

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

Authors: Saurabh Bhardwaj¹, Smriti Srivastava¹, Abhishek Bhandari¹, Krit Gupta¹, Hitesh Bahl¹ and J.R.P. Gupta¹

Netaji Subhas Institute of Technology, New Delhi 110078, India.

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. The results are promising.

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: *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 the art identification systems use 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. Sub-section 2.1 deals with the process of Feature Extraction. Sub-section 2.2 illustrates the classifier used, that is, the HMM. The proposed algorithm is described in section 3. The results are demonstrated in section 4.

2 Modules for Speaker Recognition:

All speaker recognition systems contain two main modules, feature extraction and feature 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 digital 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. 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 depends upon the population size and decreases with increasing population size. (1)

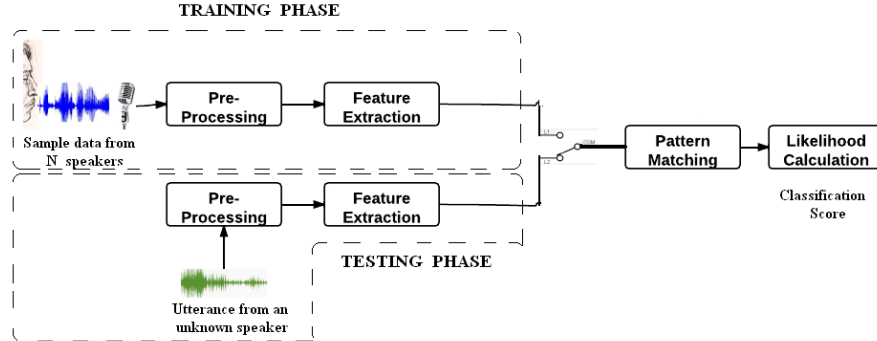


Fig.1 Block Diagram of a Speaker Recognition system

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 overshadowed the existence of majority of its predecessor methods as it acknowledges human sound perception sensitivity with respect to frequency, catering 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

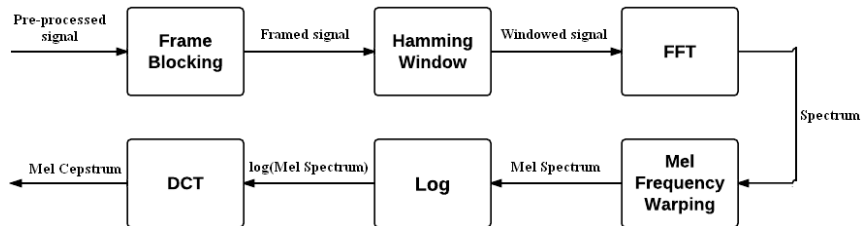


Fig.2 Block Diagram implementation of the technique

Pre-processed speech signal is frame blocked with each frame having length of 25ms with an overlap 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 our work, we have used 40 filters (6). Finally, Cepstral analysis is performed on the output of Mel filter banks using only 13 coefficients out of 40, thereby taking logarithm followed by the Discrete Cosine Transform (DCT) of the Mel spectrum to obtain a set of feature vectors (one vector corresponding to each frame) are 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, the Continuous Density HMMs (CDHMMs) operate on continuous data and their emission matrices are the distribution functions. HMM Consists of the following parameters

$O \{O_1, O_2, \dots, O_T\}$:	Observation Sequence
$Z \{Z_1, Z_2, \dots, Z_T\}$:	State Sequence
T	:	Transition Matrix
B	:	Emission Matrix/Function
π	:	Initialization Matrix
$\lambda(T, B, \pi)$:	Model of the System
\mathcal{Q}	:	Space of all state sequence of length T
$m \{m_{q1}, m_{q2}, \dots, m_{qT}\}$:	Mixture component for each state at each time
$c_{il}, \mu_{il}, \sum_{il}$:	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 $\{O_1, O_2, O_3, \dots, O_T\}$ 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 $Z \{Z_1, Z_2, \dots, Z_T\}$,

The third problem is the choice of the model parameters $\lambda(T, B, \pi)$, 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 $O = \{O_1, O_2 \dots O_T\}$ be the observation sequence and $Z = \{Z_1, Z_2 \dots Z_T\}$ 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

$$Q(\lambda, \lambda') = \sum_{q \in \rho} \log P(O, z | \lambda) P(O, z | \lambda') \quad (2)$$

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:

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

M Step:

$$\lambda' = \arg \max_{\lambda} [Q(\lambda, \lambda')] + \text{constraint} \quad (4)$$

The optimized equations for the parameters of the mixture density are

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

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

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

3 Proposed Method:

3.1 Discrete Wavelet and Wavelet Packet Transform:

For discrete wavelet transform we have:

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

Here $f[n]$, $\phi_{j_0, k}[n]$ and $\psi_{j, k}[n]$ are discrete functions defined in $[0, M-1]$, totally M points. Now,

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

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

$W_{\phi}[j_0, k]$ are called approximation coefficients while $W_{\psi}[j_0, k]$ are called *detailed coefficients*. These coefficients are obtained by using *Mallat algorithm* proposed in (9).

