

Feature Extraction Techniques in Speaker Recognition: A Review

S.B.Dhonde

Department of Electronics,
AISSMS Institute of Information Technology,
Pune -411001,India
{dhondesomnath@gmail.com,
sbdhonde@rediffmail.com}

S.M.Jagade

Department of E&TC,
TPCT College of Engineering
Osmanabad, India
{smjagade@gmail.com}

Abstract: This paper presents a brief survey on various feature extraction techniques like Linear Predictive Cepstral Coefficients (LPCC), Perceptual Linear Prediction Coefficients (PLPC), and Mel-Frequency Cepstral Coefficients (MFCC) to extract the effective and efficient features from speech signal for speaker recognition. We also discuss the problems associated with well-known methods of feature extraction. We conclude with the some future areas in which the work can be done in order to extract efficient speech features to increase the accuracy of speaker recognition system.

Keywords – *Speaker Recognition, Cepstral Analysis, Linear Predictive Cepstral Coefficients (LPCC), Perceptual Linear Prediction Coefficients (PLPC), Mel-Frequency Cepstral Coefficients (MFCC)*

I. INTRODUCTION

The feature extraction is a process of retaining useful information from speech signal while discarding unwanted signal such as noise. The feature extraction transforms raw acoustic signal into compact representation [1]. A sequence of feature vectors which represents compact speech signal is computed by feature extraction method [2]. Feature vectors which are extracted from raw signal in feature extraction module emphasize speaker specific properties and suppress statistical redundancies. Using feature vectors of the target speaker, we train the speaker model. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. When performing analysis of complex data one of the major problems stems from the number of variables involved. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy.

In speaker recognition, accuracy and recognition rate degrades because of various aspects like, variability from speaker; utterance provided by speaker (may change every time because of emotions and illness). Also, variability from environment, noise in speech signal (because of transmission channel), background noise, and reverberation corrupts the input speech signal in testing mode. The time function feature is ineffective because it changes significantly when same speaker speaks same utterance. The features which will provide correct information and is robust against noise should to be computed in such cases. The features can be classified into short-term spectral features, voice source features, spectro-temporal features, prosodic features and high-level features [3]. Generally, the feature extraction schemes for speaker recognition can be categorized into Linear Predictive Cepstral Coefficients (LPCC), Perceptual Linear Prediction Coefficients (PLPC), and Mel-Frequency Cepstral Coefficients (MFCC).

This paper is organized as follows. In section 2, we discuss feature extraction schemes Linear Predictive Cepstral Coefficients (LPCC), Perceptual Linear Prediction Coefficients (PLPC), and Mel-Frequency Cepstral Coefficients (MFCCs). The section 3 discusses the

comparison of these schemes and problems associated with these schemes. We conclude with some future scopes for the improvement in accuracy of speaker recognition system.

II. FEATURE EXTRACTION TECHNIQUES

1.1 Linear Predictive Cepstral Coefficients (LPCC)

LPC is used to calculate spectrum of the signal [4]. It approximates speech samples as a linear combination of past samples. This scheme minimizes the sum of squared difference between past samples and linearly predicted samples over some finite interval. The unique set of predictor coefficients can be determined by minimizing such difference.

The pre-emphasis of speech signal is the first step for flattening the spectrum of speech signal. Pre-emphasis boosts the higher frequencies in the signal. The next step is to frame the signal and multiply it by window function in order to reduce spectrum leakage in speech frame. The vocal tract model can be represented by all-pole model. It provides the set of auto regression coefficients called Linear Prediction Coefficients (LPCs). The vocal tract transfer function can be given as,

$$V(z) = \frac{1}{1 - \sum_{k=1}^p a_k z^{-k}}$$

In last step, cepstrum is calculated by means of cepstral analysis. Cepstral coefficients can also be calculated from the LPC via a set of recursive procedure.

Though LPC features emphasizes on formant structure, it ignores details like nasal, piriform fossa which are useful in speaker recognition. Another disadvantage of LPC is that it does not capture spectral valleys. LPC is not so good features for identification of speakers. However, it is good for speech recognition.

