GMM-Based Speaker Identification

Stefan Mijalkov
ECE Undergraduate

Lohith Muppala
ECE Undergraduate

***Abstract*—Creating a system for Speaker Identification using Gaussian Mixture Models (GMM).**

I. Introduction

The focus of this project is to create a Speaker Identification system with Gaussian Mixture Models. The system is capable of recording voices, creating models and identifying speakers with accuracy of 90%.

II. Method Description

Gaussian Mixture Models is a common technique for speaker identification. To get a fully functional system, we prompt the user to record 5 voice recordings, each 10 seconds long. Mel Frequency Cepstral Coefficients (MFCC's) are extracted from the voice and used to model a distinct GMM model for each speaker. The models are saved in a directory and used in the recognition phase. A pre-trained garbage model is used for result improvements.

III. Experiment Results and Evaluation

With the initial run, the program creates the required directories: *audio_database* and *gmm_models.* The stored audio files in the directory *garbage* are read, a garbage model is created and saved in the *gmm_models* directory. The program then prompts the user to chose between 3 options: *Add a new person to the database, Identify person,* and *Exit* as shown in Fig 1.1:

```
################################################################
###            GMM-based speaker identification system        ###
################################################################
Directory name already exists
Directory name already exists
c:\Users\hazar\Desktop\SpeakerIdentification\garbage\a0014.wav
c:\Users\hazar\Desktop\SpeakerIdentification\garbage\a0221.wav
c:\Users\hazar\Desktop\SpeakerIdentification\garbage\a0315.wav
c:\Users\hazar\Desktop\SpeakerIdentification\garbage\a0499.wav
c:\Users\hazar\Desktop\SpeakerIdentification\garbage\b0159.wav
c:\Users\hazar\Desktop\SpeakerIdentification\garbage\q44.wav
c:\Users\hazar\Desktop\SpeakerIdentification\garbage\quiet0.wav
c:\Users\hazar\Desktop\SpeakerIdentification\garbage\quiet1.wav
c:\Users\hazar\Desktop\SpeakerIdentification\garbage\quiet3.wav
c:\Users\hazar\Desktop\SpeakerIdentification\garbage\quiet4.wav
+ Modeling completed for speaker: garbage  with data point =  (12224, 40)
------------------------------------------------------------

1. Add a new person to the database
2. Identify person
3. Exit
```

Fig 1.1 – User prompts

Training phase:

The user is required to speak for 50seconds in total, (5 recordings, 10 seconds long). Once done, the recordings will be saved in the *audio_dataset's* subdirectory with the speaker's name. The program reads the saved audio files and extracts Mel Cepstral Coefficients (MFCC's). The process is shown below in figure 1.2.

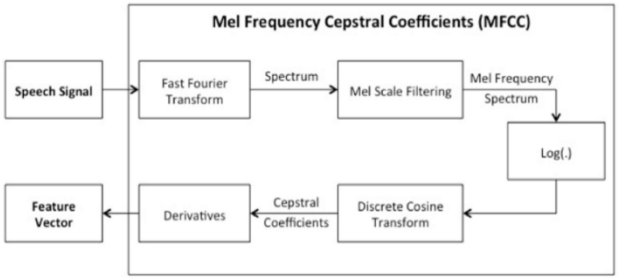

Figure 1.2 – Extracting MFCC

The extracted MFCC vectors are used to model a GMM. The models are saved in the *gmm_models* directory containing the name of the speaker.

garbage.gmm	11/29/2020 9:42 PM	GMM File
josh.gmm	11/29/2020 9:44 PM	GMM File
simona.gmm	11/23/2020 8:39 PM	GMM File
stefan.gmm	11/24/2020 3:54 PM	GMM File
tatjana.gmm	11/23/2020 8:50 PM	GMM File

Fig 1.3 – Saved GMM models

Testing phase:

In the testing phase, the user is required to record his/her voice for 10 seconds. The MFCC's are extracted from the current speech recording and fed into the preexisting models. Log-likelihoods are calculated for each model, and the speaker with maximum log-likelihood is the predicted speaker. Figure 2.1 shows the predicted speaker with Stefan's voice and Fig 2.2 shows the predicted speaker with random non-speech noise.

```
Speakers found in database:  ['garbage', 'simona', 'stefan', 'tatjana']
Press ENTER to record a 10sec long audio...
* recording
* done recording
        Detected as -  stefan
        Log likelihoods: [-40.65423493 -41.20790652 -36.61155513 -50.60601547]
------------------------------------------------------------
```

Fig 2.1 – Predicted speaker with recorded speech.

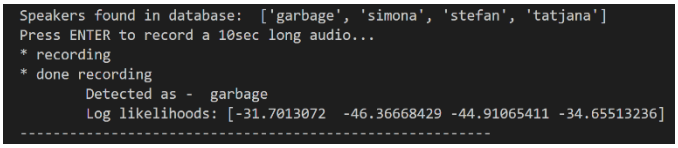

Fig 2.2 – Predicted speaker with recorded noise

Accuracy:

To test the accuracy, a dataset [4] of 34 distinct speakers was used containing 5 recordings per speaker. We achieved accuracy of 91.18%.

```
Correct!        Detected as -  belmontguy
belmontguy-20110426-geu\wav\b0156.wav
Correct!        Detected as -  belmontguy
belmontguy-20110426-geu\wav\b0157.wav
Correct!        Detected as -  belmontguy
belmontguy-20110426-geu\wav\b0158.wav
Correct!        Detected as -  belmontguy
Correct guesses:  155
Wrong guesses:  15
Accuracy:  91.18 %
```

Fig 3.3 – Accuracy and performance

IV. Conclusions

GMM modeling proves to be an effective approach for speaker identification with small data. Accuracy of 91.18% is very high and acceptable for applications that are not dealing with highly advanced security systems. To get a higher accuracy we could have included the pitch period which will make a big difference when distinguishing between male and female speakers. To get accuracy of >98% Deep Neural Networks can be used.

V. Contributions

Both group members worked equally on the project. Stefan mainly worked on the code that deals with extracting MFCC coefficients and modeling GMM's. Lohith mainly worked on the structure of the program, the directory creation, and the part that deals with voice recording and storage.

Both group members participated in each part of the project and put equal effort.

V. References

[1] Kumar, A. (2019, July 01). Spoken Speaker Identification based on Gaussian Mixture Models : Python Implementation. Retrieved November 26, 2020, from https://appliedmachinelearning.blog/2017/11/14/spoken-speaker-identification-based-on-gaussian-mixture-models-python-implementation/

[2] Voice recording using pyaudio. Retrieved November 26, https://stackoverflow.com/questions/40704026/voice-recording-using-pyaudio

[3]http://cs229.stanford.edu/proj2017/final-posters/5143660.pdf

[4] https://www.dropbox.com/s/87v8jxxu9tvbkns/development_set.zip?dl=0

V. Apendix