

A Wavelet Based Approach for Speaker Identification from Degraded Speech

A. Shafik, S. M. Elhalafawy, S. M. Diab, B. M. Sallam and F. E. Abd El-samie

Department of Electronics and Electrical Communications, Faculty of Electronic Engineering
Menoufia University, Menouf, Egypt

Emails: {mero431, saidelhalafawy, dr_salah_diab, b_m_salam and fathi_sayed}@yahoo.com

Abstract: This paper presents a robust speaker identification method from degraded speech signals. This method is based on the Mel-frequency cepstral coefficients (MFCCs) for feature extraction from the degraded speech signals and the wavelet transform of these signals. It is known that the MFCCs based speaker identification method is not robust enough in the presence of noise and telephone degradations. So, the feature extraction from the wavelet transform of the degraded signals adds more speech features from the approximation and detail components of these signals which assist in achieving higher identification rates. Neural Networks are used in the proposed method for feature matching. The Comparison study between the proposed method and the traditional MFCCs based feature extraction method from noisy speech signals and telephone degraded speech signals with additive white Gaussian noise (AWGN) and colored noise shows that the proposed method improves the recognition rates computed at different degradation cases.

Keywords: Speaker identification, Wavelet transform, MFCCs, Neural networks.

1. Introduction

In 1970s, the key technologies for pattern recognition models were developed with the introduction of linear prediction methods for spectral representation. In the 1980s, speaker identification based on statistical methods with a wide range of networks for handling language structures was introduced [1,2]. The key technologies introduced during this period were the Hidden Markov models (HMMs) and the stochastic language model, which together enabled for handling virtual and continuous speech recognition [3]. Another technology that was introduced in the late 1980s was the idea of Artificial Neural Networks (ANNs) [4,5]. These technologies have served in the current progress in this area.

In speaker identification systems, the two major operations performed are feature extraction and classification [2]. The feature extraction can be considered as a data reduction process that attempts to capture the essential characteristics of the speaker with a small data rate. There are various techniques for extracting speech features in the form of coefficients such as the linear prediction coefficients (LPCs), the Mel-Frequency Cepstral Coefficients (MFCCs) and the Linear Prediction Cepstral Coefficients (LPCCs) [2]. Classification is a process having two phases; speaker modeling and speaker matching. In the speaker modeling step, the speaker is enrolled to the system using features extracted from the training data. When a sample of data from some unknown speaker arrives, pattern matching techniques are used to map the features from the input speech sample to a model corresponding to a known speaker. The combination

of a speaker model and a matching technique is called a classifier. Classification techniques used in speaker identification systems include Gaussian Mixture Models (GMMs), Vector Quantization (VQ), HMMs and ANNs [2,6,7].

The MFCCs are the most popular acoustic features used in speaker identification. The use of MFCCs for speaker identification provides a good performance in clean environments, but they are not robust enough in noisy environments. They are based on the known evidence that the information carried by low frequency components of the speech signal is more than that carried by high frequency components. The MFCCs assume that the speech signal is stationary within a given time frame and may therefore lack the ability to analyze the localized events accurately [2, 7-9]. Recently, a lot of research has been directed towards the use of wavelet based features [10-12]. 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 identification 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 or telephone degradations by introducing more features from the signal wavelet transform. The rest of the paper is organized as follows. Section 2 gives an overview on the structure of any speaker identification system. Section 2 discusses the process of feature extraction. Feature matching is discussed in section 4. In Section 5, the proposed speaker identification method is introduced. Section 6 gives the experimental results. Finally, Section 7 summarizes the concluding remarks.

2. Speaker Identification System

An Automatic speaker identification system comprises two stages; a feature extraction stage and a classification stage as shown in Fig.(1). This system operates in two modes; training and recognition modes. Both of them include a feature extraction step which is sometimes called the front end of the system. The feature extractor converts the digital speech signal into a sequence of numerical descriptors called the feature vector [2]. The features exploited in this paper are

the MFCCs and some polynomial coefficients which model the shape of the time waveform of the MFCCs.

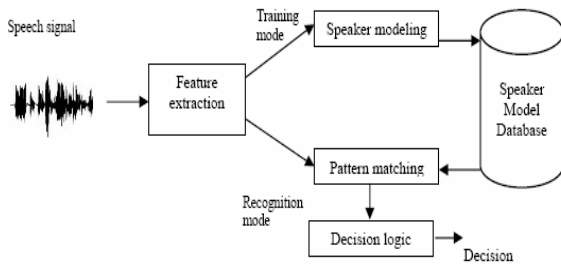


Figure 1. Automatic speaker identification system.

