Speaker Identification System using Wavelet Transform and VQ modeling Technique

Shailaja S Yadav
ME (Pursuing)
Dept of Electronics and Telecomm. Engg
Rajarshi Shahu College of Engg. Tathawade,
Pune, India-411033

D.G. Bhalke
Professor
Dept of Electronics and Telecomm. Engg
Rajarshi Shahu College of Engg. Tathawade,
Pune, India-411033

ABSTRACT
In this paper, speaker identification system based on the wavelet transform is introduced. The proposed system identifies the speakers by their acoustic characteristics in speech signal of speakers. In this system, pre-processing of speech signal is used to remove silent part of speech signal. Discrete Wavelet Transform is used to decompose signal at two levels. DWT based Mel frequency cepstral coefficients (MFCC) and Traditional MFCC are used as a feature for speaker identification system. The similarity between the extracted features and set of reference features is calculated by Vector Quantization to determine speaker identity. TIMIT Database of different 15 speakers is used.

General Terms
Speaker Recognition, Feature Extraction, Multiresolution Analysis (MRA), MFCC.

Keywords
Speaker Identification System, Feature Extraction, Discrete Wavelet Transform, MFCC, Vector Quantization.

1. INTRODUCTION
Speech signal contains message as well as auxiliary information such as identity of speaker, characteristics of room, handset etc. The goal of speech recognition is to extract the message while ignoring such auxiliary information and goal of speaker recognition is to get information about speaker without giving much more importance to message. Speaker Recognition is divided into speaker identification and speaker verification, depending upon the application.

Speaker identification is the task of determining who is talking from a set of known voices or speakers. In the speaker identification system, an unknown speaker's speech signal is analyzed and compared with known speaker's speech models. The unknown speaker is identified as the speaker whose model best matches the input speech signal. Thus speaker identification involves determining identity of speaker belonging to closed set.

Speaker identification can be characterized as text-dependent or text-independent methods. In text dependent method, user is expected to say predetermined text- a voice password. In this method of speaker identification system, speaker provide speech signal of sentences or key words for both training and testing mode and same text is used to identify speaker. In text- independent method, the system relies only on the voice characteristics of the speaker, not rely on specific text being spoken.

Feature extraction is one of the important tasks in speaker identification system as the success of these systems is dependent on the robust features extracted from speech signal. A human speech varies from person to person by pitch and formant. A lot of research work has already been done in the area of speaker identification. Most popular techniques are MFCC and LPC; however MFCC based system have good performance in speaker identification [11]. Wavelet transform based approach is proposed in [12], in which 9th level wavelet transform was obtained for each sampled speech input. Wavelet transform provides Multiresolution analysis with dilated windows. In this system features are extracted using discrete wavelet transform so as to develop robust speaker identification system.

The organization of the paper is as follows: In section 2, Methodology is explained .The detail explanation about features used in this paper is discussed in section 3 and 4. In section 5 explanations about vector quantization is given. Experimental setup & results are discussed in section 6. The paper conclusion is given in section 7.

2. METHODOLOGY
The proposed speaker identification system is divided into two modes, training mode and testing mode. In training mode, speech signal is pre-processed and after pre-processing the wavelet transform is applied to extract features to develop speaker model. In testing mode, speech signal is pre-processed and similar proposed features are extracted to build speaker model. With the help of vector quantization classifier the speaker identity will be decided.

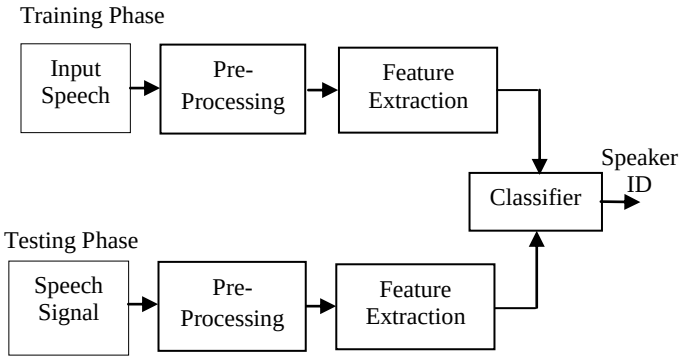

Fig. 1: Speaker Identification System.

