Recognition and Annotation of Speech Involving Different Speakers

Avrosh Kumar; *Department of Music Technology, Georgia Institute of Technology*

Georgia Tech School of Music College of Architecture

Georgia Tech Center for Music Technology College of Architecture

Motivation

Speech processing and speaker recognition technologies are a part of many electronics and software we use today. Speech processing is widely used for enabling voice commands and speech-to-text transcription. Speaker recognition technologies are used used in audio forensics and biometrics to identify and verify a speaker by analyzing his speech.

These applications are focused around individual speakers. My motivation for this project is to identify more than one speakers in a conversation. Speech conveys information that we easily perceive not just as spoken language but also use it to understand underlying emotions and to identify people by differences in their voices. I am performing spectral analysis of speech to look for the artifacts that can be used to differentiate between speakers in the same clip of audio.

I propose to use this to transcribe speech-to-text assigned to every speaker present in a speech recording. This application has practical applications in situations like conference meetings to generate speaker-based log of the text which can be used to generate minutes of the meeting.

Background

Speaker detection or recognition is a well researched topic. Many existing papers discuss the use of spectral features to identify a speaker against a classifier trained with previously collected speech information from the same speaker.

Differentiating between different speakers is not same as what these papers describe. We need to employ methods to cluster similar data-points. Similar data-points represent speech spoken by the same speaker. So, this brings up the question of how to create data-points that represent enough information to differentiate one speaker from another. For example, an average female speaker's vocal frequency range will be much higher than an average male's.

Dataset

Obtained English Language Speech Database for Speaker Recognition (ELSDSR) [1] dataset. It has 9 sentences of variable lengths spoken by 22 different speakers consisting of 12 male and 10 female speakers belonging to age group 24-63 years.

Using a python script, I created random combinations of these sentences while annotating their start times that I used as ground truth. For example, I created a set of 50 audio files each containing sentences spoken by 4 of the randomly selected speakers from the set with at least 2 repetitions by the same speaker.

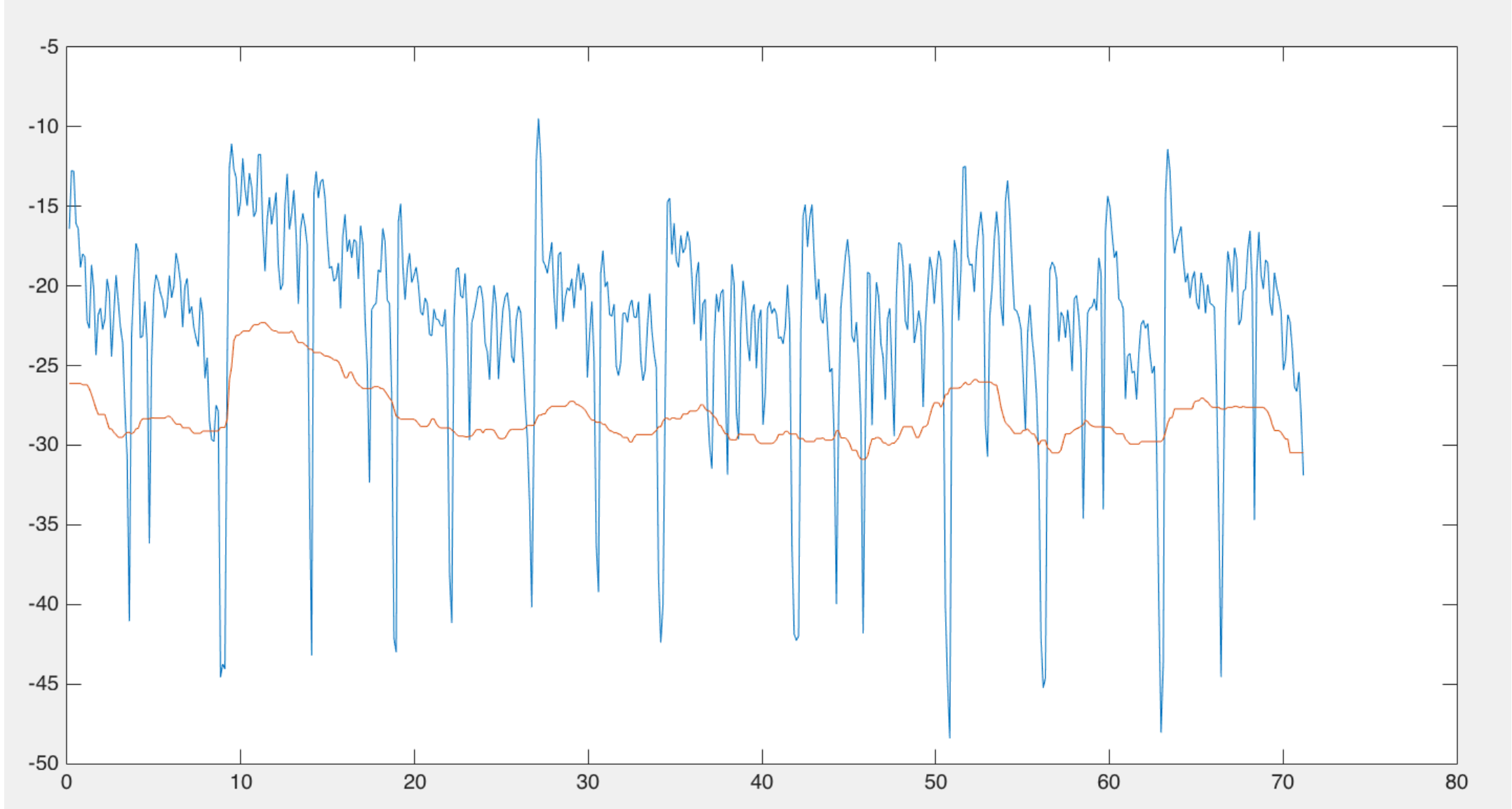
[1] ELSDSR, http://www.imm.dtu.dk/~lfen/elsdsr/index.php, Last accessed 05/04/2016

