A Review Article on Speaker Recognition with Feature Extraction

Parvati J. Chaudhary¹, Kinjal M. Vagadia²
¹PG Student, Department of Signal Processing and VLSI Technology
²Assistant Professor E&C Dept V.G.E.C, Chandkheda
Gujarat Technological University, Ahmedabad, Gujarat

Abstract-- This paper talks about speaker recognition as an ordinary process whereas speaker identification and speaker verification refer to definite tasks or assessment modes associated with this process. Here, Speaker Recognition is nothing but the computing task of validating a person's claimed identity using features extracted from the database of various voices. For the areas in which security is a foremost concern, speaker Recognition technique is one of the most useful and popular biometric recognition techniques. Various techniques for feature extraction like MFCC, RCC, LPC, LPCC, and PLPC are discussed here.

Keywords-- speaker, recognition, identification, verification, feature extraction

I. INTRODUCTION

Speaker recognition is the process of recognizing involuntarily who is speaking on the basis of entity information included in speech waves. Recognition technique uses the speaker's voice to verify their identity and provides control access to services such as voice dialing, information services, voice mail, database access services, security control for remote access to computers, confidential information areas and several fields which concern safety as the main measures.

Speech is a complex signal produced as a result of several transformations happening at several different levels: linguistic, articulatory, semantic, and acoustic. Differences in these transformations are reflected in the differences in the acoustic properties of the speech signal. Further, there are speaker related differences because of inherent anatomical differences in the vocal tract and the learned speaking habits of dissimilar persons. In speaker recognition, all these differences are taken into account.

Speaker recognition helps in the basic purpose of speaker identification which forms a difficult domain. Examples may include, security systems users having to speak a PIN (Personal Identification Number) number or to speak their credit card number to verify their identity. By checking the voice uniqueness of the input utterance with the help of Speaker recognition, the system can add an extra level of security.

The field of speaker recognition has gained immense popularity in various applications ranging from embedding recognition in a product which allows a unique level of hands-free and intuitive user interaction, automated dictation and command interfaces etc.

II. CLASSIFICATION OF SPEAKER RECOGNITION SYSTEM

Classification of speaker recognition is shown below in Fig.2.1.

Fig 2.1: Block diagram of classification

A. Open Set Vs Closed Set

This type of classification is based on the set of trained speakers available in a system.

1. Open Set: An open set system can have any number of trained speakers. We have an open set of speakers and the number of speakers is always greater than one.

2. Closed Set: A closed set system has only a specified (fixed) number of users registered to the system.

B. Identification Vs Verification

Speaker recognition is a biometric system which takes the speech samples, extracts the characteristic features and performs the computing task of validating a user's claimed identity. As shown in Fig.2.2, it is performed in two parts: Identification and verification.
‘Verification’ performs a binary decision which consists of determining whether the person speaking is the same person he/she claims to be or to put it in other words verifying their identity. On the other hand ‘Identification’ does the job of matching (comparing) the voice of the speaker (known or unknown) with a database of reference templates in an attempt to identify the speaker.

![Fig 2.2: Speaker Recognition](image)

**Fig 2.2: Speaker Recognition**

Automatic speaker identification and verification are often considered to be the most natural and economical methods for avoiding unauthorized access to physical locations or computer systems.

1. **Speaker identification**: It is the process of determining which registered speaker provides a given utterance shown in Fig. 2.3.
2. **Speaker verification**: It is the process of accepting or rejecting the identity claim of a speaker shown in Fig. 2.4.

Both the figures depict the differences between ASI (Automatic Speaker Identification) and ASV (Automatic Speaker Verification) systems. Fig 2.3 and Fig 2.4 gives the basic block diagrams of both the processes.

![Fig 2.3 Speaker Identification](image)

**Fig. 2.3 Speaker Identification**

C. **Text-Dependent Vs Text-Independent**

This type is based on the text uttered by the speaker during the identification process.

1. **Text-Dependent**: In this type, the test utterance is the same to the text used in the training phase. The test speaker has prior knowledge of the system.
2. **Text-Independent**: In this type, the test speaker doesn’t have prior knowledge about the contents of the training phase and can speak anything.

