Speaker identification system using empirical mode decomposition and an artificial neural network

Jian-Da Wu *, Yi-Jang Tsai

Graduate Institute of Vehicle Engineering, National Changhua University of Education, 1 Jin-De Rd., Changhua City, Changhua 500, Taiwan

Abstract

This paper presents a speaker identification system using empirical mode decomposition (EMD) feature extraction method and artificial neural network in speaker identification. The EMD is an adaptive multi-resolution decomposition technique that appears to be suitable for non-linear, non-stationary data analysis. The EMD sifts the complex signal of time series without losing its original properties and then obtains some useful intrinsic mode function (IMF) components. Calculating the energy of each component can reduce the computation dimensions and enhance the performance of classification. The features were used as inputs to neural network classifiers for speaker identification. In the speaker identification, the back-propagation neural network (BPNN) and generalized regression neural network (GRNN) were applied to verify the performances and the training time in the proposed system. The experimental results indicated the GRNN can achieve better recognition rate performance with feature extraction using the EMD method than BPNN.

1. Introduction

Biometric security systems are convenient for users because it is not necessary to remember passwords or carry identification cards. Further, the biometric recognition is a technology based on a feature vector derived from a specific physiological or behavioral characteristic, which is exclusive and unique, such as facial characteristics as well as the iris, retina, and fingerprints, etc. (Liu & Silverman, 2001; Pankanti, Bolle, & Jain, 2000). These methods have been applied for many purposes, however, they are often restricted to special conditions. With improved research of vocal signals, we can communicate with computers through our voice, meaning computers are able to recognize what we say and who we are; therefore, besides entering the command into the computer, the most convenient way to operate a computer is to speak to it. The particular characteristics of the sound and the way of talking of each person are unique. It is easy to distinguish the sound of people using computers, even though the sounds we hear cannot be distinguished by our ears. Thus, we can precisely recognize each person by analyzing their voice using computers. Because of its non-contact characteristic, speaker identification system can be effectively utilized in suspect identification, vehicle entry, door control systems, PC login, and other services.

In the present study, the execution of speaker identification is divided into two stages: the first is feature extraction and the second is speaker classification based on the extracted features (Sari-kaya, Pellom, & Hansen, 1998). In traditional techniques, the sound features are usually obtained by fast Fourier transform (FFT) (Corinthios, 1971) or short time Fourier transforms (STFT) (Portnoff, 1980), the FFT is the most useful method for frequency domain feature extraction. Nevertheless, this method loses some information in the time domain while the signals are converted into the frequency domain, therefore, the STFT was developed to overcome the above drawback and evaluate the sinusoidal frequency. However, they adopt signals fixed within a given time frame and may therefore lack the ability to analyze events correctly (Avci & Akpolat, 2006). In contrast to the Fourier transform, the characteristic of wavelet transform (WT) gives short time intervals for the high-frequency bands and long time intervals for the low-frequency band. The main advantage of WT is it has an adjustable window size. In 2006, Wu and Chen used the continuous wavelet transforms (CWT) for fault signal diagnosis of an internal combustion (Wu & Chen, 2006). In 2009, Wu and Ye proposed a continuous wavelet transforms method for driver identification (Wu & Ye, 2009). Unfortunately, it usually generates an immense amount of wavelet coefficients, and therefore will be highly redundant in CWT. Thus, the discrete wavelet transform (DWT) was developed to improve on CWT, and it can avoid generating redundant information. The DWT permits the systematic decomposition of a signal into its sub-band levels. It can be performed with minimum distortion of the signal, even for stationary signal analysis (Seker & Ayaz, 2003). The wavelet packet decomposition (WPD) is able to decompose the signal into the entire time–frequency range.

Keywords:
Biometric security systems
Speaker identification
Empirical mode decomposition
Back-propagation neural network
Generalized regression neural network

Article info

Article history:
Received 8 April 2010
Received in revised form 15 July 2010
Accepted 21 July 2010
Available online 28 July 2010

Corresponding author.
⇑E-mail address: jdwu@cc.ncue.edu.tw (J.-D. Wu).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.11.013
and then provides more accurate resolution than the DWT. However, the choice of suitable wavelet functions is still a principal disadvantage. Therefore, the empirical mode decomposition (EMD), a signal processing technique particularly suitable for non-linear and non-stationary series, has recently been proposed (Huang et al., 1998) as a new tool for data analysis. The EMD method is able to decompose a complex signal into a series of intrinsic mode functions (IMF) and a residue in accordance with different frequency bands (Coughlin & Tung, 2004). These are signal components highlighting distinct time-scales (frequencies) of the input time-series. The IMF represents the natural oscillatory mode embedded in the signal. EMD is self-adaptive because the IMF works as the basis functions determined by the signal itself rather than what is pre-determined. Therefore, EMD is highly efficient in non-stationary data analysis.

