Text Independent Automatic Speaker Recognition System with M-GMM

Vaishali¹, Jubby Rani², Amit Kumar³

¹,² M.Tech Student, ECE Department, KITM, KUNJPURA (Karnal), Kurukshetra University, Kurukshetra
³ Asst. Professor ECE Department, KITM, KUNJPURA (Karnal), Kurukshetra University, Kurukshetra

Abstract—Speaker Recognition Systems are a newly emerging field in the biometric security systems. Recognition on the basis of voice is a difficult task as even humans are not very good at recognizing people with similar voice textures or that of mimicry artists that can change their voice texture accordingly. Further, there is a great application of speaker recognition systems, like in high security regions and with non-invasive biometric detection; speech is the only way to get biometric information of a person. Thus, speaker recognition systems have a wide research scope. Many different methods have been proposed to perform speech recognition and most of them have weak detection rate in certain circumstances. In this thesis, we have developed an improved Speaker Recognition system based on a text independent modified Gaussian Mixture Model (m-GMM). The modified GMM has a polynomial modeling scheme along with DCE based Mel-frequency Cestrum Coefficients (MFCCs) which is an improved model of MFCC for noisy scenario. The DCE is applied to reduce the correlation in a similar speaker scenario (not same) and thus allow a much robust speaker recognition in a complex mimicked speaker scenario also. The experimental code is developed in Matlab and tested by taking different database size and the results are presented.

I. INTRODUCTION

1.1. Speaker Recognition

Speaker recognition is the identification of a person from characteristics of voices (voice biometrics). It is also called voice recognition. There is a difference between speaker recognition (recognizing who is speaking) and speech recognition (recognizing what is being said). Speaker recognition involves the speaker identification to output the identity of the person most likely to have spoken from among a given population or to verify a person’s identity who he/she claims to be from a given speech input. While finger prints and retinal scans have been usually considered to be reliable ways of authenticating people, voice identification has the convenience of easy data collection over telephone. Extraction of optimum features depicting the variations in speaker characteristics also influence the accuracy.

1.1.1 How Speaker Recognition Works

Speaker Recognition has been a simpler task, since it only requires comparison between test pattern and one reference template and involves a binary decision of whether to accept or reject a speaker.

Basically identification or authentication using speaker recognition consists of four steps:

1. voice recording
2. feature extraction
3. pattern matching
4. decision (accept / reject)
Depending on the application a voice recording is performed using a local, dedicated system or remotely (e.g. telephone). The acoustic patterns of speech can be visualized as loudness or frequency vs. time. Speaker recognition systems analyze the frequency as well as attributes such as dynamics, pitch, duration and loudness of the signal.

During feature extraction the voice recording is cut into windows of equal length, these cut-out samples are called frames which are often 10 to 30 ms long.

Pattern matching is the actual comparison of the extracted frames with known speaker models (or templates), this result in a matching score which quantifies the similarity in between the voice recording and a known speaker model. Pattern matching is often based on Hidden Markov Models (HMMs), a statistical model which takes into account the underlying variations and temporal changes of the acoustic pattern. Alternatively Dynamic Time Warping is used; this algorithm measures the similarity in between two sequences that vary in speed or time, even if this variation is non-linear such as when the speaking speed changes during the sequence.

1.2. Technology

The various technologies used to process and store voice prints include frequency estimation, hidden Markov models, Gaussian mixture models, pattern matching algorithms, neural networks, and matrix representation, Vector Quantization and decision trees. Some systems also use "anti-speaker” techniques, such as cohort models, and world models.

Ambient noise levels can impede both collections of the initial and subsequent voice samples. Noise reduction algorithms can be employed to improve accuracy, but incorrect application can have the opposite effect. Performance degradation can result from changes in behavioral attributes of the voice and from enrolment using one telephone and verification on another telephone ("cross channel"). Integration with two-factor authentication products is expected to increase. Voice changes due to ageing may impact system performance over time. Some systems adapt the speaker models after each successful verification to capture such long-term changes in the voice, though there is debate regarding the overall security impact imposed by automated adaptation.

Capture of the biometric is seen as non-invasive. The technology traditionally uses existing microphones and voice transmission technology allowing recognition over long distances via ordinary telephones (wired or wireless).

Digitally recorded audio voice identification and analogue recorded voice identification uses electronic measurements as well as critical listening skills that must be applied by a forensic expert in order for the identification to be accurate.

