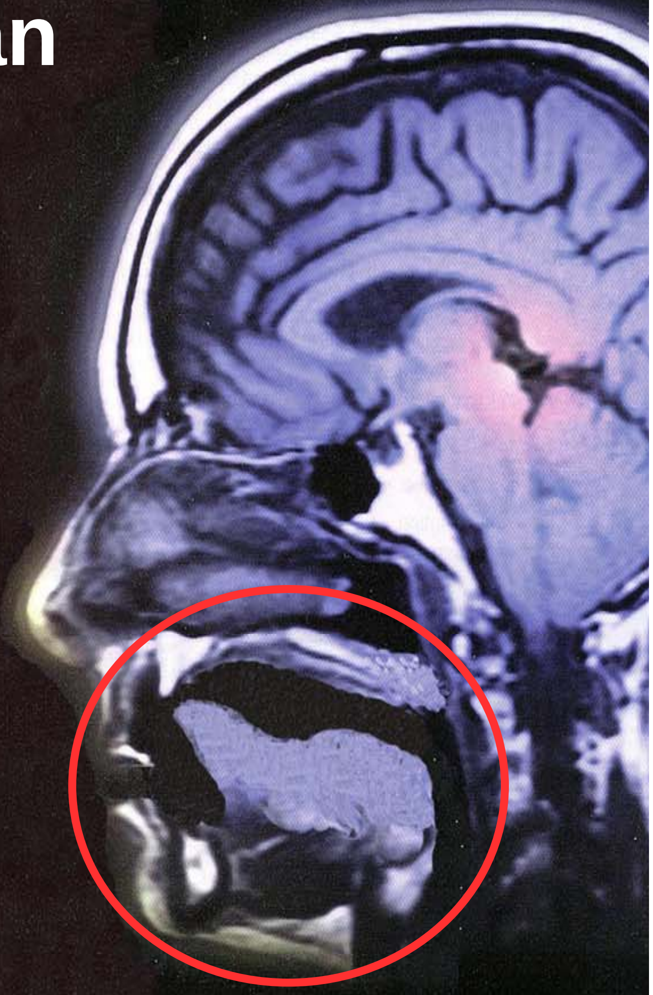
Speaker Recognition

Phonetic Discriminative Power of American English /r/



Ryan J. Amundsen

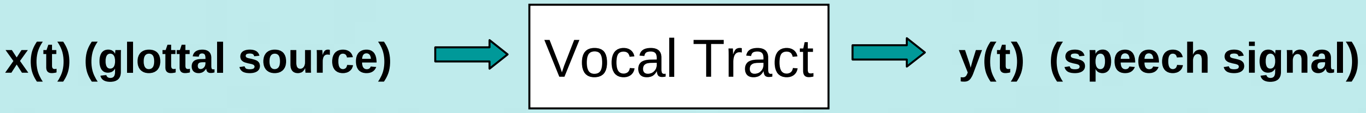
Dr. Carol Espy-Wilson

Daniel Garcia-Romero

