Scientific Journal of Impact Factor(SJIF): 3.134

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2,Issue 5, May -2015

Mel-Frequency Cepstral Coefficients for Speaker Recognition : A Review

[1] S.B.Dhonde , Department of Electronics, AISSMS Institute of Information Technology, Pune -411001,India
[2] S.M.Jagade, Department of E&TC, TPCT College of Engineering
Osmanabad, India

Abstract :*This paper gives a brief survey on feature extraction technique Mel-Frequency Cepstral Coefficients (MFCCs) to extract the effective and efficient features from speech signal. In this paper, we have studied the different type of modifications suggested by different researchers in MFCC technique. It aims to understand all the state-of-the art in the field of MFCC such that researchers will be able to understand the modifications carried out in MFCC technique.*

Keywords *– Speaker Recognition, Feature Extraction, Cepstral Coefficients*

1. INTRODUCTION

Speaker recognition is the process of identifying person from known set of voices. The input to speaker recognition system is speech signal uttered by a speaker. Speech signal contains speaker specific information. Speaker recognition can be text-dependent and text-independent. Text-dependent speaker recognition system requires a fixed phrase as a password whereas, text-independent system has no such a constraints [1] [2]. The speaker recognition system includes training and testing mode. In training mode, speaker specific features are computed and stored in the database in the form of speaker model. In testing mode, unknown speaker's features are compared with the speaker model stored in the database in training mode. The speaker is identified based on minimum distance calculated in the testing mode. Feature extraction and feature matching are the important steps in speaker recognition.

In feature extraction, speaker specific properties are extracted from speech signal. It de-emphasizes on all other information which is not useful in speaker recognition, for example, it does not consider what is being said. Feature extraction represents raw acoustic signal into compact representation [3]. A sequence of feature vectors provides the compact representation of raw speech signal[4] is used to train the speaker model. This speaker model is stored in the database in training phase. The speaker specific properties such as resonance of the vocal tract is categorized as short-term spectral features. The short-term spectral features computed using short frames of speech signal which is of 20-30 ms duration.Mel-frequency cepstrum coefficients (MFCCs) are used to represent the spectrum of speech signal in speaker recognition. The properties of human auditory system are modelled in the case of MFCCs across frequency.

In this paper, we review feature extraction scheme Mel-Frequency Cepstral Coefficients (MFCCs). The section 2 discusses the MFCC feature extraction scheme. In section 3, we have reviewed the state-of-art of MFCC scheme for improving its accuracy. In the last section, we discuss some future scopes for the improvement in accuracy of MFCC technique.

2. MEL- FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

MFCCs are widely used in speaker recognition [3] [5]. MFCCs are based on known variations of human ear's critical bandwidth with frequencies. Mel-filters are spaced linearly at low frequencies and logarithmically at high frequencies to capture the phonetically important characteristics of speech. The Figure 1 shows MFCC extraction procedure.

a) **Pre-emphasis**
Pre-emphasis is performed to enhance the higher frequencies of the spectrum [6] [7]. Pre-emphasis flattens the signal making it less susceptible to finite precision. The following FIR filter is applied to the input speech signal.

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

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

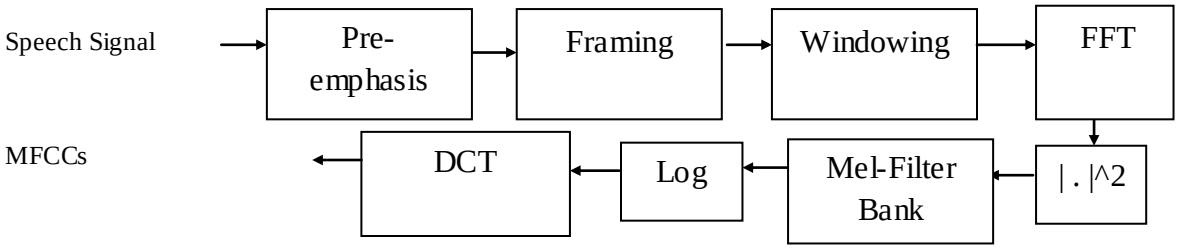

Figure 1 Block Diagram of MFCC Technique

b) Framing

The speech signal is quasi-periodic in nature. The speech signal is divided into number of frames of duration 20-30 msec. over which speech signal assumed to be stationary [6][8]. There is 50% overlap between two successive frames in order to avoid any loss of information.

c) Windowing

The discontinuities of the signal are minimized by tapering the beginning and end of each frame to zero. This is performed using windowing technique. Generally, hamming window is preferred. Hamming window has better side lobes suppression. The following hamming window is multiplied with each frame.

