S.M.A.R.T. Project: Streaming Multi-speaker Automatic Real-time Transcription

By Y.F. Wang, Y.W. Zhang

Department of Electrical and Computer Engineering, Boston University

Introduction

Focus: Achieve the automatic caption for streaming services.

- Use multi-speaker detection, speech-to-text transcription to generate automatic transcripts.
- Use Natural-language API to achieve accurate multi-language transcription.
- Use windows console application to let user directly use our product to get transcript of user specified audio file.

Current Stage: Developed a program to be used for implementing speech-to-text and multi-speaker recognition for attach speaker tag identified to speech caption. Designed and implemented UI for app.

Future developments

Detach API:

- Design and train our own model for multi-speaker recognition and speech-to-text transcription.

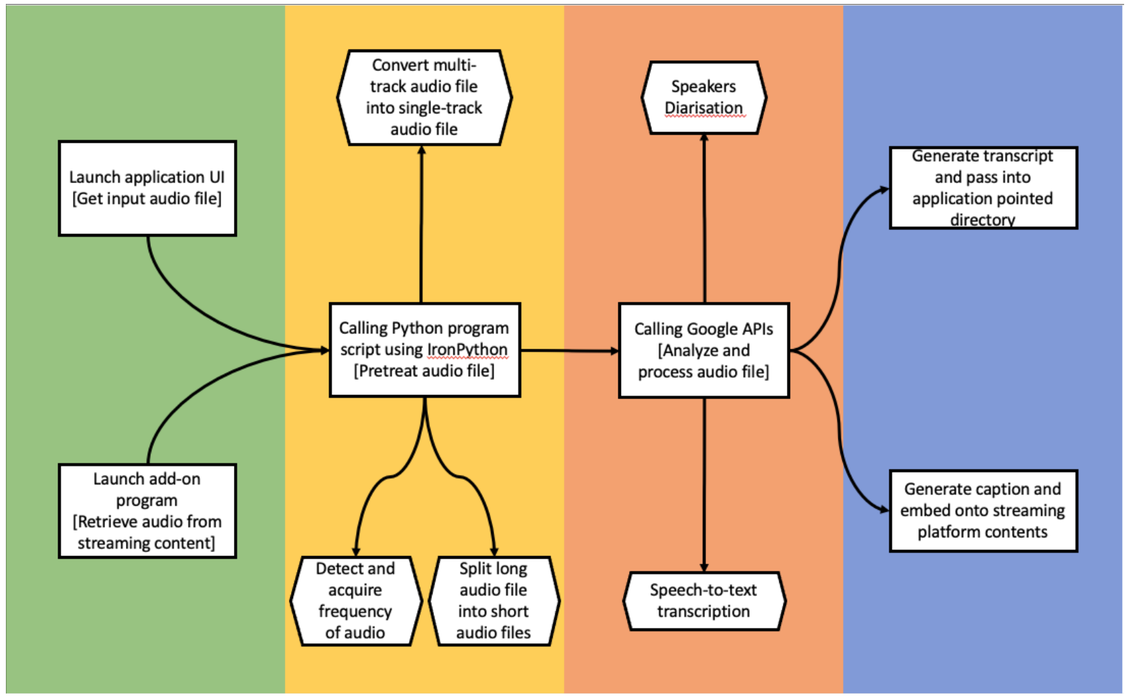
With Google APIs:

- Design and implement timer shaft to embed caption output into audio file.
- Design and implement translation use Google natural language API.
- Polish application and add-on extension version for our program.

System Components

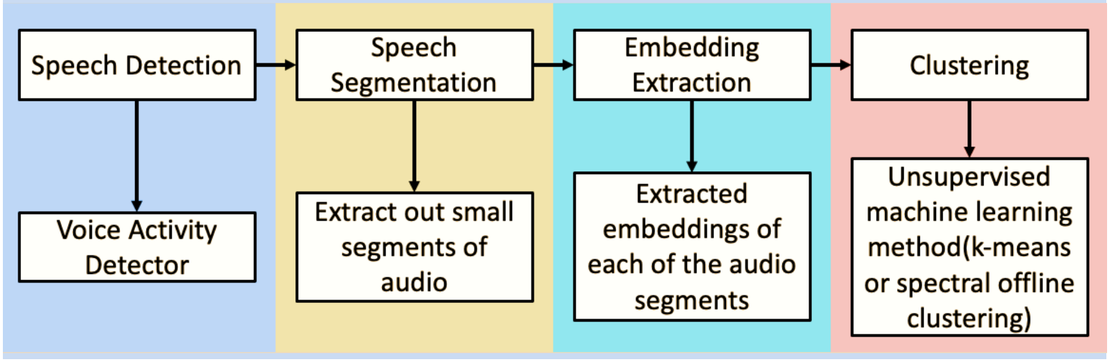
- C#
- Python
- IronPython
- Google Cloud Services
- Google Speech APIs
- Google Natural Language API
- Visual Studio

System diagram (Windows Console Application and Embed Add-on)



Method diagram

**Description of how our current using method works from input to output (arrows specified orders)*



Results

Combination of *Resemblyzer* and *Spectral Cluster* Library

```
(base) C:\ec601\Resemblyzer-master>python speaker_diarization.py
Loaded the voice encoder model on cpu in 0.03 seconds.
(156, 256)
[4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6
 6 6 6 6 5 5 5 5 5 5 5 5 5 5 5 5 5 5 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3
 3 3 3 3 3 3 3 3]
[('4', 0, 1880.0), ('0', 1880.0, 3440.0), ('1', 3440.0, 4160.0), ('6', 4160.0, 5
0.0, 7280.0), ('2', 7280.0, 9080.0), ('3', 9080.0, 10100.0)]
```

Reasons:

- Current using model training database not large and spread enough to get higher accuracy and better modulation for handling and analyzing audio.
- The characteristic vocal prints tracing by trained model are not enough for fully distinct different voices from speakers, need more data sets to get better pattern tracing model.
- Each team member trained model using different dataset to get better result, will compare and select after finishing initial training. [So far Wang's model has processed 312 data sets from 1500 sets]

**Decided to use google speech APIs for better recognizing speakers in conversations and getting more accurate results.*

Windows Console Application UI