Cocktail Party Problem

EE 516
Introduction to Digital Signal Processing

Imani Muhammad-Graham 50262878
Shreshta Shekar 50485117
Uday Kiran Manduri 50447368

Introduction

Idea

The cocktail party effect refers to the ability of the brain to focus on a single speaker while filtering out other voices and background noise. Humans perform very well at the cocktail party problem.

This project shows how to use a deep learning network to separate individual speakers from a speech mix where one male and one female are speaking simultaneously.

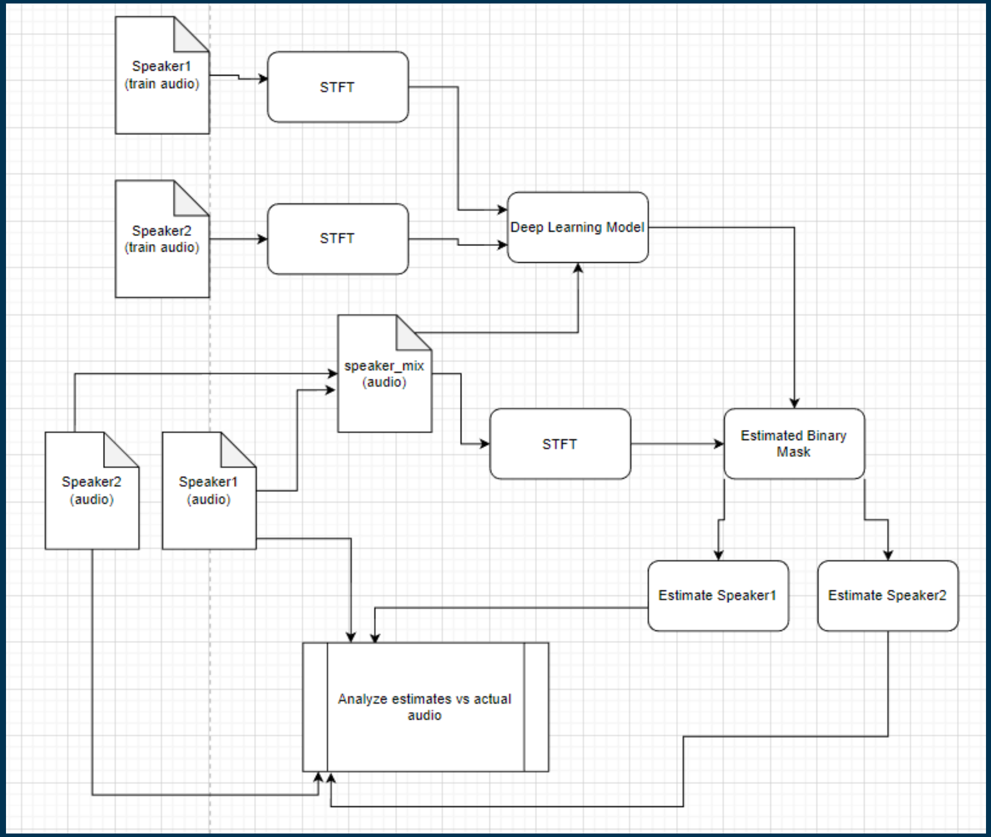
We have used deep learning to first train the model and then separate the male voices from the female voices

Idea

- With respects to DSP, the cocktail party problem is a blind source separation problem
- Our implementation will attempt source separation but will not be blind

Given

- Original speech from two different sources one "male" one "female"
  - These can be any two sources, for this example we are using male and female
- Training speech used to train the deep learning model



Implementation

Pre-Processing

Load	Load audio files containing male and female speech sampled at 4 kHz. Listen to the audio files individually for reference.
Combine	Combine the two speech sources. Ensure the sources have equal power in the mix. Scale the mix so that its max amplitude is one.
Visualize	Visualize the original and mix signals. Listen to the mixed speech signal.
TF representation	Use stft to visualize the time-frequency (TF) representation of the male, female, and mix speech signals. Use a Hann window of length 128, an FFT length of 128, and an overlap length of 96

Source Separation

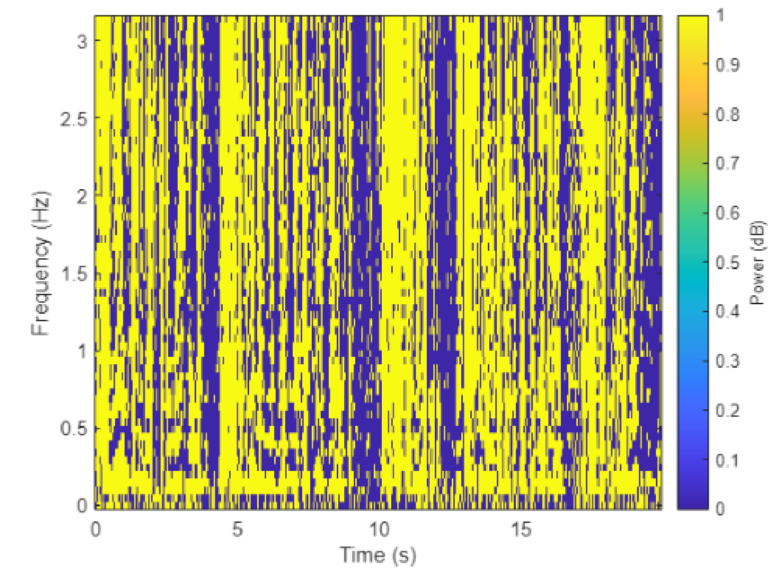
Binary Mask	Compute the ideal binary mask for the male speaker and then visualize it
Source Separation	Estimate the male speech STFT by multiplying the mix STFT by the male speaker's binary mask. Estimate the female speech STFT by multiplying the mix STFT by the inverse of the male speaker's binary mask.
Estimate Sources	Estimate the male and female audio signals using the inverse short-time FFT (ISTFT). Visualize the estimated and original signals. Listen to the estimated male and female speech signals
Evaluate	Assess binary mask efficiency by amount of noise in audio, and comparing estimate sources to the original.

