Stimulus Speech Decoding from Human Cortex with Generative Adversarial Network Transfer Learning

R. Wang, X. Chen, A. Khalilian, Z. Chen, L. Yu, A. Flinker, Y. Wang

New York University, Tandon School of Engineering
New York University, Langone School of Medicine

ISBI 2020

IEEE ISBI 2020
International Symposium on
Biomedical Imaging

NYU

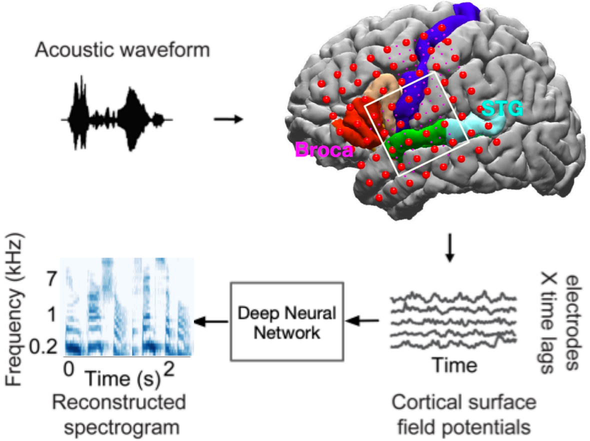
Outline

- **Introduction**
  - Auditory stimulus decoding
  - Previous work
  - In this talk
- **Method and Model**
  - Proposed framework
  - Generator network structure and pre-training
  - Encoder network structure and fine-tuning
- **Experimental Results**
  - Stimuli reconstruction
  - Attention mask visualization
- **Conclusion and future work**

NYU

Introduction - auditory stimulus decoding

- Reconstruct the speech stimuli from intracranial Electrocorticographic (ECoG) recordings.
- Study the activity of cortical regions during hearing.
- Decoding auditory stimulus from neural activity enables
    - neuroscience studies via model interpretation.
    - better understanding of how speech is processed in our brain.
    - neuroprosthetics for neurological conditions leading to loss of communication.
    - brain computer interface application.



Introduction - previous work

Stimulus Decoding

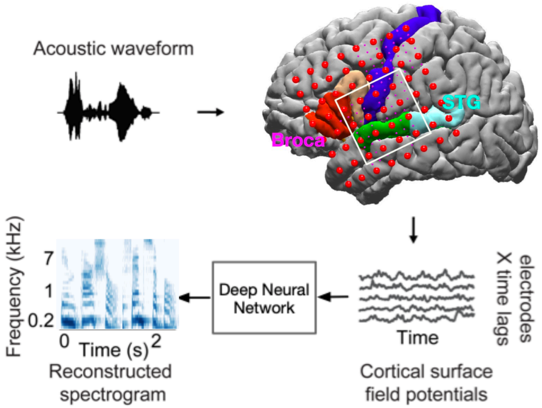
- Linear models have been used to quantitatively demonstrate STG cortical representations [1].
    - provided a means to study how the STG area reacts to speech stimulus.
    - intelligibility of the recovered speech was limited.
- WaveNet-like network for stimulus decoding from ECoG recordings in the STG area [2].
    - obtained significant improvement over linear models

[1] Brian N Pasley, et al., "Reconstructing speech from human auditory cortex," *PLoS Biology*, 2012.

[2] Ran Wang, et al., "Reconstructing speech stimuli from human auditory cortex activity using a WaveNet approach," IEEE SPMB, 2018.

