Speaker Verification System Using MFCC and DWT

Sabitha V, Prof. Janardhanan.P
(Electronics-Digital Signal Processing, Calicut University, Kerala, India)
(ECE, KMCT College of Engineering, Kerala, India)

ABSTRACT : ***This paper aims at providing a brief overview of speaker verification system. Our aim is to implement text dependent speaker verification system that is insensitive to noise. In this paper, a novel family of windowing technique is used to compute Mel Frequency Cepstral Coefficient (MFCC) for automatic speaker recognition from speech. This method is based on fundamental property of discrete time Fourier transform (DTFT) related to differentiation in frequency domain. Classical windowing scheme such as hamming window is modified to obtain derivatives of discrete time Fourier transform coefficients. This paper, presents an effective approach to improve the accuracy of speaker recognition by combining the modified MFCC and spectral information (discrete wavelet coefficients).***

Keywords - *Differentiation in frequency, DWT, Mel frequency cepstral coefficients (MFCC), feature extraction, speaker verification*

1. Introduction

Voice comes under the category of biometric identity. Anatomical structure of the vocal tract is unique for every person .So the voice information available in the speech signal can be used for speaker recognition. Speaker recognition classified in to speaker verification and speaker identification. Speaker identification is the process of determining which registered speaker provides a given utterance. Speaker verification, on the other hand, is the process of accepting or rejecting the identity claim of a speaker. Speaker verification is a broad subject, but the commercial and personal use implementations are rare. One of the main problems of speaker verification in general is the impact of noise [1].

In case of speaker verification the similarity is computed only between the input signal and the stored patterns of the other recorded speakers. If the result is less than profound threshold, then the speaker will be accepted otherwise will be considered as an imposter and discarded [1].Speaker verification is a popular biometric identification technique used for authenticating and monitoring human subjects using their speech signal. The objective of this work is to design an efficient system for human speech recognition that is able to verify human speech more accurately. This work presents a technique of text-dependent speaker verification system. The combined MFCC and discrete wavelet coefficients are used as the features, which will be inputs to the GMM classifier.GMMs have unique advantages compared to other modeling approaches because their training is relatively fast and the models can be scaled and updated to add new speakers with relative ease[2].

2. Related Works

Many researchers have been done on the feature extraction of speech. The linear predictive cepstral coefficients (LPCC) were used because of their simplicity and effectiveness in speaker/speech recognition [3, 4]. Other widely used feature parameters, namely, the mel-scale frequency cepstral coefficients [5] are the most popular acoustic features used in speaker recognition. The use of MFCCs for speaker verification provides a good performance in clean environments, but they are not robust enough in noisy environments. Recently, a lot of research has been directed towards the use of wavelet based features [6-8]. The discrete wavelet transform (DWT) has a good time and frequency resolution and hence it can be used for extracting the localized contributions of the signal of interest. Wavelet denoising can also be used to suppress noise from the speech signal and it can lead to a good representation of stationary as well as non-stationary segments of the speech signal.

In this paper, a new method for speaker verification is presented. This method is based on the extraction of the MFCCs from the original speech signal and its wavelet transform. Then, a new set of features can be generated by concatenating both features. The objective of this method is to enhance the performance of the MFCCs based method in the presence of noise by introducing more features from the signal wavelet transform. Speaker verification systems have been developed for a wide range of applications. Although many new techniques were invented and developed, there are still a number of practical limitations because of which widespread deployment of applications and services is not possible. Still it is very true that humans can recognize speech and speaker more efficiently than machines.There is now an increasing interest in finding ways to reduce this performance gap.

We have organized this paper as follows: First, in section 3, briefly explain the proposed system. In section 4, we explained the experimental set up and result and finally, the paper is concluded in Section 5.

3. Proposed system

After an extensive study of various features of human speech and the model of speech production, we have decided to use following techniques to develop a robust, credible speaker verification system.

- Estimation of MFCC
- Spectral analyses using discrete wavelet transform.

