Performance of attention based encoder

In order to generate spectrograms and waveforms from the ECoG inputs, we substitute the encoder layers of the decoding framework introduced in chapter 3. The pretraining/fintuning of the generator and the entire decoder follow the same strategy and hyperparameters. To test the performance of the proposed attention based encoder, we preliminary use one of the patient data from the Hybrid dataset as is described in section 3.6.1 for training and testing. The hybrid electrode nodes have overall 10 mm spacing with particular regions inserted by 5 mm spacing sub-grid electrodes.

The convolutional encoder of the previous chapter pad zeros for the “missing” electrodes in the grid region. For the attention based encoder, we just treat all electrodes as input nodes to the encoder and ignore the “missing” electrodes. The rest steps of preprocessing follow the same methods introduced in section 3.6.1.

For the subject we test on, 50 words are used as training samples and the rest 50 words as test samples. Figure 4.3 illustrate some reconstructed samples from the decoding network with attention based encoder.

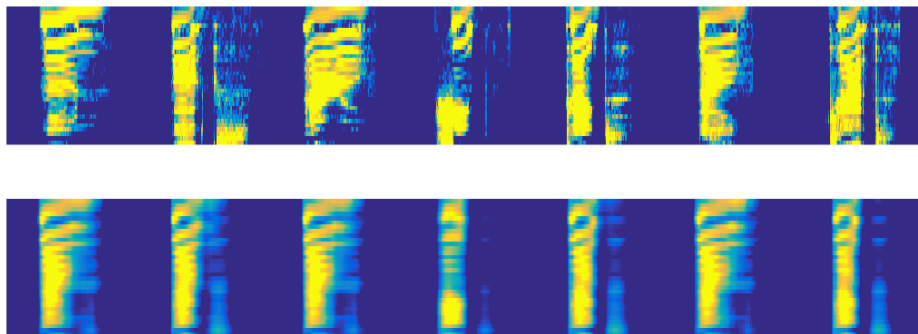


Figure 4.3: Reconstructed samples from the decoding network with attention based encoder. Upper row: ground truth spectrogram, lower row: reconstructed spectrogram

Table 4.1 compares convolutional and attention based encoder performance quantitatively. The results show very close-up performance between convolutional and attention based encoder with partially grid input. It is reassuring that the proposed attention based model achieves similar performance as convolution based model. This encourages the usage of attention based model on entirely non-grid data where convolution on space is difficult to

perform.

MSE/ CC		
	Convolutional decoder	attention based decoder
HB	0.50/ 0.72	0.50/ 0.71

Table 4.1: Quantitive comparison between convolutional and attention based decoder for spectrogram reconstruction.

In the future works, we are going to further test the performance of attention based model on other hybrid and non-grid data.

Chapter 5

Generate Produced Speech with Differentiable Speech Synthesizer

In Chapter 3, the nature of generator training limits the ECoG decoder to output a fixed length vector. This makes decoding speech with variable length difficult. Also, generating a fixed length vector without physical meaning is less interpretable. In this chapter, we solve this issue by introducing a novel neural decoding framework that is constructed with a differentiable speech synthesizer.

5.1 Speech decoding framework

The backbone of our neural decoding framework is constructed by an ECoG decoder and a speech synthesizer (Fig. 5.1a). During testing, from the high gamma components of the ECoG signal, the ECoG decoder generates a set of speech parameters that drive a differentiable speech synthesizer to generate speech spectrograms (and corresponding waveforms by Griffin-Lim algorithm [33]). During training, besides being trained to match with the speech synthesizer to output spectrograms matching the target spectrograms, the ECoG decoder is also trained to match its output with a set of reference speech parameters. This reference matching training strategy provides a more direct gradient to the ECoG decoder so that it converges faster and is less prone to overfit.

The reference speech parameters are derived from a pre-trained speech encoder. During

pre-training, the speech encoder and the speech synthesizer fulfill an auto-encoding task (i.e., mapping the input spectrogram to the speech parameters and back to the spectrogram) (Fig. 5.1b). When such speech-to-speech reconstruction is accurate, the parameters generated by the speech encoder should provide physically meaningful speech parameters. Since the pre-training is unsupervised and the subject speech audio data is easy to collect, obtaining the reference speech parameters is not laborious. Note that the speech-to-speech autoencoder and the reference parameters are only used for the training of the ECoG decoder. Once the ECoG decoder is trained, the trained decoder and the speech synthesizer can be used to convert ECoG signals to speech, without the need for the reference parameters.

- **ECoG Decoder.** The decoder maps the ECoG signals to a set of speech parameters (describing both the voiced and unvoiced components) which are then synthesized to speech spectrograms (Fig. 5.1). The ECoG decoder architecture is based on recent advances in convolutional neural networks leveraging the ResNet approach [37]. We construct a modified ResNet model with nine layers which treats the cortical input as a spatiotemporal three-dimensional tensor (two dimensions for the electrode array and one for time). The decoder is trained such that its output parameters match the reference parameters derived from a speech encoder (which is learnt separately in an unsupervised manner). Furthermore, our approach ensures that the speech spectrogram derived from these parameters, and constructed by the speech synthesizer, matches with the actual speech spectrogram. We use this approach rather than an end-to-end training as it is more data-efficient and allows us to train on a small set of samples for each patient.
- **Speech Parameters.** Our speech representation is motivated by the vocoders used for low-bit-rate speech compression dating back to the 1980's. We model speech signals as a mixing between voiced and unvoiced components, with the voiced component described by a source-filter model (dynamically filter harmonic signals) [21] and the unvoiced component generated by white noise broadband filtering. In addition to the mixing parameter, our representation includes speech formant information (frequency, bandwidth, etc) and loudness (i.e., energy of speech). See Fig 5 for details.

- **Synthesizer.** Rather than training our speech synthesizer we use a set of signal processing equations (such as harmonic oscillation, noise generation, filtering, etc.) to synthesize the spectrogram from the speech parameters. By limiting the number of speech parameters and using differentiable signal processing equations, we are able to train the ECoG decoder with a limited amount of training data. In addition, we are able to augment training by using a direct guidance approach where speech parameters (output of the decoder) are matched to the reference parameters derived from the actual speech (see Speech Encoder). This matching process (i.e., “match” in Fig. 5.1) allows for an accurate estimation of speech parameters and decoded spectrograms based on the ECoG signals. It is noteworthy that the equations we use are differentiable which allows for backpropagation from the spectrogram to the actual learning of the decoder.
- **Speech Encoder.** The speech encoder is pre-trained using an independent unsupervised approach prior to the ECoG decoder training. The encoder is trained to generate a set of speech parameters from a given spectrogram, from which the aforementioned speech synthesizer can reproduce the spectrogram. This pre-trained encoder generates reference speech parameters from actual speech signals, used for the training of the ECoG decoder. The unsupervised process can be easily used to train the speech encoder from any set of speech signals including patient specific speech. Importantly, this process constrains the speech parameter space to optimize our decoding learning and the parameters can directly drive a speech synthesizer based on differential equations.

In the rest of this section, we will first introduce the structure of the speech synthesizer (Fig. 5.1e). Then we will describe the ECoG decoder (Fig. 5.1c) and Speech encoder (Fig. 5.1d).

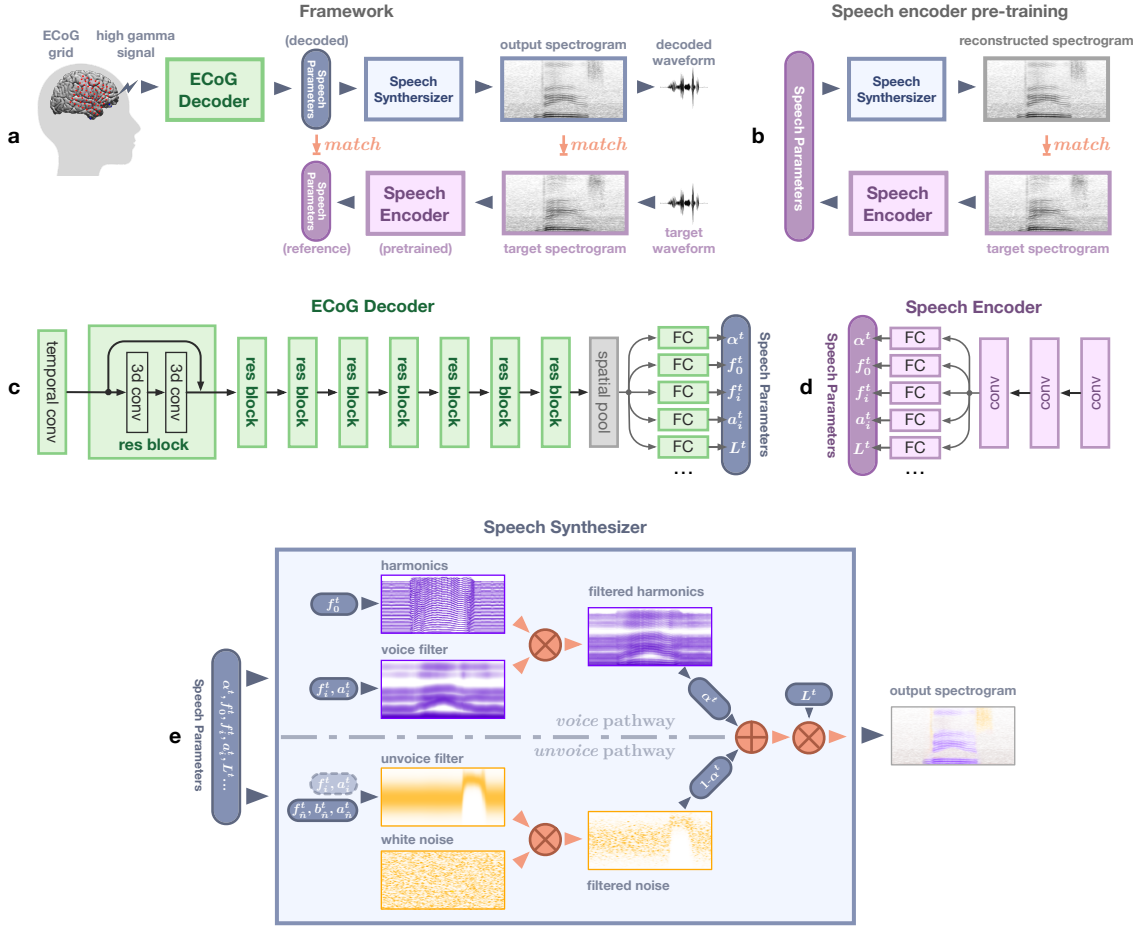


Figure 5.1: Structure of the decoding framework. (a) the overview of the overall network architecture. An auto-encoder structure (b) is used to pretrain the speech encoder by training a speech encoder to generate proper speech parameters that can reconstruct input spectrograms through the speech synthesizer. (c) The ECoG decoder is a modified three-dimensional residual network. After an initial temporal convolutional layer and eight residual blocks (constructed by three-dimensional convolutional layers), multiple convolutional layers (each has temporal kernel size of 1) generate speech parameters separately. (d) The speech encoder in (a) is constructed by a convolutional network, with three convolutional layers backbone and the same multihead output structure as in (c). (e) Illustrates the processes within the speech synthesizer. The harmonics (in voice pathway) and white noise (unvoice pathway) are generated and filtered (multiplication in spectrogram domain) by voice and unvoice filters, respectively. The filtered results are then weighted averaged according to a mixing parameter and then amplified by the loudness parameter.

