International Journal of Engineering Research and General Science Volume 5, Issue 2, March-April, 2017
ISSN 2091-2730

Review on Feature Extraction and Classification Techniques in Speaker Recognition

Swathy M S[1] , Mahesh K R[2]
[1]PG Scholar, [2]Assistant Professor
Dept of ECE, Thejus Engineering College
Email: [1]ms.swathy.ms@gmail.com

Abstract— This paper provides a brief survey on human speech production, feature extraction techniques and classification techniques in speaker recognition. It also discussed about the basic theories of speaker recognition and applications of speaker recognition which is increasing day by day like authentication , surveillance and forensic speaker recognition, etc. LPC, MFCC and WT are the important feature extraction techniques and GMM, HMM and SVM are the important classification techniques. The main aim of this paper is to summarize different feature extraction and classification techniques used for speaker recognition.

Keywords—   Speaker Recognition, Feature Extraction, LPC, MFCC, WT, Classifier, GMM, HMM, SVM

1. INTRODUCTION

Speech is the vocalized form of communication based upon the movements of articulatory organs. Each spoken word is a combination of a limited set of vowel and consonant. Speech is the most efficient way of communication. Each individual has their own unique voices, or we can say that voice keeps the identity of each person. This uniqueness mainly due to the length of the vocal tract , sharp and precise movement of articulatory organs and differences in their speaking habits. Actually speaker recognition is a complex task. Speaker recognition is used to identify a person. The person's identification process is carried out from the characteristics obtained from voices, It is also called voice recognition. Speaker recognition can be classified into two speaker identification and speaker verification.

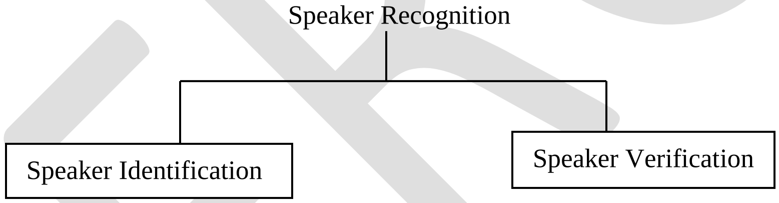

Fig 1. Classification [2]

Speaker identification identifies who is speaking. Speaker verification is the process of accepting or refusing a particular speaker [2] . Speaker recognition mainly consists of two steps one is feature extraction and feature classification [1] . Widely used Feature Extraction techniques are LPC, MFCC, WT. Preprocessing has another name called front end processing. Preprocessing mainly removes the unwanted information present in the speech signals. The main steps of preprocessing are end point detection, pre emphasis filtering, frame blocking and windowing. Wavelet transform used for reducing the noise present in the speech signal. Feature extraction techniques are mainly used to obtain the information which is embedded in the speech signal. Techniques used for feature extraction are LPC, LPCC, MFCC, WT. After extracting the features it has given to the classifier section. Widely used classifier techniques are DTW, HMM, GMM, SVM.

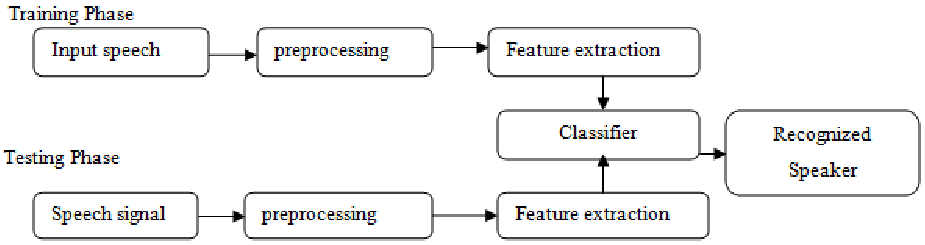

Fig 2 . Main blocks of speaker recognition [1]

II. HUMAN SPEECH PRODUCTION

Speech can be defined as a communication tool. Speech allows the communication between human beings. Speech production in human beings is a daily mechanism. We think that it is a simple mechanism, but its internal mechanism is very complex. Breathing is the first step of speech production. Breathing consists of two processes one is inhaling and the other is exhaling. During inhaling air enters into the lungs. Exhaling is the flow of air out of the organism. During exhaling air flows through the lungs, trachea, larynx, vocal folds, mouth, lips, nasal cavity etc. Movement of these organs called articulatory movement. Articulatory movement control is called motor control and it is done by human brain [3] .