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

where

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

d) FFT

The frequency content of the windowed signal is estimated using Fast Fourier Transform. It converts each frame from time domain to the frequency domain. The short-term frequency content of the signal is estimated using FFT of each frame. Power spectrum is computed using squared magnitude of a windowed signal.

e) Mel-Filter Bank

The windowed signal is multiplied with mel-filter bank. The Mel scale is an auditory scale similar to frequency scale of human ear (similar to how human ear perceives sound) [7]. The scale is roughly linear below 1 kHz and logarithmic above 1 kHz. The relationship between linear frequency and mel scale is given by following formula,

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

f) DCT

In the last step, logarithm operation is performed followed by discrete cosine transform to decorrelate the log energies [9]. The DCT compresses the signal. The cepstrum is calculated using discrete cosine transform (DCT) or inverse Fourier transform to obtain MFCCs [3] [5] [10] [11].

The difference from the real cepstral is that a nonlinear frequency scale is used, which approximates the behaviour of the auditory system. Additionally, these coefficients are robust and reliable to variations according to speakers and recording conditions. MFCC is an audio feature extraction technique which extracts parameters from the speech similar to ones that are used by humans for hearing speech, while at the same time, deemphasizes all other information.

3. STATE-OF-ART IN MFCC

Though MFCC is popular feature extraction technique in speaker recognition, its performance is directly proportional to noise. This scheme is very sensitive to noise interference, which tends to drastically degrade the performance of recognition systems because of the mis-matches between training and testing [12]. Also, channel mismatch problem in speaker verification arises because of different types of channels or handsets used in training and testing phase [13].

It is observed that there are few areas where performance of this technique degrades. The most of the researchers made effort in the field to extract robust features from MFCC technique. The following table describes the well-known existing algorithm.

Table 1 Contribution of various researchers for robust features from MFCC technique

Sr. No.	Author	Contribution	Modifications in MFCC	Performance Improvement
1.	Zunjing, Zhigang, 2005 [12]	Use of speech enhancement methods spectral subtraction and median-filter with modified MFCC Transformation.	Replacement of log function in standard MFCC with a combination of power law function and log function.	Recognition error rate is reduced over a wide SNR range. Improved system performance is obtained using TIMIT database.
2.	Murty, Yegnanarayana, 2006 [10]	Importance of complementary information (residual phase) for speaker recognition.	Use of Residual Phase Information by combining with MFCC features. Phase features were complementary to that of MFCC	An equal error rate of 10.5% is obtained by combination of MFCCs and residual phase on NIST-2003 database.It represents that speaker specific information is present in residual phase.
3.	Ajmeara et al., 2011 [6]	Used Radon transform and Discrete Cosine Transform based features for speaker identification.	Radon transform is substituted for mel-filter bank.	Limited Radon projections provides computationally efficient low dimensional feature vector. The proposed Radon based features are robust to channel and session variations.
4.	Hanilci et al., 2012 [13]	Recognition accuracy is improved by Regularized All-Pole Models (RLP, RWLP, and RSWLP).	DFT in MFCC is replaced by regularized linear prediction (RLP) method.	Experiments carried on NIST 2002 corpus indicates that proposed method performs better in noisy conditions. It outperforms DFT and LP techniques in terms of factory and babble noises.
5.	Selva Kumari et al., 2012 [14]	Inverted MFCC is used as complementary information.	Fused Inverted MFCC with conventional MFCC.	The performance is evaluated on a part of TIMIT database and a maximum identification efficiency achieved is 93.88%
6.	Nakagawa et al., 2012 [11]	Modified phase information extraction method.	The normalization method for phase information is proposed and combined with MFCCs.	Experiments are performed using NTT database. Speaker identification rate of 98.8% is achieved and Speaker verification equal error rate is reduced to 0.45%
7.	Kinnunen et al., 2012 [5]	The variance of spectrum estimate is reduced by low variance Multitaper MFCC features	The windowed DFT is replaced with multitaper spectrum estimate.	Experiments carried out on different corpora, NIST 2002, NIST 2008 showed consistent improvements. On the interview-interview condition, MinDCF is improved over baseline windowed DFT by relative 1) 20.4% for GMM-SVM classifier 2) 13.7% for GMM-JFA classifier On the telephone data, MinDCF is reduced by relative 18.7% for GMM-JFA classifier
8.	Ajmera and Holambe, 2012	Used fractional Fourier transform based features for	The Fourier transform in conventional MFCC scheme	The proposed scheme is robust against additive noise, session and