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 (9). 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).

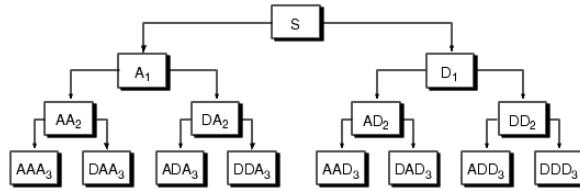


Fig.3 Wavelet packet decomposition tree

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 particular time instant with a limited precision, with the resolution according to the Heisenberg Uncertainty principle dependent on the size of the analysis window.

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

Narrower windows provide better time resolution while wider ones provide better frequency resolution (10). 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.

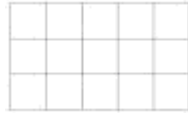


Fig.4 Time-frequency plane uniformly partitioned in STFT

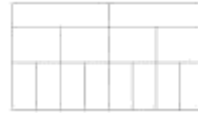
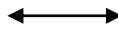


Fig.5 Time-frequency plane non-uniformly spaced (constant area) in wavelet transform

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 characteristic. But, WPT increases the computational burden and is time consuming and conventional wavelet packet transform mechanisms do not warp the frequencies according to the human auditory perception system. So, in our 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 Method:

3.3.1 Wavelet Packet based Mel Frequency Cepstral Features:

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

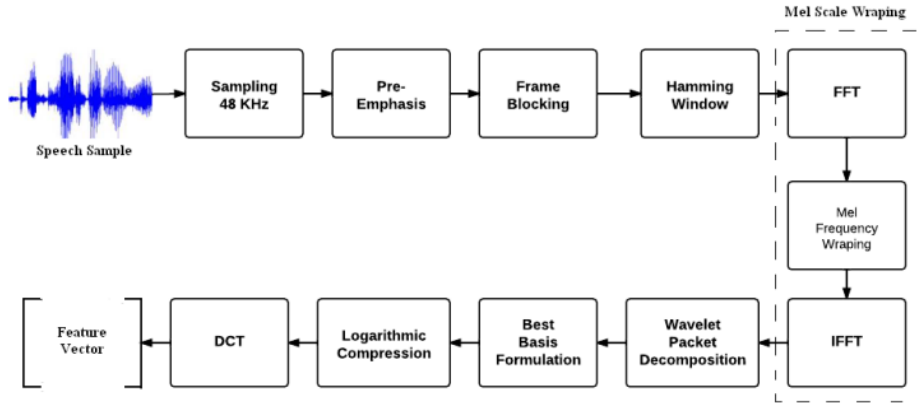


Fig.6 Block diagram representation of proposed method

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.
- A framing window was utilized next. The frame size was kept fixed to 25 milliseconds, a skip rate of 10 milliseconds was chosen 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.

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

- 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-Wrapped 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 wavelet packet transform corresponds to a maximum frequency of 31.25 Hz giving us 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 then it was 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,

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

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

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

3.3.2 Speaker Identification:

After extracting the features we have used HMM or single state HMM called GMM (Gaussian Mixture Model) 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 $\log P(X|\lambda)$.

From 'N' HMMs $\{\lambda_1, \lambda_2, \dots, \lambda_N\}$ corresponding to 'N' speakers, the speaker can be identified with a test sequence using:

$$P(X | \lambda_{required}) = F[P(X | \lambda_1), \dots, P(X | \lambda_N)] \quad (15)$$

Where $F()$ is the maximum of the likelihood values of the model

$(\lambda_1, \lambda_2, \dots, \lambda_N)$. The model corresponding to the highest Log-Likelihood value is selected as the identified speaker.

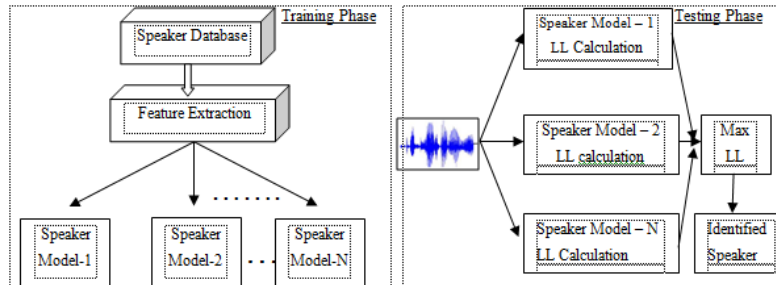


Fig.7 Procedure for the classification algorithm

4 Experimental Results:

Having acquired the appropriate test samples from the free online English speech database site (11), we created the database containing speech samples of 30 distinct speakers with 10 non-identical utterances each. Then, tabular (table 1, 2) as well as pictorial (fig 8, 9) representation of the results we obtained. The number of states (Q) was kept constant in each case whereas the number of mixtures (M) was varied in each case.