3. WAVELET TRANSFORM

Wavelets are families of functions $\psi_{jk}(t)$ generated from a single base wavelet, called the mother wavelet, by dilations and translations, i.e,

$$\Psi_j k(t) = 2^{j/2}\Psi(2^j t - k) \quad j,k \in Z \qquad (1)$$

Where *Z* is the set of all integers, *j* is the dilation (scale) parameter and *k* is the translation parameter. In order to use the idea of multi-resolution, first define the scaling function and then define the wavelet in terms of it. First, define a set of scaling functions in terms of integer translates of the basic scaling function by

$$\Phi_k(t) = \Phi(t-k) \quad k \in Z \ , \Phi \in L2 \qquad (2)$$

Continuous Wavelet Transform, Discrete Wavelet Transform and Daubechies Wavelet Transform are the different categorize of Wavelet Transform, Discrete wavelet transform is a special case of the wavelet transform that provides a compact representation of a signal in time and frequency. Wavelet transform provides multi-resolution framework, making it possible to analyze a signal at several levels of resolution. Discrete wavelet transform decomposes a signal into approximation and detail function and these functions can be realized by using a pair of FIR filters h and g called low-pass and high-pass filters, respectively[2].The low frequencies have narrow bandwidths and the high frequencies have wide bandwidths. The wavelet decomposition for second level is shown in figure 2.

In the decomposition, detail function is kept as it is and only approximation function is decomposed further. The output of high pass filter and low pass filter can be represented by equation 3 and 4.

$$Y^{low}[k] = \sum^{n} X[n]h[2k-1] \qquad (3)$$

$$Y^{high}[k] = \sum^{n} X[n]g[2k-1] \qquad (4)$$

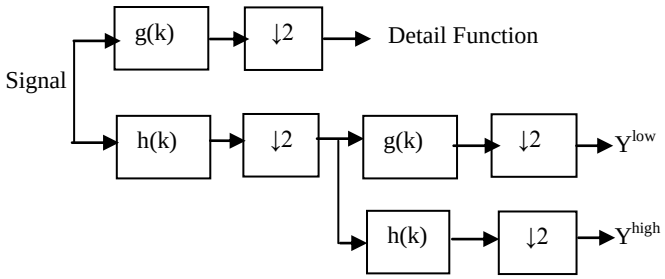

Fig. 2: Wavelet decomposition of speech signal

4. FEATURE EXTRACTION

The features carry the characteristics of the speech which are different from one speaker to another. As this goal is to identify the speaker, these features will play the important role in this identification system. It helps to make a decision on how much system is successful in extracting useful information from the speech so as to differentiate between speakers and identify them according to their features.

Frequency Band Analysis, Formant Frequencies, Pitch Contour and Harmonic Features are the certain features used in speaker recognition system. Commonly speech features are extracted using method such as MFCC, LPC, LPCC, RCC, LFCC, EFCC,CFCC and phase information etc. [1]-[7].

Among above feature extraction methods the mostly used ones are the MFCC and LPCC. Traditional MFCC coefficient is affected by noise equivalently in MFCC feature vector. In practical, each coefficient is affected different by noise so there is a need to weight each coefficient depends on the noise level effect to it. In sub-band recognizers, the weight of each coefficient for each sub-band is the same, therefore the sub band weighting algorithm is most important [8]. In this system, wavelet transform is used to extract feature from speech signal. The wavelet analysis is performed by passing the signal into successive high pass and low pass filter in order to overcome restricts of traditional MFCC. Then extracted wavelet coefficients are further applied to MFCC so as to increase accuracy of system. Block diagram of MFCC is shown in fig.3.