Xinhui Zhou

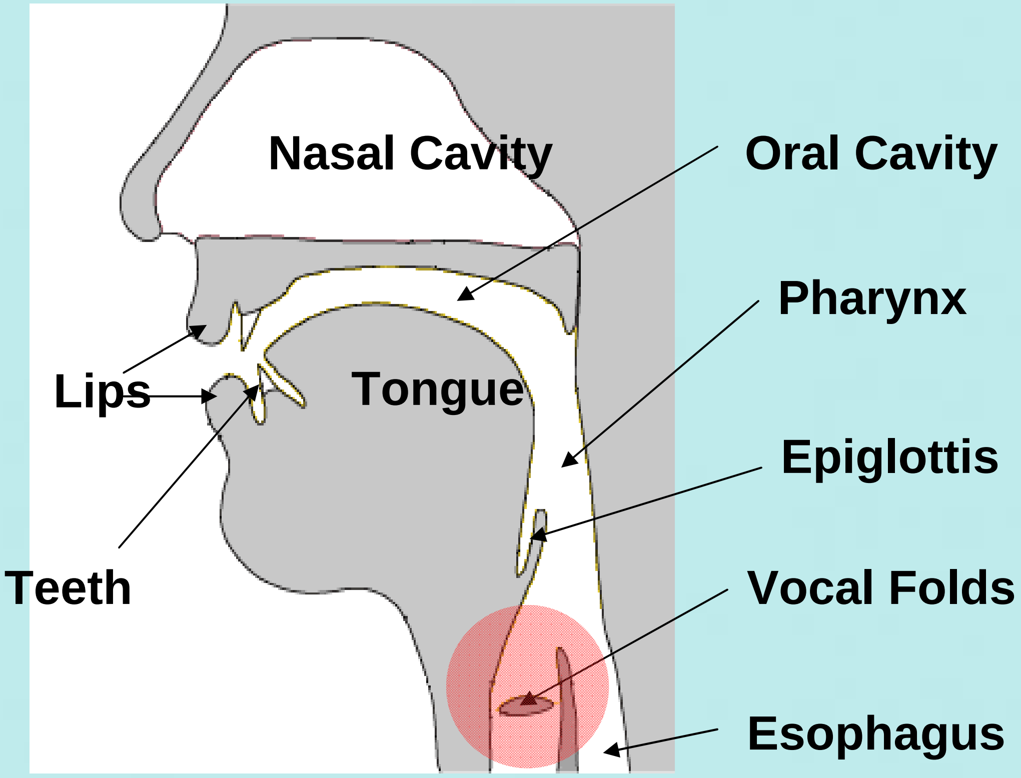
Background

Speech Production System

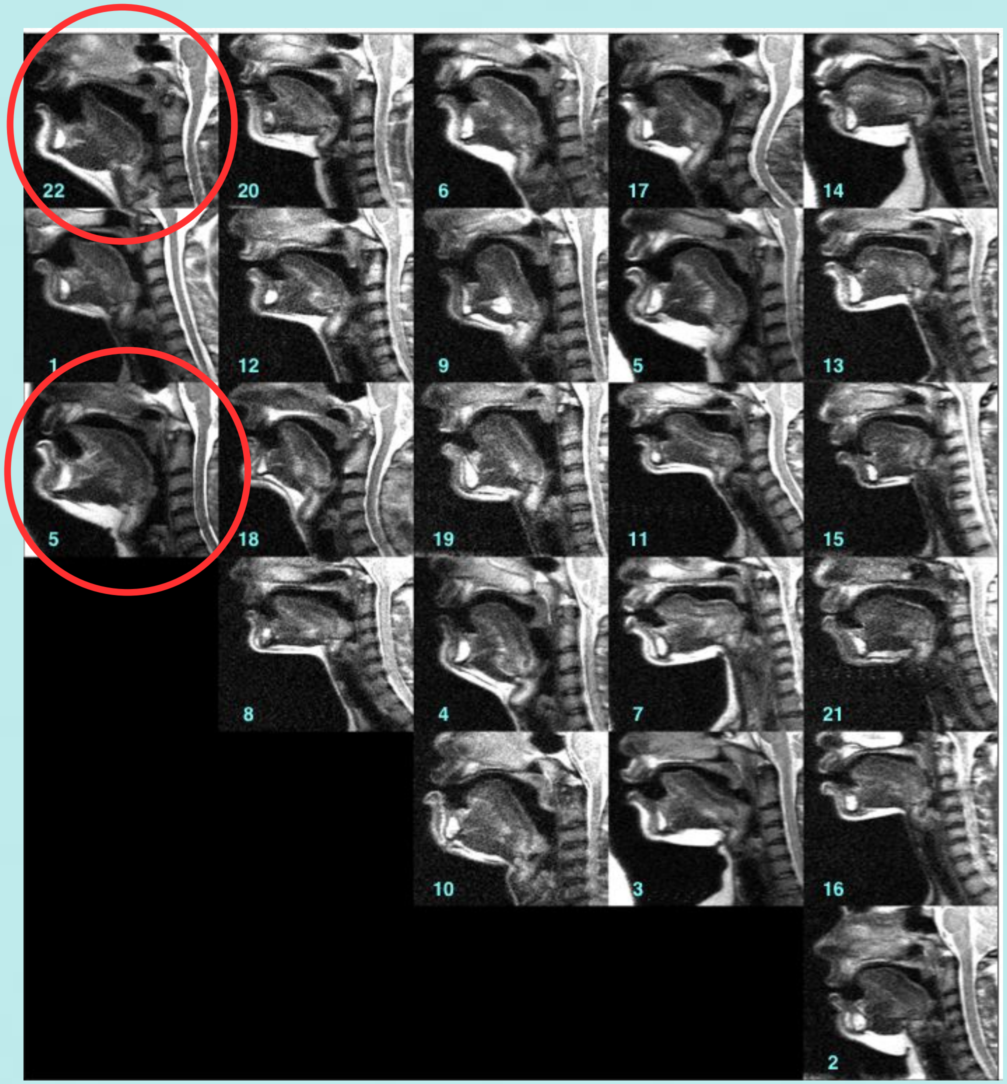


x(t) (glottal source) → Vocal Tract → y(t) (speech signal)

- Speaker specific information comes