Speaker Recognition Using Discrete Wavelet Transform and Artificial Neural Networks

Amin A. Abdul Fattah¹* and Layla M. Salih²

¹Department of Electrical Engineering, College of Engineering, Salahaddin University-Erbil, Iraq
²Department of Electrical Engineering, College of Engineering, University-Erbil, Iraq

ARTICLE INFO

Article History:
Received: 18/06/2015
Accepted: 12/10/2015
Published: 16/5/2016

Keywords:
Speaker recognition, text-dependent, feature extraction, DWT, ANN

*Corresponding Author:
Amin A. Abdul Fattah
Email: amab_Fattah@yahoo.com
University of Kirkuk, Iraq

ABSTRACT

In recent years biometrics has emerged as applied scientific discipline with the objective of automatically capturing personal identifying characteristics that distinguish one individual from another and using the measurements for security, surveillance, and forensic application. Speaker recognition is the process of automatically recognizing who is speaking based on individual information included in the speech waves. This paper presents the speaker identification method based on Discrete Wavelet Transform (DWT) and Artificial Neural Networks (ANN). In this study the DWT is used to extract a speaker's discriminative features from the mathematical representation of the speech signal. These feature vectors are used to train a feedforward neural network which is used to model the speakers and make the decision task. A database of 20 speakers (10 male and 10 female) has been used with a vocabulary of Kurdish words. The system led to 100% identification rate for text-dependent and 80% identification rate for text-independent.

1. Introduction

The speech signal contains many levels of information. Primarily a message is conveyed via the spoken words. At other levels, speech conveys information about the language being spoken, the emotion, gender, and the identity of the speaker. While speech recognition sets its goals at recognizing the spoken words in speech, the aim of automatic speaker recognition is to identify the speaker by extraction, characterization and recognition of the information contained in the speech signal. There are two main factors that make speaker recognition a compelling biometric; (1) Speech is a natural signal to produce that is not considered threatening by the users to provide, and (2) the telephone system provides a familiar network of sensors for obtaining and delivering the speech signal (Singh, 2003). Speaker-specific variations in speech signal are partly due to the anatomical differences in speech-producing organs, and partly due to idiosyncrasies of the speaker, such as speaking habits and emotional state (Guruprasad et al, 2003).

Speaker recognition encompasses verification and identification. Automatic speaker verification (ASV) is the use of a machine to verify a person's claimed identity from his voice. In automatic speaker identification (ASI), there is no a priori identity claim, and the system decide who the person is, what group the person is a member of, or (in the open-set case) that the person is unknown (Campbell, 1997).

A generic speaker recognition system is shown in figure 1, where the desired features are first extracted from the speech signal. These features are then used as input to a
classifier, which makes the final decision regarding verification or identification (Farrell et al, 1994).

Speaker recognition methods can also be divided into text-independent and text-dependent methods. In a text-independent system, speaker models capture characteristics of somebody's speech which show up irrespective of what one is saying. In a text-dependent system, on the other hand, the recognition of the speaker's identity is based on his or her speaking one or more specific phrases, like passwords, card numbers, PIN codes, etc. (Reda and Aoued, 2005). The current approach is aimed at text-dependent and text-independent speaker identification.

Some of the recent works on speaker recognition depend on classical features including formant frequencies (Daqrouq et al, 2009), Mel Frequency Cepstral Coefficients (MFCC) (Kaur and Sharma, 2014), MFCC and inverted MFCC (Singh and Rajan, 2011), LPC-Residual and MFCC (Curipe and Camacho, 2013), Linear Prediction Cepstral Coefficient (LPCC) and MFCC (Kumar and Ladhukar, 2015), LPC and MFCC (Subhashini and Pratap, 2014), LPC, LPCC, and RASTA-PLP (Visalakshi and Dhanalakshmi, 2014).