5.1.1 Differentiable speech synthesizer

In traditional vocoder, speech is generated by switching between voiced content and unvoiced content. Each content comes from an autoregressive system driven by a certain excitation signal that is either a harmonic signal or a white noise signal [16]. Inspired by such a process, we construct our speech synthesizer shown in Fig. 5.1(e). It consists of two pathways. The *voice pathway* generates a voiced component by driving a harmonic excitation with time varying fundamental frequency (i.e., pitch) f_0^t through a voice filter consisting of N formant filters, each described by a center frequency f_i^t and an amplitude $a_i^t, i = 1, 2, \dots, N$. The *unvoice pathway* generates an unvoiced component by driving a white noise through an unvoice filter described a center frequency f_n^t , bandwidth b_n^t and amplitude a_n^t (in addition to the N formant filters for the voice pathway). These two components are adaptively combined with a time-varying mixing factor α^t , controlling the relative contribution between voiced sounds (for sonorant phonemes including vowels and nasals) and unvoiced sounds (for voiceless plosives and fricatives such as /p/, /s/). The voiced plosives and fricatives (such as /b/, /z/) can be generated as a combination of voiced and unvoiced components. Finally, the combined signal is amplified by a loudness parameter L_t . In our study, we used $N = 6$ formants. The synthesizer is driven by a total of 18 time-varying speech parameters, including the fundamental (or pitch) frequency f_0^t , the mixing factor between the two pathways α^t , the 12 parameters for the voice filter (f_i^t, a_i^t) and the 3 parameters for the unvoice filter f_n^t, b_n^t, a_n^t , and the loudness L^t . Given the parameter values at each time sample, the synthesizer can generate a spectrogram sample. The spectrogram is a differentiable function of the speech parameters so that we can back-propagate the gradient of the training loss in terms of the predicted spectrogram to the speech parameters, which can then be backpropagated to either the speech encoder or the ECoG decoder parameters. Specifically, let the $V^t(f)$ represent the spectrogram of the voicing component, $U^t(f)$ that of the unvoicing component, and $\alpha^t \in [0, 1]$ the mixing factor. The combined spectrogram can be written as $S^t(f) = \alpha^t V^t(f) + (1 - \alpha^t) U^t(f)$. (1) Finally, the synthesized speech spectrogram is $\tilde{S}^t(f) = L^t \circ S^t(f)$, where L^t is the loudness that modulates the signal cross time.

Formant filters in the voice pathway The filter in the voice pathway consists of multiple formant filters, corresponding to the multiple formants associated with vowels. The formant filter shape over frequency, which is related to the resonance property of the vocal tract, is closely related to the timbre of speakers’ voice [46]. We have found that a predefined analytic form such as generalized Gaussian cannot cover all feasible filter shapes . Instead, we learn a speaker-dependent prototype filter for each formant based on the speaker’s natural speech. We represent the prototype filter ($G_i(f)$) for the i -th formant as a piecewise linear function, linearly interpolated from $g[m], m = 1 \dots M$, the amplitudes of the filter at M uniformly sampled frequencies up to f_{max} . We restrict the resulting filter $G_i(f)$ to be unimodal (with a single peak of value 1) by properly constraining $g[m]$. Given $g[m], m = 1 \dots M$, the peak frequency f_i^{proto} and the half-power bandwidth b_i^{proto} can be determined. The actual formant filter at any time can be written as a shifted and scaled version of $G_i(f)$. Specifically, at time t , given an amplitude (a_i^t), a center frequency (f_i^t), and a bandwidth (b_i^t), the i -th formant filter is given by

$$F_i^t(f) = a_i^t \cdot G_i \left(\frac{b_i^{proto}}{b_i^t} \cdot (f - f_i^t) + f_i^{proto} \right) \quad (5.1)$$

Then the filter for the voice pathway with N formant filters can be written as $F_h^t(f) = \sum_{i=1}^N F_i^t(f)$. We learn the parameters $g[m], m = 1 \dots M$ for $G_i(f)$ during the unsupervised pre-training of the speech encoder, which does not require neural data. Fitting such a prototype filter is not data-hungry even with a relatively large M . We used $M = 20$ in our experiment. Although two formants ($N=2$) have been shown to suffice for intelligible reconstruction [11], we use $N=6$ in our experiments for more accurate synthesis. We denote the parameter set for the voice filter at time t by $\mathcal{S}^t = \{(f_i^t, a_i^t, b_i^t) | i \in \{1, \dots, N\}\}$. As explained later, the bandwidth b_i^t parameters are not independent speech parameters, rather functions of the center frequencies f_i^t .

Unvoice filter For the unvoice pathway, we add a broadband filter described by $\{(f_n^t, a_n^t, b_n^t)\}$. The shape of this filter $F_n^t(f)$ follows equation 5.1 but with the filter coefficients $(\alpha_i^t, f_i^t, b_i^t)$ replaced by $(\alpha_n^t, f_n^t, b_n^t)$. The bandwidth is constrained to satisfy $b_n^t > 2000\text{Hz}$, following the

broadband nature of obstruent phonemes. We also keep the multiple formant filters in the voice filter described by \mathcal{S}^\sqcup . This is motivated by the fact that human beings differentiate consonants with similar sounds such as /p/ and /d/, not only by the immediate burst of these sounds, but also the development of the following formant frequency until the next vowel [49]. To encode such formant transitions, we use the same formant filter parameters for modeling the narrow passbands in both the voiced component and the unvoiced component. The parameter set for the unvoiced component is thus $\mathcal{T}^\sqcup = \mathcal{S}^\sqcup \cup \{(f_n^t, a_n^t, b_n^t)\}$. The overall filter for the unvoice pathway is: $F_n^t(f) = F_n^t(f) + \sum_{i=1}^N F_i^t(f)$.

To further reduce the dimension of the parameter space, we model the bandwidth b_i^t of a formant filter as a piecewise linear function of the center frequency f_i^t , inspired by the statistics reported in the literature [36]. Specifically, we assume

$$b_i^t = \begin{cases} a(f_i^t - f_\theta) + b_0, & \text{if } f_i^t > f_\theta \\ b_0, & \text{otherwise} \end{cases}$$

where threshold frequency f_θ , slope a and baseline bandwidth b_0 , are three parameters that can be learnt during unsupervised pre-training, shared among all formant filters.

Harmonic excitation In the voice pathway, the voice filter is applied on the harmonic excitation. This pathway models the human production of vowels and nasals, which results from the voice excited by the vocal cord shaped by the vocal tract. The excitation is constructed by sinusoidal harmonic oscillations with a time varying fundamental frequency f_0^t . Inspired by the formulation in [19], we define the harmonic excitation h^t as: $h^t = \sum_{k=1}^K h_k^t$, where K is the total number of harmonics ($K=80$ in our experiment). Assuming the initial phase is 0, each harmonic resonance h_k^t at time step t has an instant phase that is the accumulation of resonance frequency in the past. Specifically, the k -th resonance at time step t is $h_k^t = \sin(2\pi \sum_{\tau=0}^t f_k^{(\tau)})$, where $f_k^{(t)} = k f_0^{(t)}$. Denoting the spectrogram of h^t as $H^t(f)$, the spectrogram of the voice component is the multiplication of $H^t(f)$ and the voice filter, i.e., $V^t(f) = H^t(f) \circ F_h^t(f)$.

Noise excitation The unvoice pathway models consonants like plosives and fricatives, where the vocal tract and human mouth filter the airflow through the mouth. It follows a similar process as in the harmonic counterpart. The major difference is that the excitation being filtered becomes stationary white Gaussian distributed noise $\hat{n}(t) \sim \mathcal{N}(0, 1)$, with a corresponding spectrogram $N^t(f)$. The filtered noise spectrogram (i.e., the unvoice component) is $U^t(f) = N^t(f) \circ F_n^t(f)$.

5.1.2 ECoG decoder and speech encoder

The ECoG decoder is constructed by a three-dimensional ResNet that treats time-varying signals on an ECoG grid array as spatiotemporal three-dimensional tensors (width \times height \times time). As is depicted in Fig. 5.1c, after an initial temporal convolutional layer (with 128 feature map filters and kernel size of $1 \times 1 \times 9(72ms)$), the signal passes through eight residual blocks. Each block contains two three-dimensional convolutional layers (with 128 feature map filters, each has kernel size of $3 \times 3 \times 5(50ms)$). The output of the residual blocks creates a shared latent representation consisting of 128 feature maps (each is a one-dimensional temporal signal by average pooling the two spatial dimensions), which is then fed into different output heads (each applies one-dimensional convolution with temporal kernel size of 1) to generate speech parameters. The overall temporal receptive field for generating one speech parameter is 73 temporal samples of 584 ms.

The speech encoder network architecture we choose is as simple as possible to demonstrate the effectiveness of the speech synthesizer design. In the experiment, we use three layers of temporal convolution (we treat the frequency axis of the spectrogram as the feature dimension) to generate a latent representation (Fig. 5.1d). Each convolutional layer has 128 feature maps and temporal kernel size of 3 frames (24ms). To output the speech parameter, we apply the same multi-head structure to the latent representation as in the last layer of the ECoG decoder.

5.2 Loss and training hyper-parameters

The speech encoder is trained with a weighted average of the mixed spectral loss and the parameter loss. The mixed spectral loss [19] is defined as:

$$L_{MSS}(\tilde{S}^t(f), S^t(f)) = L_{\text{lin}}(\tilde{S}^t(f), S^t(f)) + L_{\text{mel}}(\tilde{S}^t(f), S^t(f)),$$

in which,

$$L_{\text{lin}}(x, y) = \|x - y\|_1 + \|\log x - \log y\|_1$$

$$L_{\text{mel}}(x, y) = \|x_{\text{mel}} - y_{\text{mel}}\|_1 + \|\log x_{\text{mel}} - \log y_{\text{mel}}\|_1$$

where $S^t(f)$ and $\tilde{S}^t(f)$ denote the ground truth and reconstructed spectrograms, respectively, subscript *lin* means that the frequency is in the linear scale while the subscript *mel* means the frequency is in the mel scale.

Let's denote the j -th reconstructed speech parameter as \tilde{P}_j^t and its reference P_j^t , the overall training loss for the ECoG decoder becomes:

$$\begin{aligned} L &= L_{\text{spectrogram}} + L_{\text{speechparameters}} \\ &= \lambda_0 L_{MSS}(\tilde{S}^t(f), S^t(f)) + \sum_j \lambda_j (\|\tilde{P}_j^t - P_j^t\|_2) \end{aligned}$$

where λ_j balance the contribution from different loss terms since they have different physical meanings and scales.

Both the speech encoder and ECoG decoder are fitted by Adam optimizer with hyper-parameters: $lr = 10^{-3}, \beta_1 = 0.9, \beta_2 = 0.999$. We train an individual ECoG decoder and speech encoder per patient. The pre-training of the speech encoder and the training of the ECoG decoder share the same training/testing set partition. During the training, we augment the data by randomly selecting a one second segment from the production period of each trial and repeating this multiple times. The segment is cropped so that it contains

the entire vocalized word.

