Wavelet Transform Based Automatic Speaker Recognition

S. Malik [1] and Fayyaz A. Afsar[2]
Department of Computer and Information Sciences
Pakistan Institute of Engineering and Applied Sciences
Islamabad, Pakistan
E-mail: [1] sidraa.malik@gmail.com, [2] afsar@pieas.edu.pk

Abstract—**An effective feature extraction technique for speaker recognition is presented in this paper. It uses multi resolution property of wavelet transform and Mel-Frequency Cepstral Coefficients (MFCCs) for analyzing the speech signal. For individual speaker, first the speech signal is decomposed using Discrete Wavelet Transform (DWT) into approximations and details coefficients. Approximation coefficients are then used to compute MFCCs. Experimental results were computed on PIEAS Speech Database for text independent speaker identification. The proposed method gives very good recognition rate i.e. 96.25% for non telephonic and 86.77% for telephonic speech data. In addition to this, analysis for choosing the appropriate number of MFCCs, the appropriate number of decomposition levels and wavelet type has also been performed.**

Keywords- ***speaker recognition, MFCC, wavelet, LPC***

I. INTRODUCTION

Speech is man's most commonly used way of communication. The primary type of the spoken information is the word, which the speaker tries to pass to the listener but other types that are also included in the speech are information about language being spoken, emotions, gender and most important, identity of the speaker.

The area of speaker verification is most natural and realistic way of authentication in biometrics. On account of this fact, speaker recognition is widely used in areas of access control. Telephonic Speaker recognition can provide an alternative or supplemental means of voice entry for authenticated provision of telephone based services. Generally the speaker recognition task is divided into speaker verification and speaker identification. Speaker verification is simply a task of accepting or rejecting an identity claim of an unknown speaker whereas speaker identification is a task of determining which of the registered speakers' best matches the identity claim. Next, the speaker recognition task can be either text-dependant or text-independent. The former depends upon the speaker to utter the same text for recognition as it was provide during training, but the latter does not require any specific text. The performance of Speaker recognition system highly depends upon the various techniques being selected for extracting speech features and modeling a speaker.

For the past two decades, various methods have been proposed in order to bring improvement in field of speaker recognition. For feature extraction, Linear Predictive Coding (LPC) [1, 6] and Mel- Frequency Cepstrum Coefficients (MFCC) [2] are conventional methods which were later followed by Linear Predictive Cepstral Coefficients (LPCC) and their regression coefficients. LPC is based upon basic principles of sound production but it has the drawback of performance degradation in the presence of noise. LPC cepstral coefficients (LPCC) and their residues are methods which are known to give smoother spectral envelope and hence give stable representation than LPC [1]. MFCC is known to be a widely used method which is based on filter bank approach in which filters are spaced in a pattern similar to that of human ear's which does not follow a linear scale. However, these conventional methods have been proven good for feature extraction in speech recognition but they are not necessarily good in speaker identification [3]. Recently, wavelets have been used [3, 4] as a tool for compression and denoising speech. According to [4] use of wavelets can reduce the influence of noise interference in feature extraction using LPCCs and has proven to give better performance rates over telephonic speech. When transmitted through a telephone network, speech is compressed using lossy techniques which adds noise into the signal. Moreover, poor-quality microphones can degrade speaker recognition performance significantly. Thus, noise reduction has been considered as an important aspect in later researches specially those involving telephonic speech. Then, considering speaker modeling techniques, Vector Quantization (VQ) techniques have been used [4] to compress data efficiently.

In this paper, we propose a new feature extraction method which uses multi resolution property of wavelets in order to reduce noise in speech signal. Mel filter bank is used in order to extract features of an individual speaker which on the whole gives satisfactory results on telephonic and non telephonic speech. The rest of the paper gives a short description of implemented technique, which divides into feature extraction and speaker modeling. Then, conclusions are drawn based on results section.

978-1-4244-4873-9/09/$25.00 ©2009 IEEE

II. Description of Implemented Technique

The general architecture of speaker recognition system is given in Fig.1. It is based on feature extraction, speaker modeling and classification. Feature extraction is the first phase carried out to obtain compact and speaker dependant information. After extracting features, we transform these features to create a model for each speaker and store it. In Patten matching or classification, for an unknown speaker, we match the model for the unknown speaker to stored templates. Decision is based on how closely model for an unknown speaker matches with the stored ones.

A. Feature Extraction

The feature extraction technique consists of three phases i.e. voice activity detection, use of discrete wavelet transform (DWT) and extraction of Mel-coefficients. Voice Activity Detection is used to extract out speech data from a speech signal which also contains some non speech data. Approximation coefficients from DWT of a signal are obtained in order to denoise speech signal and Mel filters are later used due to their perception of speech which is similar to that of human ear.