from the glottis and the vocal tract.



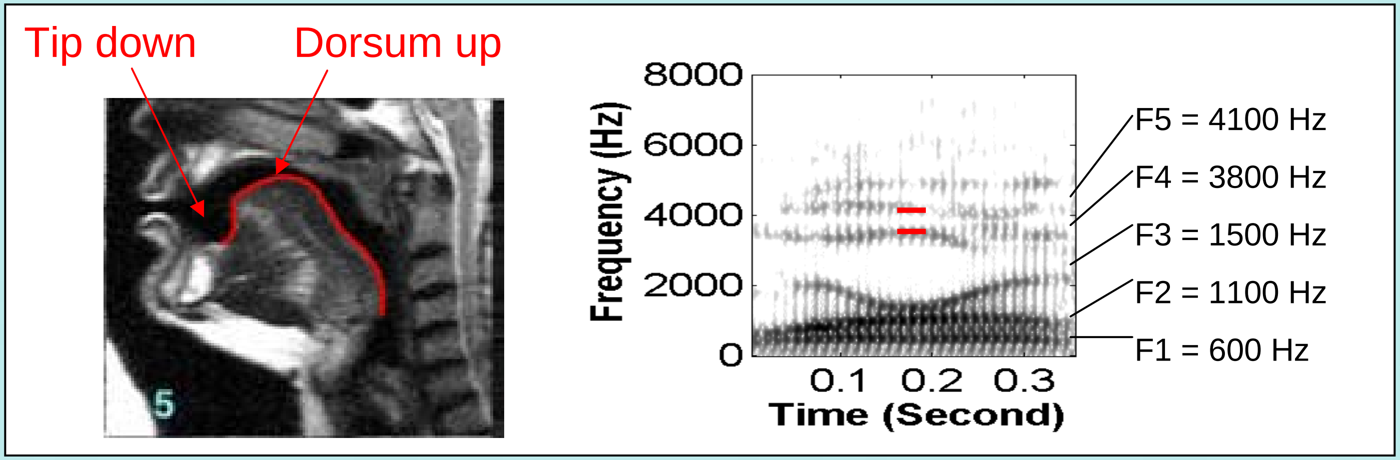
Background



- Different speakers use different tongue postures to annunciate American English /r/
- Examples: /r/ in “red”, “arrow”, or “Ryan”