For successful classification, each speaker is modeled using a set of data samples in the training mode, from which a set of feature vectors is generated and saved in a database. Features are extracted from the training data essentially stripping away all unnecessary information in the training speech samples leaving only the speaker characteristic information with which speaker models can be constructed [2]. When a sample of data from some unknown speaker arrives, pattern matching techniques are used to map the features from the input speech sample to a model corresponding to a known speaker.

3. Feature Extraction

The concept of feature extraction contributes to the goal of identifying speakers based on the low-level properties. The extraction produces sufficient information for good speaker discrimination and captures this information in a form and size which allows efficient modeling. Thus, feature extraction can be defined as the process of reducing the amount of data present in a given speech sample while retaining speaker discriminative information. In the following subsections, an explanation for the extraction of the MFCCs and the polynomial coefficients is presented.

3.1. Extraction of MFCCs

The MFCCs are commonly used features for speaker identification. They are extracted from speech signals through cepstral analysis. The human speech production process involves an excitation source which is a pulse stream or uncorrelated noise and the vocal tract which is modeled by a linear time invariant filter. The idea of cepstral analysis is to separate components of the excitation and the vocal tract, so that the speech or the speaker dependent information can be obtained. The cepstral analysis is a tool used to separate the redundant pitch information from the more important vocal tract information [2]. The MFCCs are also based on the human perception of the frequency content which emphasizes low frequency components more than high frequency components.

Calculation of the MFCCs proceeds similar to the cepstral transformation process shown in Fig.(2). The input speech signal is first framed and windowed, the Fourier Transform is then taken and the magnitude of the resulting spectrum is warped by the Mel-scale. The log of this spectrum is then

taken and the discrete cosine transform is applied [2]. The Mel is a unit used to measure the perceived pitch or frequency of a tone. The Mel-scale is therefore a mapping between the real frequency scale in Hz and the perceived frequency scale in Mels. The Mapping is virtually linear below 1 kHz and logarithmic above as given by the following relation [2]:

$$f_{Mel} = 2595 \log_{10} \left(1 + \frac{f_{Linear}}{700} \right) \quad (1)$$

The calculation of the MFCCs is based on the short term analysis, and thus for each frame, the MFCCs vector is computed. In this process, the speech signal is pre-emphasized to remove glottal and lip radiation effects. The pre-emphasis is implemented by a first order finite impulse response (FIR) filter of the form [13]:

$$H(z) = 1 - az^{-1} \quad (2)$$

where $0.9 \leq a \leq 0.99$.

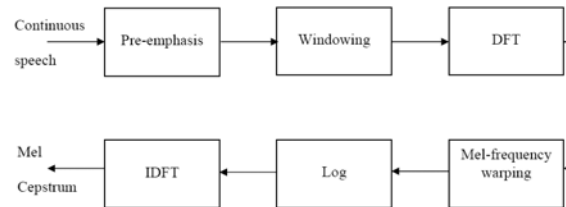


Figure 2. Cepstral transformation of a speech signal.

The speech signal must first be broken up into small sections, each of N samples. These sections are called frames and the motivation for this framing process is the quasi-stationary nature of speech. That is the characteristics of the speech signal are time varying, however if we examine the signal over discrete sections which are sufficiently short in duration, then these sections can be considered as stationary and exhibit stable acoustic characteristics [2]. Typically a frame size of 20ms to 40ms is used where the number of samples per frame N will depend on the sampling rate of the data. To avoid loss of information, frame overlap is used. Each frame begins at some offset of L samples with respect to the previous frame where $L \leq N$.

For each frame, a windowing function is usually applied to increase the continuity between adjacent frames. Common windowing functions include the rectangular window, the Hamming window, the Blackman window and flattop window. Windowing in time domain is a pointwise multiplication of the frame and the window function. According to the convolution theorem, the windowing corresponds to a convolution between the short term spectrum and the window function frequency response. A good window function has a narrow main lobe and low side lobe levels in its frequency response. The most commonly used window function in speech processing is the Hamming window which defined as [2]:

$$w_H(n) = 0.54 - 0.46 \cos\left(\frac{2n\pi}{N-1}\right) \quad (3)$$

where $n=0, 1 \dots N-1$.

