Effect of Varying MFCC Filters for Speaker Recognition

Amol A. Chaudhari
Department of Electronics Engineering
A.I.S.S.M.S. Institute of Information Technology, Pune, India

S.B. Dhonde
Assistant Professor, Department of Electronics Engineering
A.I.S.S.M.S. Institute of Information Technology, Pune, India

ABSTRACT
This paper presents speaker recognition system with emphasis on MFCC feature extraction scheme. The optimum number of MFCC filter selection is necessary for the performance of speaker recognition. In this paper, the number of MFCC filters are varied. The effect of varying filters on computational time required for training and testing phase is provided in this paper. The experimental results have been evaluated on the developed database of 75 speakers. The recognition rate achieved is 96% in case of 30 MFCC filters with approximate computational time (testing phase) of 79.66 seconds.

General Terms
Speech Signal Processing

Keywords
Speaker Recognition, Feature extraction, MFCC, LBG algorithm, Euclidean distance

1. INTRODUCTION
Speaker recognition refers to recognize persons from their voice. Speaker recognition can be classified into speaker identification and verification. Speaker identification is to identify person from known set of voices and speaker verification is to determine whether the person is who he/she claims to be [1]. Feature extraction is an important stage in speaker recognition system followed by feature matching. The features representing speaker specific properties are computed in feature extraction. In training mode, these feature vectors are stored in the database and compared with unknown speaker's speech features in testing mode.

Feature extraction is a process of retaining speaker specific information. Feature vectors which are compact, less redundant are formed in feature extraction process. The features can be classified into short-term spectral features, voice source features, spectro-temporal features, prosodic features and high-level features [2] [3]. LPCC and MFCC are most commonly used feature extraction techniques for speaker identification. MFCC scheme is generally preferred because of its robustness [3]. The spectrum of the windowed speech signal is integrated with mel filter bank followed by log and discrete cosine transform to obtain MFCC. The extraction of effective and efficient speech features is necessary to tackle the problem of channel variation and additive background noise. In literature there has been many attempts for robustness of speaker recognition system. The radon transform based speech features have been proposed in [3]. Wavelet packet transform with irregular decomposition for speaker identification has been proposed in [4]. Admissible wavelet packets based features for speaker identification have been proposed in [5]. Fused mel feature set for speaker identification has been proposed in [6]. The importance of coefficient order for speaker recognition has been demonstrated in [7]. Wavelet analysis and the effect of number of MFCC features on recognition accuracy has been proposed in [8]. The effect of feature vector size of MFCC on identification accuracy is presented in [9]. The number of choices in creating feature vectors from MFCC has been assessed in [10].In this paper, the effect of number of MFCC filters on recognition rate is studied. This paper is organized as follows. The feature extraction scheme is discussed in section 2. Speaker modeling and feature matching are described in section 3. Experimental set-up is presented in section 4. Results and discussion are presented in section 5 followed by conclusion in section 6.

2. FEATURE EXTRACTION
It is necessary to extract effective and efficient features which emphasize on speaker specific properties for speaker recognition. The quality of subsequent steps depends on feature extraction as it is first step in the speaker recognition [11]. In this section, we have discussed MFCC feature extraction scheme. The figure 1 shows the block diagram of MFCC scheme.

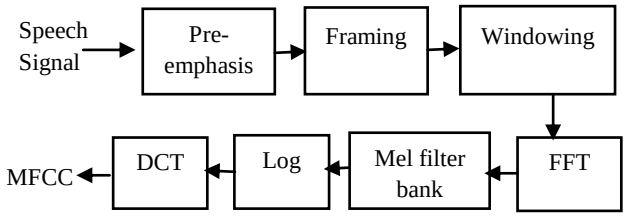

Fig 1: Block diagram of MFCC

The first step in MFCC scheme is pre-emphasis. Pre-emphasis is performed to boosts higher frequencies which are diminished during speech production mechanism. High-pass filter is usually preferred for pre-emphasis of speech signal. This filter is given by, $H(z) = 1-az^{-1}$ where, $0.9 \leq a \leq 1$. The speech signal is then divided into frames of duration 10-30 milliseconds with 25-50% overlap in the step called as framing [12]. Over this short duration, the speech signal is assumed to remain stationary. The overlapping is to avoid any loss of information. The each frame is then multiplied with window function in order to smooth the signal. Hamming window given by, $w(n) = 0.54 - 0.46 \cos(2\pi n/(T-1))$ is preferred as it provides better side lobes suppression [12]. The fast Fourier transform of windowed signal is calculated to obtain the spectrum. The windowed spectrum is multiplied with mel-filter bank. The mel filter bank is based on mel scale given by, $f(mel)=2595\times \log_{10}(1+f_{Hz}/700)$. This scale is based on how human ear perceives the sound. It is roughly linear upto 1 kHz and logarithmic above 1 kHz. The log operation is performed to separate the vocal tract response from excitation signal thereafter followed by discrete cosine transform which compacts the signal.