II. PROBLEM FORMULATION

During the past few decades, the verification problem has been attracting considerable research attention in the speech research community. The verification problem encompasses all problems which require a binary answer: yes or no. In speech technology, speaker verification and utterance verification are two most active areas due to their increasing importance in many practical
applications. In speaker verification, based on a user’s voice, the goal is to make the decision of whether to accept or to reject the identity claimed by the speaker.

It is found that most of the researches are concerned with speaker identification systems and speaker dependent speech recognition systems. In both cases the efficiency of the SRS is concerned with training the system by the particular user’s speech samples to be recognized by them. This kind SRS’s can be operated by only users who had trained the system previously, even after the success of training, the correct recognition is impacted due to the noisy environment. To overcome these problems speaker independent speech recognition system is a solution, which can break the limitations of the users after the one time proper training the SRS. This will also increase the success rate of speaker identification system and speaker dependent speech recognition systems.

Biometric technology [3] is fast gaining popularity as means of security measures to reduce cases of fraud and theft due to its use of physical characteristics and traits for the identification of individuals. The earliest methods of biometric identification included fingerprint and handwriting while more recent ones include iris/eye scan, face scan, voice print, and hand print. Biometric voice recognition and identification technology focuses on training the system to recognize an individual’s unique voice characteristics (i.e., their voice print).

The application developed will be speaker independent and it provides accurate results under noisy environments also. Following are some objectives we have decided to reach in order to measure the performance of the system.

III. OBJECTIVES

The work has been focused to achieve the following objectives:

1. Implementation of base system with MFCC features.
2. To extract voice features using GMM (Gaussian Mixture Model).
3. To classify the input data using PLDA (Probabilistic Linear Discriminate analysis).
4. Comparison of both the systems to show improved accuracy.
IV. METHODOLOGY

- Speech Database Acquisition
- Implementation of Pre-processing Unit
- Implementation of GMM+PLDA Modeling Unit
- Implementing Testing Unit
- Testing the System on different Metrics
- Presenting the Findings

V. PRESENT WORK

For the implementation of the objectives we selected for our research work we have followed the following method Fig 4.1. We have designed a system in matlab which takes voice/audio signal as input.

5.3.1. Speech Acquisition

Analog speech signals acquired through the microphone which gives digital representation of speech signal. Speech recording is the first step of implementation. Recording has been done by native speaker.
5.3.2. Speech Pre-processing

To enhance the accuracy and efficiency of the extraction processes, speech signals are normally pre-processed before features are extracted. There are four steps in Pre-processing.

5.3.2.1. Background Noise Removal

First step in the signal processing is Removal of Background Noise. By this process, background noise is removed from the data so that only speech samples are the input to the further processing.

5.3.2.2. Pre-emphasis

The digitized speech waveform has a high dynamic range and suffers from additive noise. In order to reduce this range and spectrally flatten the speech signal, pre-emphasis is applied. First order high pass FIR filter is used to pre-emphasize the higher frequency components.

5.3.2.3. Frame Blocking

The speech signal is split into several frames such that each frame can be analyzed in the short time instead of analyzing the entire signal at once. The frame size is of the range 0-20ms. Then overlapping is applied to frames. Overlapping is done because on each individual frame, hamming window is applied. Hamming window gets rid of some of the information at the beginning and end of each frame. Overlapping reincorporates this information back into our extracted features.

5.3.2.4. Windowing

Windowing is performed to avoid unnatural discontinuities in the speech segment and distortion in the underlying spectrum. The choice of the window is a tradeoff between several factors. In speaker recognition, the most commonly used window shape is the hamming window.

5.3.3. Feature Extraction

We have used GMM (Gaussian Mixture Model) for feature selection after we extract the feature. Better selection of vectors in GMM (Gaussian Mixture Model) is important from the training point of view. The robustness of the system in speech samples of considerable diversity of noise is important. GMM is a probabilistic model used for density clustering and estimation. GMM helps differentiation of feature vectors from the parent set of feature vectors. GMM is based on the hypothesis that all vectors are independent.

5.3.4. Pattern Classification

In the classification process we are using Eigenvector values in PLDA (Probabilistic linear Discriminant analysis) for better classification. The frame work that we are using in which Gaussian Probabilistic linear Discriminant analysis is used for transformation of eigenvectors of features, reduces the undesired parts in the vectors. Using Gaussian Probabilistic linear Discriminant analysis clearly helps us to get better speaker discrimination.

