Speaker Identification Over Audio and Video

Anita Kumari Rawat
2021aim1002@iitrpr.ac.in
Department of CSE
IIT Ropar

Srinivas P
2021csm1014@iitrpr.ac.in
Department of CSE
IIT Ropar

Dr. Mukesh Saini
mukesh@iitrpr.ac.in
Department of CSE
IIT Ropar

Abstract— **We aim to generate unsupervised speaker models in the audio domain and face models in the video domain from videos. Such models can later be utilized for speaker identification in the audio domain (answering the question ”Who was speaking and when”) and face recognition (answering the question ”Who was seen and when”) for given videos of speakers presentation. Our major goal is to construct unsupervised speaker models in the audio domain and face models in the video domain, so that the models may later be used for audio-video person recognition. In contrast to most other existing systems, our method generates slightly more speaker-model candidates than the actual number of speakers. The rationale for this is that it is more vital that no two separate speakers have the same identity than that each speaker has only one.**

Index Terms— **audio-video speaker identification, audio speaker recognition, and face recognition.**

I. OVERVIEW

In this project, we have divided the idea into two parts, i.e., video and audio.

A. Video

The human face is a crucial factor in identifying a person, i.e., even identical twins have distinctive looks. So, to remember one another, facial recognition and identification is necessary. The human face is a crucial factor in determining a person. Even identical twins have distinctive looks. [1] To remember one another, facial recognition and identification are necessary.

Therefore to verify someone’s identification using biometrics, a facial recognition system is used. Nowadays, a lot of applications, including phone unlocking systems, criminal identification systems, and even home security systems, use face recognition as a standard technique. [2] Due to the fact that this method just requires a face image instead of other dependencies like a key or token, it is more secure. The two steps of a person recognition system are face detection and face identification.

The approach used for face detection:

- Training the model
- Face detection
- Face encoding

B. Audio

There is a growing demand for approaches that automatically evaluate digital content since audio information is more crucial. One of the critical areas of voice signal-based research is speaker identification. Other significant areas of its application include Speech Recognition, Speech-to-Text Conversion, and so on. [3] A crucial element in accomplishing Speaker Identification is the Mel Frequency Cepstral Coefficient (MFCC).

The most often used model for training on our data is the Gaussian Mixture Model (GMM), but many valuable models may be used for the training job; one such model is Hidden Markov Model (HMM). Recently, deep learning, particularly artificial neural networks, has been used for the majority of the model training phase for a speaker recognition project (ANN). We have mainly used the pairing of MFCC with GMM. As the primary feature, we took into account MFCC with ”tuned parameters,” while the second feature was delta-MFCC, and finally, we used GMM to train our model with a few tweaked parameters.

The approach used for face detection:

- Noise reduction and Silence Removal
- Feature Extraction
- Training the model
- Testing the model