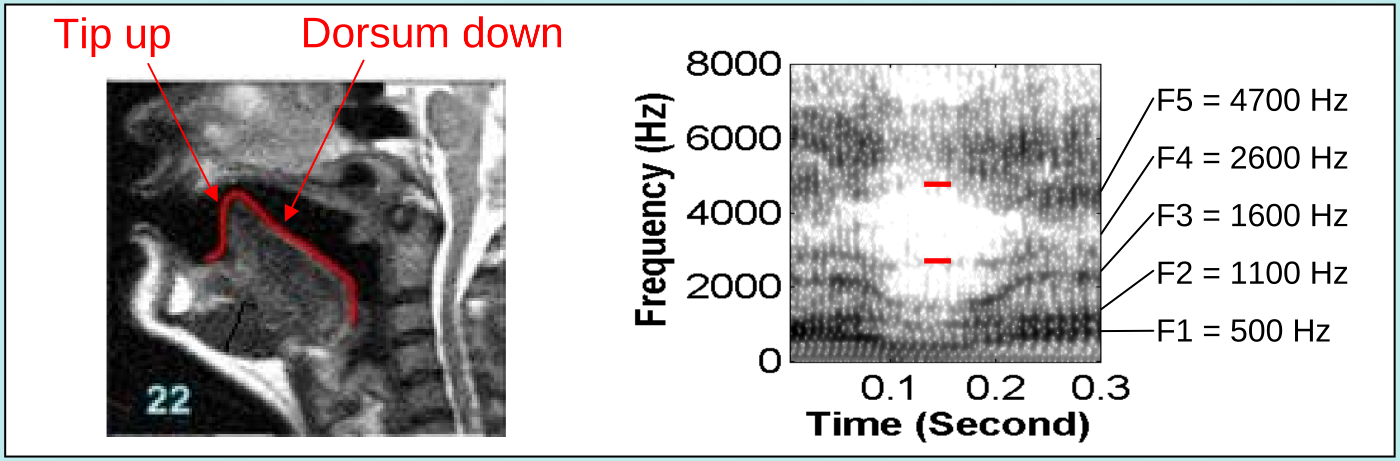
Background

“Bunched” /r/



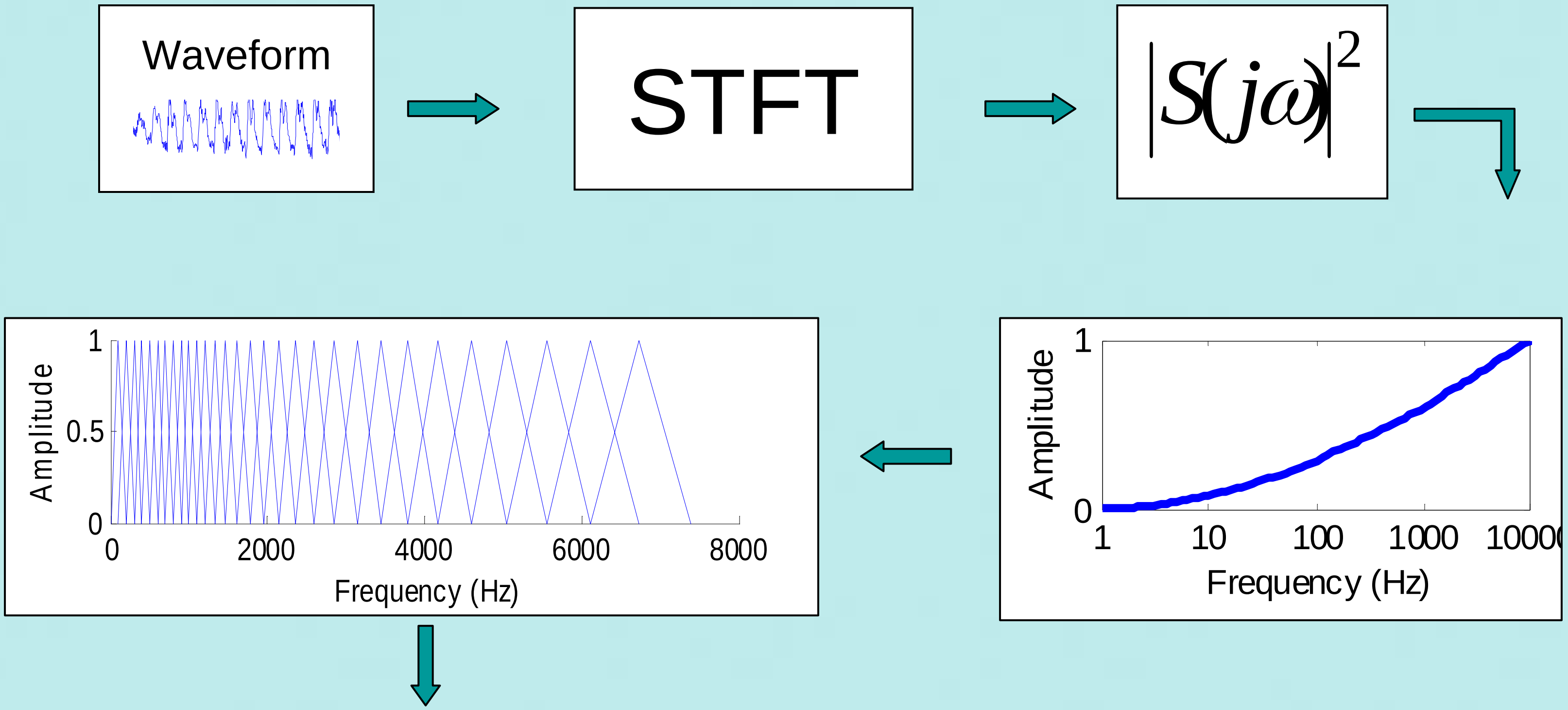
vs.