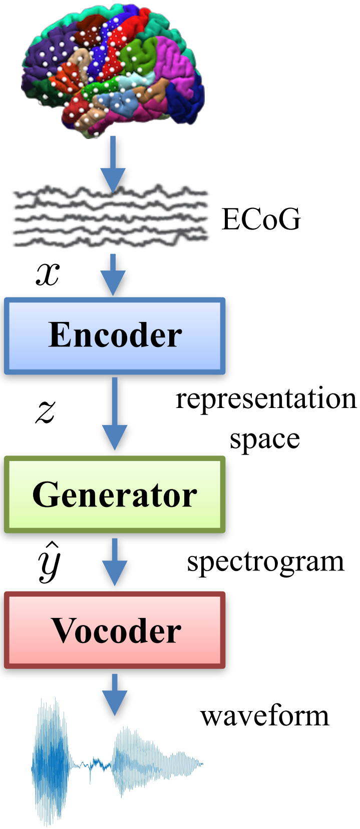
Introduction - in this talk

- Goal of this project
  - leverage deep learning to decode intelligible audio stimuli from ECoG recordings of the cortical regions including the STG area.
  - study the developed model to understand speech perception in cortical networks.
- Major challenges
  - scarcity of the training data: limited ECoG and speech stimuli pairs.
  - variability of electrode density and placement across subjects
  - limited number of subjects and limited words per subject.



Introduction - in this talk

- To overcome the challenges:
  - use available natural speech datasets.
  - use generative network and transfer learning ideas.
  - pre-train parts of a network using a large corpus of natural speech data.
- Proposed methods are applicable when training data is limited.



ECoG

x

Encoder

z

representation space

Generator

$\hat{y}$

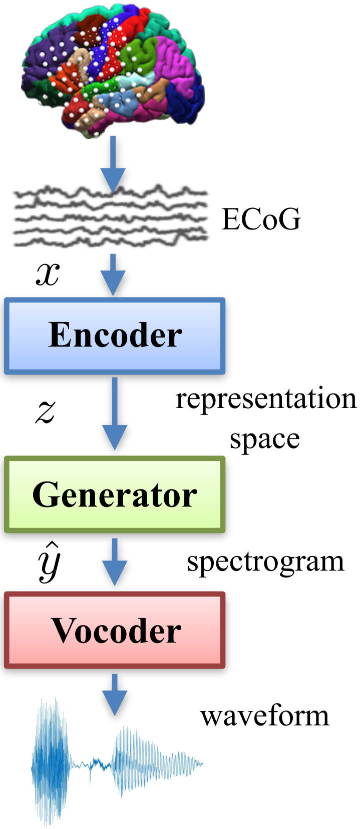
spectrogram

Vocoder

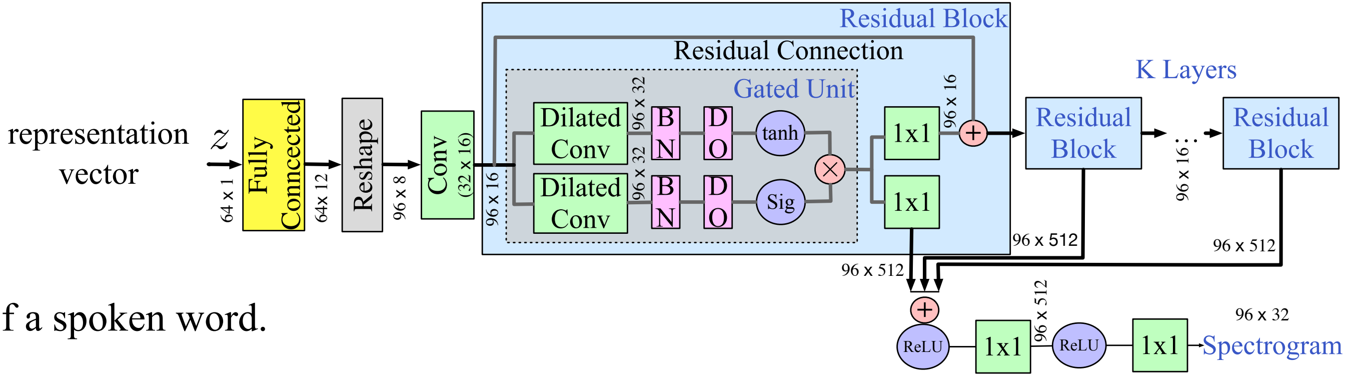
waveform

Proposed framework

- Encoder maps an ECoG signal to a representation space with a prescribed distribution.
- Generator constructs a spectrogram from the representation vector.
- Vocoder converts the spectrogram to sound waveform.
- Pre-train the generator and vocoder using large corpus of speech data.
- Refine the encoder and generator using the paired ECoG and stimuli data.