However, these methods have some disadvantages. These methods accept signal stationary within a given time frame and may therefore lack the ability to analyze the localized events correctly. Moreover, the LPC method accepts a particular linear (all-pole) model of speech/word production which strictly speaking is not the case (Avci, 2006). Literature on various studies reveals that in case of the above parameters, the feature vector dimensions and computational complexity are higher to a great extent.

Some researchers (Woo et al, 2001; Zheng and Ching, 2004; Avci, 2006; Al-Ani et al, 2007; Shafik et al, 2009; Abdalla and Ali, 2010; Daqrouq et al, 2009; Pawar et al, 2014) are concentrating on the wavelet transform for speaker feature extraction stage.

Wavelets have emerged as a new and powerful tool for nonstationary signal analysis. One of the main properties of this technique is its ability to analyze signal with different levels of resolution. Fast transients can be analyzed with short windows, while slowly varying phenomena can be observed with longer time windows (Greenberg et al, 2004).

In this paper, we consider a hybrid approach for isolated word speaker identification, where the Discrete Wavelet Transform (DWT) is used as a feature extraction component. Also, in order to further reduce the dimensionality of the extracted feature vectors, statistics over the set of wavelet coefficients are used. Artificial Neural Network (ANN) is well-known as a technique that has the ability to classify nonlinear problem. An (ANN) is used for pattern recognition.

2. The Database

A database is created for Kurdish language using 20 speakers (10 male, 10 female). The database was taken from Salahaddin University / Engineering College (students and staff). Each speaker utters 9 Kurdish words twice, one for the training phase and the other for the recognizing phase.

3. Discrete Wavelet Transform (DWT)

Wavelet analysis is a windowing technique with variable-sized regions, which allows the use of long time intervals where we want more precise low-frequency information, and shorter regions where we want high-frequency information (Misiti et al, 2000)

For many signals, the low-frequency part contains the most important part. It gives an identity to a signal. Consider the human voice, if, we remove the high-frequency components, the voice sounds different, but we can still tell what’s being said. In wavelet analysis, we often speak of approximations and details. The approximations are the high-scale, low-frequency components of the signal. The
details are the low-scale, high frequency components. The DWT is defined by the following equation:

$$W(j, k) = \sum_j \sum_k x(k) 2^{-j/2} \psi(2^{-j} n - k)$$

(1)

Where $\psi(t)$ is a time function with finite energy and fast decay called the mother wavelet. The DWT analysis can be performed using a fast, pyramidal algorithm related to multirate filterbanks. As a multirate filterbank the DWT can be viewed as a constant Q filterbank with octave spacing between the centers of the filters. Each subband contains half the samples of the neighboring higher frequency subband. In the pyramidal algorithm the signal is analyzed at different frequency bands with different resolution by decomposing the signal into a coarse approximation and detail information. The coarse approximation is then further decomposed using the same wavelet decomposition step. This is achieved by successive highpass and lowpass filtering of the time domain signal and is defined by the following equations (Krishnan and Anto, 2009):

$$y_{\text{low}}[n] = \sum_{k=-\infty}^{\infty} x[k] h[2n - k]$$

$$y_{\text{high}}[n] = \sum_{k=-\infty}^{\infty} x[k] g[2n - k]$$

(2)

Where $y_{\text{low}}[n]$, $y_{\text{high}}[n]$ are the outputs of the lowpass (h) and highpass (g) filters, respectively after subsampling by 2. Because of the downsampling the number of resulting wavelet coefficients is exactly the same as the number of input points. A variety of different wavelet families have been proposed in the literature (Tzanetakis et al, 2001). In this study the daubechies (Db2 - Db5) and Haar wavelet are used to encode the speech signals. Figure 2 shows the decomposition and the reconstruction of three levels by using DWT.

4. Artificial Neural Networks

Artificial Neural Networks (ANNs) are computational models which attempt to mimic the learning function of the brain. ANNs are composed of a set of neurons or processing units which are connected together by means of connecting weights. ANNs which are structured to learn and generalize so that the network may learn by continuous adjustment of the weights of the connections (Topping et al, 1998).

