International Journal of Advanced Technology & Engineering Research (IJATER)
www.ijater.com

ANALYSIS OF DIFFERENT FEATURE EXTRACTION TECHNIQUES FOR SPEAKER RECOGNITION SYSTEM: A REVIEW

Mr.Yoghesh Dawande [1,] Dr. Mukta Dhopeshwarkar[2]
Department of Computer Science and IT [1, 2]
Dr.Babasaheb Ambedkar Marathwada University, Aurangabad (MS) 431004s, India
yogeshdawande7089@gmail.com
mukata_d@gmail.com

Abstract:

This paper gives an overview of various methods and techniques used for feature extraction in speaker recognition. The research in speaker recognition have been evolved starting from short time features reflecting spectral properties of speech (low-level or physical traits) to the high level features (behavioral traits) such as prosody, phonetic information, conversational patterns etc.
In this paper, we first give a brief overview of Speech Recognition and then describe some feature extraction technique. We have compared MFCC, LPCand PLP feature extraction techniques.

.

Introduction

Human voice conveys information about the language being spoken and the emotion and gender for the identity of the speaker. Speaker recognition is a process where a person is recognized on the basis of his voice signals [1, 2]. The Objective of speaker recognition is to determine which speaker is present based on the individual's utterance. This is in contrast with speaker verification, where the objective is to verify the person's claimed identity based on his or her utterance. Speaker identification and speaker verification fall under the general category of speaker recognition [3, 4].In speaker identification there are two types, one is text dependent and another is text independent. Speaker identification is divided into two components: feature extraction and feature classification. In speaker identification the speaker can be identified by his voice, where in case of speaker verification the speaker is verified using database.

Historically, all speaker recognition systems have been mainly based on acoustic cues that are nothing but physical traits extracted from spectral characteristics of speech signals. So far the features derived from the speech spectrum have proven to be the most effective in automatic systems, because the spectrum reflects the geometry of system that generates the signal. Therefore the variability in the dimensions of the vocal track is reflected in the variability of the spectra between the speakers. However, studies [5] have proved that there is a large amount of information suitable for speaker recognition being the top part related to learned traits and the bottom part to physical traits.

Speech Recognition Process:

The goal of speech recognition is for a machine to be able to "hear," understand," and "act upon" spoken information. The earliest speech recognition systems were first attempted in the early 1950s at Bell Laboratories, Davis, and Biddulph andBalashek developed an isolated digit Recognition system for a single speaker [6]. The goal of automatic speaker reorganization is to analyze, extract characterize and recognize information about the speaker identity. The speaker Reorganization system may be viewed as working in a four stages

- Analysis
- Feature extraction
- Modeling
- Testing

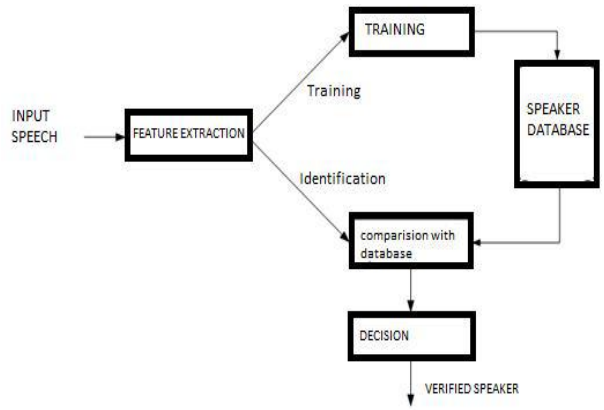

Fig.1. Block diagram of Speech recognition process

Feature Extraction is the most important part of speech recognition since it plays an important role to separate one speech from other. Because every speech has different individual characteristics embedded in utterances. These characteristics can be extracted from a wide range of feature extraction techniques proposed and successfully exploited for speech recognition task. But extracted feature should meet some criteria while dealing with the speech signal such as:

1) Easy to measure extracted speech features
2) It should not be susceptible to mimicry
3) It should show little fluctuation from one speaking environment to another
4) It should be stable over time

5) It should occur frequently and naturally in speech

From human speech production mechanism, it is possible to identify individual using the speech data. Speech contains speaker specific information due to vocal track and excitation source. Larynx is the major excitation source, whereas vocal track is the major resonant structure. Speaker information is due to particular shape, size and dynamics of vocal track and also the excitation source. These features related to physiological nature of human speech production are called physical traits, which are used in state-of-art systems. However human speaker recognition relies on other sources of information like speaking style, pronunciation etc. Such features are referred to as behavioral traits. Further, the behavioral traits like how the vocal tract and excitation source are controlled during speech production are also unique for each speaker. The information about the behavioral trait is also embedded into the speech signal and can be used for speaker recognition. Thus the information present in speech signal carries the identity of speaker at different levels. To properly represent speech data, it is necessary to analyses it using suitable analysis techniques. The analysis techniques aim at selecting proper frame size and shift for analysis and also at extracting the relevant features in the feature extraction stage [7].