5.3 Produced speech decoding performance

We first demonstrate that our approach produces accurate speech decoding with detailed acoustic features. The model’s decoded spectrogram preserves the spectro-temporal structure of the original speech and reconstructs both vowels, consonants (Fig. 5.2a) as well as the overall spectral energy distribution. These acoustic details result in a reconstruction which preserves the speakers’ timbre and leads to naturalistic voice decoding. Our model’s speech parameters which include loudness, formant frequency and the mixing parameter (i.e., voiced or unvoiced) are decoded accurately with correct temporal alignment of each word onset and offset (Fig. 5.2b, c). The overall accuracy of the fundamental frequency (i.e., pitch) and the first two modeled formants (i.e., F1-F2), together with other parameters, are a major driving force for accurate speech decoding as well as naturalistic reconstruction which mimics the patient’s voice.

In order to evaluate the performance and quality of speech we use several objective metrics, including the correlation coefficient (CC) between the decoded spectrogram and actual produced speech [3, 4, 38], an objective measure for speech intelligibility, Short Time Objective Intelligibility (STOI) [4, 73], and a measure of spectral distortion, mel-cepstral distortion (MCD) [5, 53]. Across all participants and metrics our neural decoding results performed much above chance (Fig. 5.2d in grey; estimated using shuffled data) and approached an upper bound of performance based on the unsupervised auto-encoder (i.e., speech-to-speech) which does not use neural data. The performance range across metrics and our participants are equal and often better than the current literature [3–5, 38]. Critically, all these models represent the non-causal case (Fig. 5.2d) which uses data both from the past (feedforward) and the future (feedback), which is currently a common practice [2–5, 55] with the exception of a nominal few [38].

In order to directly assess the performance of the causal (predicting using only the past) and anti-causal (predicting using the future feedback) models and compare them with the non-causal (using past and future) model, which is standard in the field, we trained three

separate models varying the temporal convolution direction. Our results (Fig. 5.2e) show a slight decrease in performance with the causal model, however it performs close to the other models while providing a more accurate and constraint interpretation because it is causal and only uses past signals to predict the future speech. This is encouraging, as it suggests that, with additional improvement in the decoder design and training, it is possible to design practically applicable neuroprosthetic speech synthesizers. Also, comparable performance between causal, anticausal, and noncausal approaches indicates a similar amount of information contained by feedforward and feedback signals, and that both causal and anticausal models are accurate for fair feedforward-feedback analysis.

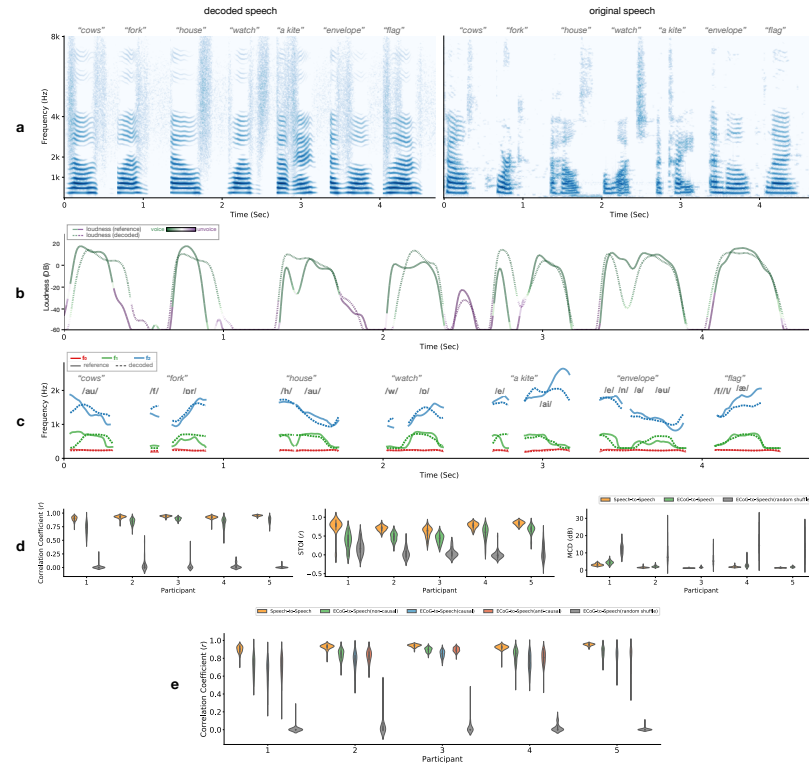


Figure 5.2: Comparison of original and decoded speech produced by the model. (a) Spectrograms of decoded (left) and original (right) speech exemplar words. (b) Decoded loudness parameter with the voiced (mostly vowel) or unvoiced (mostly consonant) mixing parameter color coded over the loudness curves. (c) Frequencies of the first two formants (F1, F2) and the pitch (F0). The matching between decoded (dashed) curves and reference (solid) curves in both averaged frequencies during each phoneme as well as the overall temporal dynamic leads to intelligible and naturalistic decoding of voiced sounds. (d) Evaluation of the decoded speech quality in objective metrics. The correlation coefficient of spectrogram (CC, left), short-time objective intelligibility (STOI, middle), and Mel cepstral distortion (MCD, right) are used for the evaluation. Note that lower MCD values represent better performance. Both the reconstructed speech from the speech auto-encoder (yellow boxes) and the speech decoded by the ECoG decoder (green boxes) are reported. Besides, the performance of a model trained on a shuffled dataset (trained by matching the decoded spectrogram from a neural signal in a given duration to randomly selected segments of spectrograms during the entire recording session) is also reported as a control. (e) Comparison of CC metric among noncausal (green), causal (blue), and anticausal (red) models. Compared to the shuffled model (the same shuffled model as in Fig. 5.2d), the close performance among noncausal, causal, and anticausal models demonstrates adequate information for decoding speech in both feedforward and feedback signals during speech production.

5.4 Perceived speech decoding performance

Besides decoding produced speech, we also investigated decoding speech stimuli with the same speech synthesizer and framework following the aforementioned network design. The network decodes ECoG signal during perception period to generate perceived stimuli. Figure 5.3 reports the decoded samples. The spectro-temporal structure of the original speech are well preserved by the model and both vowels and consonants are reconstructed (Fig. 5.3a). The speech parameters are decoded accurately (Fig. 5.3b, c). The correlation coefficient reported in Fig. 5.3e (purple boxes) demonstrate high performance of the decoded stimuli in objective metrics. This result shows that our proposed speech synthesizer is capable of decode both produced and perceived speech. Participants 4 and 5 have damaged STG. The average CC among the first three participants are 0.88 (perception) and 0.84 (production). And the fact that accuracy of decoded speech for perception is higher than production further supports that decoding produced speech is a more difficult task than decoding perceived speech. Comparing to the results yielded by trans-GAN approach introduced in Chapter 3 where decoding achieves 0.72 (perception) average CC and 0.59 (production) average CC, on the same dataset, the method proposed in this chapter has a significant improved performance with 0.88 (perception) and 0.84 (production) average CC, respectively. Since participants 4 and 5 have damage to the STG, we exclude them for average CC computation.

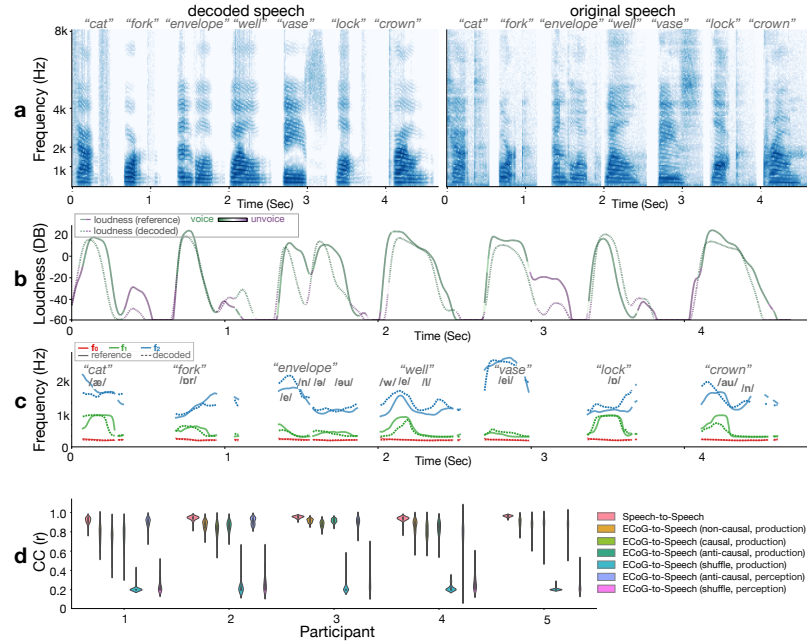


Figure 5.3: Comparison of original and decoded speech produced by the model for stimulus decoding. (a) Spectrograms of decoded (left) and original (right) speech exemplar words. (b) Decoded loudness parameter with the voiced (mostly vowel) or unvoiced (mostly consonant) mixing parameter color coded over the loudness curves. (c) Frequencies of the first two formants (F1, F2) and the pitch (F0). The matching between decoded (dashed) curves and reference (solid) curves in both averaged frequencies during each phoneme as well as the overall temporal dynamic leads to intelligible and naturalistic decoding of voiced sounds. (d) Evaluation of the decoded speech quality in objective metrics. The correlation coefficient of spectrogram (CC). The speech decoded by the ECoG decoder (purple boxes) are reported. Besides, the performance of a model trained on a shuffled dataset (magenta boxes, trained by matching the decoded spectrogram from a neural signal in a given duration to randomly selected segments of spectrograms during the entire perception period) is also reported as a control.

5.5 Discussion

Recently, a growing number of studies have leveraged deep neural networks for cortical speech decoding. Both convolutional neural networks (CNN) [2, 4, 78, 79] and recurrent neu-

ral networks (RNN) [5] have been employed to map ECoG signals into speech as well as text [55]. However, our approach greatly diverges from these studies. *Firstly*, we develop a novel differentiable speech synthesizer that can generate natural speech from a compact set of interpretable speech parameters based on several signal processing equations. This rule-based synthesizer allows for unsupervised pre-training of meaningful encoded representations (reference speech parameters), as well as reduces the capacity of the entire model and increases training data efficiency. Our approach provides a direct mapping to a patient's voice and eliminates the need for labeled articulatory data that maps speech to articulatory dynamics as proposed by Anumanchipalli et al. [5]. Also, we demonstrate that our proposed model is capable of decoding both produced and perceived speech. *Secondly*, our compact speech representation leverages an interpretable decomposition of speech into voiced and unvoiced components. This decomposition is biologically necessary, has been reported in neural representations across frontal and temporal cortices [12, 45] and stands in contrast to other traditional speech synthesizing approaches [21, 22]. *Lastly*, the speech neural decoding models to date mostly employ non-causal operations. Since such decoders require both past and future information for decoding, they are not applicable for real time speech prosthetic application. Furthermore, these mixed operations hinder disentangling feedforward and feedback cortical contributions. In addition to providing a causal model which directly translates to practical speech prosthetics, our approach provides one of the first reports which can dissociate feedforward and feedback cortical contributions during speech production (see Chapter 6).