Binary Masks

- Defines a region of interest, and gives 0 or 1 based on given condition
- In our case the regions of interest are each window of the mix's STFT
- The condition is if the Male speech's power >= Female's
- This mask is then applied to the mix to separate male and female sources

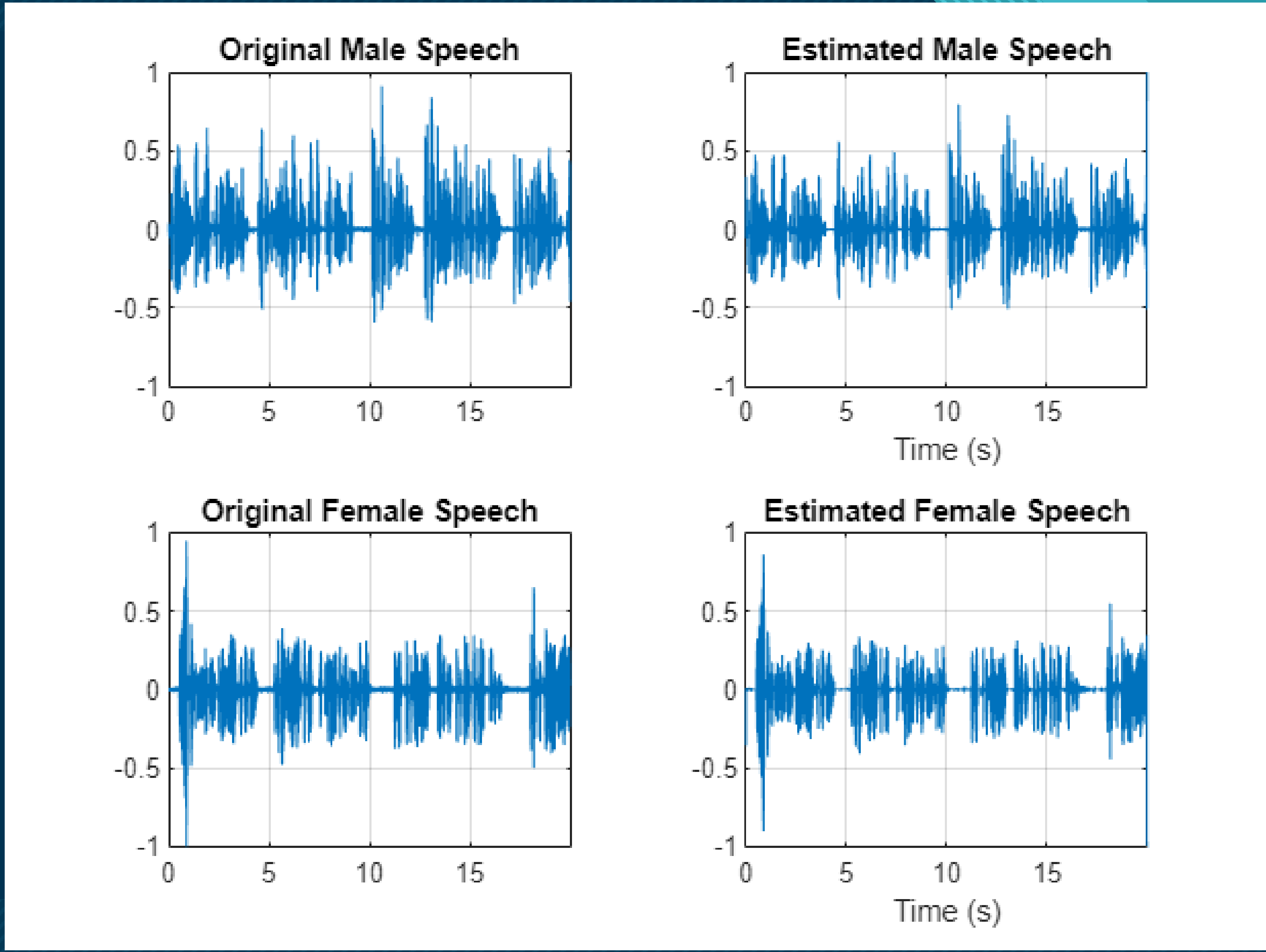
```
P_M = stft(mSpeech,Window=win,OverlapLength=overlapLength,FFTLeng
P_F = stft(fSpeech,Window=win,OverlapLength=overlapLength,FFTLeng
[P_mix,F] = stft(mix,Window=win,OverlapLength=overlapLength,FFTLe

binaryMask = abs(P_M) >= abs(P_F);
```




```
P_M_Hard = P_mix.*binaryMask;
P_F_Hard = P_mix.*(1-binaryMask);
```