“Retroflex” /r/



Methodology

Step 1: Mel-Scale Filter Bank Energies



Waveform

STFT

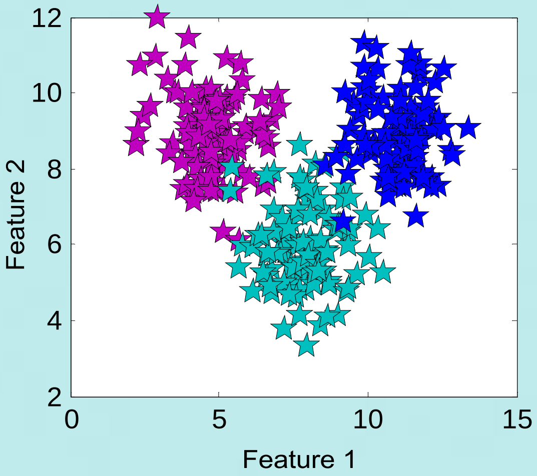
$|S(j\omega)|^2$

- Output = 31 average energy values per frame

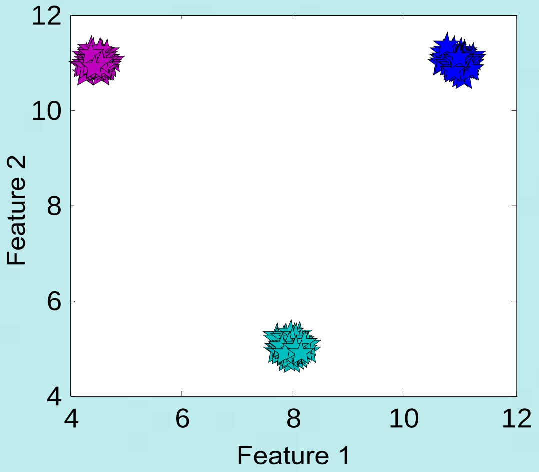
Methodology

Step 2: Discriminative Power

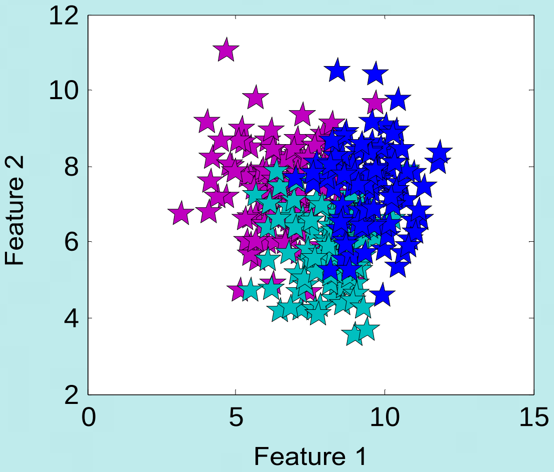
$$\text{Discriminative Power} = \frac{\text{Between-Class Variance}}{\text{Within-Class Variance}}$$