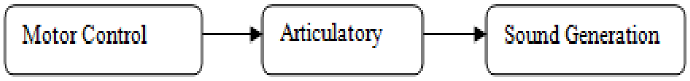

Fig 3: Block diagram of human speech production mechanism

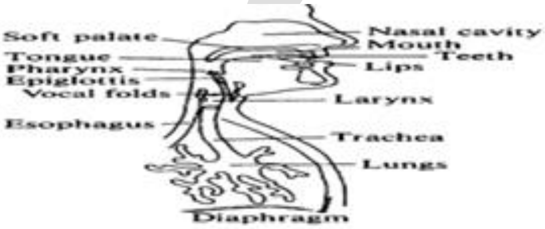

Fig 4 . Human speech production [3]

Pitch, mean, signal to noise ratio, quality, variance, etc. Are some of the basic parameters of speech.

III. FEATURE EXTRACTION TECHNIQUES

Feature extraction plays an important role in speaker recognition. It is very difficult to obtain the data that embedded in the speech signal, so we go for different feature extraction techniques for extracting features from the speech signal. LPC, LPCC, MFCC,WT are some of the feature extraction techniques used today [6] .

A)LINEAR PREDICTIVE COEFFICIENTS (LPC)

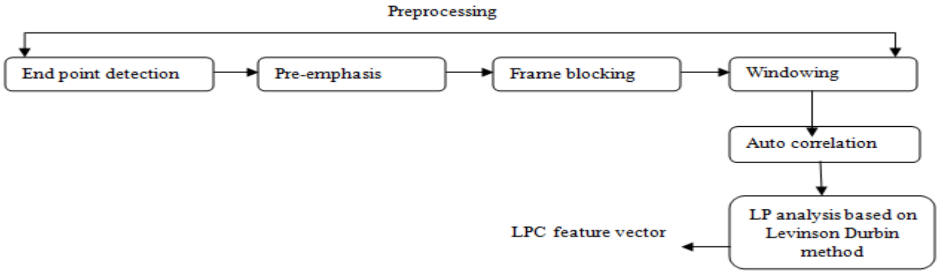

Fig 4. Steps involved in LPC [7] , [6]

LPC (Linear Predictive Coding) is widely used in speech processing. Which features can be obtained as predictive coefficients. From the name it is clear that it predicts the values. In LPC predict the future values of the future values [6]. In LPC linear prediction is the basic principle. Redundancy in the speech signal exploited in LP. The prediction of a present value is determined from the combination of 'm' previous samples. Predicted sample is s_1 (n). 'm' is called the prediction order of the LP [7].

$s_1(n) = \sum_{K=1}^{m} a_{KS}\,(n-k)$……………………(1)

a_ks are the linear prediction coefficients. S (n) is the widowed speech.

S (n) = x (n). w (n)………………………(2)

Prediction error can be calculated as, e (n) = s (n) - s_1 (n). Each frame of autocorrelations is converted into LPC parameter set by using Levinson Durbin's method.

B) LINEAR PREDICTIVE CEPSTRAL COEFFICIENTS (LPCC)

LPCC is one of the predetermined techniques used for extracting features from the speech signal. It is an extension to the above mentioned LPC technique [6] . In LPCC also a linear prediction is the basic principle. The current value is predicted from the previous value. In LPCC coefficients are represented in Cepstrum domain.

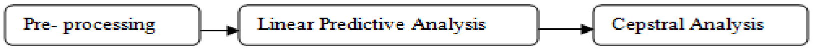

Fig 5 . Steps involved in LPCC [6]

C) MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCC).

The audio signal is a non stationary signal, which is its frequency and amplitude response variable with respect to time. We consider a small duration with an assumption that in that small time interval signal not varying much. We usually take 20 ms to 40 ms frames. MFCC can be defined as a short term power spectrum of the human voice. Now a days MFCC has greater significance in speech processing. It approximates the human system response accurately. After applying MFCCC to the speech signal we get features as in the form of Cepstral coefficients [7].

1. Frame the signal
2. Take FFT of the signal
3. Multiply each FFT magnitude of the corresponding Mel frequency filter value.
4. Take the log of filter bank energies.
5. Take DCT on Mel log energy values (Cepstrum).
   Mel (f) = 2595*$\log_{10}$ (1+f/700) [6]

Mel scale is defined as the perception scale of pitch. Two main advantages of MFCC are highest identification rate and least false rejection rate. By taking Mel scale and log value MFCC approximates the human perception system.