3.1 Estimation of MFCC

Feature extraction algorithm steps are:

1) Speech signal converted to windowed frames.The size of window depends on the input speech signal frequency.Thus the system becomes insensitive to speaking rate.

2) In this work, we apply a simple time domain processing of speech after it is multiplied with a hamming window. The processing is based on well-known difference in frequency property of discrete time Fourier transform [9], and it can be easily integrated with standard window during DFT computation [10].

Let x(n) be a windowed speech frame of length N and its DTFT is given by, X($e^{j\omega}$). From differentiation in frequency property that DTFT of nx(n) can be written as,

$$: \hat{X}(e^{j\omega}) = j\frac{dX(e^{j\omega})}{d\omega} \qquad (1)$$

As DFT coefficients X(k) are samples of DTFT at ω = $\frac{2\pi k}{N}$, DFT of nx(n) are discrete samples of $\hat{X}$ ($e^{j\omega}$) at ω = $\frac{2\pi k}{N}$. Therefore, $\hat{X}$(k) = $\hat{X}$ ($e^{j\omega}$) are the DFT coefficients of nx(n).Since x(n) is a windowed speech frame, it can be represented as w(n)s(n), where s(n) is raw speech frame and w(n) is window function. Proposing a new window function as $\hat{w}$(n) = nw(n). The windowed speech frame is then represented as $\hat{x}$(n) = $\hat{w}$(n) s(n). From generalization of differentiation in frequency property, for an integer τ, DTFT of n^{τ}x(n) is $j^{\tau}\,\frac{d^{\tau}X(e^{jw})}{d^{\tau}w}$. Therefore, the window function of τ –th order window can be written as n^{τ}w(n). The window functions are shown in Fig.3.1.1 for first and second order along with hamming window. Assume that power spectrum of Hamming windowed signal is given by P(ω), and power spectrum of the modified window is $\hat{P}$(ω). Therefore, P(ω) = H^2(ω) = $\left|X(e^{j\omega})\right|^2$and $\hat{P}$(ω) = $\hat{H}^2$(ω) = $\left|\frac{dX(e^{j\omega})}{d\omega}\right|^2$, where H(ω) and $\hat{H}$(ω) magnitude spectrum of two signals respectively. The mathematical connection between power spectrum of new windowed speech frame and power spectrum of original Hamming windowed speech frame[10].

$$: \hat{H}^2(\omega) = \frac{1}{4P(\omega)}\left[\frac{dp(\omega)}{d\omega}\right]^2 \times sec^2[\Phi(\omega) - \varphi(\omega)] \qquad (2)$$

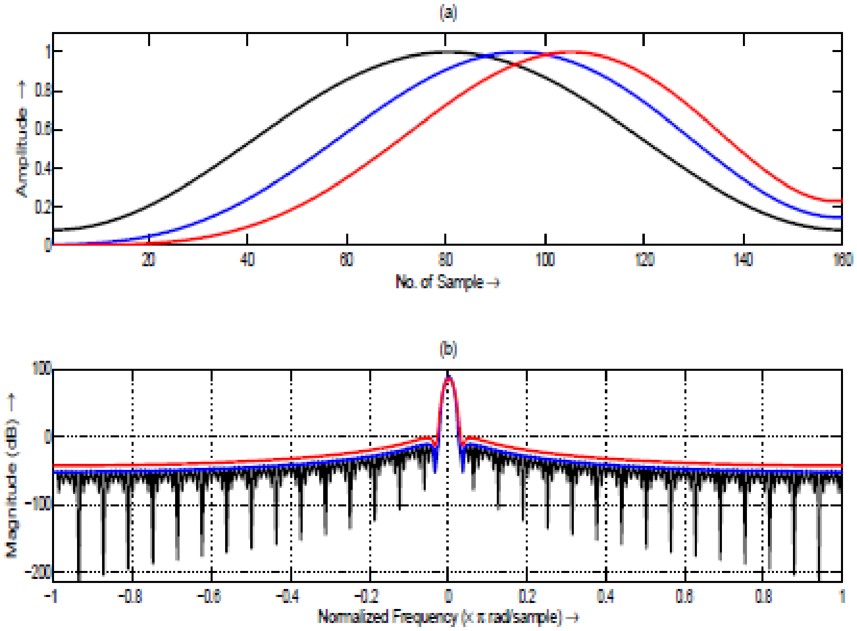


Figure.3.1.1 Comparison of Hamming window (black) with first (blue) and second (red) order differentiation based window in (a)time domain and (b)frequencydomain for a window of size 160 samples