1.2 Perceptual Linear Prediction Coefficients (PLPC)

PLP is used to calculate power spectrum of the speech signal. It modifies the spectrum of speech signal by several transformations. The basic idea is to obtain the auditory spectrum and approximate it by all pole model. This scheme first computes a power spectrum estimate. The power spectrum is integrated using a Bark-scale filter bank, which models the critical band frequency selectivity inside the human cochlea. The relationship between Bark scale and linear frequency is given by,

$$f_{bark} = 6 * \ln \left[\frac{f}{600} + \left(\left(\frac{f}{600} \right)^2 + 1 \right)^{0.5} \right]$$

The bark scale filters are trapezoidal in shape [5]. The pre-emphasis of the signal is done using equal loudness curve after frequency integration [1]. Next, cube root of power spectrum is taken as the perceived loudness is approximately the cube root of intensity. The power spectrum is compressed in this step. [5]. After this step, inverse Fourier transform is applied to the filter outputs to obtain autocorrelation sequence and Linear Predictive analysis is performed to smooth the spectrum. The final features are also obtained using cepstral recursion from the LP coefficients [1]. PLP model is identical to LPC model except that in PLP spectral characteristics are transformed to match characteristics of human auditory system. DFT and LP techniques are merged in PLP scheme.

1.3 Mel-Frequency Cepstral Coefficients (MFCC)

MFCCs are popular in speaker recognition [1] [6] [7]. This scheme models human auditory system in which mel-filters are spaced linearly at low frequencies and logarithmically at high frequencies to capture the phonetically important characteristics of speech. The higher frequencies of the spectrum are enhanced by using following FIR filter which is applied to the input speech signal. [8] [9]. The process of pre-emphasis flattens the signal making it less susceptible to finite precision.

$$y(n) = x(n) - \alpha x(n-1)$$

where $x(n)$ is the input speech signal and $0.9 \leq \alpha \leq 1$

The speech quasi-periodic signal is divided into number of frames of duration 20-30 msec. over which speech signal assumed to be stationary with 50% overlap between two successive frames in order to avoid any loss of information[8]. After this step, hamming window is generally applied to each frame to minimize the discontinuities in the signal. The following hamming window is multiplied with each frame.

$$x_a = y_a(n) \cdot w(n) \quad a = 1, 2, 3, \dots, T$$

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{M-1}\right)$$

The Fast Fourier transform of each frame is computed in the next step which converts the each frame to frequency domain representation. The magnitude of windowed signal is computed to obtain the power spectrum. In next step, windowed signal is multiplied with non-linear frequency scale called mel-scale which is roughly linear below 1 kHz and logarithmic above 1 kHz. The relationship between linear frequency and mel scale is given by following formula,

$$mel(f) = 2595 * \log_{10}\left(1 + \frac{f}{700}\right)$$

The mel-filter bank is triangular in shape [5]. For each energy, logarithm operation is performed and in the last step, cepstrum is calculated using discrete cosine transform (DCT) to decorrelate the log energies [10] or inverse Fourier transform to obtain MFCCs [1] [7] [11] [12]. The DCT compresses the signal.

III. DISCUSSION

In the feature extraction techniques discussed above, LPC is good for speech recognition but it is not so good for speaker recognition. This scheme does not capture speaker specific information as it is not suitable for modelling anti-resonance generated by piriform fossa. However, the fundamental frequencies called formants can be effectively deduced from LPC [4]. The PLP and MFCC schemes are somewhat related to each other. Both PLP and MFCC techniques models human auditory system with non-linear scale. The non-linear scale used in PLP technique is Bark-scale and in MFCC, it is Mel-scale. It has been viewed over as, among these techniques MFCC comparatively performs better. However, the performance of MFCC is proportional to noise. Its accuracy degrades in the presence of noise and channel mismatch problems [6] [13]. In literature, many researchers have performed modifications in MFCC to improve its accuracy. Such efforts are discussed in [6]-[15]. The speaker recognition accuracy can be increased by modelling the nasal cavity. The robust features can be computed from combination existing techniques with new information such as high-level features, complementary information.

IV. CONCLUSION

We have reviewed feature extraction techniques and found that Mel-Frequency Cepstral Coefficients (MFCC) is most widely used technique for in speaker recognition. The factors channel mismatch, background noise affects the performance of MFCC technique. Many other sources of information from speech signal such as high-level information, complementary information can be used to improve accuracy of speaker recognition technique. Also, LPC technique can be a good scheme for speech recognition. The schemes PLP and MFCC are based on non-linear behaviour of human auditory system whereas LPC is linear in nature. The extraction of effective speech features is necessary to increase the accuracy of speaker recognition. The efforts can be made in a way to combine low-level features and high-level features.

