Unsupervised Speaker Identification

Thivakkar Mahendran
University of Massachusetts Amherst
tmahendran@umass.edu

Varun Prasad
University of Massachusetts Amherst
@umass.edu

1. 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 identify which person is speaking at that specific time and what they said.

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

For example, if two speakers were talking in a meeting we would like out tool to output like:
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 speaker's voices have not been trained before but still the program would identify the two speakers.

2. Problem Statement

2.1. Dataset

The dataset that we 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 (1) virtualspeech.com, 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. At the end of the project, we are planning to experiment on the seconds of each audio clip to see its effect on the accuracy of the model.

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. To start with, the dataset includes 7 celebrities (Obama, Hillary, Ivanka, Trump, No Speaker, Modi, Xi-Jinping, and Chadwick-Boseman) and 1 "no speaker" class which includes multiple background noises. 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	Indian	32.4 mins
Xi-Jinping	Male	Chinese	Chinese	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

We are planning to add more diverse speakers to make the model predict better. We will experiment with the dataset to see the impact on the model. For example, we will train the model on all male voices or only english as the language to see the impact on the model.

2.2. Expected Result and Evaluation

The project is split into two parts, the first part would be training the model to recognize and predict the 8 differ-

ent speakers from the dataset. The second part would be to modify the model to recognize different speakers who have not been trained before. We want the model to differentiate different speakers and label each speaker as "speaker#". A python script has been created to get a live stream of audio data from the computer, process every 1 second audio clip by extracting features on the audio and then using the features to predict the speaker by using the model. So every 1 second, the program will output the current speaker or "no speaker" if no one is talking.

For the first part of the project we will be performing the five-fold cross validation to train and evaluate the model. We will be evaluating the model by getting the accuracy, precision, and recall. We would like the model to perform at least 90% accuracy on the test dataset so the model would work better in the second part of the project.

Evaluating the second part of the project would be slightly harder because we don't have a labeled dataset of speakers talking at every timestamp. We either need to build our own dataset to make it work for this specific use case or we have to manually evaluate the model. One possible way to evaluate is to play an interview video between two or more people and physically check if the model predicts the speaker speaking at that very point of time.

We are expecting several factors to impact the accuracy of the model. For example, the speaker's gender, age, race, language, dialect of a language could affect the accuracy of the model. Multiple speakers speaking at the same time might also impact the accuracy of the model. These are some things that we are interested in evaluating. These results would help us improve the model in the future based on its weakness.

3. Technical Approach

4. Preliminary Results

References