1) Voice Activity Detection

Voice Activity Detection (VAD) is used to separate speech and non-speech data from a speech signal. Non- Speech data include pre-utterances, post-utterances and silence between words [5]. According to the literature, VAD algorithms are based on energy estimation methods, Entropy based techniques [5] and wavelets. Energy estimation phenomenon among all is extensively used due to its simplicity. These methods work effectively in areas where noise varies slower over the signal and the speech segment energy is considered greater than noise level.

In the proposed feature extraction technique, threshold mechanism is used after estimating short time energy of a signal. Threshold value (λ) estimates the noise level in speech. When the energy of a speech signal is greater than λ, speech is detected and when the energy is lower than λ, noise or pause is detected.

2) Discrete wavelet Transform

Second step is to compute discrete wavelet transform of a signal. The DWT of a signal *x* is calculated by passing it through a series of filters. First the samples are passed through a low pass filter with impulse response *g* resulting in a convolution of the two given in (1).

$$y[n] = (x * g)[n] = \sum_{k=-\infty}^{\infty} x[k]g[n-k] \qquad (1)$$

The signal is also decomposed simultaneously using a high-pass filter *h*. The output gives the detail coefficients (from the high-pass filter *h*) and approximation coefficients (from the low-pass *g*). These two filters are related to each other and are known as Quadrature Mirror Filters.

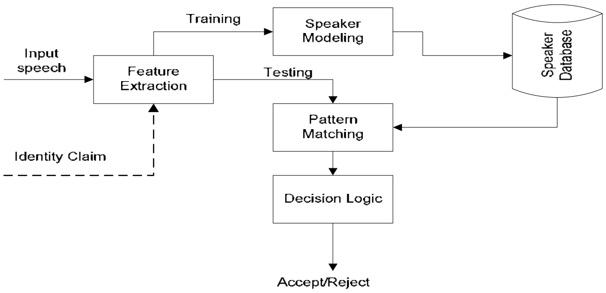

Figure 1. General Structure of Speaker Recognition

Approximation coefficients give characteristics of lower frequencies in the signal, whereas details give information about higher frequency characteristics. The Approximation coefficients at each level can be used for another level decomposition (after a down-sampling of 2) and this can be extended to multiple levels in order to get more frequency resolution. This is shown in Fig 2.

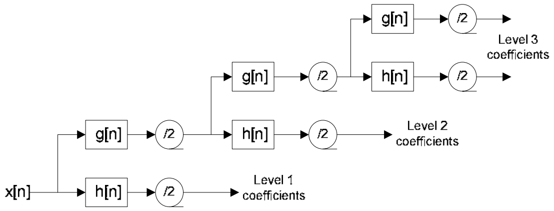

Figure 2. A 3-level decomposition

After getting approximation coefficients, we use these coefficients to model the speech signal. Details coefficients contain high frequency signal data which is affected by noise and contains little information about the identity of the speaker as it varies greatly with change in the text spoken and recording/acquisition conditions. Therefore details coefficients are not used in speech signal modeling.

An analysis of different wavelets for speaker authentication has been performed as a part of this work. The wavelet level giving optimized results with symlet-7 wavelet is one in which case we can use a simple filtering stage analogous to low pass filtering with a filter based on the symlet-7. This will reduce the computational load of the technique further.

3) Mel-Frequency Cepstral Coefficients

Mel-cepstrum is one of the most commonly used feature extraction technique used in both speech and speaker recognition.

MFCC technique is based on the known variation of the human ear's critical bandwidth frequencies with filters that are spaced linearly at low frequencies and logarithmically at high frequencies to capture the important characteristics of speech. MFCC is composed of five phases as shown in Fig. 3. First phase is of framing speech signal in order to analyze speech signal in shorter frames due to its non stationary nature. Frame size is 256 in this case. The next step involves windowing of each frame which minimizes the discontinuities at start and end of each frame. Then windowed speech signal is converted from time domain to frequency domain by taking Fast Fourier transform (FFT) which gives insight to frequencies present in that speech signal. Once converted to frequency domain, the signal is passed through Mel-frequency wrapping block. The purpose of the Mel-bank is to simulate the critical band filters of the hearing mechanism. Mel-Filters emphasize on low frequencies and ignore higher frequencies just like human ear behaves. Fifth step is to take log of the spectrum and compressing it by discrete cosine transform, DCT. The resultant matrices are referred to as Mel-Frequency Cepstrum Coefficients. This spectrum provides a fairly simple but unique representation of the spectral properties of the voice signal. Important property of cepstral coefficients is that they are fairly uncorrelated with each other [6] which give edge to them over performance when compared to LPC coefficients which are highly correlated.