Methodology I

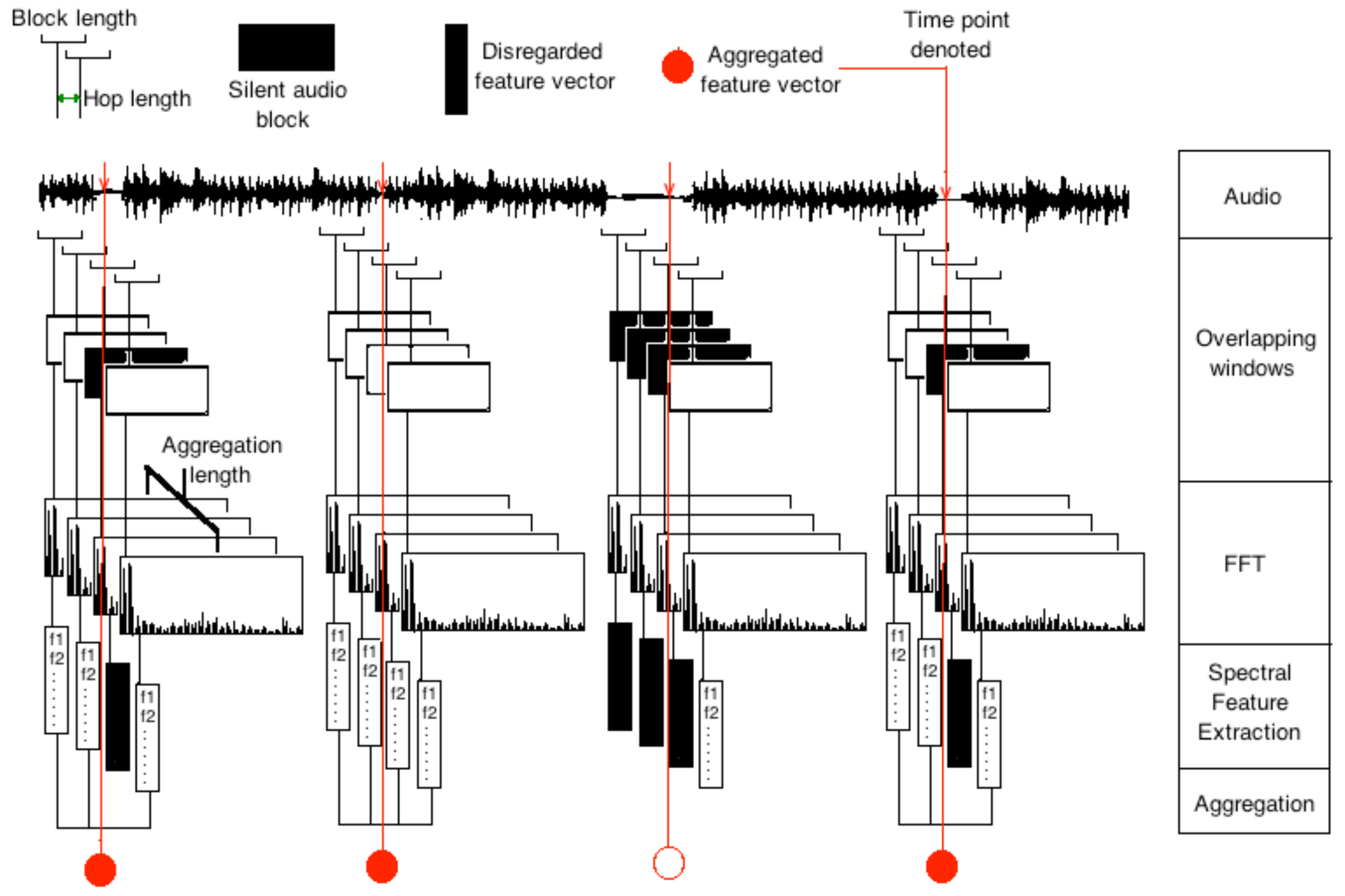
First set a block length to process audio in smaller blocks with overlap. Silent blocks do not represent any speaker so they may induce inaccuracies in clustering. Using the RMS values obtained for each block decide whether the block represents silence or not based on an adaptive median threshold.