Neural network applications to recognition have been reported in recent years (Lung, 2006; Xiang & Berger, 2003). They are widely applied in data analysis and speaker recognition. The main advantage of the artificial neural network is the transfer function between the input vectors and the target matrix does not have to be predicted in advance. In this paper, the investigation of speaker identification using EMD and the neural network is proposed. The GRNN algorithm compared the performance with the traditional BPNN in this paper. They were implemented to classify thirty-six speakers of speech output signal. In the following sections, the proposed methods and performance of the speaker identification system will be described.

2. Feature extraction of empirical mode decomposition

2.1. Intrinsic mode function

A multi-resolution decomposition technique is presented (Huang et al., 1998), empirical mode decomposition (EMD), which is adaptive and appears to be suitable for non-linear and non-stationary signal analysis. It was carried in the time domain to form the basis functions adaptively. The major advantage of EMD is the basis functions can be directly derived from the signal itself. The intrinsic modes are not necessarily sinusoidal functions. Compared with Fourier analysis, EMD analysis is adaptive while the basis functions of Fourier analysis are linear combinations of fixed sinusoids. In fact, IMFs can be both amplitude and frequency modulated. Apparently, EMD is empirical, intuitive, direct, and adaptive. The EMD decomposes the original signal into a definable set of adaptive basis of functions called the intrinsic mode functions. Each IMF must satisfy two basic conditions (Huang et al., 1998): (1) in the whole data set, the number of extrema and the number of zero crossings must either equal or differ at most by one. Note, either local minima or local maxima are extrema. Moreover, a sample \( S_i \) in a time-series is a local maximum if \( S_i > S_{i-1} \) and \( S_i > S_{i+1} \), and a sample \( S_i \) is a local minimum if \( S_i < S_{i-1} \) and \( S_i < S_{i+1} \), where \( i \) is a discrete time; and (2) at any point, the mean value of the envelope, one defined by the local maxima (upper envelope) and the other by the local minima (lower envelope) is zero. Since sifting is a recursive process, a sifting stopping rule is required. The definition above is empirical and currently there are no definite equations for estimating IMFs. Therefore, any arbitrary time series satisfying conditions 1 and 2 is an IMF.

2.2. Sifting process

In the cause of the separate components called IMFs, we perform a process called sifting. The purpose of sifting is to subtract the large-scale features of the signal repeatedly until only the fine-scale features remain. First, the original speech signal, \( x(t) \), should be enclosed by the upper and lower envelope in the time domain. Using a cubic spline, the local maxima is connected forming the upper envelope \( u_i(t) \) and the local minima is connected forming the lower envelope \( l_i(t) \). The two envelopes cover all the data points. The envelope mean \( m(t) \) is determined as follows, \( m(t) = (u_i(t) + l_i(t))/2 \). The first component is described as

\[
h_1(t) = x(t) - m(t)
\]

(1)

The component \( h_1(t) \) is now examined to see if it satisfies the conditions to be an IMF. If \( h_1(t) \) doesn’t satisfy the conditions, \( h_1(t) \) is regarded as the original data, the sifting process would repeat, obtaining the mean of the upper and lower envelopes, which is designated as \( m_1 \); therefore,

\[
h_{11}(t) = h_1(t) - m_{11}(t)
\]

(2)

Then, repeat the procedure until \( k \) is an IMF, i.e.,

\[
h_{1k}(t) = h_{1(k-1)}(t) - m_{1k}(t)
\]

(3)

After \( k \) siftings, we explained the first intrinsic mode to be

\[
c_1 = h_{1k}
\]

(4)

Finally, \( c_1 \) revealed the higher frequency component of IMF. To obtain enough physical definitions of IMF, the sifting stop criteria, namely, the stop conditions, are of great importance and are found by the following equation:

\[
SD = \sum \frac{|h_{1(k-1)}(t) - h_{1k}(t)|^2}{|h_{1(k-1)}(t) + h_{1(k-1)}(t)|}
\]

(5)

The typical values of SD are 0.2 and 0.3. To obtain the second and subsequent intrinsic mode functions, the residual signal can be calculated as

\[
x(t) - c_1(t) = r_1(t)
\]

(6)

\( r_1 \) considers the original data, and by repeating the above procedures, \( x(t) \) could be obtained by the second IMF component \( c_2 \). The procedure as described above is repeated for \( n \) times, then the \( n \)-IMFs of signal \( x(t) \) could be obtained

\[
r_1(t) - c_2(t) = r_2(t)
\]

(7)

\[
r_{n-1}(t) - c_n(t) = r_n(t)
\]

The decomposition procedure can be stopped when the residue, \( r_n(t) \), becomes a constant, a monotonic function, or a function containing only a single extrema, from which no more IMF can be extracted. By summing Eqs. (6) and (7), the original signal can be reconstructed as follows:

\[
x(t) = \sum_{i=1}^{n} c_i(t) + r_n(t)
\]

(8)

where \( n \) is the number of IMF’s extracted from \( x(t) \), \( r_n(t) \) denotes the final residue, and \( t \) is time. The IMFs \( c_1, c_2, \ldots, c_n \) are comprised of different frequency bands ranging from high to low. The frequency components included in each different frequency band, while \( r_n \) represents the central tendency of signal \( x(t) \). Fig. 1 indicates the entire flow chart of empirical mode decomposition.