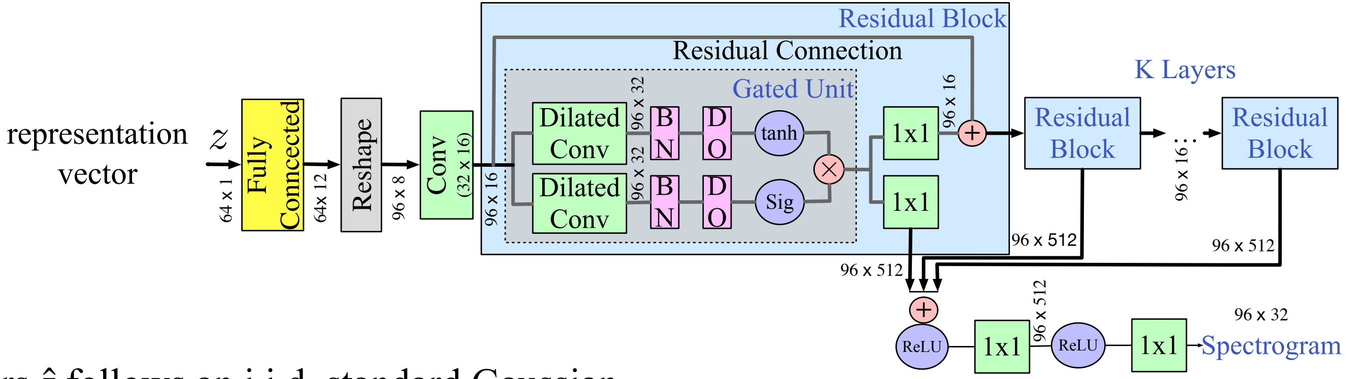
Generator network structure



- Generates a spectrogram of a spoken word.
- Can be pre-trained on a larger corpus dataset.
- Proposed structure is inspired by WaveNet [1].
- Different dilation rates allow filters to span small to large temporal duration.

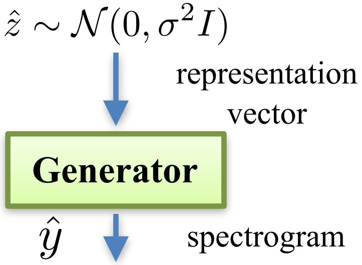
[1] Akira Tamamori, et al., “Speaker-dependent WaveNet vocoder,” Interspeech, 2017.

Generator network pre-training

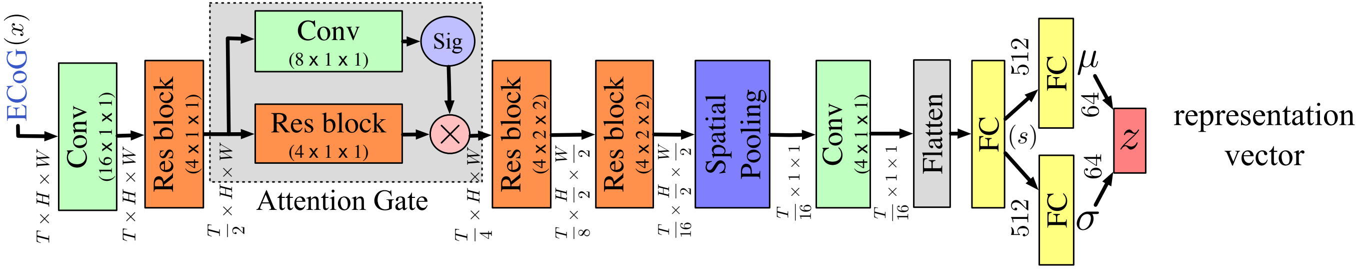


- Pre-training: random vectors $\hat{z}$ follows an i.i.d. standard Gaussian distribution

- Gaussian distribution since
  - encoder output distribution is not known ahead of time.
  - it maximizes the capacity of the representation space.