D) WAVELET TRANSFORM (WT)

Wavelets are finite length waves. It simplifies the task of feature extraction. It provides multi resolution and multi scale analysis. Scaling and shifting are the two important operations in wavelet transform. It comes under sub band coding [7] . In WT we divide the speech signal into two sub bands high frequency and low frequency bands. Mainly low frequency component provides the identity of the signal and high frequency component mainly contains the noise part of the signal. But sometimes it may contain the useful information. Main classification of WT is a DWT (Discrete Wavelet Transform).

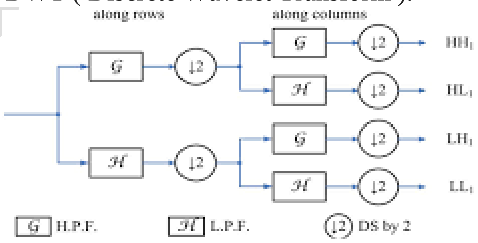

Fig 6. Decomposition tree of DWT [19]

During each level of decomposition we can remove the noise present in the speech signal. Another classification of DWT is WPD and DWPD. In WPD we decompose a high frequency signal also. DWPD is also known as hybrid wavelet transforms. In DWPD, DWT is applied in the low frequency band and WPD applied in the high frequency band. DWPD combines the features of both DWT and WPD.

IV. FEATURE CLASSIFICATION TECHNIQUES

After feature extraction step we obtain the features. Then these features apply to the classifier section. Classifier compares the obtained features with stored features [8] . Based upon this comparison classifier recognizes the particular speaker. Classifiers can be mainly classified into two supervised and unsupervised. If a classifier requires training data, then it comes under supervised otherwise unsupervised.

A) DYNAMIC TIME WARPING (DTW)

DTW is mainly developed for speaker recognition. It is mainly used to find out the similarities between two times based sequences. For example, similarities in walking could be detected by using DTW. Here calculate the time normalized distance, for finding the similarities between the sequences. In speaker recognition, sequences are in the form of speaker information. Each speaker's information sequence is compared to the reference sequence. Then find out the time normalized distance between the sequences. Speaker with minimum time normalized distance is taken as the authenticated speaker.

B) HIDDEN MARKOV MODEL (HMM)

HMM is similar to a Markov model. We can call HMM as a stochastic process. In Markov model future states depend only on the current state, not on the past states [8]. The states are directly visible to the observer. But in Hidden Markov model, the state is not directly visible to the observer. But the output dependent on the state is visible. HMM follows the Markov Property.

C) VECTOR QUANTIZATION (VQ)

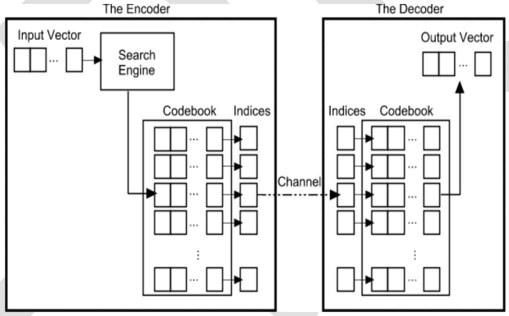

Fig 9. VQ encoder and decoder [10]

VQ comes under lossy compression technique. In vector quantization first the signal is divided into vectors. Then apply quantization to each vector. VQ provides multidimensional representation. Main steps involved in VQ are given below,

1) Construct codebook which is composed of code vector.
2) For encoding calculate the minimum Euclidian distance between each input vector with vectors in the code book.
3) After obtaining the minimum distance replace the vector by the index in the codebook.

Decision boundaries and reconstruction levels are two important terms comes under VQ. LBG algorithm is used for finding the decision boundaries.

D) SUPPORT VECTOR MACHINE (SVM)

SVM is used for classification. SVM can be classified into binary SVM and multi SVM. In binary SVM, we can determine whether the person is recognized or not. Binary SVM compares the features of two speakers. But multi SVM compares the features of more than two speakers. It comes under supervised classifier. Basic of SVM is to create a hyper plane. This hyper plane differentiates the features [8]. In binary SVM features are classified into two classes, each class for recognized and non recognized speaker.

