Speaker Recognition with various Feature extraction and classification Techniques: A review

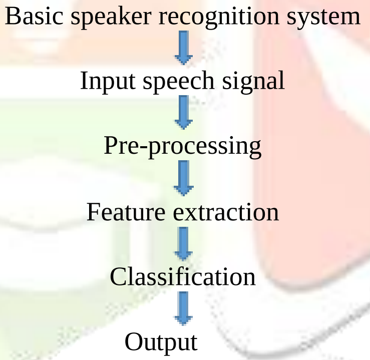
Mrs. Suhasini S Goilkar
Assistant Professor
Department of Electronics and Telecommunication Engg.
Finolex Academy of Management and Technology, Ratnagiri, Maharashtra, India 415639

Abstract: Speech is very natural form of human communication. Speech processing is one of the significant application area of digital signal processing. In speech processing the research developments like speech recognition, speaker recognition, speech synthesis, speaker identification, speech extraction and speech coding. In speaker recognition process who is speaking, is recognized automatically on the basis of individual information provided in speech wave. In speaker recognition technique the speaker's speech is used to verify their identity and recognize using feature extraction techniques. The objective of this review paper is to summarize various feature extraction and classification techniques.

Keywords— Analysis, segmentation, Feature extraction, speaker identification, matching, LPC, MFCC, RASTA filtering

I Introduction

From last fifty decades speech recognition is the research area. Many of the developments have been made in speech recognition. But still a one complete system is not developed which gives an accurate results. Speech signal contain different levels of information. Speaker recognition is used in many speech processing applications especially in security and authentication. Now a day's security is a major requirement [1]. Speaker and speech recognition are very closely related systems but these two systems are different. Speech recognition is the process of recognizing what is being said and speaker recognition is the process of recognizing who is speaking [2].


Basic speaker recognition system

Input speech signal

Pre-processing

Feature extraction

Classification

Output

Preliminary signal processing is carried out to improve the characteristics of the signal such as reducing inserted distortions and adjustment of the frequency range. In pre-processing the silent period of speech signal is removed. In feature extraction features are extracted using different techniques. In classification the different classifiers are used for classification.

II Speaker Recognition Techniques

Different speaker recognition techniques are given in the table and discussed below.

TABLE. 1

Analysis	**Feature extraction**	**Modelling**	**Matching**	**Classification**
Segmentation	LPC	Speaker Recognition	Whole-word matching	DTW
Super-segmentation	LPCC	Speaker identification	Sub-word matching	VQ
Sub-segmentation	MFCC	Speaker independent		HMM
	RASTA filtering			GMM

***A.** Analysis*

When speaker speaks, the speech includes different type of information that helps to identify a speaker. Information of each speaker is different because of the vocal tract, the source of extraction, behavior feature [8].

***B.** Segmentation*

Speaker segmentation is used to detect the speaker change boundaries in a speech stream. It is performed in two steps algorithm which includes potential change detection and refinement. Splitting an audio stream in to acoustically homogeneous segments, so as every segment ideally contains only one speaker. Here the testing to extract the information of speaker is done by utilizing the frame size as well as the shift which is in between 10 to 30 ms in range.

***C.** Sub-segmentation*

In this analysis the testing to extract the information of speaker is done by utilizing the frame size as well as the shift which is in between 3 to 5 ms range.

***D.** Supra-segment*

The analysis to extract the behavior features of the speaker is done by utilizing the frame size as well as the shift size that ranges in between 50 to 200 ms [4].

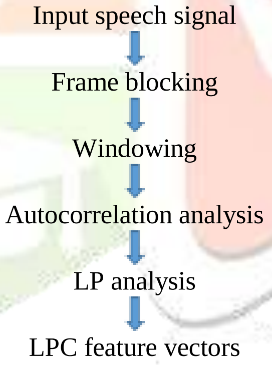
III EXTRACTION

Feature extraction is the main part of the system called heart of the system. In feature extraction extract the features from the input speech signal that helps the system to identify the speaker. Speech signal can be represented by a sequence of feature vectors in order to application of mathematical tools without the loss of generality. In practical real life systems, several of these features are used in combinations. Feature extraction compresses the magnitude of the input signal without causing any harm to the power of speech signal [10]. Different feature extraction techniques

A. Linear predictive coding

This method is widely used in speech coding, synthesis, verification, storage, speech recognition and speaker recognition. This method provides extremely accurate estimates of speech parameters. It is a predominant technique for determining the basic parameters of speech and to provide the precise estimation of speech parameters and computational model of speech. Speech sample can be approximated as a linear combination of past speech samples is the basic idea behind linear predictive coding.