The DFT of a windowed frame of speech is computed to obtain the magnitude spectrum. The DFT is mathematically defined as [2]:

$$S(k) = \sum_{n=0}^{N-1} s(n) e^{-j2\pi kn/N} \quad (4)$$

where $s(n)$ is a time sample of the windowed frame. The IDFT is defined as [2]:

$$s(n) = \frac{1}{N} \sum_{k=0}^{N-1} S(k) e^{j2\pi kn/N} \quad (5)$$

The magnitude spectrum is frequency warped in order to transform the spectrum into the Mel-frequency scale. The Mel-frequency warping is performed using a Mel-filter bank composed of a set of bandpass filters with constant bandwidths and spacings on the Mel-scale. The bank consists of one filter for each desired Mel-frequency component, where each filter has a triangular filter bandpass frequency response. The triangular filters are spread over the entire frequency range from zero to the Nyquist frequency. The number of filters is one of the parameters which affect the recognition accuracy of the system. The last stage involves performing a discrete cosine transform (DCT) on the log of the Mel-spectrum. This replaces the IDFT stage in practice for increasing the computational efficiency.

If the energy of the m^{th} Mel-filter output is $\tilde{S}(m)$, the MFCCs will be given as follows [2]:

$$c_j = \sqrt{\frac{2}{N_f}} \sum_{m=1}^{N_f} \log(\tilde{S}(m)) \cos\left(\frac{j\pi}{N_f}(m-0.5)\right) \quad (6)$$

where $j=0, 1, \dots, J-1$, J is the number of MFCCs, N_f is the number of Mel-filters and c_j are the MFCCs. The number of the resulting MFCCs is chosen between 12 and 20, since most of the signal information is represented by the first few coefficients. The 0th coefficient represents the average log energy of the frame.

3.2 Extraction of Polynomial Coefficients

The MFCCs are sensitive to mismatches or time shifts between training and testing data. Thus, there is a need for other coefficients to be added to the MFCCs to reduce this sensitivity. Polynomial coefficients are used for this purpose. These coefficients can help in increasing the similarity between the train and the test utterances if they are related to the same person. If each MFCC is modeled as a time waveform over adjacent frames, polynomial coefficients are used to model the slope and curvature of this time waveform for each MFCC. Adding these polynomial coefficients to the

MFCCs vector will be helpful in reducing the sensitivity to any mismatch between the training and testing data [13].

To calculate the polynomial coefficients, the time waveforms of the cepstral coefficients are expanded by orthogonal polynomials. The following two orthogonal polynomials can be used [13]:

$$P_1(i) = i - 5 \quad (7)$$

$$P_2(i) = i^2 - 10i + 55/3 \quad (8)$$

To model the shape of the MFCCs time functions, a nine elements window at each MFCC is used. Based on this window assumption, the polynomial coefficients can be calculated as follows [13]:

$$a_j(t) = \frac{\sum_{i=1}^9 P_1(i) c_j(t+i-1)}{\sum_{i=1}^9 P_1^2(i)} \quad (9)$$

$$b_j(t) = \frac{\sum_{i=1}^9 P_2(i) c_j(t+i-1)}{\sum_{j=1}^9 P_2^2(i)} \quad (10)$$

where $a_j(t)$ and $b_j(t)$ are the slope, and the curvature of c_j in the t^{th} frame. The vectors containing all c_j , a_j and b_j are concatenated to form a single feature vector.

4. Feature Matching using Artificial Neural Networks

The classification step in automatic speaker identification systems is in fact a feature matching process between the features of a new speaker and the features saved in the database. Neural Networks are widely used for feature matching. Multi-layer perceptrons (MLPs) consisting of an input layer, one or more hidden layers and an output layer can be used for this purpose [4,5]. Figure (3) shows an MLP having an input layer, a single hidden layer and an output layer. A single neuron only of the output layer is shown for simplicity. This structure will be used for feature matching because it is suitable for the problem considered in this paper.

Each neuron in the neural network is characterized by an activation function and its bias, and each connection between two neurons by a weight factor. In this paper, the neurons from the input and output layers have linear activation functions and hidden neurons have sigmoid activation function $F(u) = 1/(1+e^{-u})$. Therefore, for an input vector \mathbf{X} , the neural network output vector \mathbf{Y} can be obtained according to the following matrix equation [4,5]:

$$\mathbf{Y} = \mathbf{W}_2 * F(\mathbf{W}_1 * \mathbf{X} + \mathbf{B}_1) + \mathbf{B}_2 \quad (11)$$

where \mathbf{W}_1 and \mathbf{W}_2 are the weight matrices between the input and the hidden layer and between the hidden and the output

layer, respectively, and \mathbf{B}_1 and \mathbf{B}_2 are bias matrices for the hidden and the output layer, respectively.