PART II

Interpreting Speech Cortical Networks

Chapter 6

Feedforward-Feedback Contribution Analysis for Speech Decoding Models

6.1 Background

The human neocortex is anatomically divided by the central sulcus which separates the frontal lobe from the posterior temporal parietal and occipital cortices [50]. Traditionally, this separation has been viewed as a functional division wherein high order planning and feedforward motor execution lies in the frontal cortices in contrast to sensory and refferent feedback processing across posterior cortices for the various sensory modalities (e.g., auditory, visual, somatosensory, etc.) [27]. High order capacities such as working memory, cognitive control and decision making are often viewed as initiated by frontal cortices with direct influence on sensory cortices [29, 57, 72].

Cardinal among human high order functions is the ability to plan and execute complex speech sequences which carry semantic and linguistic meaning [11, 43]. Speech production is one of the most complex of human motor behaviors which requires precise coordination of a multitude of muscles in the mouth, larynx and respiratory systems [70]. These finely tuned motor actions then produce refferent self-produced feedback in the auditory, tactile and proprioceptive domains. The dynamic interaction between feedforward commands and the re-fferent sensory feedback is a hallmark of sensory motor systems across the animal kingdom [15]. Prevailing models in speech motor control propose a feedforward system with

causal prediction of actions and a feedback system processing the reafferent feedback and any changes in the predicted input [34,35,40–42,44]. Despite implementational differences across these models there is a consensus that the causal feedforward system is mainly supported by frontal cortices, while posterior cortices support anticausal feedback processing. However, these models do not specify the timing of recruitment for causal and anticausal processing nor do they agree on the exact set of regions supporting each function.

A growing body of literature has leveraged unique human neurosurgical electrocorticographic (ECoG) recordings in order to obtain a combined spatial and temporal resolution critical for investigating speech production. Studies have detailed the organization and dynamics by which speech is planned [26] and executed [9,12] in frontal cortices as well as the auditory feedback architecture in temporal cortices [10,24,31,32]. Furthermore, the unprecedented signal to noise ratio offered by human neurosurgical recordings has ushered deep neural network approaches to accurately decode speech represented in auditory [2,4,78,79] and sensorimotor [5] cortices. Despite these impressive leaps forward, approaches to date have not been able to disentangle feedforward and feedback contributions to speech as the causal control signals and anti-causal responses co-occur.

We directly address this issue by interpreting the produced speech decoding model that we proposed in Chapter 5. By learning neural network architectures which apply either casual, anticausal (or both) spatial-temporal convolutions we are able to analyze the overall feedforward and feedback contributions, respectively, as well as elucidate the temporal receptive fields of recruited cortical regions. Our analyses reveal a surprisingly mixed architecture of causal and anticausal processing across cortex while achieving speech decoding performance on-par or better than previously reported.

6.2 Method

6.2.1 Revealing delay-dependent contribution of different cortical regions from the trained ECoG to speech model

As described in the background, the main objective of this study is to learn the contribution of different cortical regions to speech production from the developed ECoG to speech

decoding model. We accomplish this by examining how the input signal at a particular electrode (the entire signal or the signal at a certain delay relative to the produced speech) affects the accuracy of the model prediction. There are two major approaches for this purpose: gradient approaches and occlusion approaches [54, 66, 71]. The gradient approaches calculate the gradient of the network output with respect to the input. Visualization results in computer vision studies have shown that such gradients can be noisy at the pixel level [71]. Although extensive research has been conducted to improve the robustness of gradient-based sensitivity analysis, our experiments using such gradient approaches have shown that they do not produce reliable results for the spatially sparse ECoG data. The occlusion approaches investigate the changes in the network output when certain input signals or part of the signals are occluded (set to 0 or appropriate baseline). In our analysis, we use the average value of the ECoG signal during the silent period as the baseline. We have found that this approach produces more reliable contribution analysis. The problem with using the occlusion approach for computer vision applications is the heavy computational complexity, since there are typically millions of pixels in an image. This is computationally feasible for our study, given the relatively small number of electrodes. For these reasons, we adopt the occlusion approach for visualizing the contribution of different cortical regions. Furthermore, we use the changes in the accuracy of the decoded spectrogram (measured in terms of the correlation with the ground truth spectrogram) rather than the decoded spectrogram itself, to define the contribution. Before formally defining the various contribution scores, we introduce the following notations: $A_{\text{ref}}[s]$: the reference spectrogram or speech parameter over a time duration S centered at time s , i.e., from $s - S/2$ to $s + S/2$, derived by the speech-to-speech autoencoder. $A_{\text{intact}}[s]$: the model output with ‘‘intact’’ input (i.e., all ECoG signals are used). $A_{\text{occlude}}^i[s|t]$: the model output at time duration centered at s when the i th ECoG electrode signal in the time duration centered at t from $t - T/2$ to $t + T/2$ is occluded $r(\cdot, \cdot)$: correlation coefficient between two signals. We define the contribution of i th electrode in time duration t (which indicates a duration from $t - T/2$ to $t + T/2$ for a given temporal scope of T) to the output over duration s (which indicates a duration from $s - S/2$ to $s + S/2$ for a given temporal scope of S) by the reduction in the correlation coefficient between the output signal with the reference signal over the duration s when the

ith electrode signal in duration t is occluded. Specifically:

$$C^i[s, t] = \text{Mean}\{(A_{\text{ref}}[s], A_{\text{intact}}[s]) - r(A_{\text{ref}}[s], A_{\text{occlude}}^i[s|t])\}$$

where $\text{Mean}\{\cdot\}$ denotes averaging across all testing samples.

Visualizing spatial contribution map

The contribution of the entire ith-electrode signal to the entire output signal, C^i , can be determined by setting S and T to cover the entire input and output signal duration. The causal and anti-causal contribution plots in Fig. 6.1 were generated by applying such analysis to the learned anticausal model (Fig 6.1b) and causal model (Fig 3c), respectively. The contrast of the anticausal and causal contribution (Fig 6.1e) for electrode i is the difference between the causal and anticausal contribution. And the noise level contribution analysis (Fig 6.1d) is generated from the shuffled model with the non-causal model (the shuffled model is trained on an artificial dataset with temporal misaligned input-output, models of different causality are equivalent). To generate per region feedback-feedforward barplot (Fig 6.1f), we calculate the contrast contributions averaged over electrodes of the same within-subject anatomical labels corresponding to each region.

Visualizing spatial-temporal contribution receptive field

When evaluating the contribution over a finite duration we use small temporal scope $S = T = 64\text{ms}$. To Evaluate the contribution of an electrode signal to the output with various delay, denoted by τ , we average $C^i[s, s + \tau]$ for all s in a certain duration leading to

$$\tilde{C}^i(\tau) = \frac{1}{s_1 - s_0} \sum_{s=s_0}^{s_1} C^i[s, s + \tau]$$

Here we assume the effect of delay is independent of actual output time s . When $\tau \leq 0$, $\tilde{C}_{\text{causal}}^i(\tau) = \tilde{C}^i(\tau)|_{\tau \leq 0}$ reveals the causal contribution of electrode i to the output (Fig 6.1 a,b). To investigate pre-production contribution, we restrict $s + \tau$ and s to be no later than the onset of production (vise-versa for during-production analysis). When $\tau \geq 0$ the

$\tilde{C}_{anticausal}^i(\tau) = \tilde{C}^i(\tau)|_{\tau \geq 0}$ reveals the anticausal contribution (Fig 6.1c).

Visualizing per region temporal contribution receptive field

Similar to the per region plot in Fig 6.1f, to generate a contribution temporal curve for each region (Fig 6.3), we average the spatial-temporal receptive field data (Fig 6.2) over to the same within-subject anatomical region labels. As a reference to the noise level, we also generate the curve for the shuffled model (grey curves in Fig 6.3). We omit those curves that are not significantly above noise level by Wilcoxon sign rank testing between averaged (over time) region contribution curves and the averaged (over time) noise level curve. The contrast of the anticausal and causal contribution of electrode i is defined as

$$\tilde{C}_{contrast}^i = \tilde{C}_{anticausal}^i - \tilde{C}_{causal}^i$$

6.3 Result

6.3.1 Feedforward and feedback cortical contributions to speech production

To elucidate the feedforward and feedback contribution of different cortical regions to speech production, we examined the relative contribution of each electrode to decoding speech in our models. We derived the relative contribution by quantifying how the input signal at a particular electrode affects the overall accuracy (measured by the CC) of the reconstructed speech in the causal and anticausal models, respectively. In both the causal and anticausal models, peri-sylvian electrodes were important for speech decoding; however, there was a surprising recruitment of frontal regions when decoding speech based on the feedback (anticausal model, Figure 6.1b) as well as recruitment of temporal sites when decoding speech based on the feedforward signals (causal model, Figure 6.1c). We only show significant contributions that are above a threshold derived from the shuffled model (depicted in Figure 6.1d). In order to quantify the prevalence of feedforward or feedback processing, we directly contrasted the two and projected the results onto the cortex (Figure 6.1e). To ascertain regions that contribute significantly more to feedback or feedforward

processing, we conducted a region of interest analysis, based on within-subject anatomical labels of each electrode, testing for an increase in causal or anticausal contributions across trials (non-parametric paired Wilcoxon test; Figure 6.1f). We found a surprisingly mixed distribution of causal and anticausal contributions within both temporal and frontal cortices. A majority of temporal cortex were predominantly anticausal, including caudal superior temporal gyrus (STG; Wilcoxon sign rank, $P=1.607E-15$, $Z=9.6234$) and portions of middle temporal gyrus (MTG; rostral MTG: Wilcoxon sign rank test $P=2.5108E-04$, $Z=4.9359$, and middle MTG: Wilcoxon sign rank test $P=1.5257E-13$, $Z=9.0185$) as well as supramarginal cortex (Wilcoxon sign rank test $P=1.1144E-04$, $Z=5.3919$), implicating it in processing the auditory feedback signals for speech production. However, there was also a significant causal contribution in rostral STG (Wilcoxon sign rank test $P=0.0332$, $Z=-2.9628$). Similarly, the majority of sensorimotor cortex was predominantly casual, implicating it in processing the motor speech commands including ventral precentral (Wilcoxon sign rank, $P=4.9511E-08$, $Z=-7.1409$) and postcentral gyri (Wilcoxon sign rank, $P=6.419E-04$, $Z=-4.9612$). However, the dorsal division of precentral gyrus was equally causal and anticausal (Wilcoxon sign rank, $P=0.4349$, $Z=0.6525$), implicating it in processing both feedforward and feedback information equally. Within the inferior frontal cortex, we found a striking division of function wherein pars opercularis was significantly causal (Wilcoxon sign rank test, $P=8.0693E-15$, $Z=-9.6185$) while pars triangularis was significantly anticausal (Wilcoxon sign rank test, $P=2.6715E-06$, $Z=6.3518$). Overall, these findings provide evidence for a mixed feedforward and feedback processing of speech commands and their refference across temporal and frontal cortices, in contrast to a dichotomous view.

