Wavelet Packet Based Mel Frequency Cepstral Features for Text Independent Speaker Identification

Smriti Srivastava, Saurabh Bhardwaj, Abhishek Bhandari, Krit Gupta, Hitesh Bahl, and J.R.P. Gupta

Abstract. The present research proposes a paradigm which combines the Wavelet Packet Transform (WPT) with the distinguished Mel Frequency Cepstral Coefficients (MFCC) for extraction of speech feature vectors in the task of text independent speaker identification. The proposed technique overcomes the single resolution limitation of MFCC by incorporating the multi resolution analysis offered by WPT. To check the accuracy of the proposed paradigm in the real life scenario, it is tested on the speaker database by using Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) as classifiers and their relative performance for identification purpose is compared. The identification results of the MFCC features and the Wavelet Packet based Mel Frequency Cepstral (WP-MFC) Features are compared to validate the efficiency of the proposed paradigm. Accuracy as high as 100% was achieved in some cases using WP-MFC Features.

Keywords: WPT, MFCC, HMM, GMM, Speaker Identification.

1 Introduction

Human beings possess several inherent characteristics that assist them distinguish from one another. Over the years, biometrics has emerged as the science which assimilates and tries to mimic the powers of the human brain by capturing unique personal features and consequently performing the task of human identification. Voice as a biometric tool has interested plethora of researchers as it can be easily intercepted, recorded and processed. Moreover, voice biometrics offers simple and secure mode of remote access transactions over telecommunication networks by authenticating the speaker first and then carrying out the required transactions. Hence, applications of speech processing technology are broadly classified into:

Smriti Srivastava · Saurabh Bhardwaj · Abhishek Bhandari · Krit Gupta · Hitesh Bahl · J.R.P. Gupta
Netaji Subhas Institute of Technology, New Delhi 110078, India

A. Abraham and S.M. Thampi (Eds.): Intelligent Informatics, AISC 182, pp. 237–247.
springerlink.com

Speech Recognition and *Speaker Recognition*. Speech recognition is the ability to identify the spoken words while speaker recognition is the ability to discriminate between people on the basis of their voice characteristics. Further the task of speaker recognition is dissected into two categories, *speaker identification* and *speaker verification*. Speaker identification is to classify that the test speech signal belongs to which one of the N- reference speakers whereas speaker verification is to validate whether identity claimed by an unknown speaker is true or not, consequently this type of decision is binary. Several recognition systems behave in a text-dependent way, i.e. the user utters a predefined key sentence. But, text dependent type of recognition process is only feasible with "cooperative speakers". Consider criminal investigation as an application (an unwilling speaker), here recognition can only be performed in text-independent mode. With increased applications of speech as a means of communication between the man and the machine, speaker identification has emerged as a powerful tool [1]. The phenomenon of speaker recognition has been in application since the 1970's [2]. Most of the state of art identification systems uses MFCC for front-end-processing as its performance is far superior compared to all other feature extraction mechanisms as described in [3]. The paper is organized as follows. Section 2 gives a description of the modules of speaker recognition. The proposed algorithm is described in section 3. Finally, the results are demonstrated in section 4.

2 Modules for Speaker Recognition

All speaker recognition systems contain two main modules, *feature extraction* and *feature or pattern matching*. Feature extraction is the process that extracts information from a voice signal of a speaker. Feature matching is the procedure to identify the unknown speaker by matching his features with those of known speakers. Sound pressure waves are acquired with the help of a microphone or some other voice recording device. This signal is then pre-processed. Speaker recognition using the pre-processed signal is accomplished in two stages, Enrollment or feature extraction and pattern matching or classification as depicted in fig.1.