Thus wavelet analysis is performed in order to extract the features from the speech signal. Selection of a suitable wavelet function and number of levels of decomposition is important.

In this work Discrete wavelet transform (Daubechies 5) is used to extract the features from speech signal. Features vectors are obtained from the approximation coefficients and detail coefficients at second level.

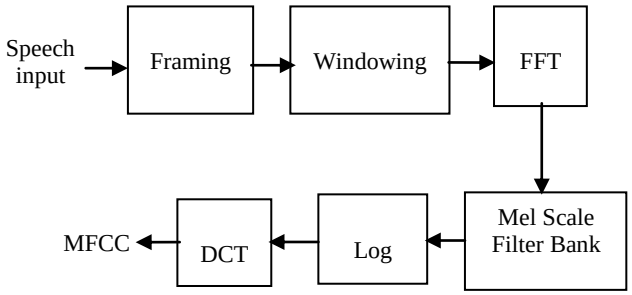

Fig. 3: Block diagram of MFCC

Step 1: Divide speech signal into a small frame with length within the range of 20 to 40 msec.

Step 2: Hamming Window is used to Discard the effect of discontinuities at edges of frame.

Step 3: Do the FFT. FFT convert each frame of N Samples from time domain into frequency domain.

Step 4: The Mel scale is based on pitch perception & uses triangular-shaped filter.

Step5: DCT is used to convert the log mel spectrum into time domain.

The result of conversion is called Mel Frequency Cepstrum Coefficient. This set of coefficient is called acoustic vectors. These acoustic vectors can be used to represent and recognize the voice characteristic of the speaker.

5. MODELING FEATURE VECTORS USING VECTOR QUANTIZATION TECHNIQUE

VQ is a quantization technique used to compress the information and manipulate the data such in a way to maintain the most prominent characteristic.

In this speaker identification system vector quantization (VQ) is used for creating a speaker model. VQ is easy to implement and it has high accuracy. In VQ, vectors are mapped from a large vector space to finite number of regions in that space. Each region is known as a cluster and it can be represented by its centre which is called as a codeword. Codebook is together collection of all codewords. Codebooks of tested speaker will compare with codebooks of trained speakers for identification of unknown speaker. The best matching result will be desired speaker.

Figure (4) shows representation of vector quantization codebook formation for speaker identification system. Only two speakers and two dimensions of acoustic space are shown in this figure. Acoustic vectors from speaker 1 and speaker 2 are represented by circles and triangles respectively. In training mode, by clustering speakers training acoustic vectors VQ codebook is generated for each known speaker using clustering algorithm. In figure (4) for speaker 1 and speaker 2 result centriod s are shown by black circles and black triangles respectively. VQ distortion is the distance from a vector to the closest codeword of a codebook. In testing mode, an input speech signal of an unknown speaker is vector quantized using each trained codebook and then VQ-distortion is computed. With the smallest total distortion, the speaker corresponding to the VQ codebook is identified as the speaker of input speech signal. [13]. Thus VQ is a process of taking a large set of feature vector and producing a smaller set of measure vector that represent Centroids of distribution.

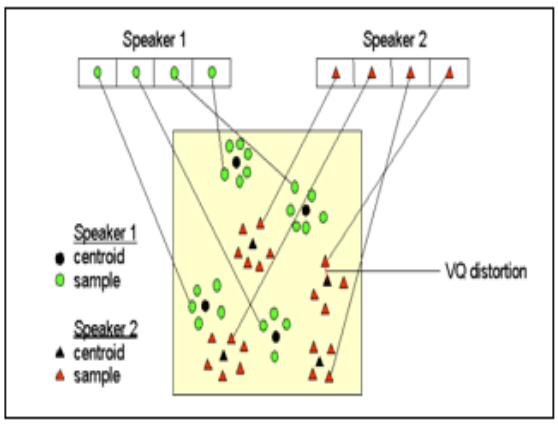

Fig. 4: Conceptual Diagram of VQ codebook

In this speaker identification system, LBG algorithm is used for clustering vectors so as to generate a codebook. In this algorithm number of centriod is directly proportional to accuracy, viz as number of centriod increases accuracy will also increase. The LBG algorithm steps are as follows:

1) Design a 1- vector codebook; that is the Centroid of the whole set of training vectors.

