Who Am I? Real Time Speaker Identification using Python

Project Brief:

Aim: The goal of this project is to develop a model that can correctly identify, in real time, the speaker (among 4 group members) by using 5 seconds of his/her audio. Group members include 3 boys (Yash, Rahil and Srihari) and 1 girl (Shikha).

Application: Includes various real-world applications such as non-biometric authentication, transcript generation (phone calls, virtual calls, etc.) , forensic analysis and more.

Methodology: For the purpose of this project, we defined a function called *'extract_features'* which converts the given waveform input file on the 'mel' scale and is used to derive MFCC, and chroma factor, stacked on top of each other horizontally. This gives rise to a numeric array, which is fit for model training. Various machine learning models were trained, tuned and tested.

The most vital step of the training process was creating a custom dataset of unique sentences recorded by the group members. Initially, a dataset was created of members speaking numbers, and singular words. However, the model was not trained well enough. Hence, the database was later shifted to contain entire sentences, 20 per person to be precise. 15 used for training, 5 for testing. Each voice sample was stored and inputted in a .wav format at a frame rate of 44.1kHz, as a mono channel input.

For creating the database, a quiet and placid environment was chosen and the recordings were made using a mobile device – 'Samsung Galaxy S20 FE', using the default voice recorder which is the Samsung Voice Recorder application. For further elimination of noise and other disturbance, the application was used in it's 'Interview' mode, which automatically detected sounds from one of the two microphones and processed it to be as clear as possible. The sounds recorder had to be further polished with software's like Adobe Audition and Audacity, where noise cancellation, compression format, frame rate and pitch equalization were performed. For each of the 4 members in the team, a total of 20 samples were collected, of which 15 were used for training purposes while the rest were used for testing. The file name convention was defined as: *speaker-speaker_code-sentence_number-.wav* which made it easier during the creation of the stacked array to be fit into models. The codes for speakers were Rahil - 1, Shikha - 2, Srihari - 3, and Yash - 4. For example, *Srihari-3-12.wav* implies that the speaker is Srihari, his code (3) and the sentence number (12 in this case).

Results: Using the Multi-layer Perceptron, we achieved an accuracy of 100% on the test set, as the model accurately identified all 20 speakers (4 people, 5 test samples each). Despite the almost perfect accuracy, challenges still persist when trying to implement the same in a 'real-time' setting. This is because the same amount of pre-processing (using professional mechanisms like Audacity and Adobe Premier Pro) can't be performed on the received audio samples, and we will have to rely on Python and its libraries to natively clean the files. This increases the possibility of a misjudged classification, but there are measures that can be taken to reduce the chances of that happening.

Firstly, using a premium quality sensor microphone, with additional tools for eliminating background sounds (such as a dead cat/ Anti- wind foam Cap) can massively elevate the quality of sound input given to the model, and thereby boosting chances of a better prediction. In addition, various python libraries can be used to further enhance the standard of sound fed to the model, which will ultimately yield better results. As an example of a real-time prediction, an instance has been shown below where the model accurately predicted the speaker.

Future Scope: Immense potential in healthcare. There is a field of research known as "vocal biomarkers" or "voice biomarkers" that explores the idea that certain characteristics of a person's voice may provide information about their health, including aspects related to various diseases.

Python code inspired by mini project on 'Speech Emotion Recognition' - DataFlair