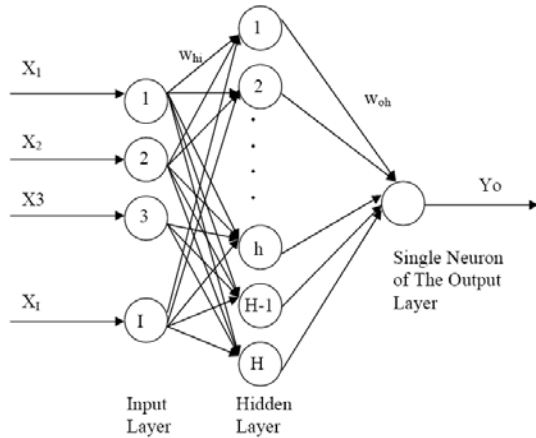


Figure 3. An MLP neural network.

Training a neural network is accomplished by adjusting its weights using a training algorithm. The training algorithm adapts the weights by attempting to minimize the sum of the squared error between a desired output and the actual output of the output neurons given by [4,5]:

$$E = \frac{1}{2} \sum_{o=1}^O (D_o - Y_o)^2 \quad (12)$$

where D_o and Y_o are the desired and actual outputs of the o^{th} output neuron. O is the number of output neurons. Each weight in the neural network is adjusted by adding an increment to reduce E as rapidly as possible. The adjustment is carried out over several training iterations until a satisfactorily small value of E is obtained or a given number of epochs is reached. The error back-propagation algorithm can be used for this task [4,5].

5. The Proposed Speaker Identification Method

In the presence of noise or telephone degradations, the speaker identification becomes a challenging task. The noise may mask the signal making the features infeasible in the identification. The telephone degradation also acts like a lowpass filter on the speech signal removing most of the characteristic features of the speaker. Thus, much more coefficients are required in the presence of noise or telephone degradations.

The discrete wavelet transform (DWT) can be a useful tool to overcome the degradation problems. Taking the one level DWT of a speech signal decomposes the signal into approximation and detail coefficients as will be mentioned in the second section. Features can be extracted from the DWT of the speech signal and added to the feature vector extracted from the signal itself to obtain a large feature vector suitable for speaker identification in the presence of degradations. Wavelet denoising can also be used to reduce the effect of noise prior to speaker identification. The proposed approach for feature extraction in the presence of degradations is illustrated in Fig.(4).

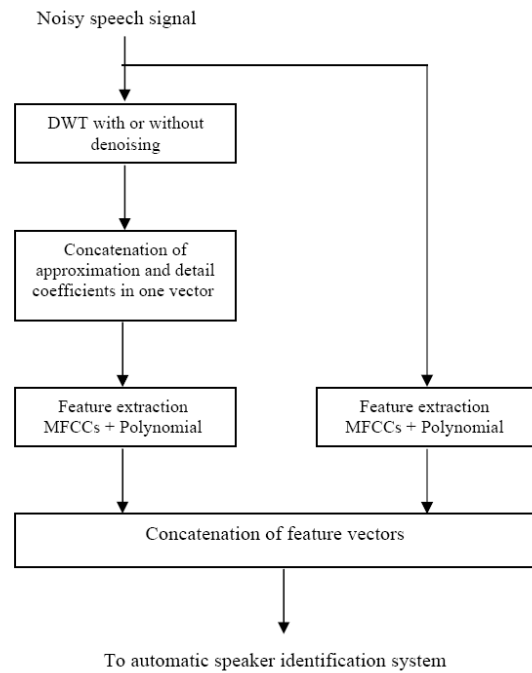


Figure 4. The proposed approach for feature extraction in the presence of degradations.

5.1. The Discrete Wavelet Transform

The DWT is a very popular tool for the analysis of non-stationary signals. It can be regarded as equivalent to filtering the speech signal with a bank of bandpass filters, whose impulse responses are all approximately given by scaled versions of a mother wavelet. The scaling factor between adjacent filters is usually 2:1 leading to octave bandwidths and center frequencies that are one octave apart [14-27]. The outputs of the filters are usually maximally decimated so that the number of DWT output samples equals the number of input samples and thus no redundancy occurs in this transform. The one level DWT decomposition reconstruction filter bank is shown in Fig.(5).

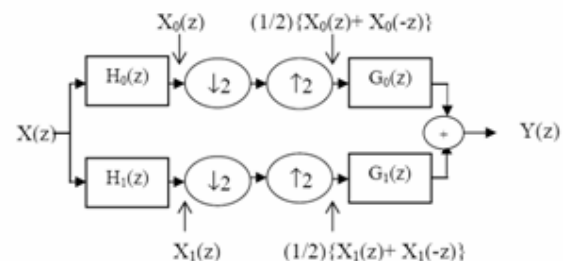


Figure 5. The two band decomposition-reconstruction wavelet filter bank.

