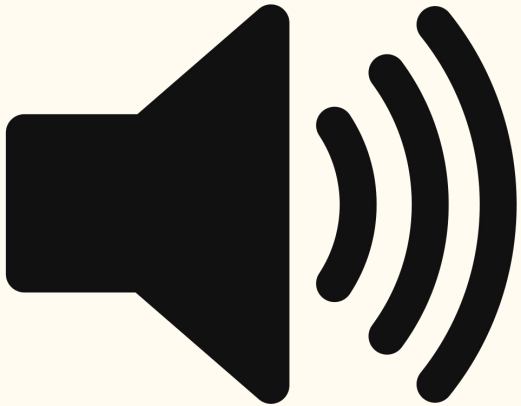
Persona - Speaker Identification System

Nicolas Stencel, Naga Sumanth, Andy Fausak, Daniel Mata

Overview

- Speaker Diarization - Segmenting audio into homogenous segments that are based on speaker identity.
- We realized that voice transcription compliments this idea perfectly and decided to pursue that as well.



[1] How are you guys doing?

[2] Can’t complain

[3] Same. What about yourself?

[1] I’m fine thank you.

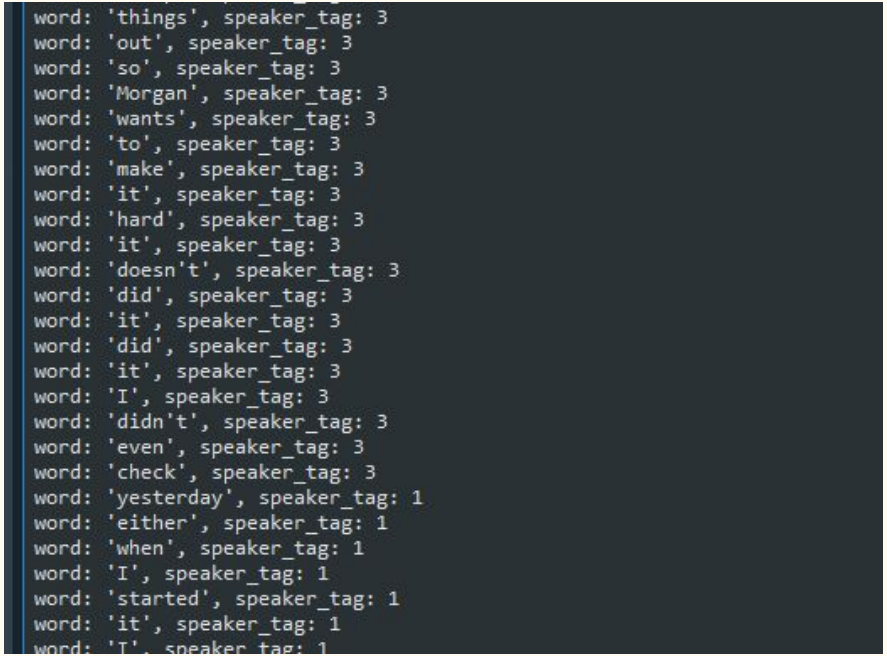
Overview

- We will be testing our system with a .wav file of a lecture that has already been transcribed. This transcription is our “true” value which we will be using to compare to the transcribed output that our system gives us.

- We are using 3 different APIs: Google, Amazon, and Azure

Google Speech to Text API

- ICSI Dataset
- Needed to enable Speaker Diarization
- Parameters we worked on:
  - Sample_rate_hertz
  - Enable_speaker_diarization
  - Diarization_speaker_count
- Default output from API



Rand Score

- Normalized output will be used as argument for the adjusted_rand_score from sklearn
- Need to deal with poor transcribing of words (“ I’m ” v.s. “ I ”, “ ‘m ”)

```
>>> from sklearn.metrics.cluster import adjusted_rand_score
>>> adjusted_rand_score([0, 0, 1, 1], [0, 0, 1, 1])
1.0
>>> adjusted_rand_score([0, 0, 1, 1], [1, 1, 0, 0])
1.0
```

Labelings that assign all classes members to the same clusters are complete be not always pure, hence penalized:

```
>>> adjusted_rand_score([0, 0, 1, 2], [0, 0, 1, 1])
0.57...
```