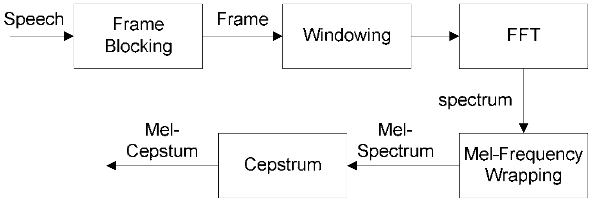

Figure 3. Steps for computing Mel-coefficients

B. Speaker Modeling

During the training phase, a model for each speaker is formed using Vector Quantization technique. It is a technique to compress information (feature space) in such a way that it maintains the most important or prominent characteristics. There are two important factors that must be considered while implementing VQ are algorithm to generate codebook and size of codebook. Here we have used K-means as clustering algorithm in reference to and codebook size of 32 [4].

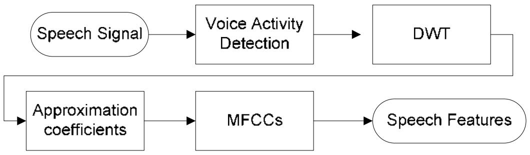

Figure 4. Stages in proposed feature extraction technique

In testing phase we need to match a testing utterance which is simply a set of non compressed features, to stored models of all the speakers. This matching is done using distortion measure [4] given in (2).

$$D = \frac{1}{N_F}\sum_{i=1}^{N_F} \min_{1 \le n \le N_{book}} [d(\overline{x_n}, x_i)] \tag{2}$$

Where N_F denotes the number of speech frames in testing utterance, x_n denotes nth reference codebook , x_i is the testing utterance, N_{book} represents total number of codebooks, d is the distance between x_n and xi and D is the overall distortion measure between test utterance and model. The identity of each speaker is then established according to average distortion for each speaker [4] and the matching speaker is the one who has minimum distortion measure.

III. Emperical Results

This section gives brief introduction about the database and results obtained on this database. Three aspects have been taken in account while analyzing results i.e. Comparison of proposed technique using MFCCs with that of LPCs, Effect of number of MFCCs, Effect of decomposition level and Effect of changing wavelet type on verification rate of speaker. Here verification is simply accepting or rejecting a speaker and verification rate percentage is how much percent of testing utterances were correctly classified by the system

A. Database Description

PIEAS speech Database is a single session database that consists of 64 speakers in total, forty-five males and nineteen females. It contains both Non-Telephonic and Telephonic speech. The database overall has an age distribution from 14 to 36 years. Each Speaker has 30 recorded samples. Out of which, 10 are recorded from Microphone (Mono), 10 from one handset (Handset-1) and 10 from other handset (Handset-2). Overall, the database consists of approximately 2000 samples. The recorded phrases include short sentences and digit strings of various lengths. Samples have been recorded at 16 KHz.

B. Comparison of MFCCs with LPCs

Results for proposed technique were compared with that in which LPCs were used. Experiments show that MFCCs along with wavelet transform perform better than LPCs with wavelet transform, especially in telephonic data as shown in Fig. 5.

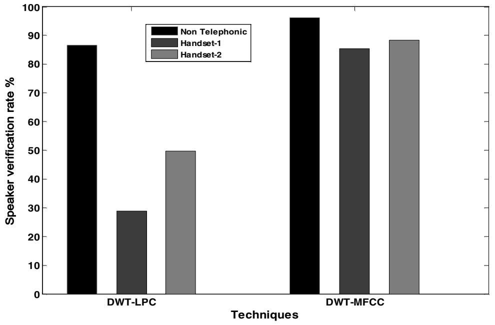

Figure 5. Comparison of recognition rate using DWT with Linear Predictive Coding and Mel-Frequency Cepstral Coefficients

C. Effect of Number of MFCCs

As previous results show that proposed technique performed better, we then analyzed effect of number of MFCCs for both telephonic and non telephonic data.

Fig. 6 shows that as the number of MFCCs are increased, recognition rate increases rapidly in start and then varies gradually. In case of non-Telephonic speech, 18 MFCCs whereas in case of Telephonic 20 MFCCs must be used. Overall, the best number of MFCCs is suggested to be 38 which gives 96% recognition accuracy on non telephonic data and 86% recognition accuracy on telephonic data.