The art of finding a good wavelet lies in the design of the set of filters, H_0 , H_1 , G_0 and G_1 to achieve various tradeoffs between spatial and frequency domain characteristics while satisfying the perfect reconstruction (PR) condition [26]. In Fig.(5), the process of decimation and interpolation by 2:1 at the output of H_0 and H_1 effectively sets all odd samples of these signals to zero. For the lowpass branch, this is equivalent to multiplying $x_0(n)$

by $\frac{1}{2}(1+(-1)^n)$. Hence $X_0(z)$ is converted to

$$\frac{1}{2}\{X_0(z) + X_0(-z)\}. \text{ Similarly, } X_1(z) \text{ is converted to}$$

$$\frac{1}{2}\{X_1(z) + X_1(-z)\}.$$

Thus, the expression for $Y(z)$ is given by [26]:

$$\begin{aligned} Y(z) &= \frac{1}{2}\{X_0(z) + X_0(-z)\}G_0(z) \\ &+ \frac{1}{2}\{X_1(z) + X_1(-z)\}G_1(z) \\ &= \frac{1}{2}X(z)\{H_0(z)G_0(z) + H_1(z)G_1(z)\} \\ &+ \frac{1}{2}X(-z)\{H_0(-z)G_0(z) + H_1(-z)G_1(z)\} \end{aligned} \quad (13)$$

The first PR condition requires aliasing cancellation and forces the above term in $X(-z)$ to be zero. Hence, $\{H_0(-z)G_0(z) + H_1(-z)G_1(z)\} = 0$, Which can be achieved if [26]:

$$H_1(z) = z^{-k}G_0(-z) \text{ and } G_1(z) = z^k H_0(-z) \quad (14)$$

where k must be odd (usually $k = \pm 1$).

The second PR condition is that the transfer function from $X(z)$ to $Y(z)$ should be unity [23]:

$$\{H_0(z)G_0(z) + H_1(z)G_1(z)\} = 2 \quad (15)$$

If we define a product filter $P(z) = H_0(z)G_0(z)$ and substitute from Eq. (14) into Eq.(15), then the PR condition becomes [26]:

$$H_0(z)G_0(z) + H_1(z)G_1(z) = P(z) + P(-z) = 2 \quad (16)$$

This needs to be true for all z and, since the odd powers of z in $P(z)$ cancel with those in $P(-z)$, it requires that $p_0 = 1$ and that $p_n = 0$ for all n even and non-zero. The polynomial $P(z)$ should be a zero phase polynomial to minimize distortion. In general, $P(z)$ is of the following form [26]:

$$\begin{aligned} P(z) &= \dots + p_5 z^5 + p_3 z^3 + p_1 z + 1 + p_1 z^{-1} + p_3 z^{-3} \\ &+ p_5 z^{-5} + \dots \end{aligned} \quad (17)$$

The design method for the PR filters can be summarized in the following steps [23]:

- 1- Choose p_1, p_3, p_5, \dots to give zero phase polynomial $P(z)$ with good characteristics.
- 2- Factorize $P(z)$ into $H_0(z)$ and $G_0(z)$ with similar lowpass frequency response.

- 3- Calculate $H_1(z)$ and $G_1(z)$ from $H_0(z)$ and $G_0(z)$.

To simplify this procedure, we can use the following relation:

$$P(z) = P_t(Z) = 1 + p_{t,1}Z + p_{t,3}Z^3 + p_{t,5}Z^5 + \dots \quad (18)$$

where

$$Z = \frac{1}{2}(z + z^{-1}) \quad (19)$$

The Haar wavelet is the simplest type of wavelets. In the discrete form, Haar wavelets are related to a mathematical operation called the Haar transform. The Haar transform serves as a prototype for all other wavelet transforms [23]. Like all wavelet transforms, the Haar transform decomposes a discrete signal into two sub-signals of half its length. One sub-signal is a running average or trend; the other sub-signal is a running difference or fluctuation. This uses the simplest possible $P_t(Z)$ with a single zero at $Z = -1$. It is represented as follows [26]:

$$P_t(Z) = 1 + Z \quad \text{and} \quad Z = \frac{1}{2}(z + z^{-1}) \quad (20)$$

Thus

$$\begin{aligned} P(z) &= \frac{1}{2}(z + 2 + z^{-1}) \\ &= \frac{1}{2}(z + 1)(1 + z^{-1}) = G_0(z)H_0(z) \end{aligned} \quad (21)$$