ARI is symmetric, so labelings that have pure clusters with members coming from the same classes but unnecessary splits are penalized:

```
>>> adjusted_rand_score([0, 0, 1, 1], [0, 0, 1, 2])
0.57...
```

If classes members are completely split across different clusters, the assignment is totally incomplete, hence the ARI is very low:

```
>>> adjusted_rand_score([0, 0, 0, 0], [0, 1, 2, 3])
0.0
```

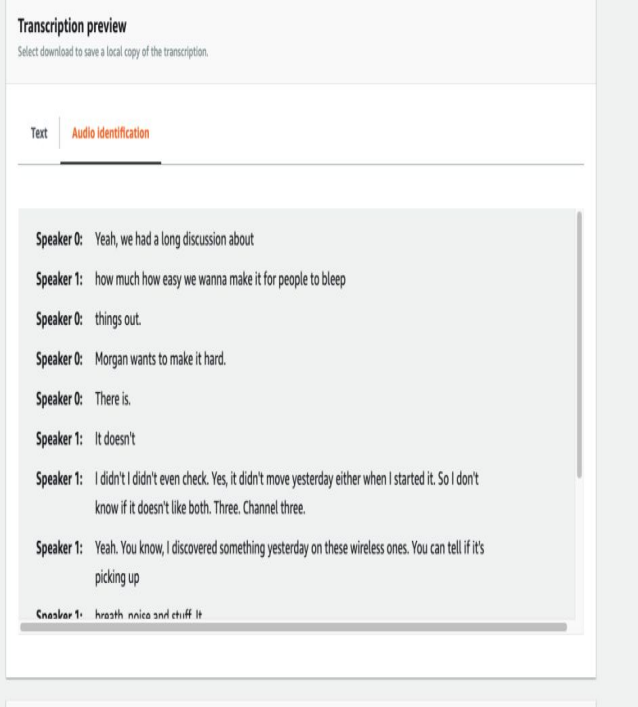
Google Speech to Text API, Normalization

- Needed to normalize our labels for Rand Score comparison
- Label based on order of initial speaker appearance

```
[3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 3, 3, 3, 3, 3, 3, 3, 3, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 3, 3, 3, 3, 3]

Normalized Labels: {3: 0, 1: 1}
[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,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0]
```

Outputs: True Label transcript, AWS Console

```
</Preamble>


<Transcript StartTime="0.0" EndTime="3021.968">

  <Segment StartTime="0.000" EndTime="3.956" Participant="me018">
    <NonVocalSound Description="mike noise"/>
  </Segment>
  <Segment StartTime="0.056" EndTime="1.861" Participant="me011">
    Yeah, we had a long discussion about
  </Segment>
  <Segment StartTime="2.674" EndTime="6.315" Participant="me011">
    how much w- how easy we want to make it for people to bleep things out. So -
  </Segment>
  <Segment StartTime="5.100" EndTime="7.881" Participant="me018">
    <NonVocalSound Description="mike noise"/>
  </Segment>
  <Segment StartTime="6.260" EndTime="6.864" Participant="fe016" CloseMic="false">
    Right.
  </Segment>
  <Segment StartTime="6.315" EndTime="7.061" Participant="me011">
    <VocalSound Description="breath"/> <NonVocalSound Description="mike noise"/>
  </Segment>
  <Segment StartTime="7.516" EndTime="8.757" Participant="fe016" CloseMic="false">
    O_K. So this is -
  </Segment>
  <Segment StartTime="8.169" EndTime="11.100" Participant="me011">
    Morgan wants to make it hard. <VocalSound Description="laugh"/>
  </Segment>
  <Segment StartTime="8.320" EndTime="13.210" Participant="me018">
    <NonVocalSound Description="mike noise"/>
  </Segment>
  <Segment StartTime="9.751" EndTime="13.549" Participant="fe016" CloseMic="false">
    The, uh, counter is not <Pause/> moving again. It -
  </Segment>
```



Todo

- Parse True labels information from ICSI dataset transcript XML file.
- We are able to get the Amazon Speaker diarization from AWS Console; we want to be able to get similar output by calling Amazon API.
- We also want to get speaker diarization output from Azure API.
- Compute Rand Score calculation between:
  1. True Label, Google API
  2. True Label, Amazon API
  3. True Label, Azure API
- Pick best 2 out of 3 cloud APIs and use it in the website.
- Website would accept *.wav* file from user, perform speaker diarization - display converted text and display Rand Score.