6.3.2 Temporal dynamics and receptive fields of speech production

Speech production includes articulatory planning and executing the motor commands, processes that recruit distinct regions of frontal cortex [26]. However, their exact temporal receptive fields remain poorly understood. Earlier, we examined the causal and anticausal cortical contributions during speech articulation. Next, we examine articulatory planning and articulation of speech production stages and derive the related temporal receptive fields across the cortex. We leverage the receptive fields to test how cortical regions contribute

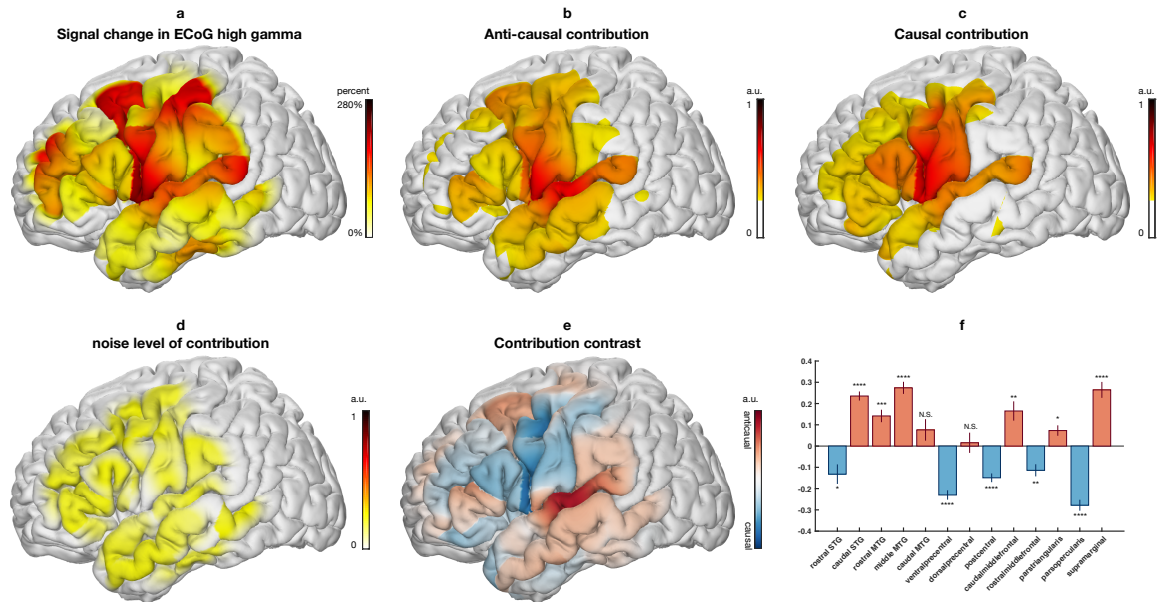


Figure 6.1: (a) averaged signal of input ECoG projected on the standardized MNI anatomical map. The colors reflect the percentage change of high gamma compared to the baseline level during the pre-stimulus baseline period. (b) shows the anticausal contribution of different cortical locations (red indicates higher contribution), while (c) illustrates the causal contribution. (d) noise level of the contribution analysis evaluated by the contributions from the shuffled model. Contributions below noise level are not shown in (b) and (c). (e) the contrast obtained by taking the difference of the anticausal and causal contribution maps (red means higher anticausal contribution, while blue means higher causal contribution). The boxplots (f) show the average difference in each cortical region (*: P-value<0.05, **: P-value<0.01, ***: P-value<0.001, ****: P-value<0.0001).

differently to speech decoding with time and how frontal cortex dynamics change when feedback is introduced (after articulation starts). Both feedforward and feedback information is processed in tandem.

We employed a similar occlusion approach to derive the temporal receptive fields as in the previous section. However, we quantified how the input signal at a particular electrode affects the accuracy of the reconstructed speech across varying delays. This approach allowed us to quantify the contribution of a specific electrode in the model as a function of delay relative to speech decoding, similarly to classical temporal receptive fields (i.e., TRF). We conducted this analysis for both causal and anticausal models during two epochs – one prior to production (-512ms \sim 0 ms; Figure 6.2a) and the other during production, which included both causal and anticausal components (0ms \sim 512ms; Figure 6.2b, c). The projection of all the temporal receptive fields onto the cortex, which were significantly above a threshold derived from the shuffled model, are plotted in Figure 6.2 as a function of delay. We found an increased frontal and MTG contribution prior to production (Figure 6.2a) compared with during production (Figure 6.2b). These processes are likely related to articulatory planning and lexical retrieval prior to speech production. During production, there was a prominent sharpening of ventral precentral gyrus receptive fields marked by a significant increase in contribution compared with pre-production (Wilcoxon sign rank test, $P=8.3979E-05$, $Z=5.4203$). While a majority of prefrontal regions engaged prior to production, there was a significant decrease in contribution across pars triangularis (Wilcoxon sign rank test, $P=1.8493E-32$, $Z=-13.6074$), middle frontal gyri (MFG; Wilcoxon sign rank test, $P=3.9177E-09$, $Z=-7.6103$ for caudal and $P=4.1581E-04$, $Z=-4.8311$ for rostral) except for pars opercularis (Wilcoxon sign rank test, $P=0.4819$, $Z=0.2066$). Similarly, to our previous results (Figure 6.1e,f), during production, we found a significant increase in anticausal contribution for caudal STG (Wilcoxon sign rank test, $P=2.6789E-17$, $Z=9.6711$), pars triangularis (Wilcoxon sign rank test, $P=0.0162$, $Z=3.9003$) and caudal MFG (Wilcoxon sign rank test, $P=0.0045$, $Z=3.9862$) compared with causal contributions. This confirms the anatomical-functional division of the inferior and middle frontal gyri as well as caudal (Wilcoxon sign rank test, $P = 2.6789E-17$, $Z = 9.6711$) and rostral separation of STG (Wilcoxon sign rank test, $p= 0.0343$, $Z= -2.9457$).

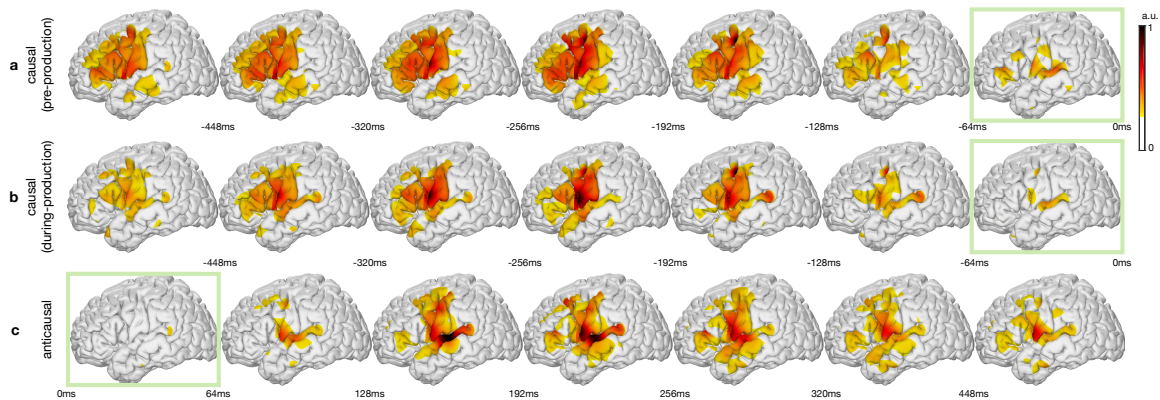


Figure 6.2: Spatial-temporal receptive fields based on decoding contribution. The contribution to decoding the current speech from cortical neural responses with certain temporal delays. (a) and (b) are the feedforward spatial-temporal receptive fields derived from the causal model by evaluating the contribution of past (negative delays) neural signals during a period before production onset (a) and after onset (b). (c) represents the feedback spatial-temporal receptive fields derived from the anticausal models that evaluate the contribution of future (positive delays) neural signals during feedback after articulation. Contributions below significance ($p < 0.05$) representing the noise level are clipped and not shown in the plots.

Next, we conducted a region of interest analysis, based on within-subject anatomical labels of each electrode, in order to derive the temporal receptive curves per region (Figure 6.3). This approach provides critical insight as to the temporal tuning and peak recruitment of various regions to feedforward processing prior to (Figure 6.3a) and during production (Figure 6.3b) as well as feedback processing (Figure 6.3c). We found a shift in receptive field tuning for the two subdivisions of precentral gyrus. Prior to production, dorsal and ventral precentral gyri were not significantly different from each other (Wilcoxon sign rank test, $P=0.454$, $Z=-0.36103$), and had close peak times (-196ms , -192ms prior to speech for ventral and dorsal precentral gyri, respectively). However, during production, these dynamics shifted and we found a significant decrease in dorsal precentral causal contribution (Wilcoxon sign rank test, $P=4.7575\text{E-}05$, $Z=-5.6272$) accompanied by a temporal separation of peaks (-208ms , -184ms for ventral and dorsal precentral gyri, respectively; Figure 6.3a,b). Within the inferior frontal gyrus, we found pars opercularis was recruited similarly both prior to production and during production for feedforward processing (Wilcoxon sign rank test, $P=0.5922$, $Z=1.7462$) at a peak delay of -248ms and -280ms , respectively. During production, pars triangularis had a selective increase in recruitment for anticausal

compared with causal contributions (Wilcoxon sign rank test, $P=0.0162$, $Z=3.9003$), implicating it in increased feedback processing (Figure 6.2c). The anticausal receptive fields during production provide evidence for feedback processing most strongly contributed by caudal STG, with the earliest peak in contributions seen in dorsal precentral gyrus (144 ms) and caudal STG (168 ms) followed by parietal (supramarginal 184ms, postcentral 192ms) and ventral precentral (280 ms) gyri. These findings suggest a preferential recruitment of prefrontal cortices in feedforward processing prior to production followed by a shift in dynamics during production when feedforward and feedback signals are jointly processed with anatomical divisions of labor.

6.3.3 Contribution analysis for speech parameters

To evaluate the contribution of brain activities to decoded speech with a finer scope, we investigate the contribution map for the speech parameters that are important to decoding accurate speeches. We report contribution maps for loudness, f_0 , f_1 , and f_2 . We observe a similar trend as in the overall spectrogram contribution analysis, where most perisylvian cortices process both feedback (Figure 6.4a) and feedforward (Figure 6.4b) signals for most reported components.

6.4 Comparing feedback cortical contribution between speech perception and production

To evaluate the function of feedback contribution, we compare contribution of feedback production and contribution during perception periods. From Figure 6.5 (a) and (b), we observe that in both perception and production periods, the contribution map involve frontal, central and temporal gyri. The difference between perception and production Figure 6.5 (c) shows that most covered cortices, especially STG are recruited more during perception than production. On the other hand, although speech-motor cortex is recruited for processing feedback signal during both listening and speaking, it contributes more during speech production.

