Feature Extraction and Classification Techniques in Speaker Recognition

Neelam Nehra[1] Pardeep Sangwan[2] Divya Kumar[3]
[1,2,3]Department of Electronics & Communication Engineering
[1,2]Maharaja Surajmal Institute of Technology, New-Delhi, India [3]IFTM University, Moradabad, Uttar-Pradesh, India

Abstract— Speech is one of the natural form to express emotion. Every person has different voice production organ like vocal tract shape, vocal fold, larynx size etc. Moreover to these differences every speaker has unique accent, fundamental frequency, rhythm, choice of vocabulary, speaking style etc. Speaker recognition is the process of verifying/identifying the speaker based on their speech sample. Feature extraction and matching algorithm are the two main process of speaker recognition .In this paper feature extraction technique Mel Frequency Cepstrum Coefficients (MFCC), Linear Predictive Coefficients (LPC) and classifiers Vector Quantization(VQ), Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) are explained.

Keywords: Speaker recognition, MFCC, GMM, VQ

I. INTRODUCTION

Every speaker has some attributes in his / her speech that is unique. No two people sound the same because they are dissimilar in their larynx sizes, vocal tract shapes and other parts of their voice production organ. Besides these physical differences, each speaker has its style of speaking, including using a particular rhythm, speech style, vocabulary selection, accent, intonation style, and fundamental frequency. Recognition of speakers refers to recognizing individuals from their voices. A training and testing phase is essential in ASR system. In training phase, a user registers by giving voice sample to the system. Then a voice model is formed of enrolled speaker by extracting particular characteristics from speech sample. In testing phase Speaker provides a speech sample that is used to measure similarity to already enrolled speaker and afterwards a decision is made [1]. Speaker recognition system can be categorized depending on trained speakers as closed and open set. In open set identification system tested speaker does not have their sample in trained database, While the speaker is identified in the closed set classification, their voice sample is obtained from the stored database In open-set identification the unidentified speaker come from common people, the system has unidentified. So the closed set SR systems are simple to execute. Identification of open set speakers is complex and precise compared to identification of open set speakers [2]. SR can also categorized by Speaker Verification and Speaker Identification. In SV there is acceptance or rejection of particular speaker it is a 1:1 match whereas SI system finds out the particular speaker from the given dataset it is a 1: N match. SI can be divided into two forms on the basis of text: speaker-dependent and speaker-independent. In text dependent in training and testing same text is used while in text independent there is no restriction on training and testing test, it can be any random phrase in both training and testing. So text independent is more challenging than that of text dependent [3].There are variability in Database collection in SR system due to age, transmission channel, transducer characteristics, environmental noise, recording instruments etc. Data base collection in Forensic Speaker Recognition is more challenging as the speaker is non-cooperative, try to change their identity [4]

II. PRE-PROCESSING

For ASR Pre-processing is the first step. The speech of speaker consists of some component that is not beneficial in the procedure of identifying the speaker. Preprocessing eliminates the undesirable component like mute part from voiced part of speech. This prepares the speech signal for feature extraction.

III. FEATURE- EXTRACTION

Pre-processing is followed by Feature Extraction. Essential speech signal attributes are collected and used for further processing. For feature extraction the extensively used techniques are Perceptual Linear Predictive Coefficients (PLPC), Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Cepstral Coefficients (LPCC), Linear Predictive coefficient (LPC), Cepstrum Analysis etc [5].

A. Mel Frequency Cepstral Coefficients

Calculating Mel-Frequency Cepstral Coefficients (MFCC) is the most important and leading method used to extract spectral characteristics. MFCC uses the frequency domain characteristics which are more accurate than time domain characteristics. MFCC uses the Mel scale which is based on human ear scale. First input signal is passed through a filter which emphasizes on higher frequencies .Energy of higher frequency signal is increased .Then it is followed by framing in which speech signal is divided into small duration of frames of 20-30 ms after that frames are multiplied by hamming window to reduce discontinuity. Then FFT is multiplied to convert into frequency domain signal from time domain. MFCCs uses Mel scale filter bank, DCT is utilized to change over log mel range into time space .Some other features can be added by using Delta cepstrum..It is most accurate and reliable technique having low bit rate. It does not provide smooth transition and satisfactory correlation [6].

B. Linear Predictive Coefficients (LPC)

LP models the human vocal tract as an infinite impulse response (IIR) system, LPC is used to compute spectrum of speech signal. As a linear combination of previous samples, it estimates speech samples. This process lowers the sum of the squared difference over a certain finite interval between past samples and linearly predicted samples. By minimizing such difference, it is possible to determine the unique set of predictor. The pre-emphasis of the speech signal is the first step towards flattening the voice signal spectrum. To increase the higher frequencies in the signal Pre-emphasis is used .The next step is to block the frame that blocks the signal into frames. In order to diminish spectrum leakage in speech signal window function is applied to remove discontinuities

at boundary of frames. In final step, cepstral analysis is used to calculate cepstrum. LPC is not such a good feature for speaker identification, although good for recognition of speech [7].