After calculating a feature vector from the spectrogram (13 MFCCs, Pitch deviation in MIDI format & RMS value) aggregate the values of blocks that fall in a larger window. I set this window size to 2 seconds so typically there will by 8 blocks in one window for a 16000 sample rate audio blocked into 4096 samples and overlapped by half the block size. This feature vector represents the timestamp of the middle block.

Methodology II



Detecting Silences using an adaptive median threshold on RMS of audio



Spectral analysis block diagram; Detection of silences; Aggregation of features; Resulting data points are clustered using GMM

Results

82% speaker detection accuracy was achieved on a set of 200 audio samples each containing variable length of spoken sentences by four different speakers.

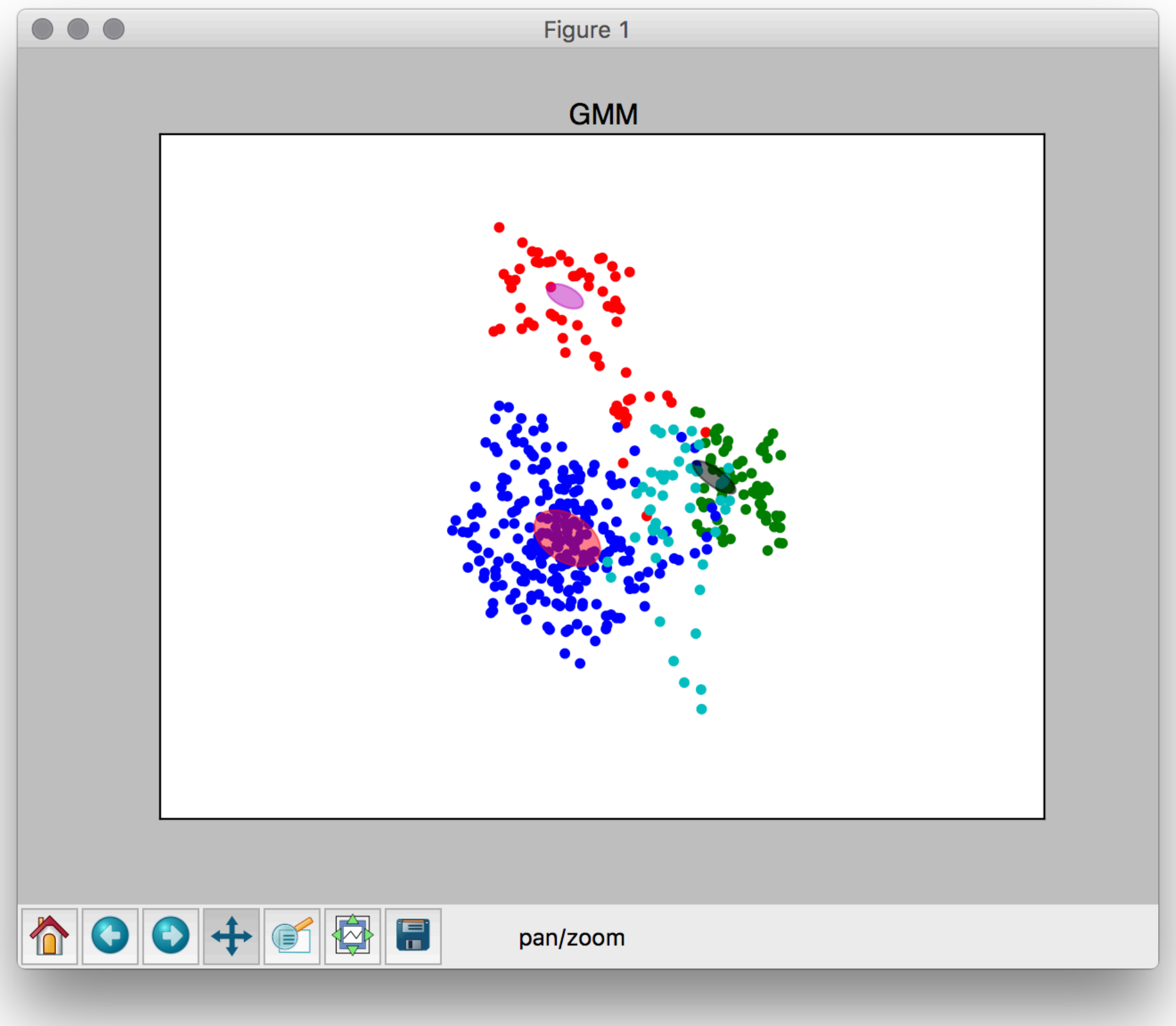
MFCCs in addition to the deviation in pitch and root mean square of audio signal gave the best separation between speakers in clustering results.

Clustering Results (plot)

The dots represent MFCC features extracted for short blocks of audio. Ground truth i.e known speakers are color-coded.