	[9]	speaker recognition	is replaced with Fractional Fourier Transform.	channel variations.
9.	Alam et al., 2013 [3]	Low-variance multitaper MFCC and PLP features are studied. Speaker verification using i-vector framework is also studied. A comparison of different weight selection for tapers is provided.	The spectral estimates obtained using a set of different tapers are averaged and then MFCC and PLP are computed.	Speaker verification results carried out on NIST 2010 SRE corpus. Improvements in recognition accuracy is provided by multitaper MFCC and PLP features. Equal error rate reduction by: a) Sine-weighted cepstrum estimator based multitaper For MFCC 12.3% For PLP 7.5% b) multipeak multitaper For MFCC 12.6% For PLP 11.6% c) Thomson multitaper For MFCC 9.5% For PLP 5.0%
10.	Sahidullahand Saha, 2013 [15]	Differentiation in frequency domain is proposed to compute MFCCs.	Hamming window technique is modified to obtain derivative of Fourier transform.	The proposed scheme is evaluated on NIST 2001, NIST 2004, and NIST 2006 databases. The performance is improved over single taper hamming window and multitaper windowing technique. The proposed method integrates complementary phase information which is applicable for speaker recognition.

4. FUTURE SCOPE

The solution for the problems of channel variations can be the combined use of short-term spectral and voice source features with high-level features to enhance recognition rate of speaker recognition. This is because short-term spectral and voice source features are easy to extract while high-level features are robust against channel mismatch effects and noise [2] [1].

The performance of speaker identification improves when complementary information extracted from high frequency part of energy spectrum of speech frame combined with conventional MFCC technique [14]. Complementary information pitch [14], prosody [14], dialectical features [14] can be combined with conventional methods like MFCC.

The speaker recognition accuracy can be increased if time-frequency representation related transform such as Fractional Fourier transform is used to substitute Fast Fourier transform [9]. Another approach can be the use of wavelet based feature for speaker recognition.

The choice of selected feature should be such that it should provide as much information as possible to identify the speaker. The reproduction of speech signal by the same speaker can be affected by speaker's health. The selection of feature in such case is very important. The feature which identifies the speaker would have low speaker variability, robust against noise, easy to compute from speech, difficult to mimic, and not to be affected by speaker's health [2].

5. CONCLUSION

We have reviewed feature extraction technique Mel-Frequency Cepstral Coefficients (MFCCs) for speaker recognition. We have also presented an overview of modifications suggested by different researchers in MFCC technique. The factors channel mismatch, background noise affects the performance of MFCC technique. The scheme for robust feature extraction is necessary. 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.

REFERENCES

[1]. 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.

[2]. 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.

[3]. Md Jahangir Alam , Tomi Kinnunen , Patrick Kenny , Pierre Ouellet, Douglas O'Shaughnessy, "Multitaper MFCC and PLP features for speaker verification using i-vectors", *Journal on Speech Communication*, Elsevier, Vol.55, Issue 2, Pages 237-251, February 2013.

[4]. 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.

[5]. 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.

[6]. 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.

[7]. 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*, Vol.6, Pages 124–129, 11 October 2011.

[8]. Mohamad Adnan Al-Alaoui, Lina Al-Kanj, Jimmy Azar, and Elias Yaacoub, "Speech Recognition using Artificial Neural Networks and Hidden Markov Models", *IEEE Multidisciplinary Engineering Education Magazine*, Vol. 3, No.3, September 2008.

[9]. 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.

[10]. 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.

[11]. 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.

[12]. 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*, Vol.10, Issue 2, April 2005.

[13]. Cemal Hanilçi, 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*, Vol.30, Pages 319–326, 13 March 2012.

[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.