We also investigate the influence of varying tasks on the contribution during perception

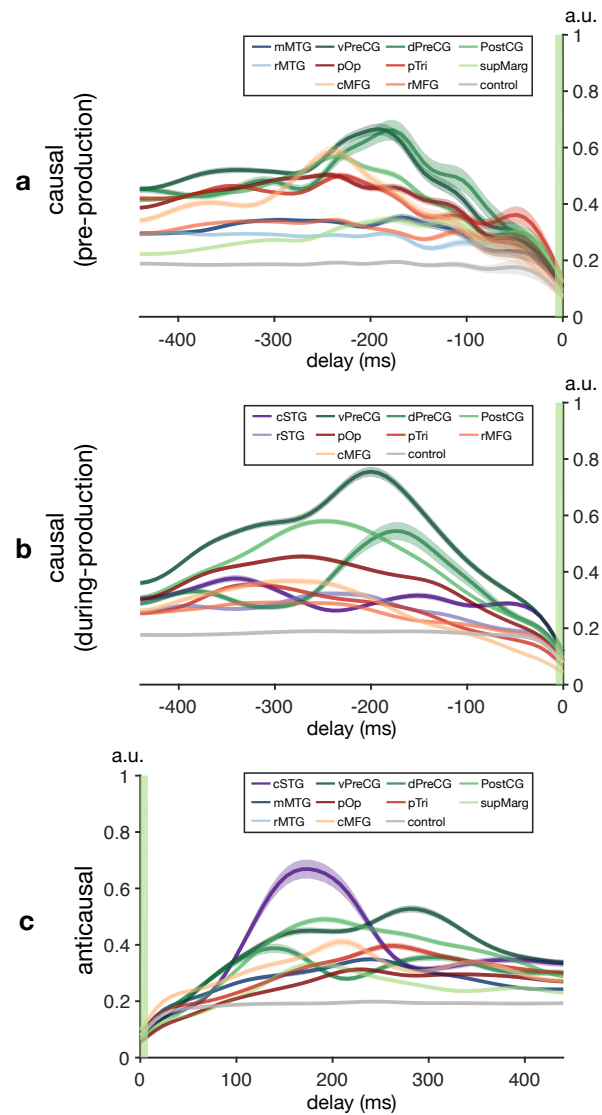


Figure 6.3: The temporal receptive field across anatomical regions. The contribution to decoding the current speech from cortical neural responses with certain temporal delays. (a) and (b) are the feedforward temporal receptive fields derived from the causal model by evaluating the contribution of past (negative delays) neural signals during a period before production onset (a) and after onset (b). (c) represents the feedback temporal receptive fields derived from the anticausal models that evaluate the contribution of future (positive delays) neural signals during feedback after articulation. The temporal propagation of the shuffled model estimates the noise level dynamics (grey curves in plots). Only regions significantly above noise level (Wilcoxon sign rank test on across-time averaged data, $P < 0.05$) are reported.

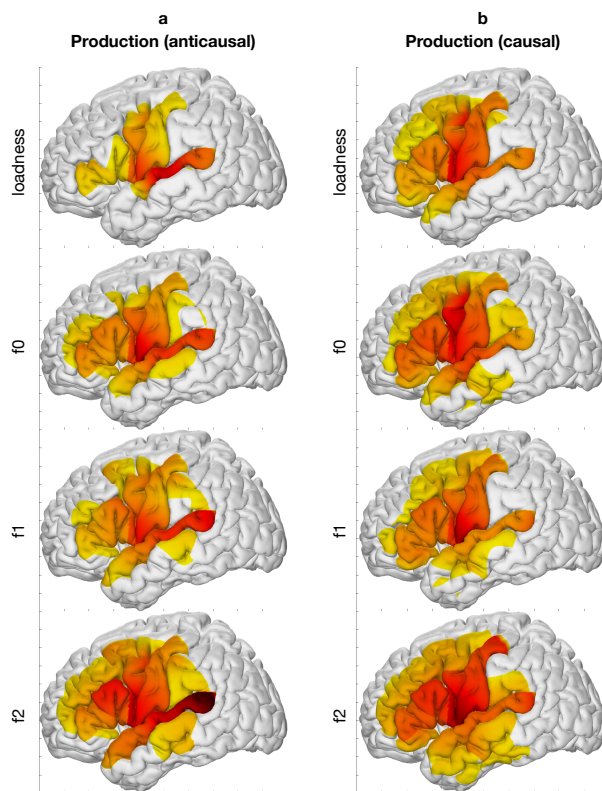


Figure 6.4: Contribution maps for speech components of (a) anticausal model and (b) causal model.

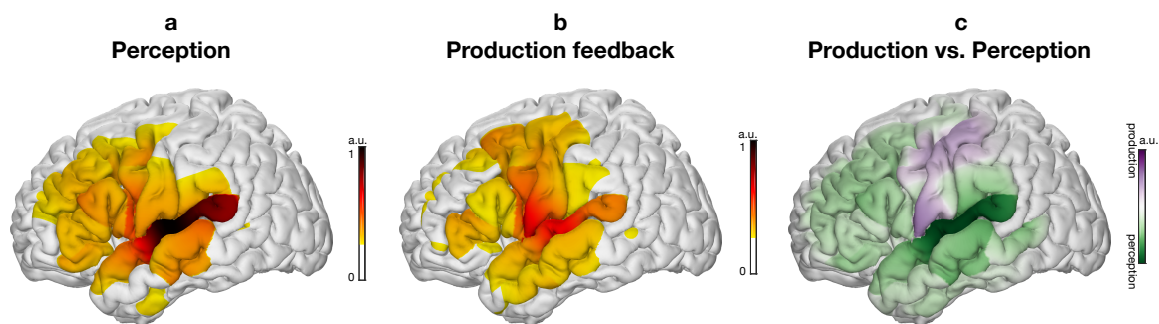


Figure 6.5: Comparison of feedback cortical contribution between perception and production periods. contribution during (a) perception and (b) production feedback process. (c) contribution contrast between (a) and (b).

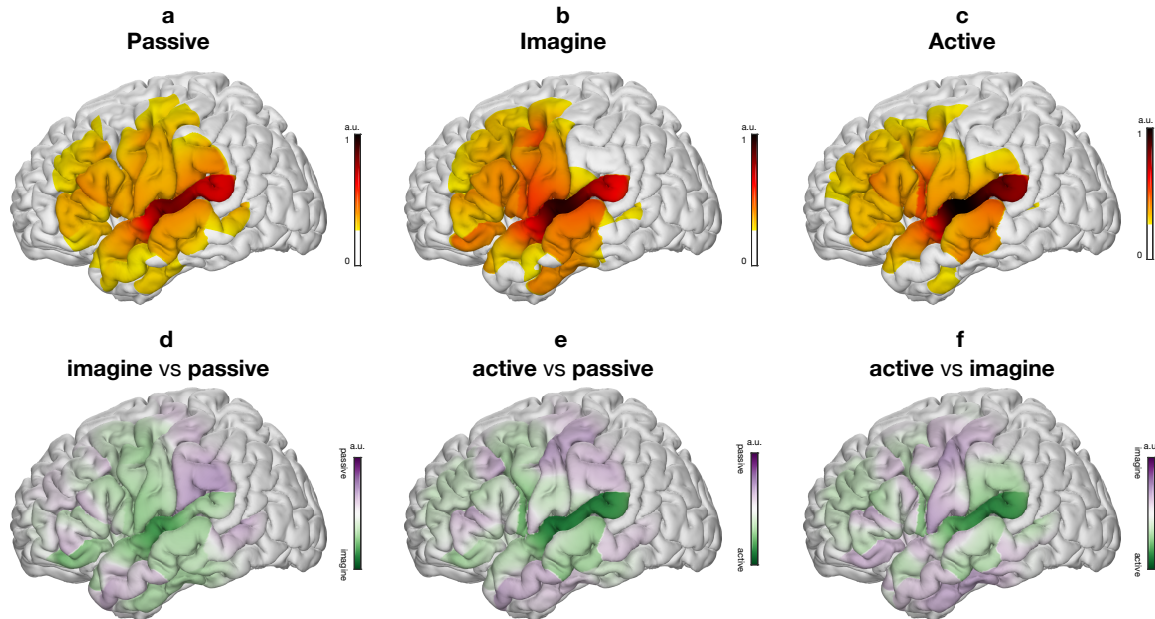


Figure 6.6: Comparison of cortical contribution between different speech tasks during perception period. Contribution map for (a) passive listening task, (b) imaginary speaking task, and (c) active speaking task. The contrast contribution map (d) between imaginary speaking and passive listening tasks, (e) between active speaking and passive listening task, and (f) between active speaking and imaginary listening tasks.

period. We designed three different tasks to complement auditory repetition. They are passive listening (the patients are not required to respond after listening to the speech stimuli), imaginary speaking (the patients are asked to repeat the listened word imaginarily without actual articulation), and active speaking (the patients are asked to repeat the listened word with actual articulation) tasks. We observe that, even though patients were listening to the same stimuli, STG and speech-motor cortex recruitment are different depending on the tasks (Figure 6.6 (a)-(c)). Both active and imaginary tasks recruit STG and speech-motor cortex more than passive task (Figure 6.6 (d) and (e)), and the active task has more STG contribution than imaginary task (Figure 6.6 (f)).

6.5 Discussion

Our study leverages a novel deep learning approach together with neurosurgical recordings and, to our knowledge, is the first to dissociate direct feedforward and feedback cortical contributions during speech production. Our neural network architecture achieves

state-of-the-art decoding of speech production, by tapping an interpretable compact speech representation and can be altered to focus on causal, anticausal and non-causal decoding. Our analyses of the cortical contributions driving the performance of these models reveal a mixed distribution of feedforward and feedback processing during speech production. This was prominent in inferior, middle frontal, and superior temporal gyri which exhibited an anatomical division between feedforward and feedback processing. Lastly, we show a change in the temporal dynamics of frontal recruitment during speech planning through production, characterized by a shift of inferior frontal and precentral gyri recruitment, processing both feedforward and feedback information at different time points and spatial locations.

During speech production, we process feedforward and feedback signals in tandem. It was previously impossible to disentangle the two. Attempts have focused on experimental manipulations which change the feedback by shifting frequency [10] or time [60]. However, these manipulations change the cortical dynamics and introduce other cognitive processes due to hearing one’s own voice altered as well as induced motor compensation. We applied convolution filters with different causality to directly train models to disentangle feedforward (i.e., causal models) and feedback (i.e., anticausal models) contributions of cortical regions. Feedforward and feedback processes are critical for driving articulatory vocal tract movement. The feedforward pathway generates an initial articulatory command and predicts sensory (auditory and somatosensory) targets; the feedback pathway compares the targets with the perceived sensory feedback and updates subsequent feedforward commands. The exact mapping between anatomical regions and their contribution to specific functional roles differ across speech motor control models ([35], [44]). Further, these findings have been developed based mostly on non-invasive studies which have low temporal (e.g., fMRI) or spatial resolution (e.g., M/EEG). Our high spatio-temporal resolution ECoG data together with advanced deep neural networks provides a fine-grained mapping of the cortical feedforward and feedback speech networks.

Consistent with the predominant speech motor control models, our results showed a dominant feedforward process in the ventral motor and pars opercularis of the inferior frontal gyrus, while posterior superior temporal and supramarginal gyri in the parietal lobe showed feedback. However, in contrast to these models, we found that cortices in the frontal lobe,

including pars triangularis and caudal middle frontal, are predominantly feedback in nature, while rostral STG appears feedforward. This feedback processing across frontal cortices became even stronger when we limited our analyses to the speech production epoch (Figure 6.2c). Additionally, most gyri (inferior frontal, caudal middle frontal, superior temporal, precentral, and postcentral cortices) had both feedforward and feedback contributions above the noise level derived from the shuffled model, suggesting the feedforward and feedback processing can mix in these regions.