3. SPEAKER MODELING AND FEATURE MATCHING

The feature vectors created in feature extraction step are used to create codebook (speaker model) in training phase. This codebook is then stored in database (.mat file). Here, vector quantization approach is used to create codebooks. Codebook is created by clustering feature vectors of speaker's training samples into M clusters. These clusters are represented by centers which are used as codebook. LBG algorithm is used for clustering of feature vectors. In testing phase, a matching score is computed between extracted feature vectors and every speaker codebook enrolled in the system. Speaker who has a model with the least matching score is preferred as an author of the test speech sample. Euclidean distance between feature vectors and speaker models in testing phase is calculated.

4. EXPERIMENTAL SET-UP

AISSMSIOIT database is recorded for experiments at sampling frequency of 44.1 kHz and 48 kHz. It consists of 75 speakers (41 male and 34 female). This database contains 150 sentences, 2 sentences spoken by each of 75 speakers. Out of two sentences, speaker reading a prepared text consists of Rainbow passage and other sentence consists of a dialog prompted by their own out of given two sentences. For training purpose, speaker reading a prepared text of approximate duration of 35-40 seconds is used and for testing purpose, dialog is used which is of approximate duration 4-5 seconds. The recording is done using Goldwave software version 6.10. The recording has been carried out in college laboratory and housing society. Out of 75 speakers, seven speaker are having in the age group of 25-45 and remaining in age group of 20-26.

The speech signal is pre-emphasized with first order high pass-filter given by equation (z) = $1\text{-}0.95z^{-1}$. The signal is divided into frames of duration approximately 23 milliseconds (in case of 44.1 kHz) and approximately 21 milliseconds (in case of 48 kHz) with 50 % overlap over which it is assumed to remain stationary followed by the hamming windowing for tapering of the signal. The spectrum of the windowed signal is calculated by fast Fourier transform (FFT). The spectrum is then multiplied by mel-filter bank followed by logarithm and discrete cosine transform (DCT) to obtain MFCCs. The speaker model is generated for each speaker from the MFCCs using vector quantization (LBG algorithm). The speaker model is then stored in the database. In testing phase, MFCCs of an unknown speaker are extracted. The Euclidean distance between MFCCs and speaker model stored in the database is calculated. The speaker is recognized on the basis of minimum Euclidean distance computed between MFCC features in testing phase and speaker model stored in database in training phase. The experiments are carried out for different number of MFCC filters i.e. 13, 20 and 30 on recorded AISSMSIOIT database. In all experiments, 0^{th} cepstral coefficient is excluded. The number of clusters of VQ were 32.

5. RESULTS AND DISCUSSION

The recognition rate in percentage of speaker recognition is carried out using following formula.

$$\text{Recognition Rate} = \frac{\text{Number of correct matches}}{\text{Total number of test speaker}} \times 100\% \qquad [3]$$

The figure 2 shows the recognition rate by varying number of MFCC filters.

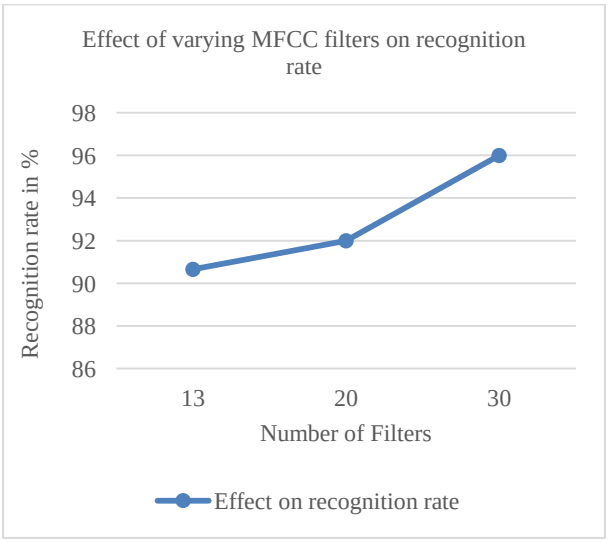

Fig 2: Effect of varying MFCC filters on recognition rate

From this figure, it is observed that recognition rate is increased as number of MFCC filters are increased. However, increasing number of MFCC filters increases the computational time required for both training as well as testing phase. This effect is shown in figure 3.