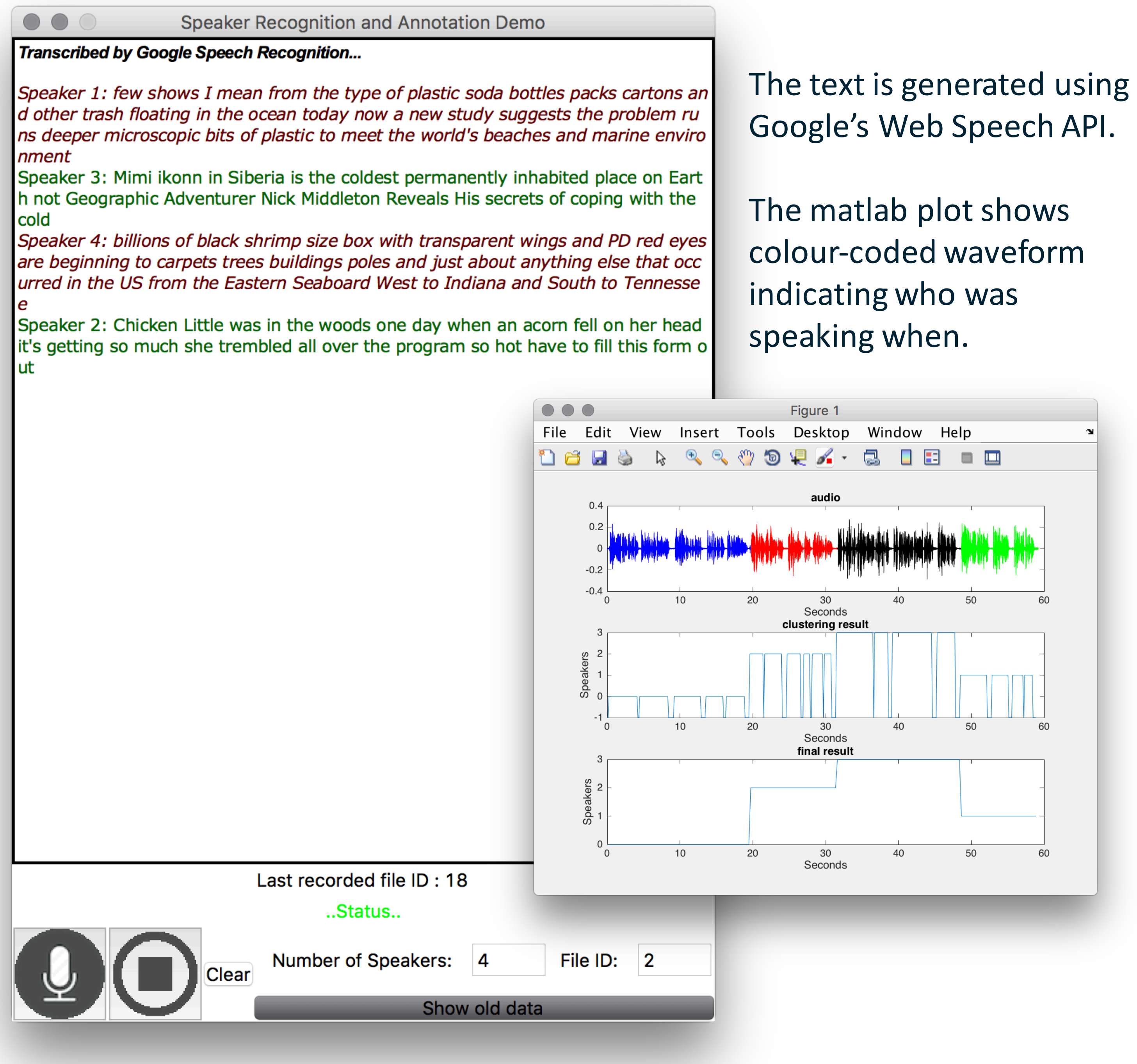
The plot shows clusters detected by Gaussian Mixture Model (using python's scikit-learn package).

Note that some speakers are well separated while there is some overlap among others.



Application

Developed a user interface in python to record audio and generate text assigned to speakers detected by the application.



Transcribed by Google Speech Recognition...

Speaker 1: few shows I mean from the type of plastic soda bottles packs cartons and other trash floating in the ocean today now a new study suggests the problem runs deeper microscopic bits of plastic to meet the world's beaches and marine environment

Speaker 3: Mimi ikonn in Siberia is the coldest permanently inhabited place on Earth not Geographic Adventurer Nick Middleton Reveals His secrets of coping with the cold

Speaker 4: billions of black shrimp size box with transparent wings and PD red eyes are beginning to carpets trees buildings poles and just about anything else that occurred in the US from the Eastern Seaboard West to Indiana and South to Tennessee

Speaker 2: Chicken Little was in the woods one day when an acorn fell on her head it's getting so much she trembled all over the program so hot have to fill this form out

Last recorded file ID : 18

..Status..

Number of Speakers: 4 File ID: 2

The text is generated using Google's Web Speech API.

The matlab plot shows colour-coded waveform indicating who was speaking when.

Discussion and Conclusion

Recognition of different speakers present in a clip of audio is possible using spectral analysis. The separation is not always clear but with normalized audio clip like a recording of a web meeting can be used to generate speaker-based text transcription using this application.

Recommendations and Future Work

- There are large differences in the way a speaker says vowels and consonants. We may be able to achieve better results by creating sub-clusters that handle that
- The results of this project are obtained from the dataset which contains normalized speech samples recorded in non-noisy environments. A lot can be learned from results obtained from real-life audio recordings which can be used to improve the methodology and get more accurate separation between speakers.
- A real time speaker recognition system
- I have to provide the number of speakers to be identified. A more challenging task can be attempted by designing a clustering algorithm to find the clusters and decide the number of speakers.
- So far we are not concerned with the natural language processing. The current speaker system can benefit from the spoken language information.

Contact Information

Avrosh Kumar
avroshk@gatech.edu

Georgia Tech