Initial Attempt - Results



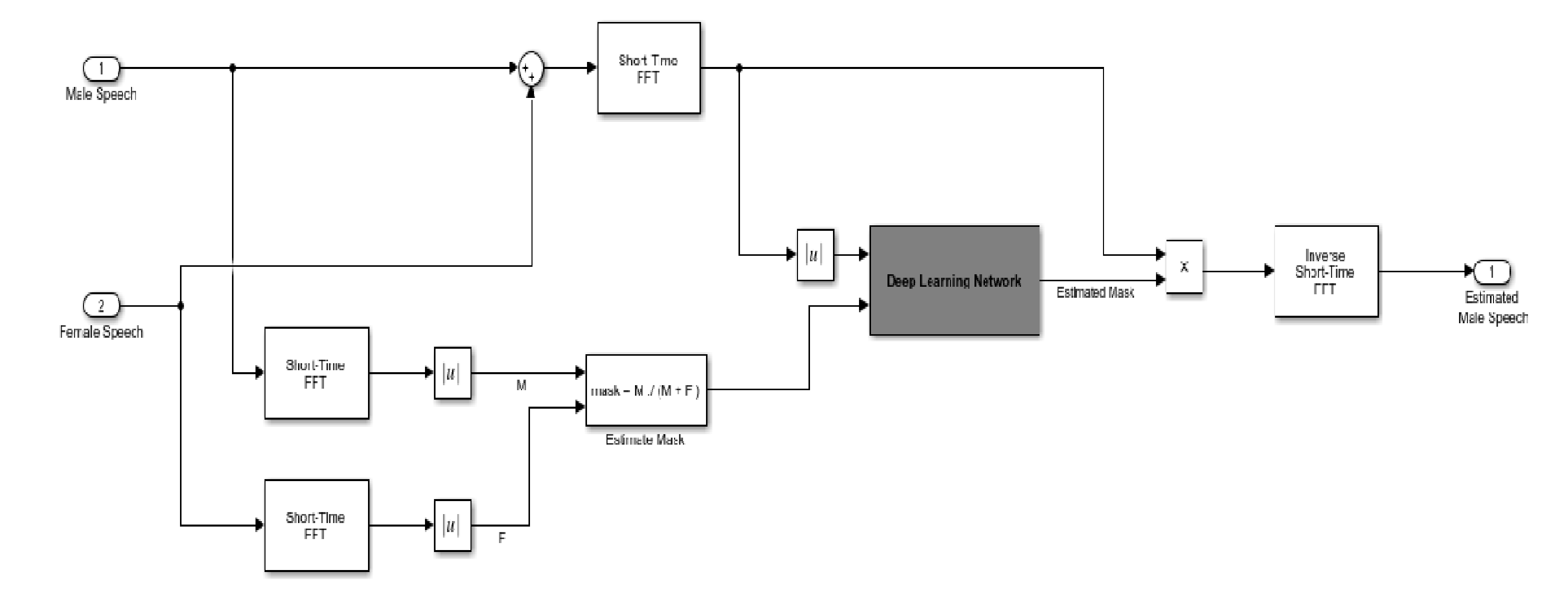
Can we improve?

Utilize deep learning model to improve mask estimation

Mask Estimation using Deep Learning

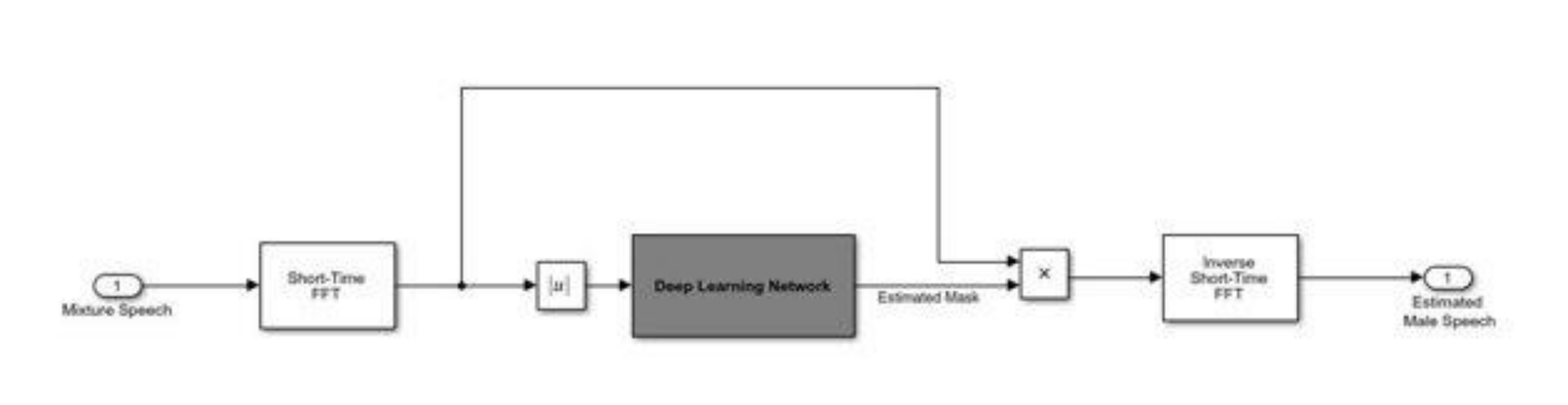
Train Deep Learning Network by using:

- Training sets from each speaker as predictors
- Sources and mix as targets



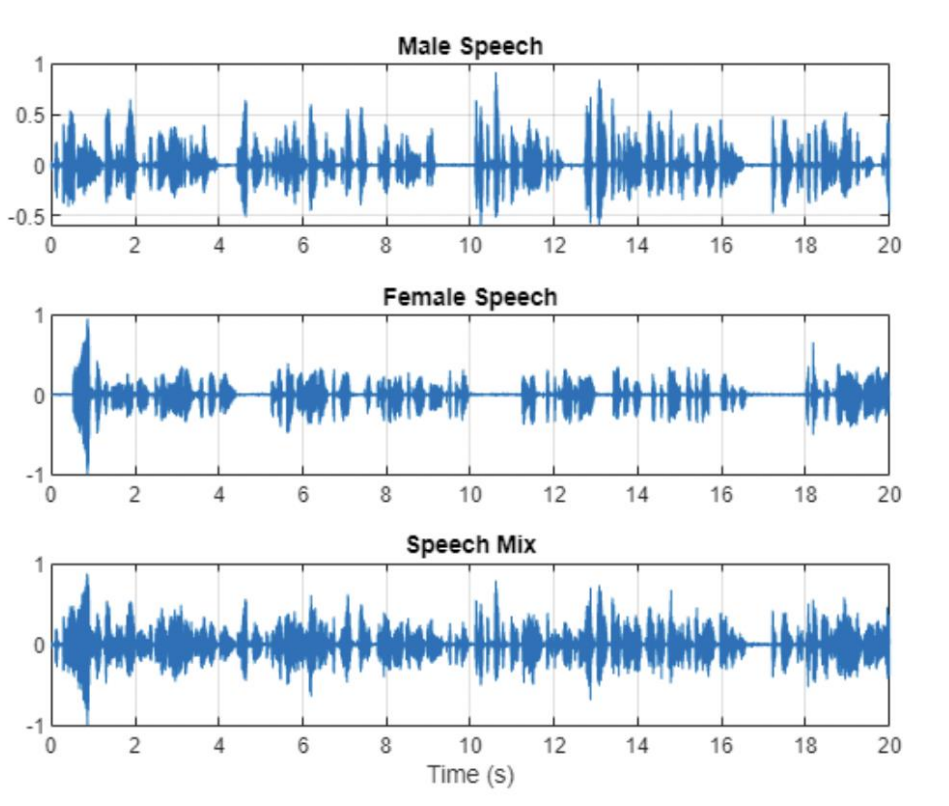
Mask Estimation using Deep Learning

Implementing Deep Learning Network Model for source separation

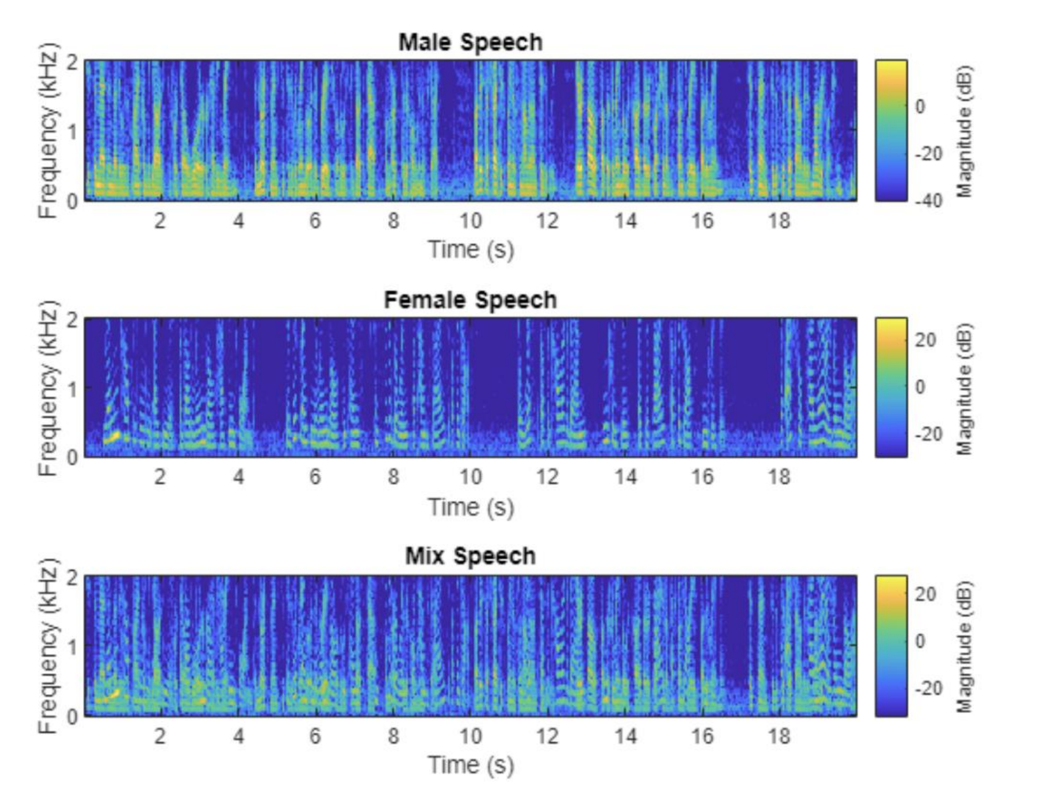


Results

Original Speech

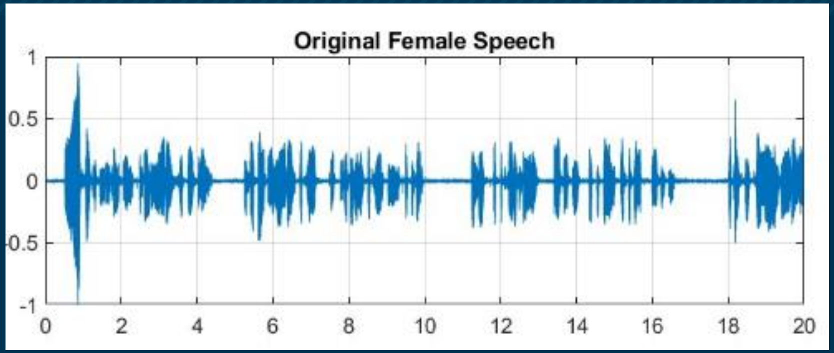