D. *Effect of Decomposition Levels*

As number of decomposition levels is increased, further information from a signal can be extracted in some cases, but increasing number of levels not only increases computational complexity but also introduces redundant data which do not contain any further information. Thus, choosing an appropriate number of decomposition levels has become a significant problem.

Fig. 7 shows results for increasing number of decomposition levels which were computed up to three levels and experiments show that highest performance is achieved on level 1. This is possibly due to the reason that speech signal loses its characteristics as the decomposition levels are increased.

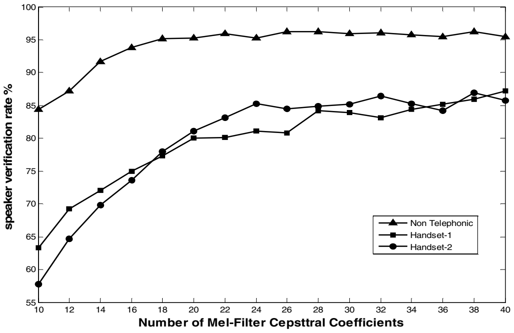

Figure 6. Effect of increasing number of Mel-Coefficients

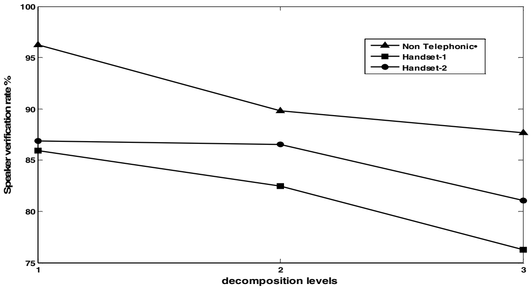

Figure 7 Effect of Decomposition Levels

E. *Effect of Wavelet Type*

Wavelet type has also been a part of analyzing. For this purpose we have used Daubechies wavelet i.e. DB1-DB3, Haar, Symlets and Discrete Meyer. Results show that Symlets has performed better in both non telephonic and telephonic speech as shown in Table 1, so we have used Symlets throughout this technique.

IV. Discussions and Conclusions

In this approach, we have presented wavelet based MFCCs as efficient feature set for speaker recognition task. The approach was motivated by multi resolution property of wavelets in denoising speech signal and MFCCs were used to mimic the behavior of human ear which emphasizes lower frequencies. Feature set for an individual speaker is constructed using approximations from wavelet decomposition and Mel-coefficients.

Recognition rate achieved is quite good i.e. 96.25% for non telephonic and 86.77% for telephonic speech. Moreover for the proposed technique, we analyzed to select best parameters and for PIEAS Speech Database, 38 MFCCs based on wavelet type Symlets 7 and decomposition level 1 have proven best. The performance of proposed method is comparable to approaches for feature extraction based on wavelet transform, LPCs and LPCCs as well as to speaker modeling techniques like VQ.

A possible future direction would be extending this classification to open-set and using a classification approach other than minimum distortion. Most important, this method can be improved to make it efficient for mismatched conditions, i.e. where the system is trained on non telephonic data and tested on telephonic data.

Table 1. Effect of changing wavelet type

Wavelet Type	Non Telephonic	Handset-1	Handset-2
Db1	95.47	86.41	87.03
Db2	95.94	85.78	85.78
Db3	95.47	84.69	87.19
Sym7	96.25	85.94	86.88
Discrete Meyer	96.09	85.47	86.88

References

[1] Sadaoki Furui, "*Recent advances in speaker recognition*", Pattern Recognition Letters vol. 18, pp. 859-872, 1997.

[2] Vergin, R., O'Shaughnessy, D., and Farhat, A., "*Generalized Mel frequency cepstral coefficients for large vocabulary speaker-independent continous-speech recognition*". IEEE Trans. Speech Audio Process, 1999.

[3] Shung-Yung Lung, "*Feature extracted from wavelet decomposition using biorthogonal Riesz basis for text-independent speaker recognition*", Pattern Recognition vol. 41, pp. 3068-3070, 2008.

[4] Chen, W.C., Hsieh C., and Lai E., "*Robust speech features based on wavenet transform with application to speaker identification*", IEEE Proceedings. Vision, image and signal processing, vol. 149 No.2, pp. 108-114, 2002.

[5] Renevey, P. and Drygajl, A., "*Entropy Based Voice Activity Detection in Very Noisy Conditions*" Proceedings of 7th European Conference on Speech Communication and Technology, EUROSPEECH'2001, pp. 1887-1890, 2001.

[6] P.K Tomi, "Spectral Features for Automatic Text-Independent Speaker Recognition", Ph.Lic. thesis, Department of Computer Science, University of Joensuu, 2003.