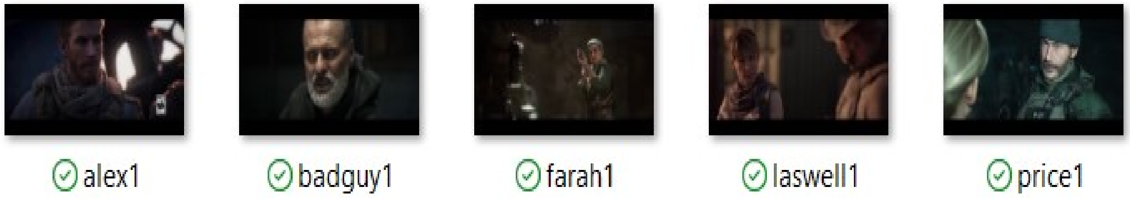


Fig. 1: Input images of the person for whom we will create a separate video clip

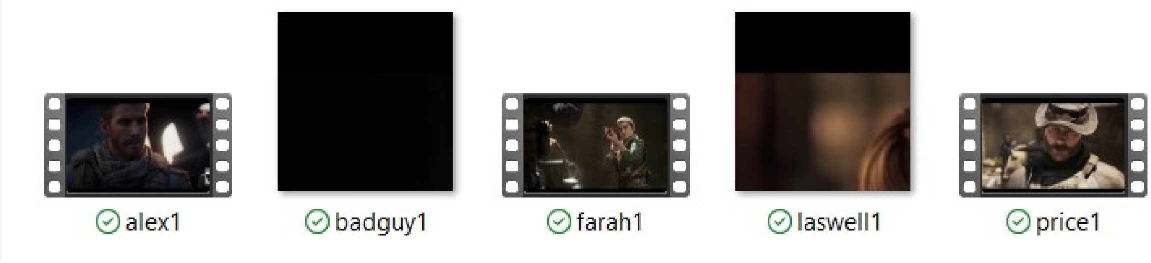


Fig. 2: output of a separate video for each person

II. Approach

A. Video

We have included a video clip and some photographs of the people who appeared in it in the video section. Our primary goal is to identify the people in the video clip so that we may create a distinct video for each of them. To finish this procedure, we first trained the people to extract each person's encoding. Then, using the built-in face recognition library, we divided the video into pictures of frames by comparing it with the person encoding values. After archiving the images, group the photos according to their names before merging all the shots of a particular individual and turning them into films.

B. Audio

Similarly, for the audio component, we have segregated a few person-specific audio samples into test and training datasets. We have mainly employed the MFCC and GMM combination. To train our model, we will first extract the MFCC features from the audio and apply GMM with a few modified parameters. The method will then be performed against the test data set to see if it matches the training models, and if it does, it will output the speaker name that does.

III. Dataset

A. Video

In order to test the effectiveness of our algorithm, we used a Cod game trailer that lasted 2 minutes and 35 seconds, which took a lot of time. So, we altered and included the 6 second video of the trailer in our algorithm. Later, we chose five random people who appeared in the video in order to extract the video clip of a specific individual from it.

- Train: 5 people who appear in the video
- Test: 6 sec video clip and ran partially on 2min 35 sec video clip

B. Audio

We've selected five speaker data sets from the Kaggle data set, which included information on about 50 distinct speakers, and used them for training the audio component. After then, we collected data for all five of these speakers (around 35 samples), not all of them equally distributed, to verify our model's accuracy and efficiency.