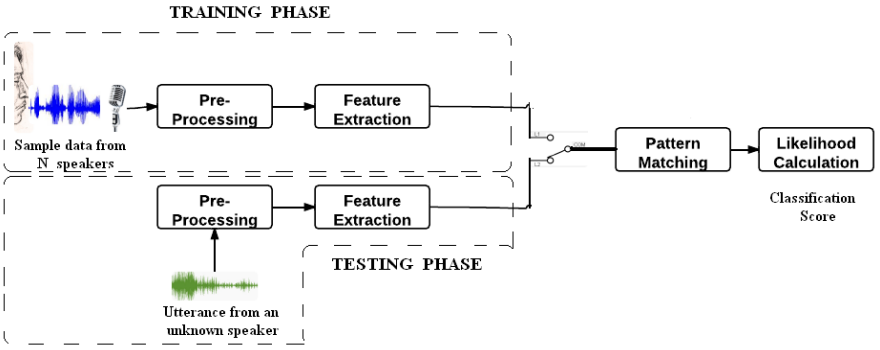

Fig. 1 Block Diagram of a Speaker Recognition system

During enrollment phase, speech sample from several speakers is recorded and a number of features are extracted using one out of the several methods available to produce individual's "voice model or template". During the next phase, pattern of an unknown utterance is compared with the previously recorded template. For speaker identification applications, speech utterance from an unknown speaker is compared with voice prints of all reference speakers. The unknown speaker is identified as that reference speaker whose voice model best matches with the model of unknown utterance. The performance of speaker identification system decreases with increasing population size. [1]

2.1 Feature Extraction

The mechanism of speech feature extraction reduces the dimensionality of the input signal by eliminating the redundant information while maintaining the discriminating capability of the signal [4]. Given the data of speech samples, a variety of auditory features are computed for each input set which constitute the feature vector. The present research proposes Wavelet Packet based Mel Frequency Cepstral feature extraction approach.

2.1.1 Mel Frequency Cepstral Coefficients

The advent of Mel Frequency Cepstral Coefficient (MFCC) technique for the task of feature extraction has over shadowed the existence of majority of its predecessor methods as it acknowledges human sound perception sensitivity with respect to frequency, providing better sound feature vectors. The most conspicuous difference between cepstral coefficients and MFCC is that the latter uses Mel filter banks to transform the frequency domain to Mel frequency domain [5]. The formula to convert f (Hz) into m (Mel) is as follows:

$$m = (2595 \log_{10} (1 + \frac{f}{700})) \quad (1)$$

The block diagram of MFCC feature extraction algorithm is as shown in fig.2.

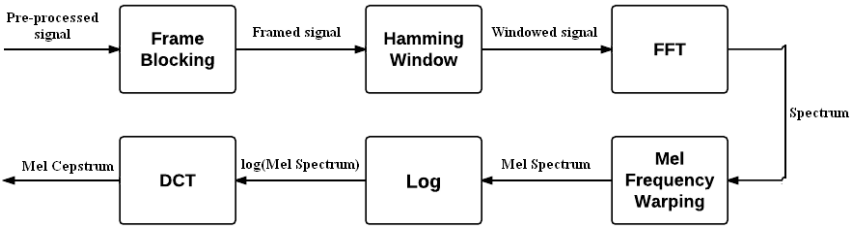

Fig. 2 Block Diagram implementation of the technique

Pre-processed speech signal is frame blocked with each frame having length of 25ms with an overlapping length of 15ms. The signal is then multiplied over short-time windows to avoid problems arising due to truncation of the signal. For

our analysis, a hamming window is utilized. For each windowed frame, spectrum is computed using Fast Fourier Transform (FFT).Spectrum is passed through Mel filter bank to obtain the Mel spectrum. In the present work, 40 filters were used [6].Finally, Cepstral analysis is performed on the output of Mel filter banks using only 13 coefficients out of 40. The logarithm followed by the Discrete Cosine Transform (DCT) of the Mel spectrum produces a set of feature vectors (one vector corresponding to each frame) which are then termed as MFCC.

2.2 Hidden Markov Model

Hidden Markov Model (HMM) [7,8] springs forth from Markov Processes or Markov Chains. It is a canonical probabilistic model for the sequential or temporal data It depends upon the fundamental fact of real world, "Future is independent of the past and given by the present". HMM is a doubly embedded stochastic process, where final output of the system at a particular instant of time depends upon the state of the system and the output generated by that state. There are two types of HMMs: Discrete HMMs and Continuous Density HMMs. These are distinguished by the type of data that they operate upon. Discrete HMMs (DHMMs) operate on quantized data or symbols, on the other hand, Continuous Density HMMs (CDHMMs) operate on continuous data and their emission matrices are the distribution functions. HMM Consists of the following parameters