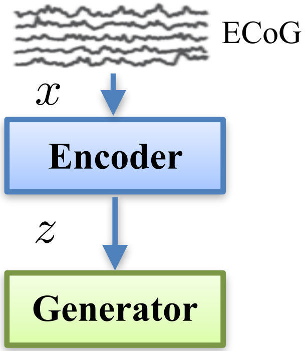


Encoder network structure

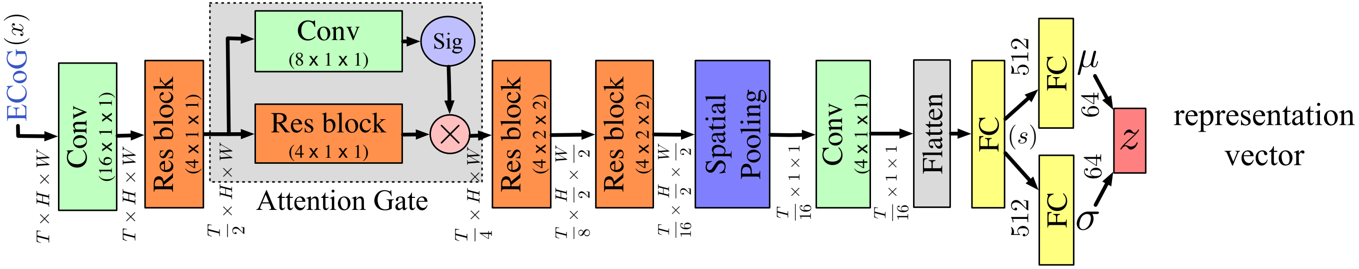


- The encoder serves as a feature extractor that maps the ECoG signals to a representation vector.

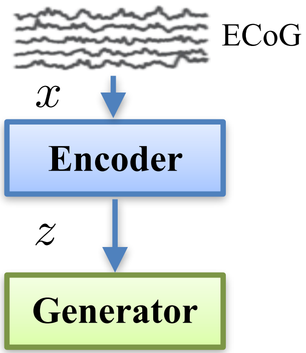
- Initial layers use temporal filtering
  - ECoG signal has less correlations between electrodes than across time.



Encoder network structure



- Attention gated unit
  - allows the network to focus only on significant electrodes at each time step.
  - improves the network accuracy by ignoring electrodes with low signal-to-noise ratio.
  - provides a way to analyze the dynamics of each cortical area.

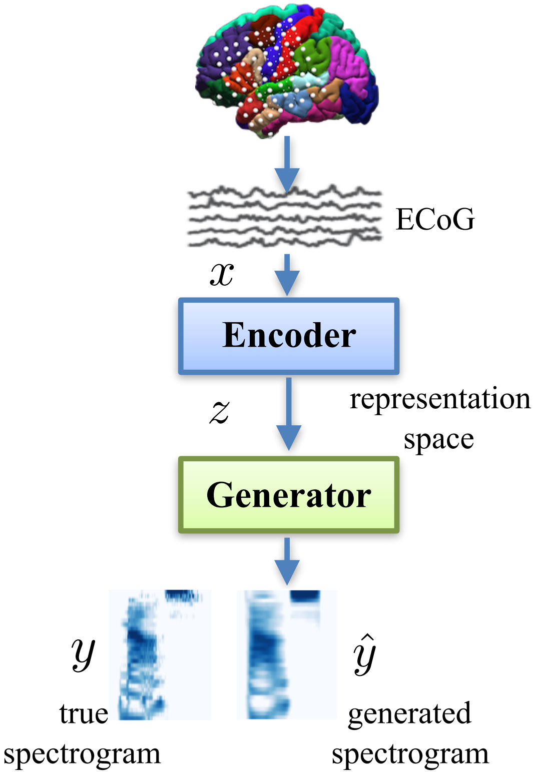


NYU

Encoder and generator fine-tuning

- Transfer the pre-trained generator and fine-tune with the encoder.
- Kullback–Leibler (KL) divergence regularizer
  - encourage the encoder output distribution to follow Gaussian.
- The loss function is:

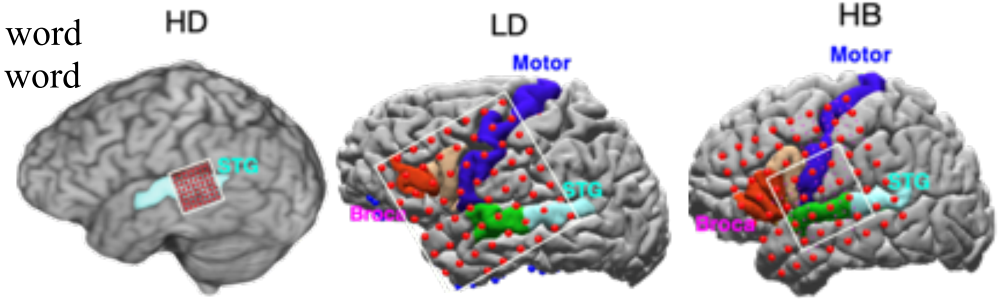
$$\text{loss} = \text{MSE}\left(y, \hat{y}(x)\right) + \lambda\,\text{KL}\left(p\left(z\right) \| \mathcal{N}\left(0, \sigma^2 I\right)\right)$$



Experimental results - dataset information

- Three different electrode density and coverage.
- The STG is sampled for all subjects, some other cortical regions are sampled in LD and HB data.
- Patients undergoing treatment for refractory Epilepsy.
- During the task, all subjects listened to speech audio of 50 different English words/pseudo-words.
- Each word repeated 4 times for HD and 2 times for LD and HB.
- One subject in the HD dataset passively listened to each word while all other subjects were required to reproduce each word after listening.

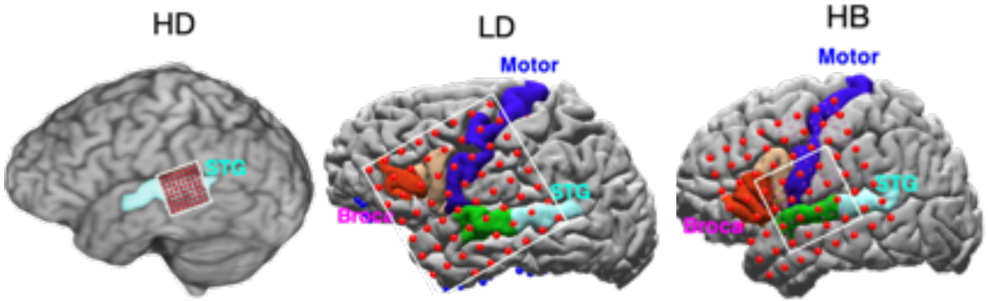
Dataset	High density (HD)	Low density (LD)	Hybrid (HB)
Spacing	4mm	10mm	10 / 5mm
Training Set	100-150 words	50 words	50 words
Number of Subjects	2	12	2



Experimental results - dataset information

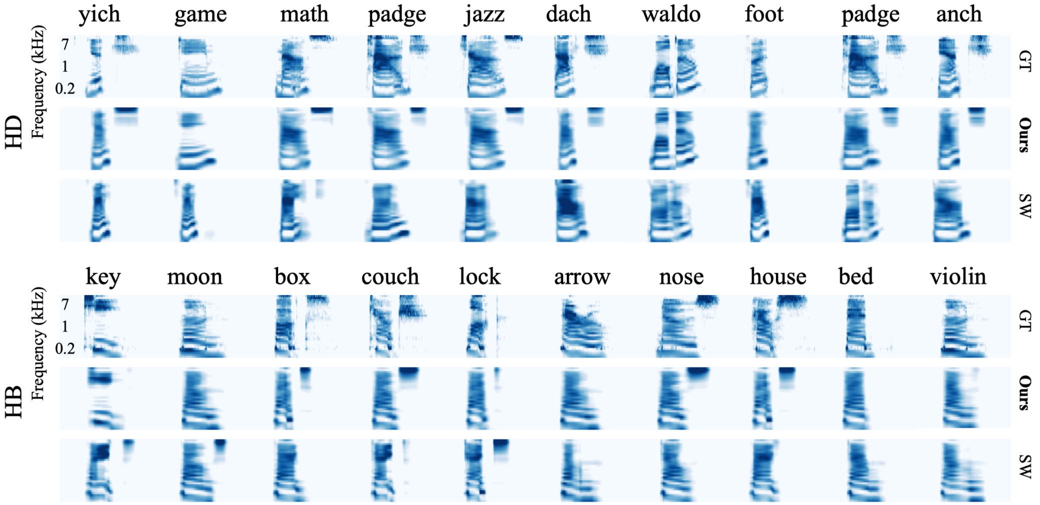
- An individual model is trained and tested for each subject.
- For each subject, 50 unique words are used for testing.
- Rest of the samples for each subject used for training.
- Results are averaged across subjects with same electrode density.