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

S.No.	GMM Parameters		No. of Speakers Recognized	
	No. of mixtures (M)	No. of states (Q)	Hybrid 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

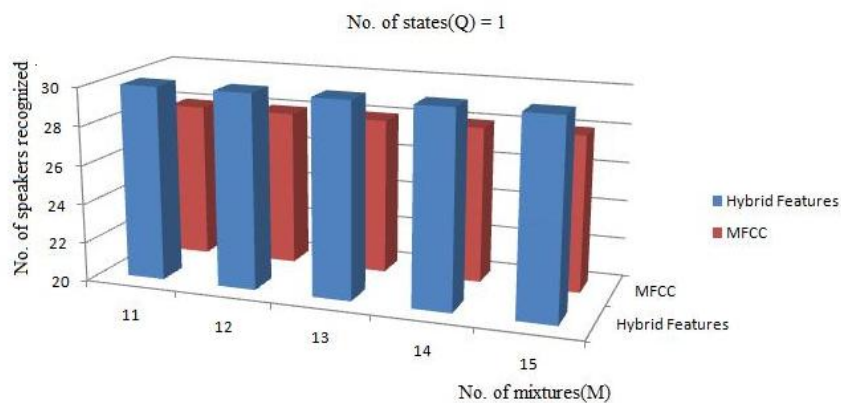
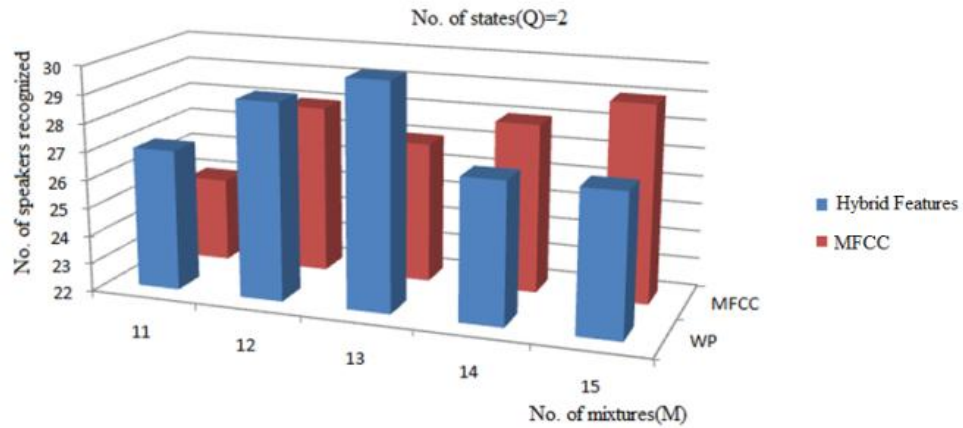


Fig.8 Output Results with no. of states (Q) = 1

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

S.No.	HMM Parameters		No. of Speakers Recognized	
	No. of mixtures (M)	No. of states (Q)	Hybrid features	MFCC features
1.	11	2	27	25
2.	12	2	29	28
3.	13	2	30	27
4.	14	2	27	28
5.	15	2	27	29

**Fig.9** Output Results with no. of states (Q) = 2

5 Conclusions:

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.

References:

1. Reynolds, D. A.(1995) Speaker Identification and Verification Using Gaussian Mixture Speaker Models. Speech Communication 17.
2. Richard H. Bolt, Franklin S. Cooper, Edward E. David Junior, Peter B. Denes , James M. Pickett, Kenneth N. Stevens (1970) Speaker Identification by Speech Spectrograms: A Scientists' View of its Reliability for Legal Purposes. The Acoustic society of America 47.

3. Douglas A. Reynolds (1994) Identification, Experimental Evaluation of Features for Robust Speaker. 4. IEEE 2.
4. Santosh K. Gaikwad, Bharti W. Gawali, Pravin Yannawar (2010) A Review on Speech Recognition Technique. International Journal of Computer Applications 10.
5. Sirko Molau, Michael Pitz, Ralf Schlüter, and Hermann Ney (2001) Computing Mel-frequency coefficients on Power Spectrum. IEEE Proceedings of IEEE 1:73-76.
6. Shi-Huang Chen, Yu-Ren Luo. Hong Kong (2009) Speaker Verification Using MFCC and Support. Proceedings of the International MultiConference of Engineers and Computer Scientists.
7. Rabiner, L. R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition 257–286.
8. Blimes, J. A. (1998) A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. International Computer Science Institute.
9. D. A. Reynolds, W. A. Campbell. Handbook of speech processing. Springer.38:
10. Mallat, S. G. (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE 111: 674-693.
11. Polikar, Robi (2012) THE ENGINEER'S ULTIMATE GUIDE TO WAVELET ANALYSIS. <http://users.rowan.edu/~polikar/wavelets/wttutorial.html>. Accessed 20 March 2012.
12. VoxForge (2012) <http://www.voxforge.org/home/downloads/speech/english>. Accessed 20 February 2012.
13. Wikipedia (2012) Mel Frquency Cepstrum. http://en.wikipedia.org/wiki/Mel-frequency_cepstrum. Accessed 15 March 2012.