Feedforward neural networks (FNN) have been widely used for various tasks, such as pattern recognition, function approximation, dynamical modeling, data mining, and time series forecasting, to name just a few. The training of FNN is mainly undertaken using the backpropagation (BP) - based learning algorithms (Yu et al, 2002).

The typical backpropagation network has an input layer, an output layer, and at least one hidden layer. There is no theoretical limit on the number of hidden layers but typically there is just one or two as shown in figure 3 (Alsmadi et al, 2009).
5. Materials and Methodology:

5.1 Preprocessing:

The sound files contained in the database are .WAV files recorded with microphone sampled at 11025 Hz with a resolution of 16 bits per sample. These sampled signals can capture all frequencies up to 5 kHz, which cover most energy of sounds that are generated by humans (Reda and Aoued, 2005). This sampling rate was chosen for recording which is twice the frequency of the original signal and follows the Nyquist rule of sampling.

The speech signal must be broken up into small frames with time length in the range of (20-40) millisecond. In this step the continuous speech signal is blocked into frames of N samples, with adjacent frames being separated by M (M < N). The human speech production is known to exhibit quasi-stationary behavior over a short period of time (20-40) msec.

Then windowing is performed on each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame (Kabir and Ahsan, 2007). In this study Hamming window was used, which has the form:

\[
w(n) = \begin{cases} 
0.54 - 0.46 \cos\left(\frac{2\pi n}{N - 1}\right) & \text{for } 0 \leq n < N \\
0 & \text{otherwise}
\end{cases}
\]

Where,

\[N = \text{number of samples in each frame}\]

The result of the windowing is given by:

\[Y(n) = X(n) \times w(n)\]

5.2 Feature Extraction

The energy in each frame in each level of decomposition is calculated with the application of equation (5) (Kotnik and Kacic, 2007). In order to further reduce the dimensionality of the extracted feature vectors, statistics over the set of the wavelet coefficients are used (Tzanetakis et al 2001). In this study, the standard deviation of the energy overall the frames for each level of the decomposition levels was calculated. The deviated energy values are collected in a vector of 9 elements. Now, the features for each speaker are calculated and the first step of recognition is done.

\[P = \frac{\sum_{i=1}^{n} S_i^2}{n}\]

Where \[S_i\] : is an element in the approximation coefficient set \[n\] : is the number of the elements in the approximation coefficient set

5.3 The Classification Process:

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. ANN is 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. 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 (Shafik et al, 2009). In this step feedforward backpropagation neural network is used with the training condition and the structure of the ANN is tabulated in table 1.
Table 1: Parameters used for the ANN

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network type</td>
<td>Feed forward backpropagation</td>
</tr>
<tr>
<td>No. of layers</td>
<td>Three layers: Input, one hidden and output</td>
</tr>
<tr>
<td>No. of neurons in layers</td>
<td>9-Input, 22- Hidden and 20 Output</td>
</tr>
<tr>
<td>Activation function</td>
<td>tansig, logsig</td>
</tr>
<tr>
<td>Performance function</td>
<td>0.0001</td>
</tr>
<tr>
<td>No. of epochs</td>
<td>1477</td>
</tr>
</tbody>
</table>

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 the following equation:

\[ E = \frac{1}{2} \sum_{o=1}^{k} (D_o - Y_o)^2 \]

Where \( D_o \) and \( Y_o \) are the desired and actual outputs of the \( o \)th output neuron. \( k \) is the number of output neurons. Each weight in the ANN 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 (Shafik et al, 2009) (see figure 4).

Figure 4: Training of the neural network using Db4 wavelet as feature extractor.