- Train: 5 speaker audio data
- Test: 35 audio samples

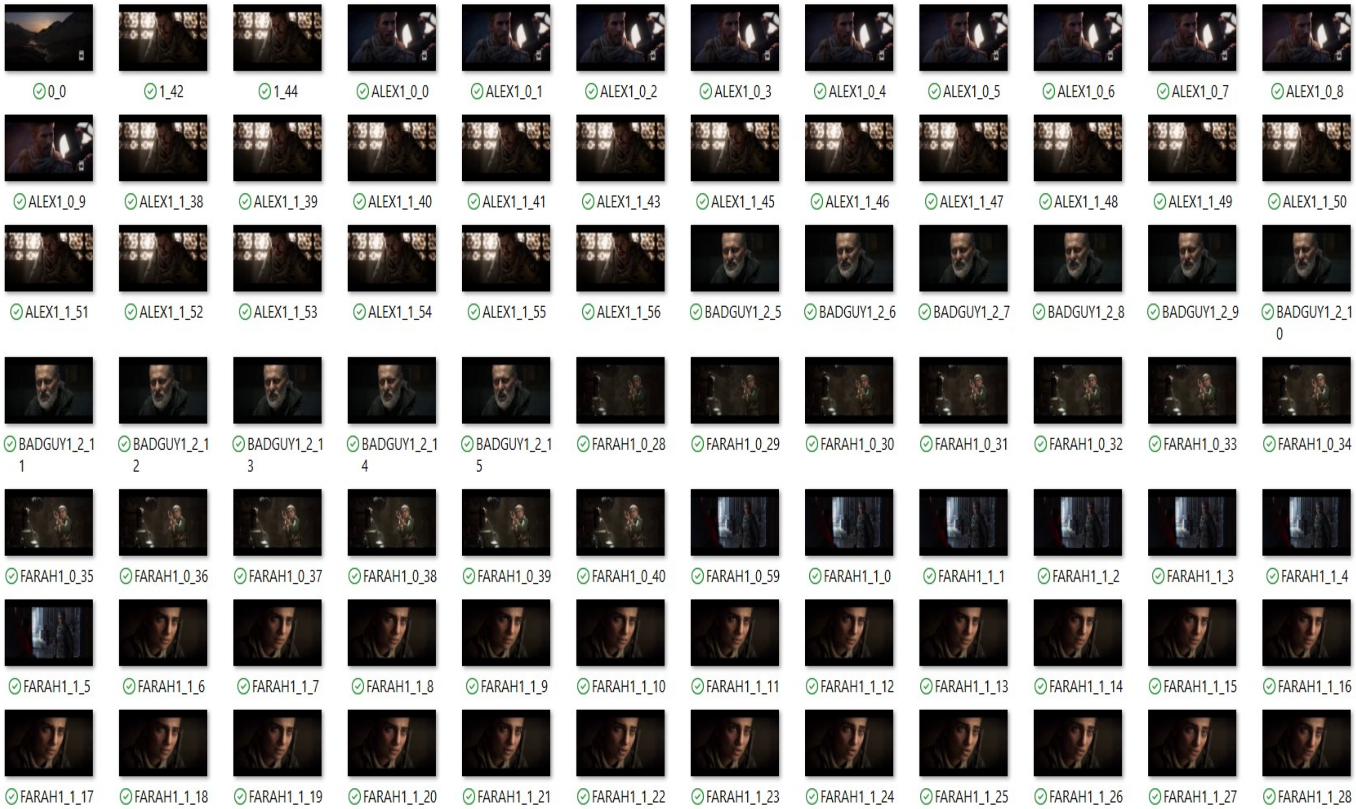
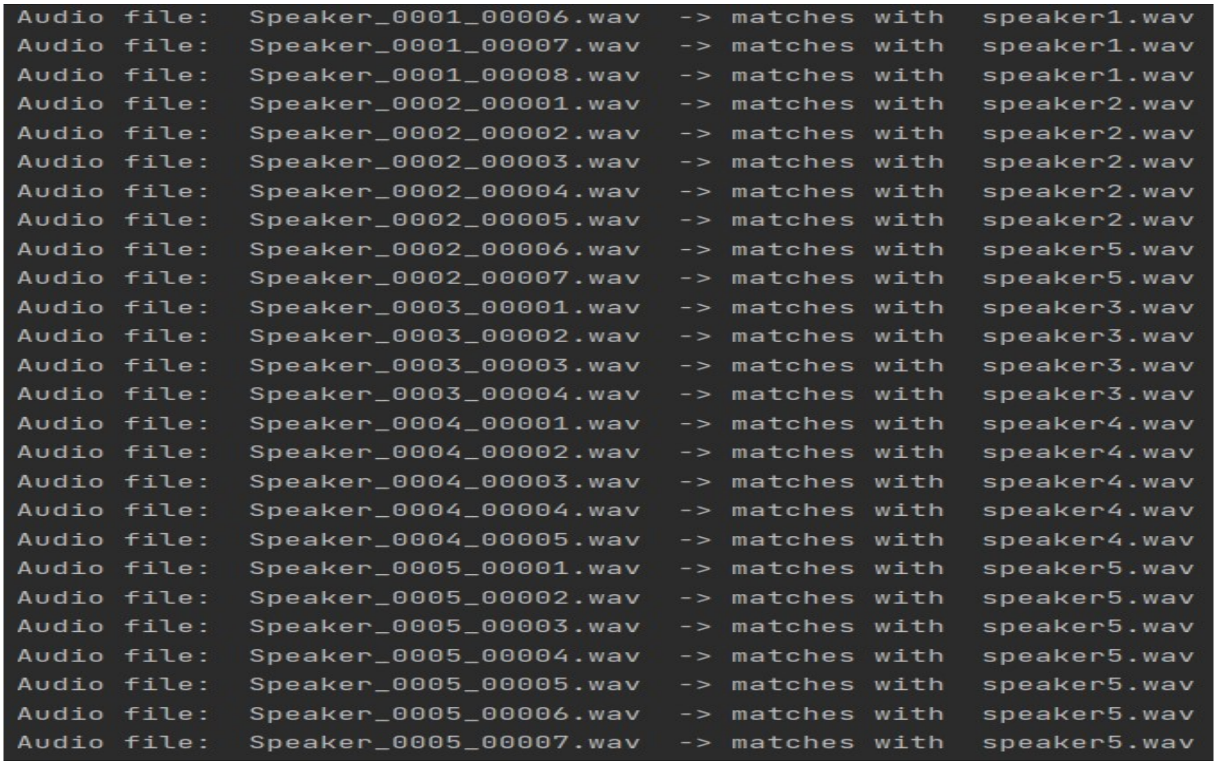