We can find $H_0(z)$ and $G_0(z)$ as follows:

$$H_0(z) = \frac{1}{2}(1 + z^{-1}) \quad (22)$$

$$G_0(z) = (z + 1) \quad (23)$$

Using Eq.(14) with $k=1$:

$$G_1(z) = zH_0(-z) = \frac{1}{2}z(1 - z^{-1}) = \frac{1}{2}(z - 1) \quad (24)$$

$$H_1(z) = z^{-1}G_0(-z) = z^{-1}(-z + 1) = (z^{-1} - 1) \quad (25)$$

The two outputs of $H_0(z)$ and $H_1(z)$ are concatenated to form a single vector of the same length as the original speech signal. The features are extracted from this vector and added to the feature vector generated from the original speech signal to form a large feature vector which can be used for speaker identification. The wavelet transformed signal vector contains both the approximation and the detail coefficients of the speech signal. So, feature extraction from this vector gives features from the lowpass as well as the highpass components of the signal which are more robust features to the presence of degradations.

5.2. Wavelet Denoising

Wavelet denoising is a simple operation which aims at reducing noise in a noisy speech signal. It is performed by choosing a threshold that is sufficiently a large multiple of the standard deviation of the noise in the speech signal. Most of the noise power is removed by thresholding the detail

coefficients of the wavelet transformed speech signal. There are two types of thresholding; hard and soft thresholding. The equation of the hard thresholding is given by [27]:

$$f_{hard}(x) = \begin{cases} x & |x| \geq T \\ 0 & |x| < T \end{cases} \quad (26)$$

On the other hand, that of soft thresholding is given by:

$$f_{soft}(x) = \begin{cases} x & |x| \geq T \\ 2x - T & T/2 \leq x < T \\ T + 2x & -T < x \leq -T/2 \\ 0 & |x| < T/2 \end{cases} \quad (27)$$

where T denotes the threshold value and x represents the detail coefficients of the DWT.

6. Experimental Results

In this section, four speaker identification experiments are carried out in the presence of different types of degradations. The degradations considered are AWGN, colored noise, telephone degradation with AWGN and telephone degradation with colored noise. Telephone degradations are simulated in our experiments by lowpass filtering of the speech signals with a small bandwidth filter.

In the training phase of the automatic speaker identification system, a database is first composed. 15 speakers are used to generate this database, each repeating a certain Arabic sentence 10 times. Thus, 150 speech samples are used to generate MFCCs and polynomial coefficients to form the feature vectors of the database. In the testing phase, each one of these speakers is asked to say the sentence again and his speech signal is then degraded. Similar features to that used in the training are extracted from these degraded speech signals and used for matching.

The features used in all experiments are 13 MFCCs and 26 polynomial coefficients forming feature vectors of 39 coefficients for each frame of the speech signal. Five methods for extracting these features are adopted in the paper. In the first method, the MFCCs and the polynomial coefficients are extracted from the speech signals only. In the second one, the features are extracted from the DWT of the speech signals. In the third method, the features are extracted from both the original speech signals and the DWT of these signals and concatenated in a single feature vector. In the fourth method, denoising is applied to the noisy signals in the testing phase only to reduce noise prior to feature extraction from the speech signals. In the last method, denoising is applied and features are extracted from both the denoised signals and the DWT of these denoised signals.

A comparison study is held between the five extraction methods for the above mentioned degradation cases and the results are given in Figs.(6) to (9). For speech signals contaminated by AWGN, it is clear from Fig.(6) that features extracted from both the speech signals and the DWT of these signals achieve the highest recognition rates at moderate and high signal to noise ratios (SNRs). At low SNRs, denoising is required.

For the case of colored noise contaminations studied in Fig.(7), the best performance is achieved by features extracted from both the speech signals and the DWT of these signals. It is also clear from this figure that denoising has no

effect in the presence of colored noise as it is mainly designed for AWGN contaminations.

For the case of telephone degradations studied in Figs.(8) and (9) for AWGN and colored noise, respectively, we notice that the performance deteriorates as the lowpass filtering removes much of the signals features. Wavelet denoising is required for both the AWGN and colored noise cases at the low SNRs. At high SNRs, features extracted from the signals and the DWT of these signals are more useful.