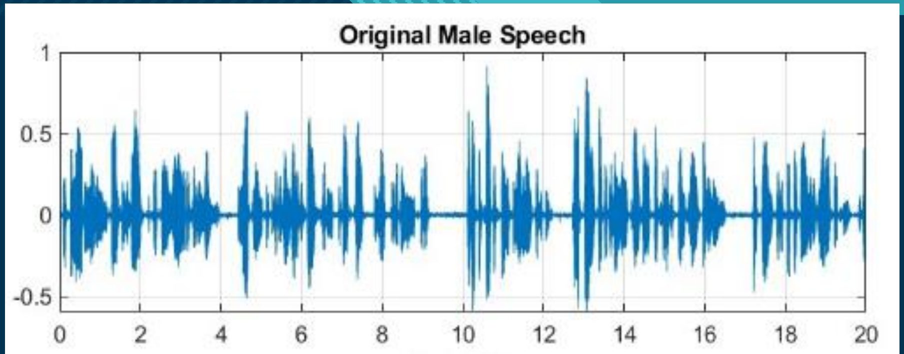
Audio Files

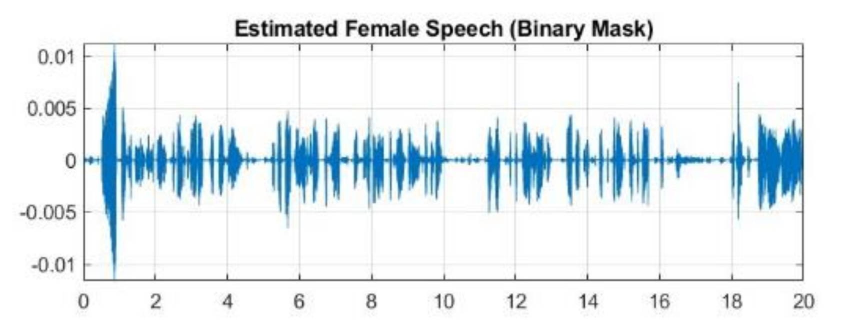


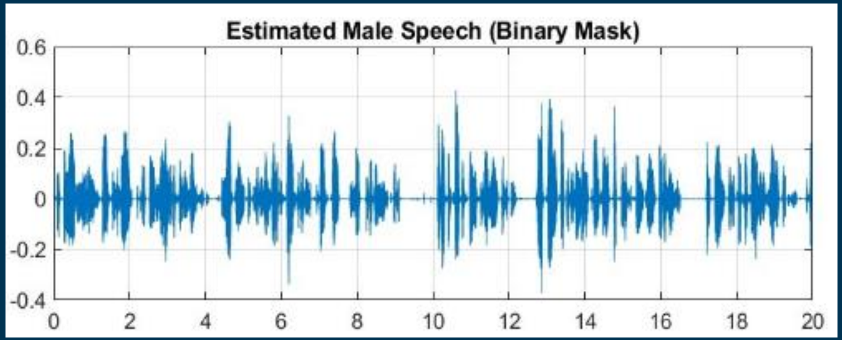
Time Frequency Representation

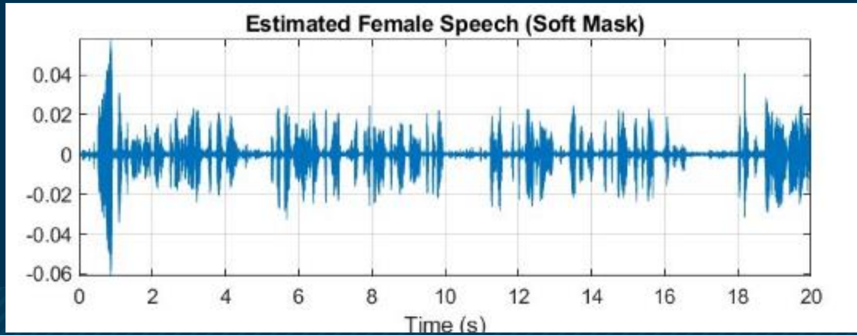
Estimated Results

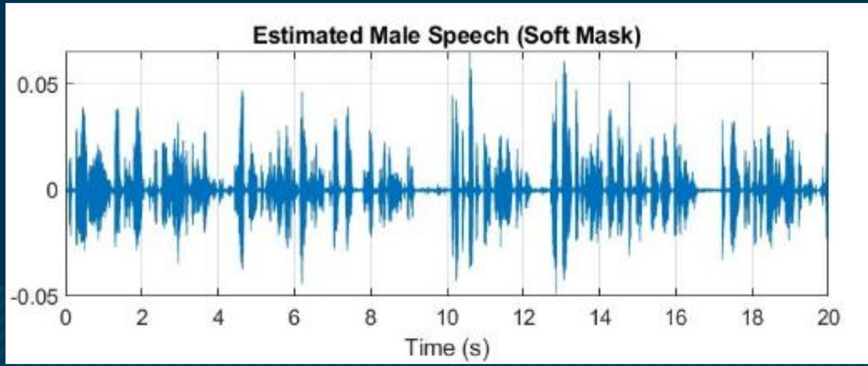




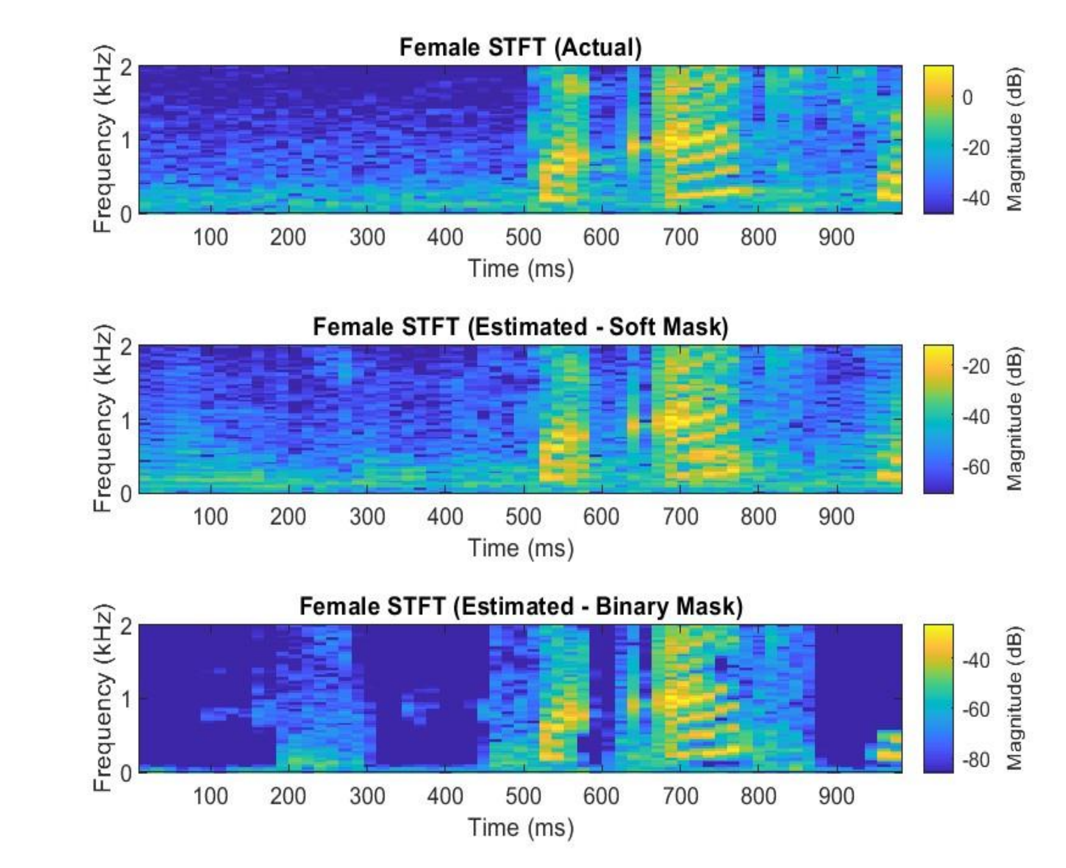