Fig. 3: face recognised and saved with their names

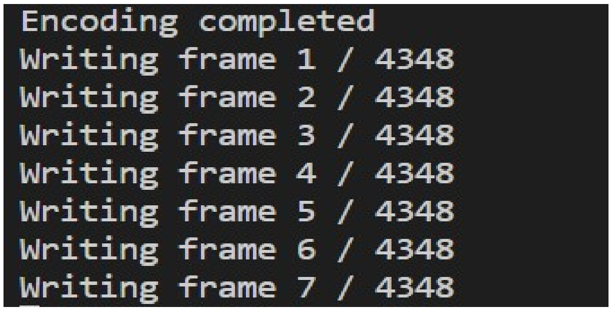


```
Audio file:  Speaker_0001_00006.wav  -> matches with  speaker1.wav
Audio file:  Speaker_0001_00007.wav  -> matches with  speaker1.wav
Audio file:  Speaker_0001_00008.wav  -> matches with  speaker1.wav
Audio file:  Speaker_0002_00001.wav  -> matches with  speaker2.wav
Audio file:  Speaker_0002_00002.wav  -> matches with  speaker2.wav
Audio file:  Speaker_0002_00003.wav  -> matches with  speaker2.wav
Audio file:  Speaker_0002_00004.wav  -> matches with  speaker2.wav
Audio file:  Speaker_0002_00005.wav  -> matches with  speaker2.wav
Audio file:  Speaker_0002_00006.wav  -> matches with  speaker5.wav
Audio file:  Speaker_0002_00007.wav  -> matches with  speaker5.wav
Audio file:  Speaker_0003_00001.wav  -> matches with  speaker3.wav
Audio file:  Speaker_0003_00002.wav  -> matches with  speaker3.wav
Audio file:  Speaker_0003_00003.wav  -> matches with  speaker3.wav
Audio file:  Speaker_0003_00004.wav  -> matches with  speaker3.wav
Audio file:  Speaker_0004_00001.wav  -> matches with  speaker4.wav
Audio file:  Speaker_0004_00002.wav  -> matches with  speaker4.wav
Audio file:  Speaker_0004_00003.wav  -> matches with  speaker4.wav
Audio file:  Speaker_0004_00004.wav  -> matches with  speaker4.wav
Audio file:  Speaker_0004_00005.wav  -> matches with  speaker4.wav
Audio file:  Speaker_0005_00001.wav  -> matches with  speaker5.wav
Audio file:  Speaker_0005_00002.wav  -> matches with  speaker5.wav
Audio file:  Speaker_0005_00003.wav  -> matches with  speaker5.wav
Audio file:  Speaker_0005_00004.wav  -> matches with  speaker5.wav
Audio file:  Speaker_0005_00005.wav  -> matches with  speaker5.wav
Audio file:  Speaker_0005_00006.wav  -> matches with  speaker5.wav
Audio file:  Speaker_0005_00007.wav  -> matches with  speaker5.wav
```

Fig. 4: output of speaker identification over audio

IV. Results

A. Video

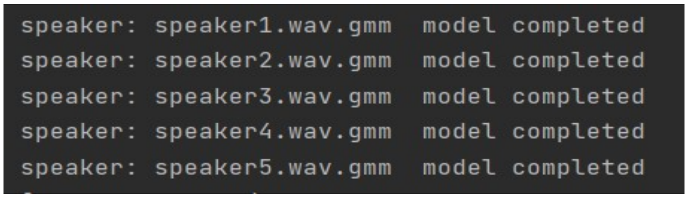


```
Encoding completed
Writing frame 1 / 4348
Writing frame 2 / 4348
Writing frame 3 / 4348
Writing frame 4 / 4348
Writing frame 5 / 4348
Writing frame 6 / 4348
Writing frame 7 / 4348
```

Fig. 5: dividing the input file into frames

B. Audio

For the audio dataset, we collected 35 samples. Thirty-three of these samples are accurately categorised. The accuracy of this model is 94.28



```
speaker: speaker1.wav.gmm  model completed
speaker: speaker2.wav.gmm  model completed
speaker: speaker3.wav.gmm  model completed
speaker: speaker4.wav.gmm  model completed
speaker: speaker5.wav.gmm  model completed
```

Fig. 6: model developed for each speaker

V. Conclusion and Future work

We developed speaker identification across audio and video independently in this project. Additionally, we may combine the two methods to extract a given speaker's output clip from any video source. For future work, in-order to increase the accuracy of the method, we can apply several algorithms and add additional parameters.

References

[1] P. Campr, M. Kunešová, J. Vaněk, J. Čech, and J. Psutka, "Audio-video speaker diarization for unsupervised speaker and face model creation," in *International Conference on Text, Speech, and Dialogue*, pp. 465–472, Springer, 2014.

[2] J. Bi and S.-C. Liu, "A speaker identification system for video content analysis," in *2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 200–203, IEEE, 2008.

[3] S. R. Arshad, S. M. Haider, and A. B. Mughal, "Speaker identification using speech recognition," *arXiv preprint arXiv:2205.14649*, 2022.