2) Twice the size of codebook by splitting each current codebook y_n according to the rule

$$y^{+}_{n} = y_n (1+\varepsilon)$$

$$y^{-}_{n} = y_n (1-\varepsilon)$$

Where n varies from 1 to the current size of the codebook and ε is splitting parameter.

3) Find the best set of Centroids for the spilt of codebook.

4) Repeat steps 2 and 3 until a codebook of size M is designed.

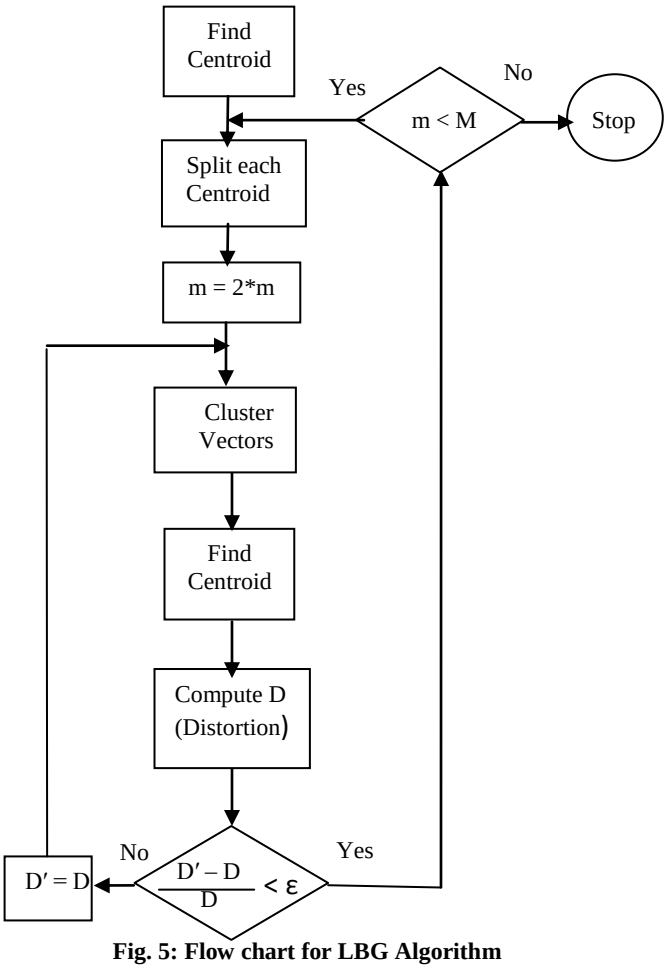

Fig. 5: Flow chart for LBG Algorithm

6. EXPERIMENTAL AND RESULT

In this speaker identification system, TIMIT Corpora database of different 15 speakers with 10 different utterances is used. For pre- processing the frame length of 32ms is used along with sampling frequency of 8000 Hz. In feature extraction 20 MFCC coefficients using wavelet transform are generated for each frame. Finally these features are used to develop VQ model to identify speaker. The outputs are shown in following figures.

Figure 6 shows the accuracy graph. Figure 7 gives the original input signal. Figure 8 gives DWT output signal. Figure 9 gives MFCC features using wavelet transform. Figure 10 gives output of VQ for value of k=64.

The results related to system performance in the form of accuracy are shown in Table 1 & 2.