O {O1,O2…OT}	:	Observation Sequence
Z {Z1, Z2…ZT}	:	State Sequence
T	:	Transition Matrix
B	:	Emission Matrix/Function
π	:	Initialization Matrix
$\lambda(T, B, \pi)$	:	Model of the System
ρ	:	Space of all state sequence of length T
m $\{m_{q1},m_{q2}….m_{qT}\}$	:	Mixture component for each state at each time
c_{il}, μ_{il}, $\sum_{il}$	:	Mixture component (i state and l component)

Single state HMM is known as GMM. For the purpose of text independent speaker identification, GMM has had a greater success over HMM [9]. There are three major design problems associated with an HMM outlined here. Given the Observation Sequence {O1, O2, O3,.., OT} and the Model $\lambda(T, B, \pi)$, the first problem is the computation of the probability of the observation sequence P $(O|\lambda)$.The second is to find the most probable state sequence Z {Z1, Z2,.., ZT}, the third problem is the choice of the model parameters λ (T, B,π), such that the probability of the Observation sequence, P $(O|\lambda)$ is the maximum. The solution to the above problems emerges from three algorithms: Forward, Viterbi and Baum-Welch [7].

2.2.1 Continuous Density HMM

Let O = {O1,O2...OT} be the observation sequence and Z {Z1, Z2…ZT} be the hidden state sequence. Now, we briefly define the Expectation Maximization

(EM) algorithm for finding the maximum-likelihood estimate of the parameters of a HMM given a set of observed feature vectors. EM algorithm is a method for approximately obtaining the maximum a posteriori when some of the data is missing, as in HMM in which the observation sequence is visible but the states are hidden or missing. The Q function is generally defined as

$$Q(\lambda, \lambda') = \sum_{q \varepsilon \rho} \log P(0, z \mid \lambda) P(0, z \mid \lambda') \tag{2}$$

To define the Q function for the Gaussian mixtures, we need the hidden variable for the mixture component along with the hidden state sequence. These are provided by both the E–step and the M-step of EM algorithm given

E Step:

$$\mathrm{Q}(\lambda, \lambda') = \sum_{\mathrm{z} \in \rho} \sum_{m \in M} \log P(O, z, m \mid \lambda) P(O, z, m \mid \lambda') \tag{3}$$

M Step:

$$\lambda' = \arg m_{\lambda} ax[Q(\lambda, \lambda')] + constrainment \tag{4}$$

The optimized equations for the parameters of the mixture density are:

$$\mu_{il} = \frac{\sum_{t=1}^{T} O_t P(z_{t=1,} m_{z_t t} = 1 \mid O_t \lambda')}{\sum_{t=1}^{T} P(z_{t=1,} m_{z_t t} = 1 \mid O_t \lambda')} \tag{5}$$

$$\sum_{il} = \frac{\sum_{t=1}^{T} (O_t - \mu_{il})(O_t - \mu_{il})^T P(z_t = i, m_{z_i t} = 1 \mid O_t \lambda')}{\sum_{t=1}^{T} P(z_t = i, m_{z_i t} = 1 \mid O_t \lambda')} \tag{6}$$

$$c_{il} = \frac{\sum_{t=1}^{T} P(z_t = i, m_{z_i t} = 1 \mid O_t \lambda')}{\sum_{t=1}^{T} \sum_{l=1}^{M} P(z_t = i, m_{z_i t} = 1 \mid O_t \lambda')} \tag{7}$$

3 Proposed Method

3.1 Discrete Wavelet and Wavelet Packet Transform

For discrete wavelet transform we have:

$$F[n] = \frac{1}{\sqrt{M}} \cdot \sum_{k} W_{\phi}[j_0, k] \phi_{j_0, k}[n] + \frac{1}{\sqrt{M}} \cdot \sum_{j=j_0}^{\infty} \sum_{k} W_{\psi}[j_0, k] \psi_{j,k}[n] \tag{8}$$

Here F[n], $\phi_{j0,k}[n]$ and $\Psi_{j,k}[n]$ are discrete functions defined in [0,M-1], a total of M points. Now,

$$W_{\phi}[j_0,k] = \frac{1}{\sqrt{M}} \cdot \sum_n f[n]\phi_{j_0,k}[n] \tag{9}$$

$$W_{\psi}[j_0,k] = \frac{1}{\sqrt{M}} \cdot \sum_n f[n]\psi_{j,k}[n] j \geq j_0 \tag{10}$$

$W_{\Phi}[j_0,k]$ are called *approximation coefficients* while $W_{\Psi}[j_0,k]$ are called *detailed coefficients*. These coefficients are obtained by using *Mallat algorithm* proposed in [10].