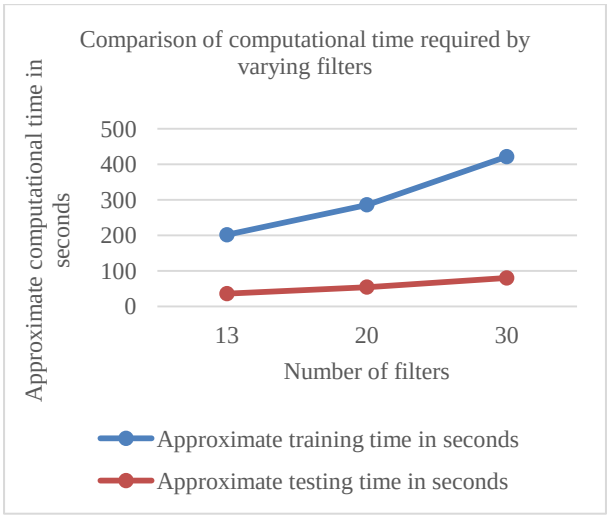

Fig 3: Comparison of computational time required by varying filters

It is observed that as number of MFCC filters increases, the % recognition rate also increases. But, increasing number of filters has impact on computational time required for training as well as testing phase.

6. CONCLUSION

In this paper, speaker recognition system is presented. This paper has mainly focused on effect of varying number of filters in recognition rate. It is found that there is a trade-off between recognition rate and computational time. Recognition rate increases as the number of MFCC filters are increased but, this also increases the computational time required for both training as well as testing phase. The recognition rate achieved is 96% in case of 30 MFCC filters with an increment in computational time as compared to other cases of MFCC filters. The number of feature vectors, their dimensionality, and number of speakers are responsible for recognition time. The optimum number of MFCC filters for speaker recognition system with acceptable computational time can be selected in the range 20-30.

7. REFERENCES

[1] Douglas A. Reynolds, "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, no. 1, pp. 12–40, 2010.

[3] 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 on Pattern Recognition*, Elsevier, vol. 44, no. 10-11, pp. 2749-2759, 2011.

[4] Jian-Da Wu, Bing-Fu Lin, "Speaker identification using discrete wavelet packet transform technique with irregular decomposition", Journal on Expert Systems with Applications, Elsevier, vol. 36, no. 2, pp. 3136–3143, 2009.

[5] Mangesh S. Deshpande, Raghunath S. Holambe, "New Filter Structure based Admissible Wavelet Packet Transform for Text-Independent Speaker Identification", International Journal of Recent Trends in Engineering, vol. 2, no. 5, pp. 121-125, 2009.

[6] 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, Journal on Procedia Engineering, Elsevier, vol. 30, pp. 319–326, 2012.

[7] R. Rajeshwara Rao, A. Prasad, Ch. Kedari Rao, "Robust Features for Automatic Text-Independent Speaker Recognition Using Gaussian Mixture Model", International Journal of Soft Computing and Engineering, vol. 1, Issue 5, November 2011.

[8] Noor Almaadeed, Amar Aggoun, Abbes Amira, "Speaker identification using multimodal neural networks and wavelet analysis", IET Journals and Magazines, vol. 4, no. 1, pp. 18-28, 2015.

[9] Ahmad, K.S.; Thosar, A.S.; Nirmal, J.H.; Pande, V.S., "A unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network," Advances in Pattern Recognition (ICAPR), 2015, pp. 1- 6, January 2015.

[10] M.Hassan Shirali-Shahreza, Sajad Shirali-Shahreza, "Effect of MFCC Normalization on Vector Quantization Based Speaker Identification", Signal Processing and Information Technology (ISSPIT), 2010, pp.250,253, December 2010.

[11] 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, no. 2, pp. 237-251, 2013.

[12] Holambe, Raghunath S., Deshpande, Mangesh S., "Advances in Non-Linear Modeling for Speech Processing", SpringerBriefs in Speech Technology, Section 6, pp. 77-82, ISBN 978-1-4614-1505-3, 2012.