The term $\frac{dp(\omega)}{d\omega}$ in (2) corresponds to the slope of the power spectrum of the Hamming windowed speech at frequency ω .Hence, as a consequence of power spectrum computation from derivative of fourier transform,obtain a modified power spectrum which is related to the slope of original power spectrum. Apart from it, the newly formulated power spectrum is also related to phase spectrum of the signal Φ(ω). There are evidences that speaker discriminating attribute is present in slope of power spectrum [11] as well as in phase information [12]. The modified DFT magnitude coefficients are nothing but the samples of $\widehat{H}(\omega)$ at $\omega = \frac{2\pi k}{N}$. Therefore, mel cepstrum computation using proposed window integrates the slope of power spectrum, phase, and of course, power spectrum of the signal. It is expected that the speech feature will be more efficient compared to the standard cepstrum which is solely based on power spectrum [10].

3) We used (3) to compute the mels for a given frequency *f* in Hz. After that taking log and DCT to get MFCC

: Mel(*f*) = 2595*log10(1 + *f*/700) (3)

3.2 Spectral analysis using DWT

The recorded speech signal contains background noise.This noise badly affects the accuracy of speaker verification.DWT reduces the noise present in input speech signal. Speech signals have a very complex waveform because of the superposition of various frequency components.By using the multi-resolution decomposing technique; one can decompose the speech signal into different resolution levels. The characteristics of multiple frequency channels and any change in the smoothness of the signal can then be detected to perfectly represent the signals [13].

The DWT is computed by successive lowpass and highpass filtering of the discrete time-domain signal as shown in fig. 3.2.2. This is called the Mallat algorithm or Mallat-tree decomposition. Its significance is in the manner it connects the continuous-time mutiresolution to discrete-time filters[14]. In the figure, the signal is denoted by the sequence x[n], where n is an integer. The low pass filter is denoted by G_0while the high pass filter is denoted by H_0. At each level, the high pass filter produces detail information, D[n], while the low pass filter associated with scaling function produces coarse approximations, A[n].

An important consideration in implementing the discrete wavelet transform is the choice of the wavelet and its associated scaling function. Once this choice is made, the implementation is a straightforward digital filtering scheme, which can be easily achieved on a general-purpose digital computer. The Haar wavelet is the simplest and is efficient in terms of computation, as the filters involved have just two taps. However, the Haar wavelet causes significant leakage of frequency components and is not well suited to spectral analysis of speech. The Daubechies family of wavelets has the advantage of having low spectral leakage and generally produces good results. There are several wavelets in this family, and we have chosen the D4 member for our application. We use three levels DWT to split the input speech into four dyadic sub-bands (0-1 kHz , 1-2 kHz, 2-4 kHz, 4-8 kHz).