3.1.1 Wavelet Packet Transform

In the DWT decomposition, to obtain the next level coefficients, scaling coefficients (low pass branch in the binary tree) of the current level are split by filtering and down sampling [10]. With the wavelet packet decomposition, the wavelet coefficients (high pass branch in binary tree) are also split by filtering and down sampling. The splitting of both the low and high frequency spectra results in a full binary tree shown in fig.3 and a completely evenly spaced frequency resolution (In the DWT analysis, the high frequency band was not split into smaller bands).

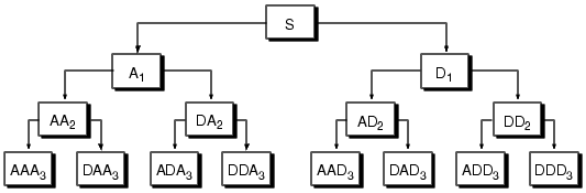

Fig. 3 Wavelet packet decomposition tree

3.2 Motivation

Speech is a "Quasi-stationary" signal. MFCC utilizes short time Fourier Transform (STFT) which provides information regarding the occurrence of a particular frequency at a time instant with a limited precision, with the resolution according to the Heisenberg Uncertainty principle dependent on the size of the analysis window.

$$Time * Frequency = \Delta t \Delta f \geq \frac{1}{4\pi} \tag{11}$$

Narrower windows provide better time resolution while wider ones provide better frequency resolution [11]. Even though STFT tries to strike a balance between the time and frequency resolution, it is admonished primarily as it keeps the length of the analysis window fixed for all frequencies resulting in uniform-partition of the time-frequency plane as shown in fig.4.

Fig. 4 Time-frequency plane uniformly partitioned in STFT

Fig. 5 Time-frequency plane non-uniformly spaced (constant area) in wavelet transform

Speech signals require a more flexible multi-resolution approach where window length can be varied according to the requirement to cater better time or frequency resolution. Wavelet Packet Transform (WPT) offers a remedy to this difficulty by providing well localized time and frequency resolution as shown in fig.5. Further, multi-resolution property of WPT makes it more robust in noisy environment as compared to single-resolution techniques and has better time-frequency characteristics. But, WPT increases the computational burden and is time consuming. Conventional wavelet packet transform mechanisms do not warp the frequencies according to the human auditory perception system. So, in this work an attempt is made for utilizing the advantages of the Mel Scale and multi-resolution wavelet packet transform to generate feature vector for the task of speaker identification.

3.3 Proposed Paradigm

3.3.1 Wavelet Packet Based Mel Frequency Cepstral Features

The block diagram for proposed approach is as shown in fig.6

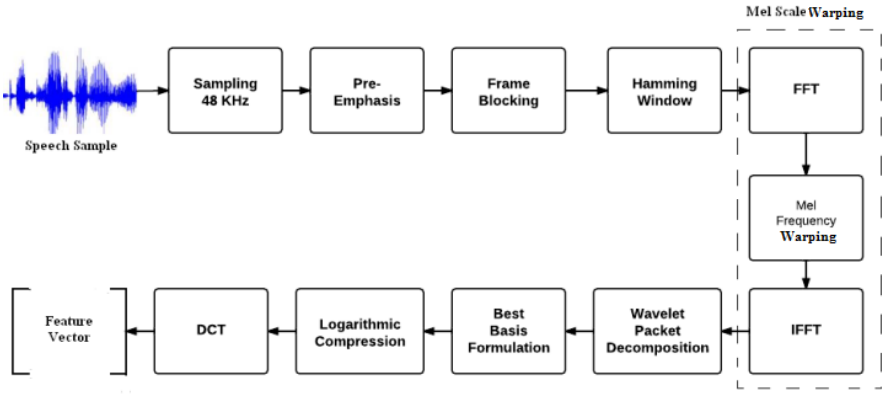

Fig. 6 Block diagram representation of proposed method

The analytical steps followed for feature extraction are as stated:

- The raw speech signal was primarily sampled at 48 kHz in order to further process it.