Feat 1	9.1988
Feat 2	2.4513



Feat 1	774.7904
Feat 2	807.2033



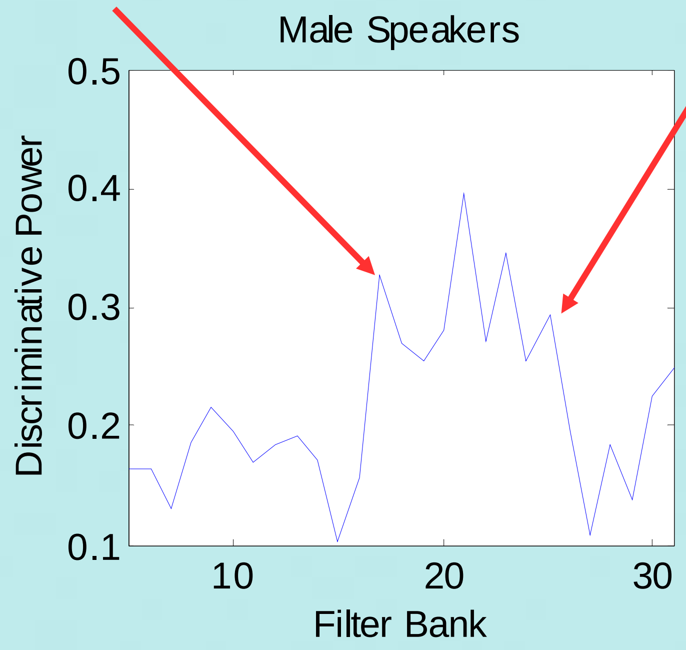
Feat 1	2.0001
Feat 2	0.6891

Analysis

Discriminative Power for All /r/s Classified by Speaker

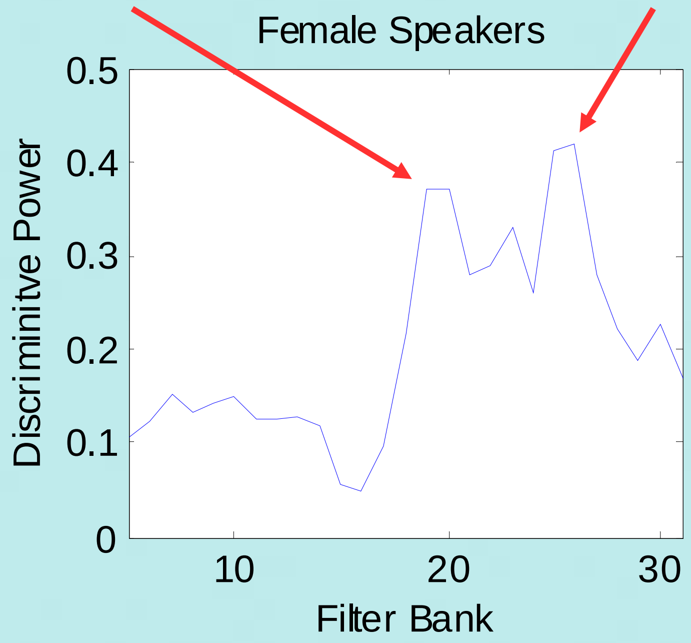
1949 Hz

3797 Hz



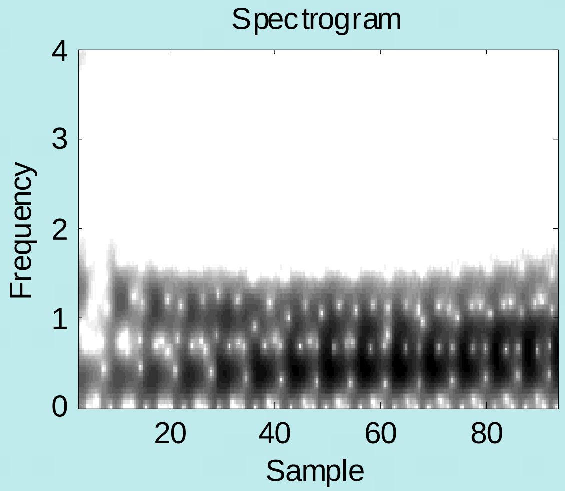
2144 Hz

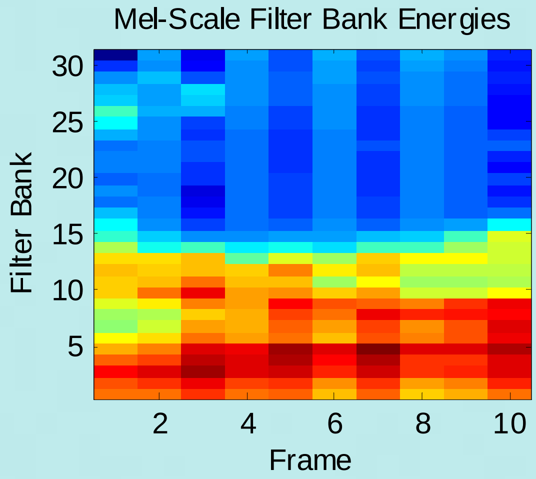
4177 Hz



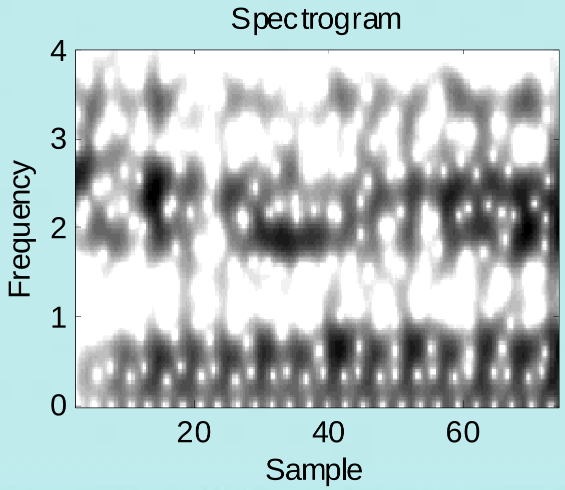
Analysis

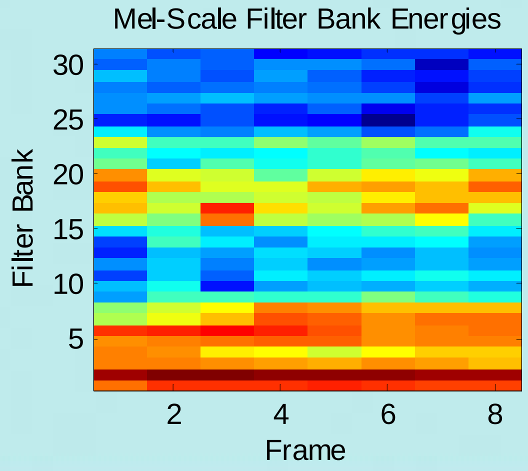
Analysis of /r/ in "destroying" for Male Speaker


Spectrogram


Mel-Scale Filter Bank Energies

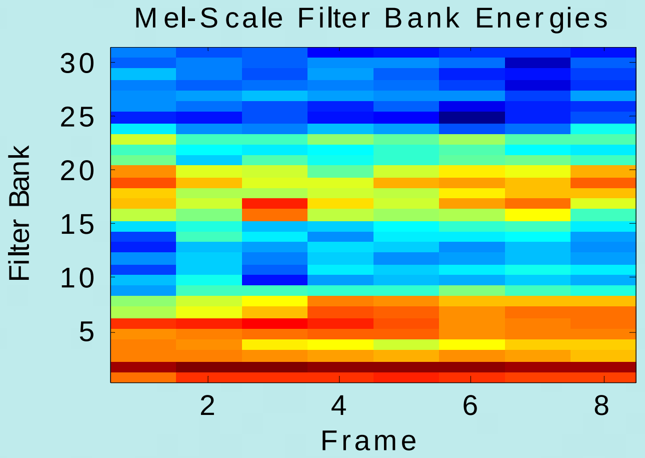
Analysis of /r/ in "injury" for Female Speaker

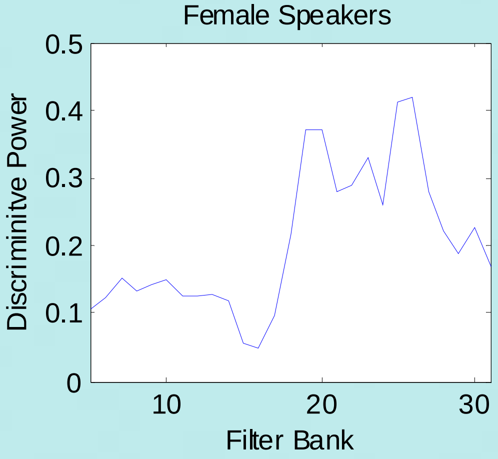

Spectrogram


Mel-Scale Filter Bank Energies

Conclusion

- Our results show that F4 and F5 have the most discriminative power for speaker ID.
- American English /r/ has most of its energy in the region of F1, F2 and F3,
- It can be inferred that there exists a strong relationship between tongue shape and F4 and F5.





Future Work

- Understand and quantify how F4 and F5 vary across different articulatory configurations of /r/
- Apply this relationship to a speaker ID algorithm