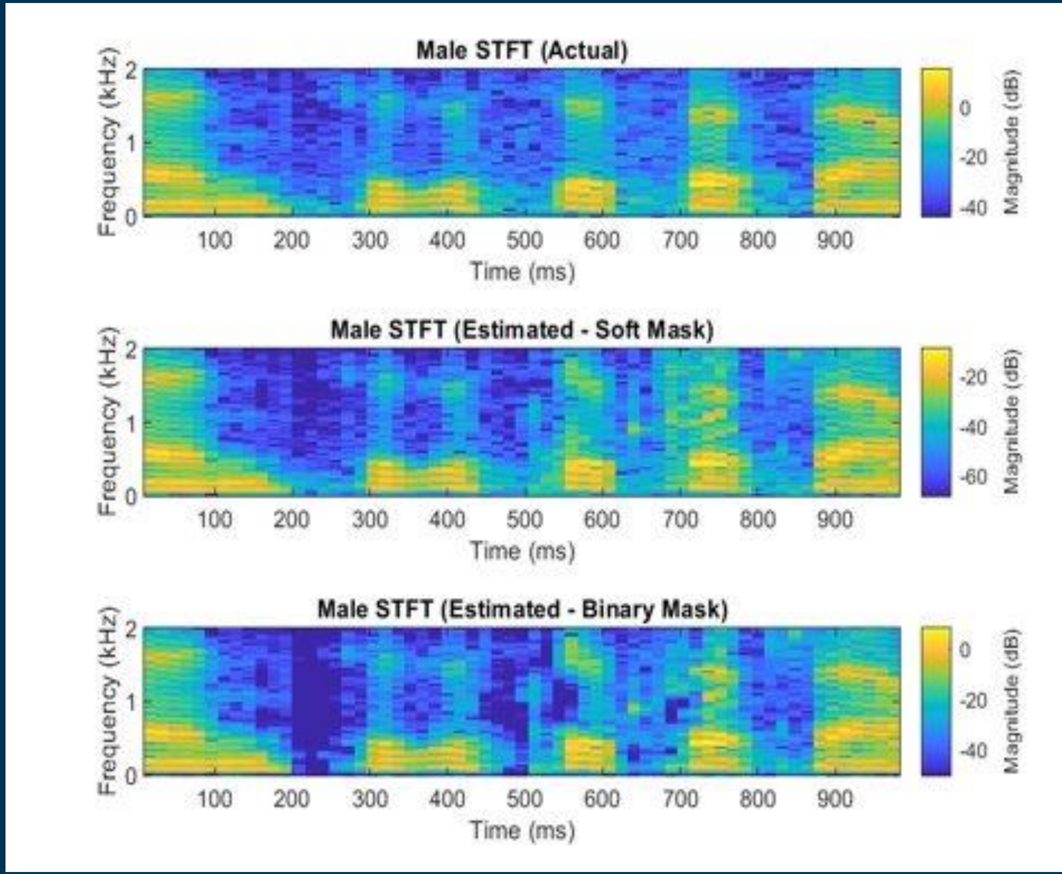






A closer look





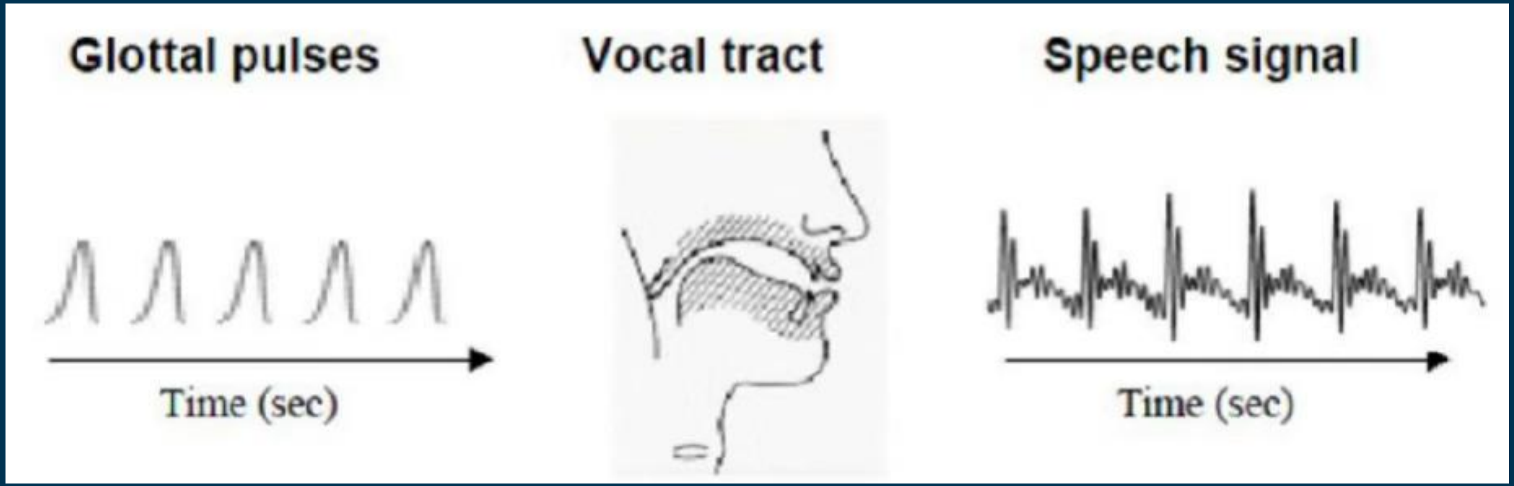
Future Improvements

Understanding Voice

- Speech can be seen as a Convolution of vocal tract frequency response with glottal pulses.

- Humans perceive frequencies logarithmically