- Next, a framing window was utilized. The frame size was kept fixed to 25 milliseconds, a skip rate of 10 milliseconds was selected to accommodate for the best continuity.
- A pre-emphasis filter as described by equation (12) was next exercised in order to improve the overall signal-to-noise ratio. A rectangular Hamming window was deployed for framing.

$$H(z) = 1 - 0.97z^{-1} \tag{12}$$

- The resultant signal was transformed from time domain to frequency domain by applying Fast Fourier Transform (FFT). Then the signal in frequency domain was Mel-Warped using Triangular Mel Filter Banks. Afterwards, signal was again transformed to time domain by applying Inverse Fast Fourier Transform for further processing of signal.
- Next, wavelet packet decomposition was applied using daubechies4 (D4) wavelet. For a full j=7 level decomposition, the WPT corresponds to a maximum frequency of 31.25 Hz giving 128 sub-bands.
- Out of 128 frequency sub-bands 35 frequency sub-bands were used for further processing since higher frequency coefficient contained paltry amount of energy and first 35 coefficients represented 99.99%. The energy in each band was evaluated, and was then divided by the total number of coefficients present in that particular band. In particular, the sub band signal energies were computed for each frame as,

$$E_j = \frac{\sum_{j=1}^{N_j} [W_j^p f(i)]^2}{N_j}, \; j = 1,....,35 \tag{13}$$

- Lastly, a logarithmic compression was performed and a Discrete Cosine Transform (DCT) was applied on the logarithmic sub-band energies to reduce dimensionality:

$$F(i) = \sum_{n=1}^{B} \log_{10} E_n Cos\left(\frac{i(n-1/2)}{B}\right), i = 1,...., r. \tag{14}$$

3.3.2 Speaker Identification

After extracting the features we have used HMM or single state HMM called Gaussian Mixture Model (GMM) for the identification. The whole procedure is as explained in fig.7. Having the WP-MFC Feature from the speech signals, CDHMMs are trained for each speaker using Baum Welch (BM) algorithm which gives the parameters of the corresponding CDHMMs. Now the identification process can be described as follows: Given a test vector 'X' the log-likelihood of the trained batches with respect to their HMM models 'λ' is computed as

$\log P(X \mid \lambda)$. From 'N' HMMs $\{\lambda_1, \lambda_2, ...\lambda_N\}$ corresponding to 'N' speakers, the speaker can be identified with a test sequence using:

$$P(X \mid \lambda_{required}) = F[P(X \mid \lambda_1), ...P(X \mid \lambda_N) \quad (15)$$

Where $F()$ is the maximum of the likelihood values of the model $(\lambda_1, \lambda_2, ...\lambda_N)$. The model corresponding to the highest Log-Likelihood value is selected as the identified speaker.

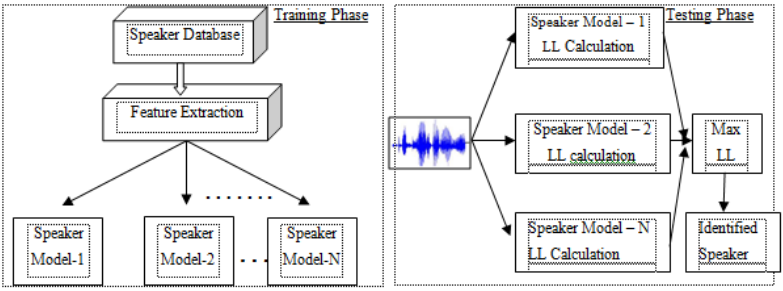

Fig. 7 Procedure for the classification algorithm

4 Experimental Results

Having acquired the appropriate test samples from the free online English speech database site [12], the database was created containing speech samples of 30 distinct speakers with 10 non-identical utterances each. Speaker models were created using 8 samples per speaker and testing was done using 2 samples of each speaker. The results of the identification process using GMM and HMM are displayed in table1, fig.8(a) and table 2, fig.8(b) respectively. The number of

Table 1 No. of states (Q) = 1 (GMM)