Our results highlight the anticausal feedback signature exhibited by sensorimotor and frontal cortices. While this goes against the canonical model of the frontal cortex in an action-perception loop [28], our findings complement a growing body of evidence showing specific responses in the frontal cortex to auditory stimuli during perception. Cheung et al. [13] found distinct auditory receptive fields as well as robust passive listening responses in ventral precentral gyrus. Similarly, the dorsal division of precentral gyrus has recently been implicated in processing auditory feedback of altered speech as well as responding robustly during passive listening [60]. However, this begs the question as to why the speech motor cortex is processing auditory information. Our feedback contribution analysis suggests that the auditory processing is specifically leveraged for anticausal processing of the reafferent signals during production. Indeed, our results show that dorsal precentral gyrus decreases feedforward processing while engaged in actual speech production (Figure 6.3b) and is recruited for feedback at an early time point together with temporal cortices (Figure 6.3c). Under this view, the auditory frontal responses seen during passive listening may constitute a representation dedicated to feedback processing when speech is produced.

To summarize, we provided a new approach to decode speech production and interrogate the recalcitrant problem of mixed feedforward and feedback processing during speech production. We were able to leverage feedforward processing only in causal models to drive neural speech prosthetics (as opposed to the literature using non-causal processing [2–5, 55]) as well as provide insights into the underpinning cortical drivers. Our results suggest a mixed cortical architecture in frontal and temporal cortices that dynamically shifts and processes both feedforward and feedback signals across the cortex in contrast to previous views associating feedforward or feedback processing of speech with primarily anterior and posterior

cortices, respectively.

Chapter 7

Attention Mask Reflects Speech Cortical Network Dynamics during Perception-Production Task

7.1 Attention Mask Visualization

In Chapter 3, we introduced a convolutional residual ECoG decoder structure 3.4. Inside, we designed a bypass convolution layer that generates a dynamic mask from the backbone features and applied back to attenuate backbone features. These attention masks help 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. Ideally, as the brain state changes over time, the attention mask also alters its attenuation pattern to emphasize more important electrodes per time step. 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. 7.1 shows the averaged evolution of the attention mask for all the test samples.

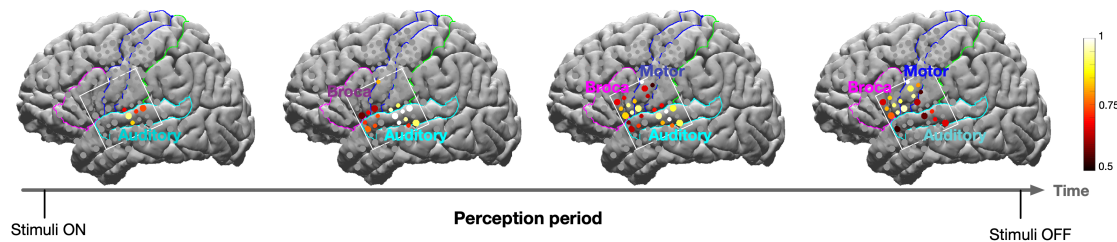


Figure 7.1: The averaged evolution of the attention mask during active listening task. The color in each electrode indicates the value of the attention mask, following the color bar. The white square shows the 8×8 grid used in the experiment. Similar dynamic is also observed in the other HB subject.

7.2 Attention Mask Reflect Perception-Production Cortical Dynamic

By observing the attention mask as in Fig. 7.1, 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 [26]. Moreover, motor cortex is active during the stimuli perception and prior to speech production. Similar phenomenon has been observed during active listening task in the literature [13]. 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 [13, 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.

Chapter 8

Impulse Response of Deep Speech Stimuli Decoding Network Reflects Phoneme Selectivity of STG

8.1 Impulse Response of a Deep Neural Network

Previous studies on pSTG cortex have shown selectivity to phonetic features for vowels and consonants [56, 61]. We were interested in evaluating whether our decoding approach also revealed similar phoneme selectivity for certain electrodes. We achieved this by investigating the “impulse response” of each electrode input for the stimuli decoding model we build in Chapter 2 . The impulse response for a given electrode can be considered as the stimuli speech (with its corresponding spectrogram with a certain temporal and spectral span) that causes a short-term response in the cortex area observed by this electrode. We generated an impulse signal for each of the 64 electrode channels. The impulse is generated between 500ms-600ms among the overall 1000ms duration. The impulse value is set to be the highest signal value of ECoG in our dataset. The rest of the temporal samples and electrodes are set to be zero. To generate the impulse response for a fitted model for each electrode, we drive the fitted model with the impulse input and obtain the decoded spectrogram for each electrode.

Electrode No.	7	8	50	54	55
Perceived phoneme	/u:	/ss	/ əu	/sh	/i

Table 8.1: Discovered phonetic features for ECoG electrodes on subject S1

Electrode No.	38	39	40	47	48
Perceived phoneme	/sh	/i	/ i:	/əu	/u:

Table 8.2: Discovered phonetic features for ECoG electrodes on subject S2

8.2 Phonetic Feature of Model Impulse Response

Figure 8.1 shows the impulse response for all 64 channels by training the model with all samples in the dataset. Since there exists an offset and orientation mismatch between ECoG arrays for different subjects, spatial distribution difference of impulse response for two models is expected. But the discovered phonetic features for both patients are robust and mostly consistent. We generated the “speech” signal from these impulse responses and were able to identify some of the phonemes, which are listed in Tables 8.1 and 8.2. We have found that the impulse responses for both subjects include the common set of /sh/, /i/ , /u:/ and /u/.

Previous studies investigating the pSTG area have revealed that certain electrode responses are sensitive to clusters of phonemes [6]. Here we further specified which phonemes are selected by certain electrodes. It is possible to uncover more electrodes with phonemes sensitivity given a bigger dataset of richer speech stimuli.

The consistency between WaveNet phonetic impulse response provides evidence for the phonetic selectivity of STG and suggests a promising direction for further study of the STG area with the discovered phonetic impulse responses.

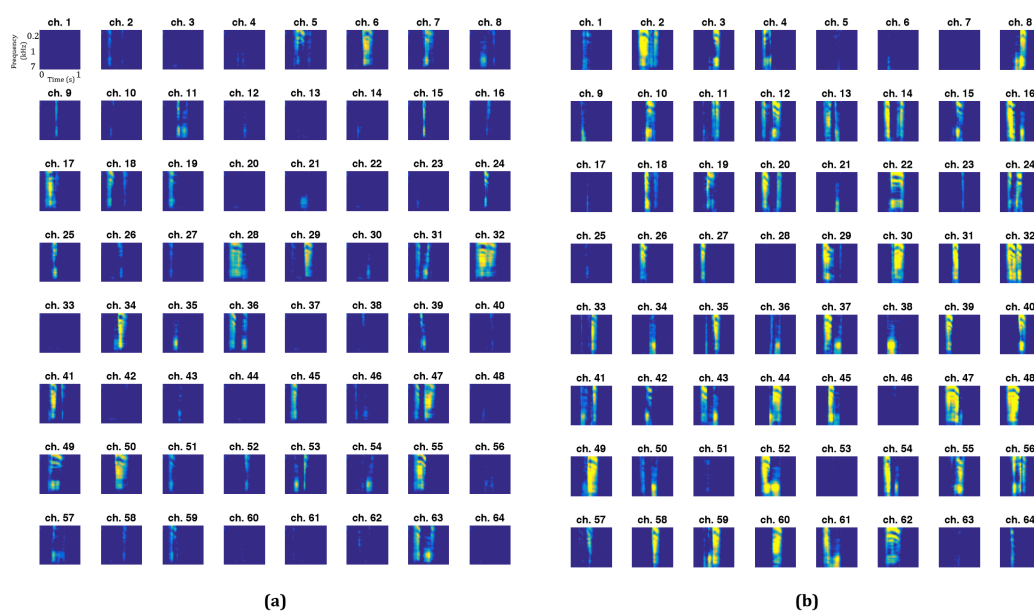


Figure 8.1: Impulse response for subject S1 (a) and S2 (b). Each subfigure corresponds to one ECoG electrode site..

Chapter 9

Summary and Future Work

9.1 Summary

In this study, we explore three approaches for decoding speech from ECoG brain activities as well as interpret the decoding networks.

In our first approach, we adapt WaveNet to decode speech from ECoG signals while listening to word stimuli. Despite a relatively small dataset, the high parameter efficiency of the adapted WaveNet is able to overcome the overfitting problem and generate reconstructed spectrograms with intelligible quality. The impulse response analysis of the fitted WaveNet model further reveals distinct phonetic encoding by some cortex areas on STG. The phonetic features discovered by these impulse responses are consistent with previous discoveries of STG area and suggest a promising direction for further analysis of the area and other auditory cortices.

For the second approach, we proposed a new framework containing a decoder to extract features from and map the ECoG signals to a representation space. A generator then converts the features to a spectrogram. 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 accurate reconstruction from cortical areas including STG. Additionally, the temporal attention mechanism introduced in the decoder allows for better generalization of the network and interpretation of the results for neuroscientific discoveries. The visualization of attention mask reflects the dynamics

among superior temporal, inferior frontal, and precentral gyri during speech perception tasks. We also proposed attentional based model which separately operates temporal and space dimensions with convolution and attention respectively. We augment the attention layer by MNI coordinate features to achieve more expressive power. The preliminary results shows similar performance of the proposed attentional model and convolutional model on partial-grid data.

Towards decoding produced speech, in the third approach, we proposed a novel differentiable speech synthesizer which generates speech spectrograms from a small set of interpretable compact acoustic speech parameters. Employing the proposed speech synthesizer, we construct a ECoG decoder that maps the ECoG signals to speech parameters with high data efficiency. The ECoG decoder can use temporal convolutions with different causalities. Analysis on the anticausal vs. causal decoders lead us towards the discovery of a surprisingly mixed feedback-feedforward processing over frontal and posterior lobe of the brain for speech production.

9.2 Future Work

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 decoders and response speech decoders, but also semantic word decoders. Our developed framework might also be useful for other medical applications where limitations in training data hinder the use of deep learning.

We are also interested in the effect of epilepsy on speech perception and production by studying the contrast contribution between patients with epileptic STG activity in and without.

Bibliography

- [1] Jonas Adler and Sebastian Lunz. Banach Wasserstein GAN. In *Advances in Neural Information Processing Systems*, pages 6754–6763, 2018.
- [2] Hassan Akbari, Bahar Khalighinejad, Jose L Herrero, Ashesh D Mehta, and Nima Mesgarani. Towards reconstructing intelligible speech from the human auditory cortex. *Scientific reports*, 9(1):874, 2019.
- [3] Miguel Angrick, Christian Herff, Garrett Johnson, Jerry Shih, Dean Krusienski, and Tanja Schultz. Interpretation of convolutional neural networks for speech spectrogram regression from intracranial recordings. *Neurocomputing*, 342:145–151, 2019.
- [4] Miguel Angrick, Christian Herff, Emily Mugler, Matthew C Tate, Marc W Slutzky, Dean J Krusienski, and Tanja Schultz. Speech synthesis from ecog using densely connected 3d convolutional neural networks. *Journal of neural engineering*, 16(3):036019, 2019.
- [5] Gopala K Anumanchipalli, Josh Chartier, and Edward F Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019.
- [6] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 214–223, 2017.
- [7] Luc H Arnal, Adeen Flinker, Andreas Kleinschmidt, Anne-Lise Giraud, and David Poeppel. Human screams occupy a privileged niche in the communication soundscape. *Current Biology*, 25(15):2051–2056, 2015.
- [8] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. *arXiv preprint arXiv:1904.09925*, 2019.
- [9] Kristofer E Bouchard, Nima Mesgarani, Keith Johnson, and Edward F Chang. Functional organization of human sensorimotor cortex for speech articulation. *Nature*, 495(7441):327–332, 2013.
- [10] Edward F Chang, Caroline A Niziolek, Robert T Knight, Srikantan S Nagarajan, and John F Houde. Human cortical sensorimotor network underlying feedback control of vocal pitch. *Proceedings of the National Academy of Sciences*, 110(7):2653–2658, 2013.
- [11] Edward F Chang, Kunal P Raygor, and Mitchel S Berger. Contemporary model of language organization: an overview for neurosurgeons. *Journal of neurosurgery*, 122(2):250–261, 2015.