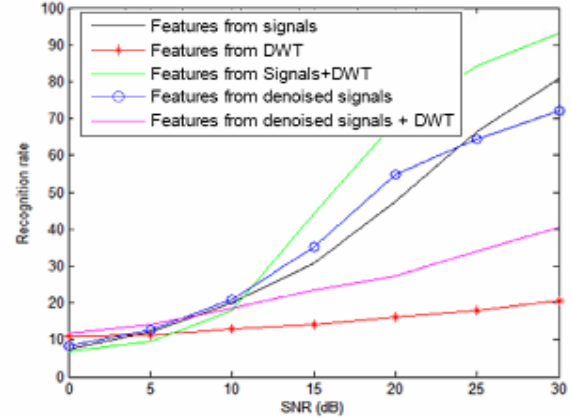


Figure 6. Recognition Rate vs. SNR for speech contaminated by AWGN.

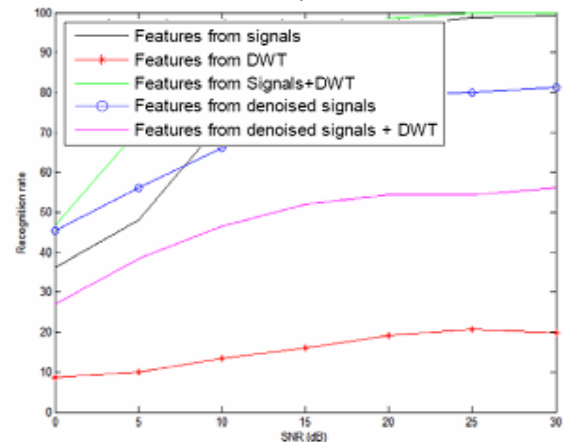


Figure 7. Recognition Rate vs. SNR for speech contaminated by colored noise.

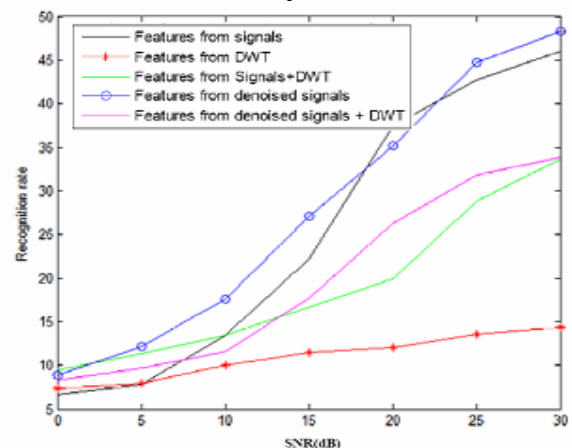


Figure 8. Recognition Rate vs. SNR in the presence of telephone degradation and AWGN.

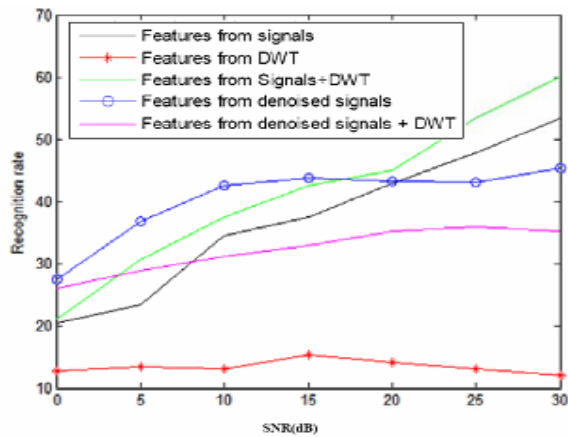


Figure 9. Recognition Rate vs. SNR in the presence of telephone degradation and colored noise.

7. Conclusions

This paper has presented a robust speaker identification method based on the wavelet transform. In this method, MFCCs and polynomial coefficients are extracted from the speech signals and the DWT of these signals and concatenated to form a large feature vector. Experimental results have shown that the proposed method is useful for feature extraction in the presence of noise contaminations and telephone degradations in the speech signals. Results have also shown that, wavelet denoising is required as a pre-processing step for speech signals at low SNRs to reduce the noise levels.