REFERENCES

- [1] Md Jahangir Alam , TomiKinnunen , Patrick Kenny , Pierre Ouellet, Douglas O'Shaughnessy, "Multitaper MFCC and PLP features for speaker verification using i-vectors", *Speech Communication, ScienceDirect*, Volume. 55, Issue 2, Pages 237-251 February 2013.
- [2] M.A.Anusuya, S.K.Katti, "Speech Recognition by Machine: A Review", *International Journal of Computer Science and Information Security (IJCSIS)*, Vol. 6, No. 3, 2009.
- [3] Tomi Kinnunen, Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors", *Journal on Speech Communication*, Elsevier, Vol.52, Issue 1, Pages 12–40, January 2010.
- [4] Yusnita M. A., Paulraj M. P., SazaliYaacobb, Nor Fadzilah M., Shahrman A. B., "Acoustic Analysis of Formants across Genders and Ethnical Accents in Malaysian English using ANOVA", *International Conference on Design and Manufacturing (IConDM2013)*, Volume 64, 13 November 2013, Pages 385–394.
- [5] Dr. Shaila D. Apte, "Speech Processing Applications", in *Speech and Audio Processing*, Chapter 3, Pages 105- 118, Wiley India Edition.
- [6] WU Zunjing, CAO Zhigang, State Key Laboratory on Microwave and Digital Communications, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China, "Improved MFCC-Based Feature for Robust Speaker Identification", *TUP JOURNALS & MAGAZINES*, Volume.10, Issue 2, April 2005.
- [7] Tomi Kinnunen, *Member, IEEE*, Rahim Saeidi, *Member, IEEE*, FilipSedlák, Kong Aik Lee, Johan Sandberg, Maria Hansson-Sandsten, *Member, IEEE*, and Haizhou Li, *Senior Member, IEEE*, "Low-Variance Multitaper MFCC Features: A Case Study in Robust Speaker Verification", *IEEE Transaction on Audio, Speech, and Language Processing*, Vol.20, No.7, September 2012.
- [8] Pawan K. Ajmera, Dattatray V. Jadhav, Ragunath S. Holambe, "Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram", *Journal onPattern Recognition*, Elsevier, Vol.44, Issue 10-11, Pages 2749-2759, 22 April 2011.
- [9] Claude Turner, Anthony Joseph, Murat Aksu, Heather Langdond, "The Wavelet and Fourier Transforms in Feature Extraction for Text-Dependent, Filterbank-Based Speaker Recognition", *Procedia Computer Science*, Volume 6, Pages 124–129, 11 October 2011.
- [10] Pawan K. Ajmera, Raghunath S. Holambe, "Fractional Fourier transform based features for speaker recognition using support vector machine", *Journal on Computers and Electrical Engineering*, Elsevier, Vol. 39, Issue 2, Pages 550-557, 13 June 2012.
- [11] K. Sri Rama Murty and B. Yegnanarayana, *Senior Member, IEEE*, "Combining Evidence From Residual Phase and MFCC Features for Speaker Recognition", *IEEE Signal Processing Letters*, Vol.13, No.1, January 2006.
- [12] Seiichi Nakagawa, *Member, IEEE*, Longbiao Wang, *Member, IEEE*, and Shinji Ohtsuka, "Speaker Identification and Verification by Combining MFCC and Phase Information", *IEEE Transaction on Audio, Speech, and Language Processing*, Vol.20, No.4, May 2012.
- [13] Cemal Haniççi, Tomi Kinnunen, Figen Ertaş, Rahim Saeidi, Jouni Pohjalainen, and Paavo Alku, "Regularized All-Pole Models for Speaker Verification Under Noisy Environments", *IEEE Signal Processing Letters*, Vol.19, No.3, March 2012.
- [14] R.Shantha Selva Kumari, S. Selva Nidhyananthan, Anand.G, "Fused Mel Feature sets based Text-Independent Speaker Identification using Gaussian Mixture Model", *International Conference on Communication Technology and System Design 2011*, Volume 30, 13 March 2012, Pages 319–326.
- [15] Md Sahidullah, *Student Member, IEEE*, and GoutamSaha, *Member, IEEE*, "A Novel Windowing Technique for Efficient Computation of MFCC for Speaker Recognition", *IEEE Signal Processing Letters*, Vol.20, No. 2, February 2013.
- [16] Douglas A. Reynolds, MIT Lincoln Laboratory, Lexington, MA USA, "An over view of automatic speaker recognition technology", *Acoustics, Speech, and Signal Processing (ICASSP), IEEE International Conference*, Vol.4, 2002