- [12] Josh Chartier, Gopala K Anumanchipalli, Keith Johnson, and Edward F Chang. Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. *Neuron*, 98(5):1042–1054, 2018.
- [13] Connie Cheung, Liberty S Hamilton, Keith Johnson, and Edward F Chang. The auditory representation of speech sounds in human motor cortex. *Elife*, 5:e12577, 2016.
- [14] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.
- [15] Trinity B Crapse and Marc A Sommer. Corollary discharge across the animal kingdom. *Nature Reviews Neuroscience*, 9(8):587–600, 2008.
- [16] Li Deng and Douglas O’Shaughnessy. *Speech processing: a dynamic and optimization-oriented approach*. CRC Press, 2018.
- [17] Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*, 2018.
- [18] John Duncan, Rüdiger J Seitz, Jonathan Kolodny, Daniel Bor, Hans Herzog, Ayesha Ahmed, Fiona N Newell, and Hazel Emslie. A neural basis for general intelligence. *Science*, 289(5478):457–460, 2000.
- [19] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. DDSP: Differentiable digital signal processing. *arXiv preprint arXiv:2001.04643*, 2020.
- [20] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1068–1077. JMLR. org, 2017.
- [21] James L Flanagan. *Speech analysis synthesis and perception*, volume 3. Springer Science & Business Media, 2013.
- [22] Mario Fleischer, Silke Pinkert, Willy Mattheus, Alexander Mainka, and Dirk Mürbe. Formant frequencies and bandwidths of the vocal tract transfer function are affected by the mechanical impedance of the vocal tract wall. *Biomechanics and modeling in mechanobiology*, 14(4):719–733, 2015.
- [23] A Flinker, EF Chang, NM Barbaro, MS Berger, and RT Knight. Sub-centimeter language organization in the human temporal lobe. *Brain and Language*, 117(3):103–109, 2011.
- [24] Adeen Flinker, Edward F Chang, Heidi E Kirsch, Nicholas M Barbaro, Nathan E Crone, and Robert T Knight. Single-trial speech suppression of auditory cortex activity in humans. *Journal of Neuroscience*, 30(49):16643–16650, 2010.
- [25] 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*, 3(4):393, 2019.

- [26] 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*, 112(9):2871–2875, 2015.
- [27] Joaquin M Fuster. The prefrontal cortex—an update: time is of the essence. *Neuron*, 30(2):319–333, 2001.
- [28] Joaquin M Fuster. Upper processing stages of the perception–action cycle. *Trends in cognitive sciences*, 8(4):143–145, 2004.
- [29] Joaquín M Fuster. The prefrontal cortex in the neurology clinic. *Handbook of clinical neurology*, 163:3–15, 2019.
- [30] Hongyang Gao and Shuiwang Ji. Graph u-nets. *arXiv preprint arXiv:1905.05178*, 2019.
- [31] Jeremy DW Greenlee, Roozbeh Behroozmand, Charles R Larson, Adam W Jackson, Fangxiang Chen, Daniel R Hansen, Hiroyuki Oya, Hiroto Kawasaki, and Matthew A Howard III. Sensory-motor interactions for vocal pitch monitoring in non-primary human auditory cortex. *PloS one*, 8(4):e60783, 2013.
- [32] Jeremy DW Greenlee, Adam W Jackson, Fangxiang Chen, Charles R Larson, Hiroyuki Oya, Hiroto Kawasaki, Haiming Chen, and Matthew A Howard III. Human auditory cortical activation during self-vocalization. *PloS one*, 6(3):e14744, 2011.
- [33] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984.
- [34] Frank H Guenther. A neural network model of speech acquisition and motor equivalent speech production. *Biological cybernetics*, 72(1):43–53, 1994.
- [35] Frank H Guenther. *Neural control of speech*. Mit Press, 2016.
- [36] John W Hawks and James D Miller. A formant bandwidth estimation procedure for vowel synthesis [43.72. ja]. *the Journal of the Acoustical Society of America*, 97(2):1343–1344, 1995.
- [37] 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*, pages 770–778, 2016.
- [38] Christian Herff, Lorenz Diener, Miguel Angrick, Emily Mugler, Matthew C Tate, Matthew A Goldrick, Dean J Krusienski, Marc W Slutzky, and Tanja Schultz. Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices. *Frontiers in neuroscience*, 13:1267, 2019.
- [39] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135. IEEE, 2017.
- [40] Gregory Hickok. Computational neuroanatomy of speech production. *Nature reviews neuroscience*, 13(2):135–145, 2012.

- [41] Gregory Hickok. The cortical organization of speech processing: Feedback control and predictive coding the context of a dual-stream model. *Journal of Communication Disorders*, 45(6):393–402, 2012. 21st Annual NIDCD-Sponsored ASHA Research Symposium (2011):Neuroplasticity in the Mature Brain.
- [42] Gregory Hickok. The architecture of speech production and the role of the phoneme in speech processing. *Language, Cognition and Neuroscience*, 29(1):2–20, 2014.
- [43] Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393, 2007.
- [44] John F Houde and Srikantan S Nagarajan. Speech production as state feedback control. *Frontiers in human neuroscience*, 5:82, 2011.
- [45] Colin Humphries, Merav Sabri, Kimberly Lewis, and Einat Liebenthal. Hierarchical organization of speech perception in human auditory cortex. *Frontiers in neuroscience*, 8:406, 2014.
- [46] Eric J Hunter, Jan G Švec, and Ingo R Titze. Comparison of the produced and perceived voice range profiles in untrained and trained classical singers. *Journal of Voice*, 20(4):513–526, 2006.
- [47] Keith Ito. The LJ speech dataset. <https://keithito.com/LJ-Speech-Dataset/>, 2017.
- [48] Aren Jansen, Manoj Plakal, Ratheet Pandya, Daniel PW Ellis, Shawn Hershey, Jiayang Liu, Channing Moore, and Rif A Saurous. Towards learning semantic audio representations from unlabeled data. *signal*, 2(3):7–11, 2017.
- [49] Jintao Jiang, Marcia Chen, and Abeer Alwan. On the perception of voicing in syllable-initial plosives in noise. *The Journal of the Acoustical Society of America*, 119(2):1092–1105, 2006.
- [50] Eric R Kandel, James H Schwartz, Thomas M Jessell, Steven Siegelbaum, A James Hudspeth, and Sarah Mack. *Principles of neural science*, volume 4. McGraw-hill New York, 2000.
- [51] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [52] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [53] John Kominek, Tanja Schultz, and Alan W Black. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In *Spoken Languages Technologies for Under-Resourced Languages*, 2008.
- [54] Sergey Korolev, Amir Safiullin, Mikhail Belyaev, and Yulia Dodonova. Residual and plain convolutional neural networks for 3d brain mri classification. In *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*, pages 835–838. IEEE, 2017.

- [55] Joseph G Makin, David A Moses, and Edward F Chang. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature Neuroscience*, 23(4):575–582, 2020.
- [56] Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F Chang. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010, 2014.
- [57] Brian T Miller and Mark D’Esposito. Searching for “the top” in top-down control. *Neuron*, 48(4):535–538, 2005.
- [58] David A Moses, Sean L Metzger, Jessie R Liu, Gopala K Anumanchipalli, Joseph G Makin, Pengfei F Sun, Josh Chartier, Maximilian E Dougherty, Patricia M Liu, Gary M Abrams, et al. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227, 2021.
- [59] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.
- [60] Muge Ozker, Werner Doyle, Orrin Devinsky, and Adeen Flinker. Cortical network underlying speech production during delayed auditory feedback. *bioRxiv*, 2021.
- [61] 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*, 10(1):e1001251, 2012.
- [62] Stavros Petridis, Themis Stafylakis, Pinghuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic. End-to-end audiovisual speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6548–6552. IEEE, 2018.
- [63] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, Dinei Florêncio, and Mark Hasegawa-Johnson. Speech enhancement using bayesian wavenet. In *Interspeech*, pages 2013–2017, 2017.
- [64] Jinfu Ren, Yang Liu, and Jiming Liu. EWGAN: Entropy-based Wasserstein GAN for imbalanced learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 10011–10012, 2019.
- [65] 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*, pages 2234–2242, 2016.
- [66] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [67] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

- [68] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018.
- [69] SHTOOKA. The shtooka project. <http://shtooka.net/>, 2019.
- [70] Kristina Simonyan, Hermann Ackermann, Edward F Chang, and Jeremy D Greenlee. New developments in understanding the complexity of human speech production. *Journal of Neuroscience*, 36(45):11440–11448, 2016.
- [71] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.
- [72] Donald T Stuss and Robert T Knight. *Principles of frontal lobe function*. Oxford University Press, 2013.
- [73] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*, pages 4214–4217. IEEE, 2010.
- [74] Akira Tamamori, Tomoki Hayashi, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda. Speaker-dependent WaveNet vocoder. In *Interspeech*, pages 1118–1122, 2017.
- [75] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002.
- [76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [77] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
- [78] Ran Wang, Xupeng Chen, Amirhossein Khalilian-Gourtani, Zhaoxi Chen, Leyao Yu, Adeen Flinker, and Yao Wang. Stimulus speech decoding from human cortex with generative adversarial network transfer learning. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 390–394. IEEE, 2020.
- [79] 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)*, pages 1–6. IEEE, 2018.
- [80] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.
- [81] 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*, 6:60478–60488, 2018.

Publication List

Ran Wang, Xupeng Cheng, Amirhossein Khalilian-Gourtani, Leyao Yu, Patricia Dugan, Daniel Friedman, Werner Doyle, Orrin Devinsky, Yao Wang, Adeen Flinker, “*Distributed feedforward and feedback processing across perisylvian cortex supports human speech*”. (Submitted to Nature Communications)

Ran Wang, Xupeng Chen, Amirhossein Khalilian-Gourtani, Zhaoxi Chen, Leyao Yu, Adeen Flinker, Yao Wang, “*Stimulus Speech Decoding from Human Cortex with Generative Adversarial Network Transfer Learning*”. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI 2020, Best Paper Award Finalist)

Ran Wang, Yao Wang, Adeen Flinker. “*Reconstructing Speech Stimuli From Human Auditory Cortex Activity Using a WaveNet-like network*”. 2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB 2018), IEEE, 2018

Ran Wang, Song Yilin, Wang Yao. “*Long-term prediction of ECOG signals with a spatio-temporal pyramid of adversarial convolutional networks*”. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, 2018

Zhiqi Chen, **Ran Wang**, Haojie Liu, Yao Wang, “*PDWN: Pyramid Deformable Warping Network for Video Interpolation*”. Open Journal of Signal Processing.

Shenghe Xu, Pei Liu, **Ran Wang**, Shivendra S. Panwar. “*Realtime Scheduling and Power Allocation Using Deep Neural Networks*”. IEEE Wireless Communications and Networking Conference (WCNC 2019)