- Cocktail party and other Voice recognition challenges can be improved with perceptually relevant methods.



Understanding Voice (cont'd)

Let :

s(t) = Speech

g(t) = glottal pulse

v(t) = vocal tract frequency response

So:

s(t) = g(t)*v(t)

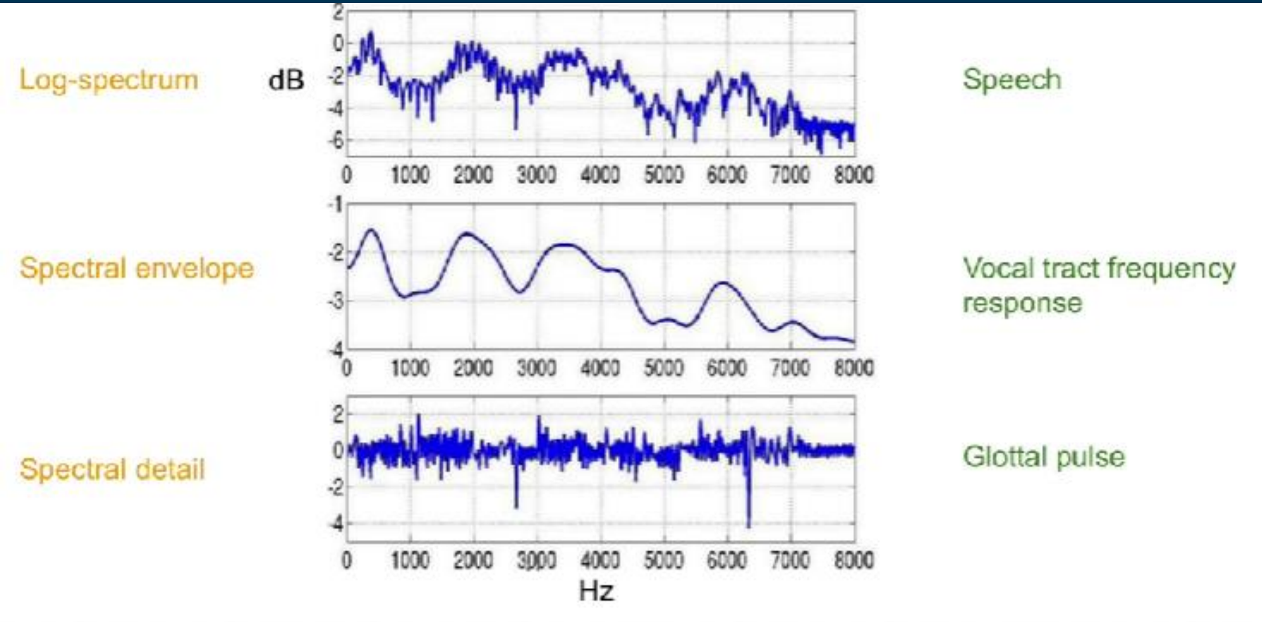
Take Fourier Transform for frequency domain