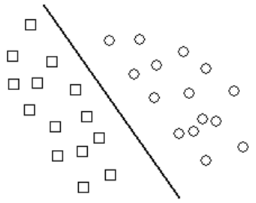

Fig 10. Binary SVM [10]

TABLE 1: COMPARISON OF DIFFERENT FEATURE EXTRACTION METHODS [6], [7]

Technique	Characteristics	Advantages	Disadvantages
LPC	Formant estimation technique, It is modeled by all pole model	Reliable, accurate & robust technique, high speed, low bit rate	Not distinguish similar vowels, Degradation
LPCC	Modeled by all pole model	Smoother & stable representation	Give detail to all frequencies
MFCC	Mimics the human auditory system, Filter bank coefficients	The accuracy is high, Low Complexity	Background noise
WT	Decomposition to sub-bands .	Time-frequency localization, MRA	Denoising, computationally fast

TABLE 2: COMPARISON OF DIFFERENT CLASSIFIERS [8],[16]

Technique	Characteristics	Advantages	Disadvantages
DTW	Unsupervised	Requires less storage space, beneficial for variable length	Cross-channel issue
HMM	Unsupervised	Rail system outputs, efficient performance	Computationally more complex, more storage space
GMM	Unsupervised	Needs less training and test data.	Compromise between DTW and HMM
VQ	Unsupervised	Computationally less complex	Real time encoding is complex
SVM	Supervised	Simple operation	Binary SVM has limitations in speaker recognition.

V. CONCLUSION

This paper has reviewed the research done in the area of speaker recognition. Different feature extraction and classifier techniques have been discussed. Each technique has its own advantages and disadvantages. After the review, we can conclude that speech production is a complex task. LPC is a very simple technique used for feature extraction. Linear prediction is the basic principle of LPC technique. MFCC approximates the human perception system more accurately due to Mel scale. MFCC is widely used today for feature extraction. Binary SVM gives result that a particular speaker is present or not in a given set of data. Multi class SVM provides classification of more than two speech signals very accurately. Main advantage of binary class SVM over Multi class SVM is that it can recognize and identify a particular speaker from a number of speakers. GMM and HMM is also widely used in speech processing. Wavelet Transform provides both time and frequency localization. Wavelet Transform comes under sub band coding. It removes the unwanted signals present in the speech. Wavelet Transform is also used for feature extraction. In future we can introduce Fusion of the techniques called hybrid techniques.

REFERENCES:

[1] Supriya Tripathis" Speaker Recognition", IEEE Explore,Third International Conference on Computer and Communication Technology 2012.

[2] S. K. Singh, Prof P. C. Pandey," Features And Techniques For Speaker Recognition",IIT Bombay.

[3] Harish Chander Mahendru," Quick review of human speech production mechanism",ISSN,Volume 9,January 2014.

[4] Masaaki Honda," Speech Production Mechanisms."2013.

T [5] Harald Hoge, Siemens AG," Basic Parameters Of Speech Signal Analysis

[6] Kirandeep Kaur, Neelu Jain ," Feature Extraction and Classification for Automatic Speech Recognition System",ISSN ,VOLUME 5,January 2015.

[7] Rekha Hebrew, Anup Vibhute," Feature Extraction In Speech Processing A Survey",IJCA,November 2014.

[8] Shubhangi S. Jarande1 , Prof. Surendra Waghmare," A Survey On Different Classifier In Speech Recognition Techniques",IJETAE,March 2014.

[9] Umer Malik1, P.K. Mishra,"Automatic Speaker Recognition Using SVM",IJSR,2013.

[10] Shreya Narang, Ms. Divya Gupta," Speech Feature Extraction Techniques: A Review"IJCSMC,March 2015.

[11] S.B.Dhonde , S.M.Jagade," Feature Extraction Techniques in Speaker Recognition: A Review",IJRMEE,May 2015.

[12] Umer Malik1, P.K. Mishra,"Automatic Speaker Recognition Using SVM",IJSR 2013.

[13] Shreya Narang, Ms. Divya Gupta," Speech Feature Extraction Techniques: A Review"

[14] Alfredo Maesa1, Fabio Garzia,"Text Independent Automatic Speaker Recognition System Using Mel-Frequency Cepstrum Coefficient and Gaussian Mixture Models" Journal of jis.2012.34041 .

[15] Md. Rashidul Hasan, Mustafa Jamil,"Speaker Identification Using Mel Frequency Cepstral Coefficients" Icece 2004.

[16] Roma Bharti , Manav rachna, "Real Time Speaker Recognition System using MFCC and Vector Quantization IJCA *May 2015*. .

[17] Aamir Khan,Muhammad Farhan ,Asar Ali "Speech Recognition:Increasing Efficiency of Support Vector Machines" IJCAVolume 35– No.7, December 2011.

[18] K.Deepak,rishispeaker"recognitionsing Support Vector Machines"issn:issue-2, Feb.-2014.

[19] Shanthini Pandiaraj and K.R. Shankar Kumar "Speaker Identification Using Discrete Wavelet Transform"journal Of Computer Science 2014.