Dataset	High density (HD)	Low density (LD)	Hybrid (HB)
Spacing	4mm	10mm	10 / 5mm
Training Set	100-150 words	50 words	50 words
Number of Subjects	2	12	2



Experimental results - stimuli reconstruction

- state-of-the-art performance in
  - averaged mean squared error (MSE)
  - correlation coefficient (CC).
- audio examples: here!

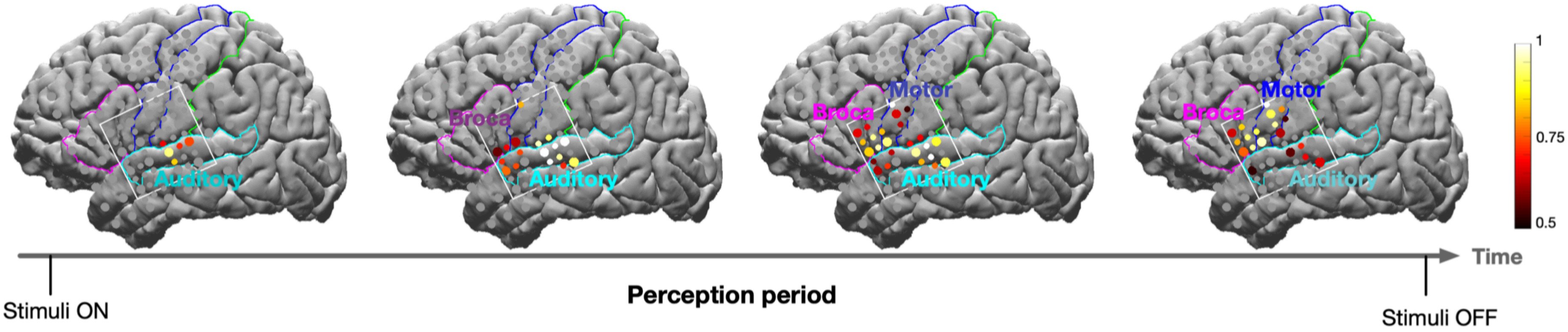


1. transfer-GAN: this work!
2. SpecWaveNet: Ran Wang, et al., “Reconstruct- ing speech stimuli from human auditory cortex activity using a WaveNet approach”, SPMB, 2018.
3. Linear model: Brian N Pasley, et al., “Reconstructing speech from human auditory cortex,” *PLoS Biology*, 2012.

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

Experimental results - attention mask visualization

- The attention mask shows attended electrodes with useful information and dynamics of different cortical areas.
- STG, Broca’s area and motor cortices are attended sequentially.
- STG is consistently attended during the perception period.
- Leant attention mask matches the findings in neuroscientific literature.



Conclusion

- Developed a framework for stimulus speech decoding from human cortex ECoG recordings.
- Proposed encoder extracts features from the ECoG signals to a representation space.
- Pre-trained generator predicts realistic spectrograms from the representation space.
- Tackled the challenge of limited training data and achieved state-of-the-art reconstruction from cortical areas including STG.
- Attention mechanism allowed for interpretation of the results for neuroscientific discoveries.

In future work:

- Use the developed techniques to study speech processing in human cortex in finer details.
- Decode the reproduced speech.
- Use the framework for other applications with limitations in training data.

Thank you!