S.No.	No. of Gaussian Mixtures (M)	No. of States (Q)	No. of Speakers Recognized	
			WP-MFC Features	MFCC Features
1.	11	1	30	28
2.	12	1	30	28
3.	13	1	30	28
4.	14	1	30	28
5.	15	1	30	28

Table 2 No. of states (Q) = 2 (HMM)

S.No.	No. of Gaussian Mixtures (M)	No. of States (Q)	No. of Speakers Recognized	
			WP-MFC Features	MFCC Features
1.	11	1	27	25
2.	12	1	29	28
3.	13	1	30	27
4.	14	1	27	28
5.	15	1	27	29

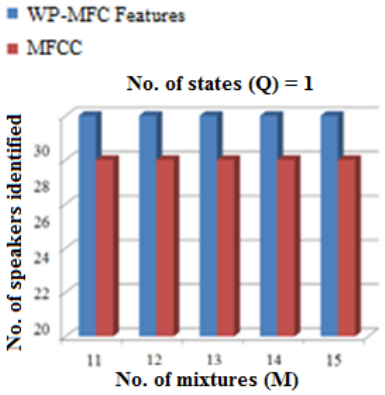

Fig. 8(a) Output Results with GMM

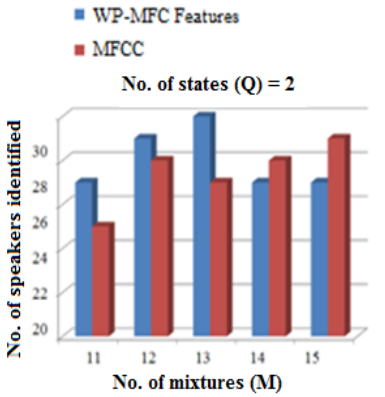

Fig. 8(b) Output Results with HMM

states (Q) was kept constant whereas the number of mixtures (M) was varied in each case.

5 Conclusion

Speaker Recognition is the use of machine to recognize a speaker from the spoken words. In this paper, we introduced a robust feature extraction technique for deployment with speaker identification system. These new feature vectors termed as Wavelet Packet based Mel frequency Cepstral (WP-MFC) Coefficients offer better time and frequency resolution. HMM and GMM were used to classify the acoustic data. Experimental results of the comparison between the performance of the proposed feature vectors and MFCC reveal the real life effectiveness of the proposed method. Also, better performance of GMM over HMM for speaker identification was confirmed.

References

[1] Reynolds, D.A.: Speaker Identification and Verification Using Gaussian Mixture Speaker Models. Speech Communication 17 (1995)

[2] Bolt Richard, H., Cooper Franklin, S., David Edward Jr., E., Denes Peter, B., Pickett James, M., Stevens Kenneth, N.: Speaker Identification by Speech Spectograms: A Scientists' View of its Reliability for Legal Purposes. The Acoustic Society of America 47 (1970)

[3] Reynolds Douglas, A.: Identification, Experimental Evaluation of Features for Robust Speaker. IEEE Transactions on Speech and Audio Processing 77, 257–285 (1994)

[4] Gaikwad Santosh, K., Gawali Bharti, W., Pravin, Y.: A Review on Speech Recognition Technique. International Journal of Computer Applications 10 (2010)

[5] Sirko, M., Michael, P., Ralf, S., Hermann, N.: Computing Mel-frequency coefficients on Power Spectrum. IEEE Proceedings of IEEE 1, 73–76 (2001)

[6] Chen, S.-H., Luo, Y.-R.: Speaker Verification Using MFCC and Support. In: Proceedings of the International MultiConference of Engineers and Computer Scientists (2009)

[7] Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition, pp. 257–286 (1989)

[8] Blimes, J.A.: A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. International Computer Science Institute (1998)

[9] Reynolds, D.A., Campbell, W.M.: Springer Handbook of Speech Processing. Text Independent Speaker Recognition. Springer (2008)

[10] Mallat, S.G.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE 111, 674–693 (1989)

[11] Robi, P.: The Engineers Ultimate Guide to Wavelet Analysis (2012), `http://users.rowan.edu/~polikar/wavelets/wttutorial.html` (accessed March 20, 2012)

[12] VoxForge (2012), `http://www.voxforge.org/home/downloads/speech/english` (accessed February 20, 2012)