References

- [1] T. Kinnunen, "Spectral Features for Automatic Text-Independent Speaker Recognition", Licentiate's Thesis, University of Joensuu, Department of computer science, Finland, 2003.
- [2] D. Pullella, "Speaker Identification Using Higher Order Spectra", Dissertation of Bachelor of Electrical and Electronic Engineering, University of Western Australia, 2006.
- [3] R. Chengalvarayan, and L. Deng, "Speech Trajectory Discrimination Using the Minimum Classification Error Learning", IEEE Transactions on Speech And Audio Processing, Vol. 6, No. 6, pp. 505-515, 1998.
- [4] A. I. Galushkin, Neural Networks Theory, Springer-Verlag Berlin Heidelberg 2007.
- [5] G. Dreyfus, Neural Networks Methodology and Applications, Springer-Verlag Berlin Heidelberg 2005.
- [6] P. D. Polur and G. E. Miller, "Experiments With Fast Fourier Transform, Linear Predictive and Cepstral Coefficients in Dysarthric Speech Recognition Algorithms Using Hidden Markov Model", IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol. 13, No. 4, pp. 558-561, 2005.
- [7] R. Gandhiraj, P.S. Sathidevi, "Auditory-based Wavelet Packet Filterbank for Speech Recognition using Neural Network", Proceedings of the 15th International Conference on Advanced Computing and Communications, pp.666-671, 2007.
- [8] A. Katsamanis, G. Papandreou, and P. Maragos, "Face Active Appearance Modeling and Speech Acoustic Information to Recover Articulation", IEEE Transactions on Audio, Speech, And Language Processing, Vol. 17, No. 3, pp.411-422, 2009.
- [9] S. Dharanipragada, U. H. Yapanel, and B. D. Rao, "Robust Feature Extraction for Continuous Speech Recognition Using the MVDR Spectrum Estimation Method", IEEE Transactions on Audio, Speech, And Language Processing, Vol. 15, No. 1, pp. 224-234, 2007.
- [10] B. C. Jong, "Wavelet Transform Approach For Adaptive Filtering With Application To Fuzzy Neural Network Based Speech Recognition", PhD Dissertation, Wayne State University, 2001.
- [11] Z. Tufekci, "Local Feature Extraction For Robust Speech Recognition in The Presence of Noise", PhD Dissertation, Clemson University, 2001.
- [12] R. Sarikaya, "Robust And Efficient Techniques For Speech Recognition in Noise", PhD Dissertation, Duke University, 2001.
- [13] S. FURUI, "Cepstral Analysis Technique for Automatic Speaker Verification", IEEE Transactions on Acoustics, Speech, And Signal Processing, Vol. ASSP-29, No. 2, pp. 254-272, 1981.
- [14] I. Daubechies, "Where Do Wavelets Come From?—A Personal Point of View," Proceedings of the IEEE, Vol. 84, No. 4, pp. 510- 513, 1996.
- [15] A. Cohen and J. Kovacevec, " Wavelets: The Mathematical Background," Proceedings of the IEEE, Vol. 84, No. 4, pp. 514- 522, 1996.
- [16] N. H. Nielsen and M. V. Wickerhauser, " Wavelets and Time-Frequency Analysis," Proceedings of the IEEE, Vol. 84, No. 4, pp. 523- 522-540, 1996.
- [17] K. Ramchndran, M. Vetterli and C. Herley, " Wavelets, Subband Coding, and Best Basis," Proceedings of the IEEE, Vol. 84, No. 4, pp. 541- 560, 1996.
- [18] P. Guillemain and R. K. Martinet, " Characterization of Acoustic Signals Through Continuous Linear Time-Frequency Representations," Proceedings of the IEEE, Vol. 84, No. 4, pp. 561- 585, 1996.
- [19] G. W. Wornell, " Emerging Applications of Multirate Signal Processing and Wavelets in Digital Communications," Proceedings of the IEEE, Vol. 84, No. 4, pp. 586- 603, 1996.
- [20] S. Mallat, " Wavelets For A Vision," Proceedings of the IEEE, Vol. 84, No. 4, pp. 604- 614, 1996.
- [21] P. Schroder, " Wavelets in Computer Graphics," Proceedings of the IEEE, Vol. 84, No. 4, pp. 615- 625, 1996.
- [22] M. Unser and A. Aldroubi, " A Review of Wavelets in Biomedical Applications," Proceedings of the IEEE, Vol. 84, No. 4, pp. 626- 638, 1996.
- [23] M. Farge, N. Kevlahan, V. Perrier and E. Goirand, " Wavelets and Turbulence," Proceedings of the IEEE, Vol. 84, No. 4, pp. 639-669 , 1996.
- [24] A. Bijaoui, E. Slezak, F. Rue and E. Lega, " Wavelets and The Study of The Distant Universe," Proceedings of the IEEE, Vol. 84, No. 4, pp. 670- 679, 1996.
- [25] W. Sweldens, " Wavelets: What Next?," Proceedings of the IEEE, Vol. 84, No. 4, pp. 680- 685, 1996.
- [26] A. Prochazka, J. Uhlir, P. J. W. Rayner and N. J. Kingsbury, Signal Analysis and Prediction. Birkhauser Inc. , 1998.
- [27] J. S. Walker, A Primer on Wavelets and Their Scientific Applications. CRC Press LLC, 1999