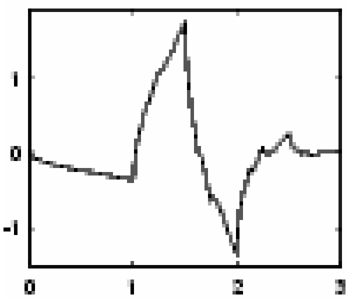
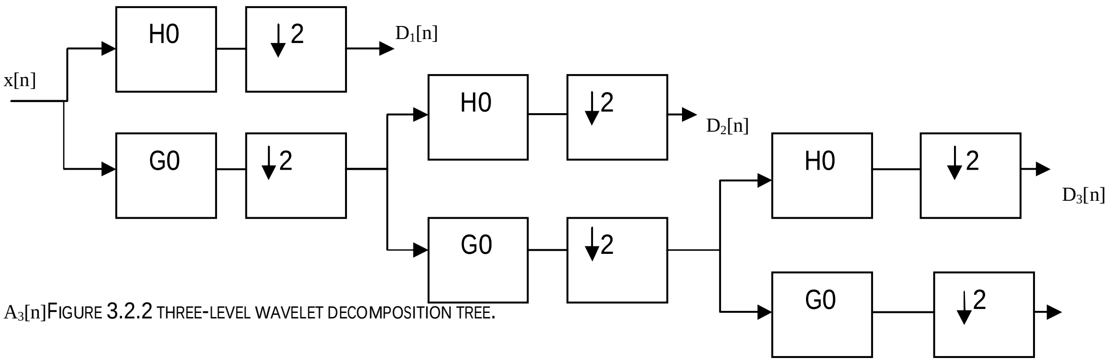


Figure.3.2.1: daubechies wavelet of order 4.



A3[n]FIGURE 3.2.2 THREE-LEVEL WAVELET DECOMPOSITION TREE.

4. Experimental set up and result

Here, we consider 5 different speakers saying 4 different words. The samples are saved in Matlab work folder. These files were recorded in Microsoft WAV format .Total 20 of the samples were taken. In the training phase of the automatic speaker verification system, a database is first composed. 5 speakers speech samples are used to generate this database.Feature vectors are extracted from each speech samples by using the above two methods and determine the threshold. For each speaker, 4 utterances were recorded at different times; the length of each utterance is approximately three seconds. For each speaker, four clean utterances provided the training utterances.The speech signals were recorded at 8kHz and 8 bits per sample. After this training step, the system would have knowledge of the voice characteristic of each (known) speaker.

In the testing phase, each one of these speakers is asked to say the words again and stored in WAV format.Then the speech signal is degraded by adding random noise in desired noise levels. Similar features to that used in the training are extracted from these degraded speech signals and used for matching. In the testing phase, the system will be able to verify the speaker.ie accept or reject the identity claim of the speaker .Here

GMM classifier used for feature matching. The performance of speaker verification is calculated by using the formula given below.

$$\% \text{ Speaker Verification} = \frac{\text{Number of correct acceptance}}{\text{Total samples used for testing}} \text{ X } 100$$

The speaker verification results are shown in Fig. 4.1. The average speaker verification performance with modified MFCC and combination of MFCC and DWT features in the clean environment is 98% and 100% respectively.