### III. **FEATURE EXTRACTION TECHNIQUES**

This module converts the speech waveform to various type of parametric illustration. According to the speaker recognition application, feature extraction is the process of retaining necessary information of the speech signal while rejecting redundant and unwanted information. This is nothing but analysis of speech signal. Sometimes while removing the unwanted information, we may lose some useful information.

Various techniques used for feature extraction are Mel-Frequency Cepstrum Coefficients (MFCC), Real Cepstral Coefficients (RCC), Linear Prediction Coding (LPC), Linear Predictive Cepstral Coefficients (LPCC) and Perceptual Linear Predictive Cepstral Coefficients (PLPC). This paper is a review of most commonly used techniques.

### A. **Mel Frequency Cepstral Coefficients (MFCC)**

MFCC is one of the most popular technique and commonly used in most of the applications of speech signal for feature extraction. It is based on the human peripheral auditory method. According to human perception, the frequency contents of sounds for speech signals, it does not follow a linear scale. It is mainly used in speaker or speech recognition systems.

Because of human perception behavior which does not follow linear scale that is above 1000 Hz, we take log scale above 1000Hz and call it as Mel Scale. This Mel scale specifies linearity up to 1000Hz and logarithmic above 1000Hz. Hence for each tone of actual frequency, a subjective pitch is measured on different scale called as Mel Scale.
Following is the formula to calculate the estimated mels for a given frequency f in Hz:

\[ \text{mel}(f) = 2595 \times \log_{10}(1 + f/700) \]

The log (mel) spectrum is converted back to time. The end result is called the Mel frequency cepstrum coefficients (MFCC). The human ear is sensitive to both the static and dynamic characteristic of a signal and the MFCC mainly concentrates on the static characteristics.

B. Real Cepstral Coefficients (RCC)

In RCCs, the signal is transformed from the time domain to the frequency domain by applying a Fast Fourier Transform (FFT) to each frame. The logarithm of the results and the inverse Fast Fourier transform (IFFT) is then applied to the signal to get the real Cepstrum of the signal, and can be written as below:

\[ \text{Real Cepstrum} = \text{IFFT} \left( \log \left( \text{FFT} \left( s(n) \right) \right) \right) \]

C. Linear Predictive Coding (LPC)

This technique analyzes the speech signal by estimating the formants. LPC also removes the effects of formants from the speech signal, and estimates the intensity and frequency of the remaining buzz. This procedure of removing the formants is called ‘Inverse filtering’, and other hand the remaining signal is called the ‘Residue’. In LPC technique, each sample of the speech signal is conveyed as a linear predictor and hence it is called as linear predictive coding.

D. Linear Predictive Cepstral Coefficients (LPCC)

This is also a well-known technique and widely used to extract the features from speech signal. For sound frames LPC parameters can effectively describe energy and frequency spectrum. The base of explaining acoustic signals spectrum, modeling and pattern recognition is set by the result of increasing logarithm which restrains the fast change of frequency spectrum, more centralized and better for short-time character and it is because of Cepstrum derived from original spectrum. One of the common short term spectral measurements currently used are LPC derived cepstral coefficients (LPCC) and their regression coefficients LPCC shows the differences of the biological structure of human vocal tract and is computed through iteration from the LPC Parameters to the LPC Cepstrum.

E. Perceptual Linear Predictive Cepstral Coefficients (PLPCC)

This is based on the magnitude spectrum of the speech analysis window. MFCC and LPC are cepstral techniques while the PLPCC is a temporal technique. The steps followed to calculate the coefficients of the PLPCC are described here.

First, compute the power spectrum of a windowed speech.

Second, for sampling frequency of 8 kHz perform grouping of the results to 23 critical bands using bark scaling.

Third, to simulate the power law of hearing, carry out loudness equalization and cube root compression.

Fourthly, perform inverse Fast Fourier Transform (IFFT).

Fifth, one is perform LP analysis by Levinson- Durbin algorithm.

And finally, convert LP coefficients into cepstral coefficients. The relationship between frequency in Bark and frequency in Hz is specified as in

\[ f (\text{Hz}) = \frac{6 \times \arcsin h(f(\text{Hz})/600)}{\text{mel (f)}} \]

IV. CONCLUSION

In this review paper, there is a discussion on classification of speaker recognition that can be used for many speech processing applications especially security and authentication. The most commonly used feature extraction techniques are discussed here among which MFCC is the commonly used.

REFERENCES