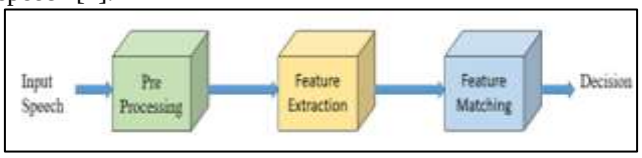

Fig. 1: Hierarchy of Speaker Recognition System

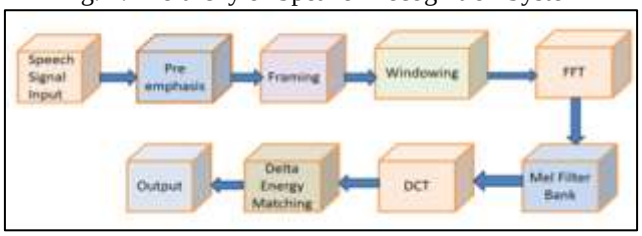

Fig. 2: Block Diagram of MFCC

IV. Feature Matching

Both training and testing use feature matching technique. Speaker model is constructed by using extracted feature vector in training phase and system is trained. While in testing phase speaker model of training phase is compared with other samples of data base that are not used in training phase. Classification involves in calculating a match score in SR system. The term match score mentions the resemblance of the input feature vectors to some model. Speaker model are made from the feature extracted from the speech signal.

A. Vector Quantization (VQ)

For data compression or coding quantization is used. VQ offers multidimensional illustration. In this technique speech signal is divided into vectors, quantization is applied to this vector. Then by using code vector a codebook is created. After that minimum Euclidian distance between each input vector and vectors in the code book is calculated then vector is replaced by the index in the codebook after obtaining the minimum distance [8]. To improve rate of recognition and solve quantization error advantage of VQ with DTW is combined.

B. Gaussian Mixture Model (GMM)

GMM is the stochastic process which uses probability density function. GMM can be used Text independent in SR system. Data is clustered in an unsupervised way (i.e., without any labeled data) in GMM. VQ model extension is GMM, with clusters overlapping and each speaker having an independent speaker mode [7].

C. Hidden Markov Model (HMM)

As compared to the conventional VQ model HMM is an efficient and advanced feature matching algorithm. HMM can model the mathematical representation in a way that speakers generate the sound. It tends to be an effective technique with better speed and accuracy, particularly for noise-degraded speech signals. A Hidden Markov Model consists of two stochastic processes. The first stochastic process is a Markov chain, not observable externally and distinguished by states and probabilities of change. The second stochastic process produces emissions that can be detected at any time, based on a probabilistic distribution depending on the state [8].

V. Conclusion

This paper explains different feature extraction techniques such as MFCC, LPC with their merits and demerits. The pre-processing of speech signals is done, in which noise and silence part from the speech signal is removed prior to extraction of features followed by classification technique. Vector Quantization (VQ), Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) classifier techniques with their features are explained in detail.

References

[1] Singh, Nilu, R. A. Khan, and Raj Shree. "Applications of speaker recognition." Procedia engineering 38 (2012): 3122-3126.
[2] Kekre, H. B., and Vaishali Kulkarni. "Closed set and open set Speaker Identification using amplitude distribution of different Transforms." In 2013 International Conference on Advances in Technology and Engineering (ICATE), pp. 1-8. IEEE, 2013.
[3] Reynolds, Douglas A., and Richard C. Rose. "Robust text-independent speaker identification using Gaussian mixture speaker models." IEEE transactions on speech and audio processing 3, no. 1 (1995): 72-83.
[4] Sangwan, Pardeep, and Saurabh Bhardwaj. "A Structured Approach towards Robust Database Collection for Speaker Recognition." Global Journal of Enterprise Information System 9, no. 3 (2017).
[5] Deshwal, Deepti, Pardeep Sangwan, and Divya Kumar. "Feature Extraction Methods in Language Identification: A Survey." Wireless Personal Communications (2019): 1-33.
[6] Sujiya, S., and Dr E. Chandra. "A Review on Speaker Recognition‖." International Journal of Engineering and Technology 9 (2017): 1592-1598.
[7] Pawar, Rupali V., Rajesh M. Jalnekar, and Janardan S. Chitode. "Review of various stages in speaker recognition system, performance measures and recognition toolkits." Analog Integrated Circuits and Signal Processing 94, no. 2 (2018): 247-257.
[8] Swathy, M. S., and K. R. Mahesh. "Review on Feature Extraction and Classification Techniques in Speaker Recognition." International Journal of Engineering Research and General Science 5, no. 2 (2017): 78-83.