Table 1 Speaker identification performance rate (%) of TIMIT Corpora database for MFCC based on DWT and traditional MFCC analysis

No. of speakers	Performance Rate (%)	
	DWT based MFCC	Traditional MFCC
5	92	90
10	88	84
15	77	74

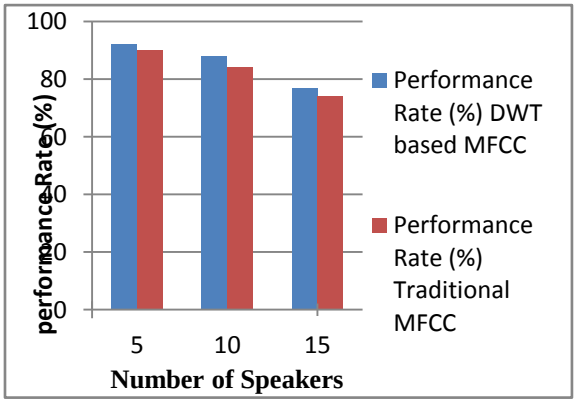

Fig. 6: Accuracy Graph

Table 2 Speaker identification performance rate for 15 speakers of TIMIT Corpora database for different codebook size

No. of Centroid (k)	Accuracy Rate		
	No. of speakers 5	No. of speakers 10	No. of speakers 15
8	82%	77%	70%
16	88%	79%	72%
32	90%	82%	75%
64	92%	88%	77%

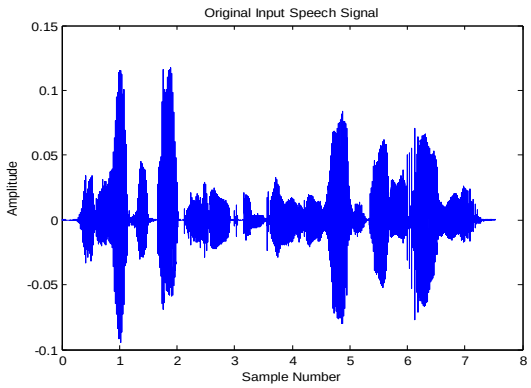

Fig. 7: Original input signal.

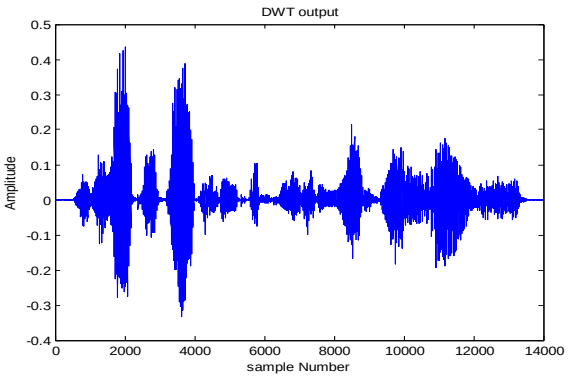

Fig. 8: DWT output signal.

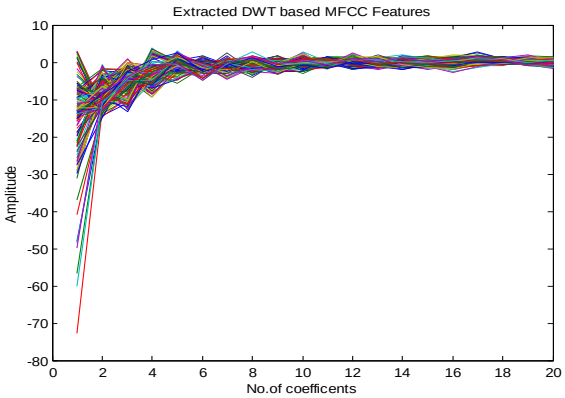

Fig. 9: MfCC features using wavelet transform.

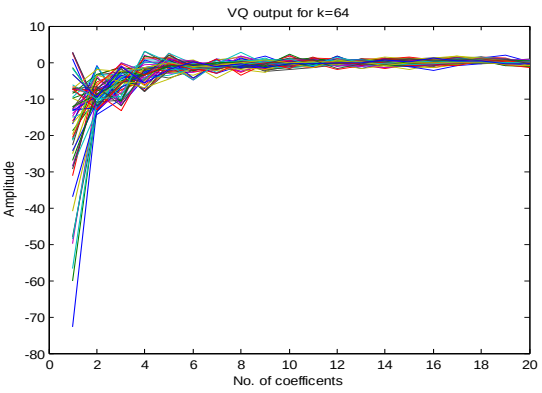

Fig .10: VQ output for value of K = 64.