Feature Extraction:

Types of Features:

A vast number of features have been proposed for speaker recognition. We divide them into the following classes:

- Spectral features
- Dynamic features
- Source features
- Suprasegmental features
- High-level features

Spectral features are descriptors of the short-term speech spectrum, and they reflect more or less the physical characteristics of the vocal tract. Dynamic features relate to time evolution of spectral (and other) features. Source features refer to the features of the glottal voice source. Super asegmental features span over several segments. Finally, high-level features refer to symbolic type of information, such as characteristic word usage. The most widely used feature extraction techniques are explained below

A. Mel Frequency Cepstral Coefficient (MFCC):

A block diagram of an MFCC feature extraction is shown (Fig. 2).This coefficient has a great success in speaker recognition application. The MFCC [8] [9] is the most evident example of a feature set that is extensively used in speech recognition. As the frequency bands are positioned logarithmically in MFCC [10], it approximates the human system response more closely than any other system. Technique of computing MFCC is based on the short-term analysis, and thus from each frame a MFCC vector is computed.

In order to extract the coefficients the speech sample is taken as the input and hamming window is applied to minimize the discontinuities of a signal. Then DFT will be used to generate the Mel filter bank. MFCC can be computed by using the formula.

Mel (f) = 2595*log10 (1+f/700) (1)

The Mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. The difference between the cepstrum and the Mel-frequency cepstrum is that in MFC the frequency bands are equally spaced on the Mel scale, which approximates the human auditory system's response. MFCCs are commonly used as features in speech recognition system. To enhance the accuracy and efficiency of the extraction processes, speech signals are normally pre-processed before features are extracted. [13]

The following figure 2 shows the steps involved in MFCC feature extraction.

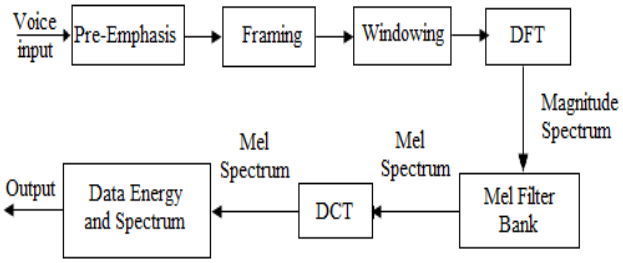

Fig.2. Block diagram of Mel frequency cepstral coefficient

B. Linear Predictive coding (LPC):

Linear prediction is a mathematical computational operation which is linear combination of several previous samples. LPC [8] [9] of speech has become the predominant technique for estimating the basic parameters of speech. It provides both anaccurate estimate of the speech parameters and it is also an efficient computational model of speech. The basic idea behindLPC is that a speech sample can be approximated as a linear combination of past speech samples. Through minimizing thesum of squared differences (over a finite interval) between the actual speech samples and predicted values, a unique set of parameters or predictor coefficients can be determined. These coefficients form the basis for LPC of speech [11]. The following figure 3 shows the steps involved in LPC feature extraction.

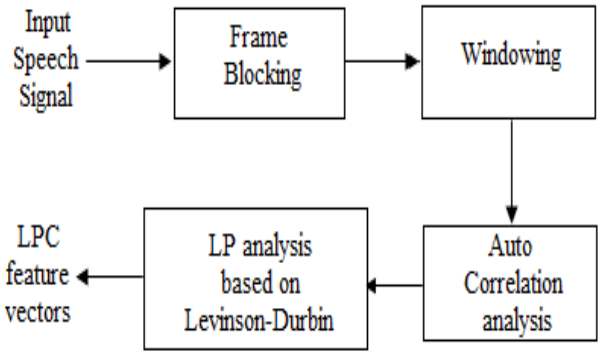

Fig.3. Block diagram of linear predictive coding

C. Perceptual Linear Prediction (PLP):

The Perceptual Linear Prediction (PLP) model developed by Herman sky 1990. The goal of the original PLP model is to describe the psychophysics of human hearing more accurately in the feature extraction process. PLP is similar to LPCanalysis, is based on the short-term spectrum of speech. In contrast to pure linear predictive analysis of speech, perceptual linear prediction (PLP) modifies the short-term spectrum of the speech by several psychophysically based transformations.[12].

