Unsupervised Speaker Identification

Thivakkar Mahendran
University of Massachusetts Amherst
tmahendran@umass.edu

Varun Prasad
University of Massachusetts Amherst
varunanantha@umass.edu

1. Abstract

Filled by Varun

2. Introduction

A voice recorder makes it super convenient to take notes or preserve the minutes of a meeting. In order to make the experience better, there are tools that can be used to automatically convert speech to text. One area where this tool currently fails at is identifying the speaker. The problem that we are trying to solve is to identify the speaker. Our ultimate goal is to build a project where we are able to record a meeting and which particular person speaks at a specific time and what they speak.

It would be cumbersome and time consuming to record each speaker's voice beforehand and train a speaker recognition model on the speakers' voice. The goal is to make this tool predict the speaker without prior training on the speaker's voice. This would be an unsupervised learning project.

For example, if two speakers were talking in a meeting we would like our tool to give an output:
Speaker1: Hey, How are you?
Speaker2: I'm doing well.
Speaker2: I was able to complete the tasks that we talked about last week
Speaker1: That's great.

These two speakers' voices have not been trained before but still the program would identify the two speakers just by getting the features from their voice.

3. Background/Related Work

Filled by Varun

4. Approach

4.1. Dataset

The dataset that we intially planned to use was VoxCeleb, which is a large scale audio-visual dataset of human speech. The dataset contains 7,000+ speakers, 1 million+ utterances, and 2,000+ hours of audio. The total size of the dataset was around 250 GB. The dataset was really huge in terms of computational complexity and also space required to store and train the model. After weeks of trying to use the dataset, we decided to build our own dataset.

We decided to create our own dataset so we could tailor the dataset to exactly suit our project. We built a pipeline to scrape audio from YouTube videos, and then split a whole audio into chunks of 1 second audio clips. We decided to create 1 second audio clips because when this model is getting used in the real world, we want to recognize the speaker instantly with as little of a delay as possible. According to the article [1] published by virtualspeech, an average person speaks about 150 words per minute, which is about 2.5 words per second. This is enough to extract useful features from the person's speech.

The YouTube videos that we chose to include in our dataset were speeches/monologues from celebrities. We thought this would be the best way to build a labeled dataset of audios of different people. The dataset includes 7 celebrities (Obama, Hillary, Ivanka, Trump, No Speaker, Modi, Xi-Jinping, and Chadwick-Boseman) and one "no speaker" class which includes multiple background noises without anyone speaking. We included the "no speaker" class so that the model can recognize when no one is speaking. We wanted the dataset to be as diverse as possible, that is why we included speakers of both genders, different races, and also different languages. The current dataset has a size of 4.1+ hours.

	Gender	**Language**	**Race**	**Length**
Obama	Male	English	Black	19.5 mins
Hillary	Female	English	White	57.5 mins
Ivanka	Female	English	White	17.9 mins
Trump	Male	English	White	41.6 mins
Modi	Male	Hindi	Asian	32.4 mins
Xi-Jinping	Male	Chinese	Asian	11.18 mins
Chadwick-Boseman	Male	English	Black	27.1 mins
No Speaker	N/A	N/A	N/A	39.6 mins

Table 1. Details of the dataset

5. Experiment

Filled by Varun

6. Conclusion

Filled by Varun

References

[1] Barnard, Dom. “Average Speaking Rate and Words per Minute.” VirtualSpeech, VirtualSpeech, 20 Jan. 2018, virtualspeech.com/blog/average-speaking-rate-words-per-minute.

[2] Sharma, Usha, et al. “Study of Robust Feature Extraction Techniques for Speech Recognition System.” 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015, doi:10.1109/ablaze.2015.7154944.

[3] “Analytical Review of Feature Extraction Technique for Automatic Speech Recognition.” International Journal of Science and Research (IJSR), vol. 4, no. 11, 2015, pp. 2156–2161., doi:10.21275/v4i11.nov151681.