Straight line – Goal, falling curve – Training

6. Experimental Results

In this study, trails were made to identify speakers by comparing a sample of their voices to a database included speech data files of 20 speakers (10 male and 10 female). The speech files consist of nine isolated Kurdish words. Each speaker repeats each word twice, one for training phase and the other for testing phase. Five of the utterances were selected for training for both text-dependent and text-independent identification. The selected data set includes the words (Bayaneet, Bash, Choneet, Ewarat, Bash, Jwana, Shereen, Haween, and Zstan), the words (Bayaneet, E, Choneet, Ewarat, and Jwana) are used as training set. Figure 5 shows the speech sample signals for the first speaker and its wavelet channels.
The performance of the speaker identification system may be affected by the type of the wavelet basis function. Haar and (Db2-Db5) type of wavelet function for feature extraction were used in this study. A feature vector of size 9 is obtained in the eight levels of the decomposition for each speech sample. The feature vectors of five words are collected to form a training set. This training set is given as an input to ANN classifier for the training purpose. The system is then tested by the same five words from the recognition set for the case of text-dependent and the remaining utterances are used for the case of text-independent. Results are shown in table 2, the graph is shown in figure 6.

<table>
<thead>
<tr>
<th>Wavelet Basis</th>
<th>Text Dependent</th>
<th>Text Independent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haar</td>
<td>70%</td>
<td>65.00%</td>
</tr>
<tr>
<td>Db2</td>
<td>82%</td>
<td>76.00%</td>
</tr>
<tr>
<td>Db3</td>
<td>84%</td>
<td>77%</td>
</tr>
<tr>
<td>Db4</td>
<td>100%</td>
<td>80%</td>
</tr>
<tr>
<td>Db5</td>
<td>85%</td>
<td>87%</td>
</tr>
</tbody>
</table>

Figure 6: The correct recognition rates for text-dependent and text-independent using different wavelet basis

The reliability of the pattern recognition system is measured by testing the system with hundreds of input speech samples of the speakers with varying quantities of noise. Noisy signals at different SNRs are used for recognizing the speakers. Results are shown in table 3, the graph is shown in figure 7.

Table 3: Details of the system performance at different levels of noise

<table>
<thead>
<tr>
<th>Training Signal Set</th>
<th>Testing Signal Set</th>
<th>Text-Dependent</th>
<th>Text-Independent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original signal set</td>
<td>Noisy signals at SNR=10dB</td>
<td>27%</td>
<td>30%</td>
</tr>
<tr>
<td>Original signal set</td>
<td>Noisy signals at SNR=15dB</td>
<td>37%</td>
<td>43%</td>
</tr>
<tr>
<td>Original signal set</td>
<td>Noisy signals at SNR=20dB</td>
<td>53%</td>
<td>58%</td>
</tr>
<tr>
<td>Original signal set</td>
<td>Noisy signals at SNR=25dB</td>
<td>84%</td>
<td>78%</td>
</tr>
<tr>
<td>Original signal set</td>
<td>Noisy signals at SNR=30dB</td>
<td>94%</td>
<td>85%</td>
</tr>
</tbody>
</table>

Figure 5: (a) Speech samples of speaker number 1, (b) A frame from the word “Choneet” and its wavelet channels
7. Conclusion

Speaker recognition is a challenging problem and there is still a lot of work that needs to be done in this area. Over the last twenty years, speaker recognition has received substantial attention from researchers in biometrics, pattern recognition, signal processing, and cognitive psychology communities. This common interest in speaker recognition technology among researchers working in diverse fields is motivated both by the remarkable ability to recognize people and by the increased attention being devoted to security application.

In this study, an automatic speaker recognition system is designed for isolated spoken words in Kurdish using DWT and ANN. 100% identification in text-dependent and 80% identification for text-independent is obtained from this study. The computational complexity and the reduction in the storage capacity of feature vector size is successfully reduced to a great extent by using the DWT. Storage reduction is an important factor when talking about applications through internet and telecommunications. Thus DWT is an elegant tool for the analysis of non-stationary signals like speech. Recognition rate can be increased by increasing the number of samples. The ANN classifier which was used in this study provides very good accuracies. Alternate classifiers like Support Vector Machines, Genetic algorithms, Fuzzy logic approaches etc. can also be used and a comparative study of these classifiers can be performed as an extension of this work.

References