Steps in linear predictive coding-



Input speech signal

Frame blocking

Windowing

Autocorrelation analysis

LP analysis

LPC feature vectors

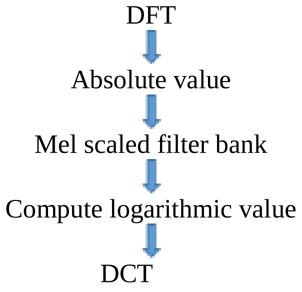
B. Linear predictive coding coefficients

This technique is just the extension of linear predictive coding. When linear predictive coefficients are represented in cepstrum domain then the obtained coefficients are linear predictive coding coefficients. It is obtaining by taking inverse DFT of the speech signal. They are more robust and reliable than linear predictive coding [3].

C. MFCC

MFCC is introduce in 1980 by Davis and Murmelstein. It is a popular technique and commonly used in most of the applications of speech signal foe feature extraction. MFCC is based on the known variations of the human ears critical bandwidth with frequencies below 1000 Hz .Main purpose of MFCC processor is to copy the behavior of human ears. It is a representation of the short term power spectrum of a sound, based on cosine transform of a log power spectrum on a nonlinear mel scale of frequency [9].

The flow of MFCC calculation as below-



DFT

Absolute value

Mel scaled filter bank

Compute logarithmic value

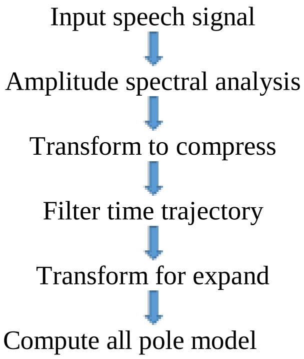
DCT

MFCC technique is considered more consistent with human hearing as compressed to LPCC.LPCC because of mel scale representation [6].

D. RASTA Filtering

RASTA is a short for Rel Ative SpecTral. It is a technique which is used to enhance the speech when recorded in a noisy environment. The time trajectories of the representations of the speech signals are band pass filtered in RASTA. Initially it was just used to lessen the impact of noise in speech signal but now it is also used to directly enhance the signal. It is widely used for the speech signals that have background noise or simply noisy speech.

The process of RASTA technique is-



Probabilistic linear discriminate analysis-PLDA- This technique is an extension of linear Probabilistic analysis. Initially this technique was used for face recognition but now it is used for speech recognition. It is based on i-vector extraction. The i-vector is one which is full of information and is a low dimensional vector having fixed length. PLDA is formulated by a generative model. Main advantage is high recognition accuracy.

IV Modelling

Modelling Technique are used to produce speaker models by making use of features extracted. Modelling techniques are further categorized in to speaker recognition and identification. Speaker recognition further classified into speaker dependent and speaker independent. Speaker identification is the task of determine who is talking from a set of known voices of speakers. It is the process of determine who has provided a given utterance based on the information contained in speech signal. The unknown voice comes from a fixed set of a known speakers, thus it is called a closed set identification. Speaker recognition method can also be divide in to text dependent and text independent methods. In text dependent methods a speaker is requires to utter a predetermined set of words or sentences. In text independent methods, there is no predetermined set of words or sentences and the speaker may not even be aware that they are being tested [5].

V Matching Techniques

The word that has been detected is used by the engine of speech recognizer to a word that is already known by making use of one of the following techniques,

A. Sub word matching

Phonemes are looked up by the search engine on which the system later performs pattern recognition. These phonemes are the sub words thus the name sub word matching. The storage that is required by this technique is in the range 5 to 20 bytes per word which is much less in comparison to whole word matching but it takes a large amount of processing.

B. Whole word matching

In this matching technique there exists a pre-recorded template of a particular word according to which the search engine matches the input signal. The processing that this technique takes is less in comparison to sub word matching. A disadvantage that this technique has is that we need to record each and every word that is to be recognized beforehand in order for the system to recognize it and thus it can only be used when we know the vocabulary of recognition beforehand. Also these templates need storage that ranges from 50 bytes to 512 bytes per word which very large as compared to sub word matching technique [1].

VI Vector Quantization

Vector Quantization is the classical quantization technique from signal processing which allows the modelling of probability density functions by the distribution of prototype vectors. It works by dividing a large set of points into groups having approximately the same number of points closest to them. Each group is represented by its centroid point. The density matching property of vector quantization is powerful, especially for identifying the density of large and high dimensioned data. Since data points are represented by the index of their closest centroid, commonly occurring data have low error, and rare data high error. Hence, Vector Quantization is also suitable for lossy data compression. By using VQ the extracted speech feature of speaker are quantized for a number of centroids. These centroid compose the codebook of that speaker. It is used for data compression and require less storage. VQ is computationally less complex. Memory requirement is achievable for real time applications. A large set of feature vectors are divided into groups having approximately the same number of points closest to them. Each group is represented by its centroid point. VQ can be defined as a mapping function that maps k-dimensional vector space to a finite set CB = {C1, C2, C3, ..…., CN}. The set CB is called codebook consisting of N number of code vectors and each code vector Ci = {ci1, ci2, ci3,……, cik} is of dimension k. The method most commonly used to generate codebook is the Linde-Buzo-Gray (LBG) algorithm. Feature vectors are extracted from input speech signal and the Euclidean distance between input speech signal and each code vector is calculated. The input vector belongs to the cluster of the code vector that yields the minimum distance [12].