5.4.2 System Architecture

Based on the literature reviews and other important documents on several techniques on Speaker Recognition, we have developed and are presenting our new architecture.
In this presented system the speech is first acquired using either a microphone or a speech file like mp3 or wav file. This speech is then pre-processed using normalization, spectral noise removal, pre-emphasis filtering and framing to create ‘N’ frames of length ‘n’ each. We have used hamming window as the windowing function for framing. Finally, these frames are used to compute MFCC features. These MFCC features are used to compute a GMM model. This model stores the Gaussian mixtures. These are transformed to eigenvectors which is then used by PLDA (Gaussian Probabilistic Linear Discriminant Analysis) block to reduce this set to a desired value. After these steps, a log-likelihood value is computed for each model stored in the database. If this value is less than a threshold value defined by testing, the speech is rejected as unknown speaker. If this value is higher than threshold, the speaker Id of maxima is returned which is the predicted speaker.

VI. RESULTS AND DISCUSSION

The code implementation of the thesis is developed in MATLAB. 52 samples of students were collected in phase one and then in phase two these is used to samples to train a GMM-UBM model. The universal background model is trained by implementing the GMM-UBM algorithm in MATLAB.

Then, each student’s model is created and labeled. This system is made to use of “structure” data type to implement all the storage as it gave us easement in tagging voice model with that of student’s id. In the current experiment, the voice sample is recorded individually in a wav file format and stored id is same as filename of the sample.

Model is created using GMM and PLDA based technique. The whole system is text independent.
In the present figure 1, the GUI of the Automatic Speaker Recognition System is shown. The system has two sections as shown “Experiment” and “Visual Direct Sound Input”. In the “Experiment” section, the system can Load the Test data, Compute MFCCs, Compute UBM, Generate each sample’s model and Test the system automatically.

In the “Visual Direct Sound Input” section, the system can load a sample and test who the user is or register a new user if it does not exist in the database.

Figure 4: The GUI Interface of the “Automatic Speaker Recognition System”

Figure 5: Automatic testing of the System with 40 voice samples.

Figure 6: Testing the System Automatically.
Figure 7: The confusion Matrix shows the Result. The red-orange colored diagonal shows the accurate classification.

Figure 8: The Accuracy of the System with 10 second sample comes out to be 98.875 %

With 10 seconds samples and 40 samples, the system can be tested and its accuracy comes out to be about 98.875% with 14.135 seconds to test the whole system which required 1600 matches of 10 seconds each.

Figure 9: The Accuracy of the System with 20 second sample comes out to be 99 %

With 20 seconds samples and 40 samples, the system can be tested and its accuracy comes out to be about 99% with 36.365 seconds to test the whole system which required 1600 matches of 20 seconds each.
Figure 10: The Sample is selected as “4.wav”. This sample is already in the model.

Figure 11: We are testing the System and see if it is recognized correctly.

Figure 12: The person is recognized as “4.wav” which is what is selected. This shows accurate recognition.

It only took 7.682 seconds to recognize the person. The recognition accuracy is very high. The system can recognized user correctly.
Figure 13: Now we have selected “53.wav” This person is not registered in the system.

Figure 14: We can see that the user is not recognized and a message shows “Unknown Person”

It only took 0.57 seconds to predict that the person is unknown.

Figure 15: We have registered this user now. It took only a fraction of second to register this new user.
Figure 16: Now the user is recognized accurately as “53.wav”.

VII. CONCLUSION

In this thesis, we have improved upon the previous work of “Speaker Recognition and Verification using i-vectors approach” and implemented a “GMM + PLDA” based system which is fast and very efficient system. We have implemented the system in MATLAB and tested the results. Our system is very fast. In the base work, the system took about 26 seconds to respond. Our system took only 7.6 seconds to respond which is fairly fast. Also, the old approach required about 50 seconds of sample to reach a viable accuracy. Our system only needed 10-20 seconds sample to accurately predict the user and also verify if the sample is from known voice or unknown voice also.

Our system tends to have a very high accuracy of 99% with only 7.6 seconds per sample response time for a 10 second voice sample.

REFERENCES


IMAGE REFERENCES