S(f) = G(f)V(f)

Take Logarithm for a Log-spectrum

Log(S(f)) = Log(G(f)) + Log(V(f)))

Log(V(f)) is known as Spectral Envelope, and contains formants, which is what gives us the identity of voice



Mel Scale

- The Mel Scale is a logarithmic transformation of a signal's frequency. The core idea of this transformation is that sounds of equal distance on the Mel Scale are perceived to be of equal distance to humans.
- It is actually much harder for humans to be able to differentiate between higher frequencies, and easier for lower frequencies. So, even though the distance between the two sets of sounds are the same, our perception of the distance is not. This is what makes the Mel Scale fundamental in Machine Learning applications to audio, as it mimics our own perception of sound.
- The transformation from the Hertz scale to the Mel Scale is the following:

$$m = 1127.108\ (1 + f/700)$$

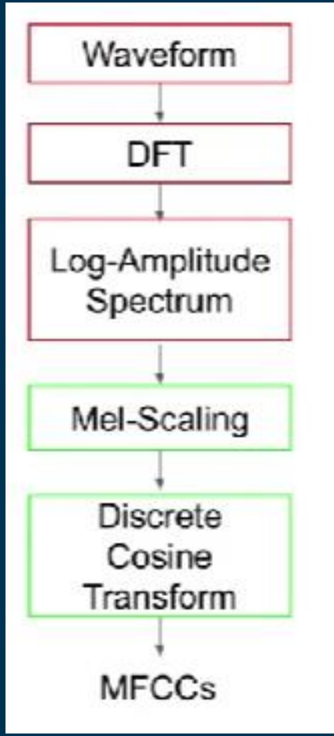
Mel Frequency Cepstral Coefficients (MFCCs)

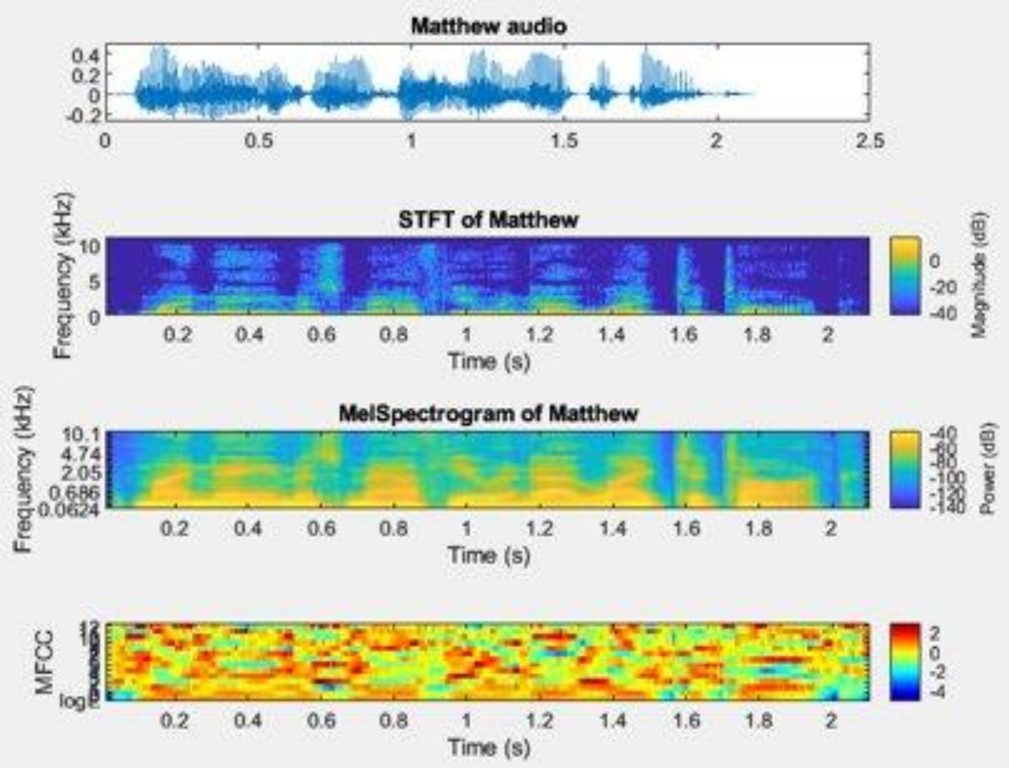
To obtain Mel Spectrogram

- Extract STFT
- Convert spectrogram to DBs
- Choose # of mel bands
- Construct mel filter banks
- Apply mel filter banks to spectrogram

To obtain MFCCs

- Use Discrete Cosine Transform to obtain real valued coefficients (formants)





Final Discussion

Source separation

- Using ideal binary masks from observation
  - Very Noisy
- Using estimated binary masks from deep learning model with training set
  - Less noisy

Future Discussion

- Possibly using Mel Spectrogram and/ or MFCCs to enhance binary mask estimation

Sources

- https://www.mathworks.com/help/deeplearning/ug/cocktail-party-source-separation-using-deep-learning-networks.html
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4111459/
- https://www.mathworks.com/help//audio/ug/speaker-diarization-using-x-vectors.html
- https://www.pathpartnertech.com/blind-source-separation-for-cocktail-party-problem
https://www.mathworks.com/help/audio/ug/speaker-identification-using-pitch-and-mfcc.html
- https://www.mathworks.com/help/audio/ref/mfcc.html
- https://www.youtube.com/watch?v=9GHCiiDLHQ4
- https://www.youtube.com/watch?v=4_SH2nfbQZ8

Thank You