VII DYNAMIC TIME WARPING

This is used specifically to deal with variance in speaking rate and variable length of input vectors because this algorithm calculates the similarity between two sequences which may vary in time or speed. To normalize the timing differences between test utterance and the reference template, time warping is done non-linearly in time dimension. After time normalization, a time normalized distance is calculated between the patterns. The speaker with minimum time normalized distance is identified as authentic speaker [12]

Dynamic time warping is an algorithm for measuring similarity between two sequences which may vary in time or speed[4].DTW has been applied to video, audio, graphic, infect any data which can be develop into a linear representation can be analyzed with DTW.

VIII GAUSSIAN MIXTURE MODEL

GMMs are commonly used as a parametric model of the probability distribution continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system. Various forms of GMM feature extraction are outlined, including methods to enforce temporal smoothing and a technique to incorporate a prior distribution to constrain the extracted parameters. Gaussian mixture models have proven to be a powerful tool for distinguishing acoustic sources with different general properties. This ability is commonly exploited in tasks like speaker identification and verification, where each speaker or group is modelled by GMM [13]. The major advantage lies in the fact that they do not rely on any segmentation of the speech signal. A fact that makes them ideal for on-line application. However this advantage means at the same time, that they are not suitable for modelling temporal dependencies but this disadvantage is of minor importance, if the focus lies on the representation of global spectral properties.

IX CONCLUSION

In this review paper there is a discussion on speaker recognition that can be used for many speech processing applications especially for security and authentication. This paper has reviewed the research done in the area of automatic speaker recognition. Different techniques for feature extraction and classification have been discussed. Each technique has got its advantages and limitations. Some techniques are preferred over others such as MFCC for feature extraction and GMM for classification. MFCC is more consistent with human hearing due to mel scale representation. MFCC can be integrated with other techniques such as wavelet decomposition and LPCC to improve the performance of the system as by integrating features more information is added to the input training data. Otherwise choice can be made depending upon certain parameters such as number of system users, storage space, classification time etc. VQ is preferred for real time systems because of its less memory requirement.

REFERENCES

[1] M. A. Anusuya, "Speech Recognition by Machine," International Journal of Computer Science and Information security, Vol.6, No.3, 2009

[2] S. J. Arora and R. Singh, "Automatic Speech Recognition: A Review, "International Journal of Computer Applications, vol60-No.9, December 2012

[3] Santosh K.Gaikward and Bharti W.Gawali, "A Review on Speech Recognition Technique," International Journal of Computer Applications, vol 10, No.3, November 2010

[4] Lindasalwa Muda, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques ", Journal Of Computing, Volume 2, Issue 3, March 2010

[5] J.P. Campbell, "Speaker Recognition: A Tutorial", Proceedings of the IEEE, vol.85, issue-9, pp. 1437-1462, September, 1997.

[6] G.R. Doddington, "Speaker Recognition – Identifying People by their Voices", Proceedings of the IEEE, vol. 73, issue-11, pp. 1651-1664, November, 1985.

[7] R. E. Wohiford, E. H. Wrench, Jr., and B. P. Landell, "A Comparison of Four Techniques for Automatic Speaker Recognition", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol.5, pp. 908-911, April, 1980.

[8] D.A. Reynolds, "Experimental Evaluation of Features for Robust Speaker Identification", IEEE Trans. on Speech and Audio Processing, vol. 2, issue-4, pp. 639-643, October, 1994.

[9] X. Zhou, D. G. Romero, R. Duraiswami, C.E. Wilson, S. Shamma, "Linear versus Mel Frequency Cepstral Coefficients for Speaker Recognition", IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Waikoloa, HI, pp. 559 – 564, 11-15 December, 2011.

[10] T. Barbu, "A Supervised Text-independent Speaker Recognition Approach", Proceedings of World Academy of Science, Engineering and Technology, International Journal of Computer, Information, Systems and Control Engineering, vol.1, issue-9, pp. 2678-2682, January, 2007.

[11] Md Sahidullah, G.Saha, “A Novel Windowing Technique for Efficient Computation of MFCC for Speaker Recognition”, IEEE Signal Processing Letters, vol.20, issue-2, pp. 149-152, February, 2013.

[12] [12] R. Schwartz, S. Roucos, and M. Berouti, “The Application of Probability Density Estimation of TextIndependent Speaker Identification,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol.7, pp. 1649-1652, May, 1982.

[13] B. Wildermoth, K.K. Paliwal, “GMM Based Speaker Recognition on Readily Available Databases”, Proceedings of the Microelectronic Engineering Research Conference, Brisbane, Australia, November, 2003.