Group member: Jiadai Yu, Shu Wang, Yulu Jiang

Research Topic

Speaker Classification (Questrom professor edition)

Predictors: Voice features, Labels: Professor names

Objective

The prediction task is to recognize the speaker through classification. Our raw data source will be the lecture recordings from the courses we had these two semesters. The model will be trained to pre-process input speech, extract audio features of professors, and perform classification tasks. We aim to develop a model that can learn the unique voice feature of each professor, then extract features of the test audio and recognize the speaking individual.

Fig. 1: Basic model of speaker classification

We primarily focus on speaker recognition, also known as speaker identification and verification in many research. We aim to identify the speaker by extraction, characterization and recognition of the audio. Our project does not aim to perform speech recognition, which is to recognize the words said in the speech and convert to text information.

Motivation

Since we have not touched upon any video and audio data type in our previous projects before, we are thrilled to undertake this challenge to discover a new aspect of machine learning. We will be using some brand new Python packages to broaden our learning.

We specifically choose the data source as Echo360 lecture recordings of Questrom professors because of high data accessibility. We will also ensure sufficient prior communication with the recording owner to address potential ethical issues.

This project is tailored to our learning experience as these speakers, i.e. professors, are someone we are used to talking with every day. We wonder if machine learning algorithms will produce different prediction results from ours when trying to classify the speaker of an audio clip. It will also be interesting to see if some professors share certain voice features in common that we may not notice.

Novelty

The innovation of our project lies in the data collection, data processing, and neural network model training. The fact that our dataset is not readily available online requires our effort to perform challenging tasks in creative means, such as web scraping.

Firstly, the data is collected from the video recordings of courses in our program. Then, we use Python to preprocess the data, converting the raw video into audio, then segment into several structured audio clips with manual labelings. We further extract the sound features from the audio clips and finally utilize

a neural network for deep learning.

We will also do extensive research on aspects of speaker recognition and audio processing. This may include the best-performing classifiers and network structure, and some features we should consider.

We aim to be flexible and adaptive in this project, especially with this novel topic. As some challenges are foreseen ahead, we can adjust our goal slightly if some issues cannot be resolved. We will also actively seek support from the professor and teaching assistant.

Implication

A direct implication for this project is that it serves as an academic tool for Questrom students to assist and improve their learning experience. With the speaker recognition technology, Echo360 could differentiate the speaker in the recording, hence helping students navigate when watching the capture.

Moving further, if we have access to a larger dataset, not limited to Questrom professors, this task will provide insights into **security authentication**. By verifying the speaker identity, the system can provide a reliable authentication that only authorized individuals are granted access to sensitive information.

Human voice characteristics, like fingerprints, are **biometrics** unique to each individual. An accurate speaker classification model can be used to confirm the identity of the speaker, which can assist criminal investigations. Based on on-site audio-visual data, this method can help identify suspects more quickly than traditional methods, such as fingerprints, hair, or blood tests.

… Also for **Natural Language Processing** and **voice-controlled interfaces**.

Data Source

Our dataset will be sourced from lecture recordings from Echo360 in our blackboard.

Ideally, we will download the lecture recordings and convert each lecture recording (normally 2-3 hours long) into short audio clips (15-20 seconds long each). Given the available recordings, we estimate our sample size being 1,000. Each audio (observation) will be matched with a label indicating the name of the professors (from 0 to 9).

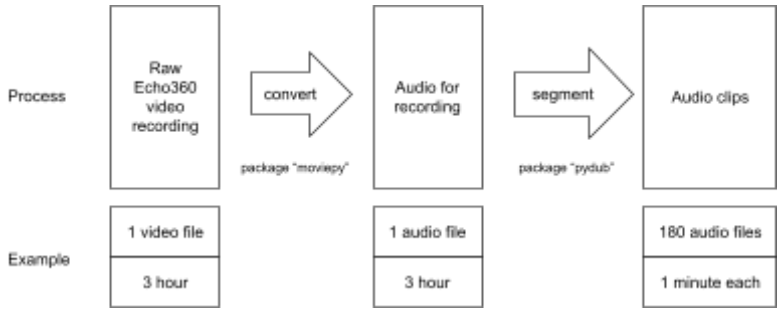

Fig. 2: Data pre-process flow chart

Data Sample

We were not able to directly download the video from Echo360 at this time due to security constraints of the platform. If allowed, we will proceed using screen recording, or contact BU IT Services for permission.