7. CONCLUSION

A wavelet transform based speaker identification approach briefly introduced in this paper. The Daubechies wavelet transform 5(Db5) has been used to obtain wavelet transform coefficients for feature extraction. For 15 speakers, from TIMIT Corpora database, MFCC based DWT results show 85% accuracy & Traditional MFCC results show 80% accuracy. Using MFCC based DWT feature extraction technique and Vector Quantization technique using LBG algorithm for modelling of features gives maximum accuracy.

In future scope, this speaker identification system can be implemented by using hybrid VQ/GMM method. This future scope implementation will helpful to improve the computation time and accuracy rate of speaker identification.

8. REFERENCES

[1] Ning Wang, P. C. Ching, , Nengheng Zheng, and Tan Lee, "Robust Speaker Recognition Using Denoised Vocal Sthisce and Vocal Tract Features, *IEEE transactions on audio, speech, and language processing, vol. 19, no. 1, January 2011*.

[2] Ching-Tang Hsieh, Eugene Lai and You-Chuang Wang, "Robust Speaker Identification System Based on Wavelet Transform and Gaussian Mixture Model'', *Jthisnal of information science and engineering 19, 267-282 (2003).*

[3] Haris B C, G Pradhan, A Misra, S Shukla, R Sinha and S R M Prasanna, "Multi-Variability Speech Database for Robust Speaker Recognition", *IEEE Conference, 2011*.

[4] Qi Li and Yan Huang, "An Auditory-Based Feature Extraction Algorithm for Robust Speaker Identification under Mismatched Conditions", *IEEE transactions on audio, speech, and language processing, vol. 19, no. 6, August 2011*.

[5] Xing Fan and John H. L. Hansen, "Speaker Identification within Whispered Speech Audio Streams", *IEEE transactions on audio, speech, and language processing, vol. 19, no. 5, July 2011*.

[6] Haris B C, G Pradhan, A Misra, S Shukla, R Sinha and S R M Prasanna, "Multi-Variability Speech Database for Robust Speaker Recognition", *IEEE Conference, 2011*.

[7] Gayadhar Pradhan and S R Mahadeva Prasanna,"Significance of Speaker Information in Wideband Speech'', *IEEE Conference, 2011*.

[8] Phung Trung Nghia1 Pham Viet Binh1 Nguyen Huu Thai Nguyen Thanh Ha2 Prayoth Kumsawat, **"**A Robust Wavelet-based Text-Independent Speaker Identification**"** *International Conference on Computational Intelligence and Multimedia Applications 2007*.

[9] Gyanendra K Verma and U.S.Tiwary, **"**Text Independent Speaker Identification using Wavelet Transform" *International Conf. on Computer & Communication Technology*.

[10] Milan Sigmund, "Automatic speaker recognition by speech signal" *Bruno university of Technology, Czech Republic*.

[11] R. Vergin, D. O'Shaughnessy, and A. Farhat, "Generalized mel requency cepstral coefficients for large-vocabulary speakerindependent continuous-speech recognition," *IEEE Transactions on Speech and Audio Processing, Vol. 7, 1999, pp.525-532*.

[12] Muzhir Shaban Al-Ani, Thabit Sultan Mohammed and Karim M. Aljebory, "Speaker Identification: A Hybrid Approach Using Neural Networks and Wavelet Transform" *Jthisnal of Computer Science 3 (5): 304-309, 2007 ISSN 1549-3636, © 2007 Science Publications*.

[13] Nitisha, Ashu Bansal "Speaker Recognition using MFCC Front End Analysis & VQ Modeling Techniques for Hindi Words using MATLAB " *International Jthisnal of Computer Applications (0975 – 8887) Volume 45– No.24, May 2012* 48