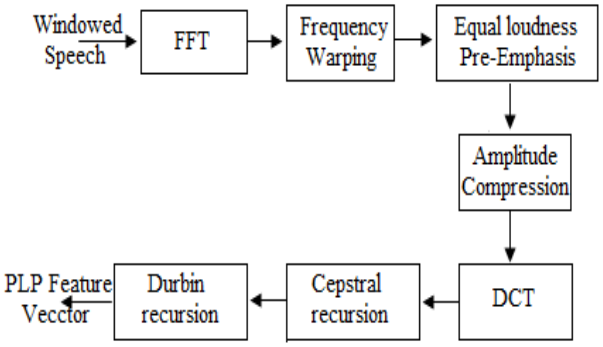

Fig.4. Block diagram of Perceptual linear prediction

Comparative Result Analysis:

Result of different feature extraction techniques MFCC, LPC, and PLP techniques. In an average the MFCC techniques give the maximum recognition rate.

MFCC	LPC	PLP
51.25%	37.5%	49.5%
86.67%	80.5%	77.4%
93.6%	76.6%	90.4%
96.5%	65.8%	78.5%

Table 1: Comparative result analysis of speech signals

Conclusion:

In this paper, we have presented an overview of the various features, the extraction methods and modeling techniques of speaker recognition. The low level features such as cepstral features work well in ideal conditions, but their performance is degraded in real time situations. Use of high level information can add complementary knowledge to improve the performance of recognition system. In practical situations many negative factors are encountered including mismatched handsets for training and testing, limited training data, unbalanced text, background noise and non-cooperative users. The techniques of robust feature extraction, feature normalization, model-domain compensation and score normalization methods are necessary. There are number of research problems that can be taken up, such as human-related error sources, real-time implementation, and forensic interpretation of speaker recognition scores. For this it is important to explore stable features that remain insensitive to variation of speakers' voice over time and are robust against variation in voice quality due to physical states or disguises. The problem of distortion in the channels and background noise also requires being resolved with better techniques.

It is known that performance of speaker recognition depends on the speaking rate of speaker and it vary for different speaker. If the test speaker's speaking rate is different from that of trained speaker's speaking rate then it affects the performance of speaker recognition with respect to time, space and computational complexity. This paper has illustrated the different feature extraction techniques of speaker identification through experimental research. MFCC is well known techniques used in speaker recognition to describe the signal characteristics.

Acknowledgments:

We author would like to thank the Department of Computer Science ant IT, and Dr.Babasaheb Ambedkar Marathwada University Aurangabad for providing the infrastructure to carry out the research.

References

[1] Campbell J.P. and Jr. "Speaker recognition: A Tutorial" Proceeding of the IEEE. Vol 85, 1437- 1462 1997.

[2] S.Furui. "Fifty years of progress in speech and speaker recognition," Proc. 148th ASA Meeting, 2004.s

[3] A. Rosenberg, "Automatic speaker recognition: A review," Proc. IEEE, vol. 64, pp. 475487, Apr.1976.

[4] G. Doddington, "Speaker recognition-Identifying people by their voices," Proc. IEEE, vol. 73, pp. 1651-1664, 1985

[5] Marcos Faundez-Zanuy and Enric Monte-Moreno, "State-of-the-art in Speaker Recognition ", IEEE A&E Systems Magazine, May 2005.

[6] R.KlevansandR.Rodman, "Voice Recognition, Artech House, Boston, London 1997.

[7] Jayanna, H.S. and S.R.M. Prasanna, 2009. Analysis, feature extraction, modeling and testing techniques for speaker recognition. IETE Technical Review, 2009, Volume 26, issue 3

[8] Corneliu Octavian DUMITRU, Inge GAVAT, "A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language", 48th International Symposium ELMAR-2006, 07-09 June 2006, Zadar, Croatia.

[9] DOUGLAS O'SHAUGHNESSY, "Interacting With Computers by Voice: Automatic Speech Recognition and Synthesis", Proceedings of theIEEE, VOL. 91, NO. 9, September 2003, 0018-9219/03 2003 IEEE.

[10] A.P.Henry Charles &G.Devaraj, "Alaigal-A Tamil Speech Recognition", Tamil Internet 2004, Singapore.

[11] N.UmaMaheswari, A.P.Kabilan, R.Venkatesh, "A Hybrid model of Neural Network Approach for Speaker independent Word Recognition", International Journal of Computer Theory and Engineering, Vol.2, No.6, December, 2010 1793-8201.

[12] Jeetkumar, Om prakashPrabhakar, Navneet Kumar sahu. "Comparative analysis of different feature extraction and classifier techniques for speaker Identification System: A Review", International Journal of Innovative Research in computer and communication engineering (An ISO 3297:2007 certified Organization) Vol. 2, Issue 1, January 2014.

[13] TejalChauhan, HemantSoni, SameenaZafar, "A Review of Automatic Speaker Recognition System", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3, Issue-4 September 2013.