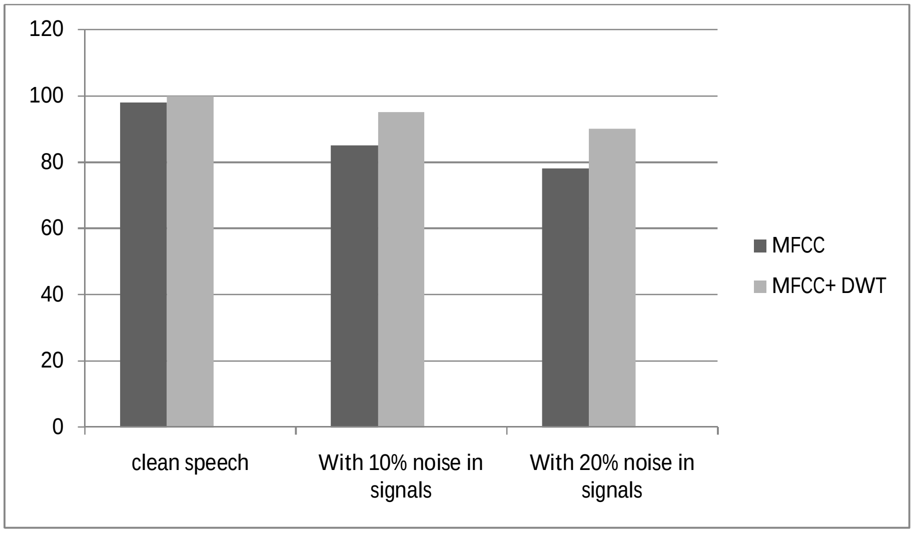


Figure 4.1: comparative performance of speaker verification system

5. Conclusion

There are considerable amount of researches going on the field of speech and speaker recognition. The results of shorter test sessions with limited users showed impressive performance in noisy and clean conditions.Here we use MFCCs because they follow the human ear's response to the sound signals and DWT reduces the noise present in input speech signal. Proposed system takes advantages of both wavelet and MFCC features and gives satisfactory results for speaker verification.Under various noise levels also,we could check the accuracy of the system. Speech may vary over a period of 2-3 years. So the training sessions have to be repeated so as to update the speaker specific database. In the future work, the proposed features can be used for text independent speaker verification and identification purpose.

6. Acknowledgement

I express deep gratitude to my guide, Prof.Janardhanan.P, for his invaluable support and guidance, which effectively contributed in successful completion of project.

References

[1] Tomi Kinnunen, Haizhou Li, An Overview of Text-Independent Speaker Recognition: from Features to Supervectors, *Speech Communication ,52(1)* ,2009

[2] Amin Fazel and Shantanu Chakrabartty,An Overview of Statistical Pattern Recognition Techniques for Speaker Verification - *IEEE Circuits and Systems Magazine , 11(2),* 2011,pp. 62-81.

[3]Atal, B., Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification, *Journal of Acoustical Society America, Vol. 55,*1974, pp. 1304-1312.

[4] White, G. M. and Neely, R. B., peech Recognition Experiments with Linear Prediction, Bandpass Filtering, and Dynamic Programming, *IEEE Trans. On Acoustics, Speech, Signal Processing, Vol. 24,* 1976, pp. 183_188.

[5] Vergin, R., O'Shaughnessy, D. and Farhat, A.,Generalized Mel Frequency Cepstral Coefficients for Large-Vocabulary Speaker-Independent Continuous-Speech Recognition,*IEEE Trans. on Speech and Audio Processing, Vol. 7,*1999, pp. 525_532 .

[6]B. C. Jong, wavelet *Transform Approach For Adaptive Filtering With Application To Fuzzy Neural Network Based Speech Recognition*, PhD Dissertation,Wayne State University, 2001.

[7] Z. Tufekci, Local *Feature Extraction For Robust Speech Recognition in The Presence of Noise*, PhD Dissertation, Clemson University, 2001.

[8]R. Sarikaya, Robust *And Efficient Techniques For Speech Recognition in Noise*, PhD Dissertation, Duke University, 2001.

[9]A. Oppenheim, S. Willsky, and S. Nawab, *Signals And Systems*(second edition ed. PHI Learning, 2009)
.

[10] Md Sahidullah,Goutam Saha, A Novel Windowing Technique for Efficient Computation of MFCC for Speaker Recognition *Signal Processing Letters, IEEE ,20(2),*Feb 2013

[11] M. Sahidullah and G. Saha, Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition, *Speech Communication*,54(4), May 2012,pp. 543–565,

[12] S. Nakagawa, L. Wang, and S. Ohtsuka, Speaker identification and verification by combining mfcc and phase information, *Audio, Speech, and Language Processing, IEEE Transactions on, 20(4),* May 2012,pp1085 –1095,

[13] A. Shafik, S. M. Elhalafawy, S. M. Diab, B. M. Sallam and F. E. Abd El-samie, A Wavelet Based Approach for Speaker Identification from Degraded Speech, *International Journal of Communication Networks and Information Security (IJCNIS),1(3),* December 2009

[14] Ching-tang hsieh, Eugene lai and You-chuang wang. Robust Speaker Identification System Based on Wavelet Transform and Gaussian Mixture Model, *journal of information science and engineering 19*, 2003,267-282.