---

**Fig. 1.** Flow chart of the empirical mode decomposition.
Data acquisition card

3. Classification of speaker identification

3.1. Back-propagation neural network

In the design of the speaker identification program, a recognition method of the features based on EMD was chosen to evaluate the effectiveness of the selected feature sets for the identification system. There are many kinds of artificial neural network (ANN) models, among which the back-propagation neural network (BPNN) model is the most widely used (Rumelhart, Hinton, & Williams, 1986; Xia, Xie, & Zhu, 1997). BPNN has many flaws such as: existence of a local minimum, improper learning rate and requiring a large number of iterations to achieve convergence. The most common configuration of BPNN is composed of three layers. Two different components make up the construct of phases: one is the feed forward phase in which the external input information at the input nodes is propagated forward to compute the output information signal at the output unit, and a backward phase in which modification to the connection strengths is made based on the differences between the computed and observed information signals at the outputs units. (Cigizoglu & Alp, 2006). In the present study, the disadvantages of the BPNN include the need for large amounts of weight calculation, the large number of iterations to converge to the desired value, a very long training time, and the existence of local minima.

3.2. Generalized regression neural network

The generalized regression neural network was presented by Specht in 1991 (Specht, 1991). Based on non-linear kernel regression in statistics, it can estimate the map inherent in any sample data, and the estimation can converge to the optimal regression surface, even with few samples. It does not require an iterative training procedure to converge to the desired solution as in the back-propagation neural network. Fig. 2 shows the structure of the GRNN. The GRNN includes 4 different layers: input layer, pattern layer, summation layer and output layer. It estimates any arbitrary function between the input and output vectors, drawing the function estimate directly from the training data. The GRNN provides the estimation of continuous variables in a linear or non-linear way, and is an optimal estimator implementing the well-known statistical concept of conditional probability. According to the definition, the regression of dependent variable $y$ on independent variable $x$ evaluates the most probable value for $y$, given $x$ and a training set. Assume $f(x,y)$ represents the known joint continuous probability density function (PDF) of a vector random variable $x$, and a scalar random variable. The conditional mean of $y$ given $X$ is given by

$$E(y|X) = \hat{y}(x) = \frac{\int_{-\infty}^{\infty} y f(x,y)dy}{\int_{-\infty}^{\infty} f(x,y)dy}$$  \hspace{1cm} (9)

In practice, the $f(x,y)$ is usually not known and, therefore, has to be estimated by a sample of observations of $x$ and $y$. The probability estimator $\hat{f}(X,Y)$ is based on the sample values $X'$ and $Y'$ of the random variables $x$ and $y$, where $n$ is the number of sample observations and $p$ is the dimension of the vector variable $x$:

$$\hat{f}(X,Y) = \frac{1}{n(2\pi)^{p/2} |\Sigma|^{1/2}} \sum_{i=1}^{n} e^{-\frac{1}{2} (x-x_i)^T \Sigma^{-1} (x-x_i)}$$ \hspace{1cm} (10)

where $d(x,x_i) = \sum_{j=1}^{p} (x_i - x_j)^2$ and $dy(y,y_i) = |y - y_i|/|\sigma_j|$ (Parzen, 1962). A physical exposition of the possibility evaluation $f(X,Y)$ it assigns the sample probability of width $\sigma$ for each sample $X'$ and $Y'$. The probability estimate is the sum of those sample probabilities, where $\sigma$ is the smoothing parameter, the scalar function $D_r^2$ defined as:

$$D_r^2 = (X - X')^2(X - X')$$ \hspace{1cm} (11)

The probability estimation is the sum of these sample probabilities. Supplanting the joint probability evaluation in Eq. (10) into the condition mean Eq. (9), and exchanging the order of integration and summation yields

$$\hat{Y}(x) = \frac{\sum_{i=1}^{n} e^{-d(x,x_i)} \int_{-\infty}^{\infty} ye^{-d(y,y_i)dy}}{\sum_{i=1}^{n} e^{-d(x,x_i)}}$$  \hspace{1cm} (12)

Assessing the two indicated integrations using $\int_{-\infty}^{\infty} ze^{-z^2}dz = 0$, yields the following:

$$\hat{Y}(x) = \sum_{i=1}^{n} ye^{-d(x,x_i)}$$ \hspace{1cm} (13)

The Eq. (5) is the weighted sum over of all the training patterns. Each training pattern is weighted exponentially according to smoothing factors, and its Euclidean distance to the unknown pattern $x$. The smoothing factor with each feature with a small value should be distributed to insignificant features so they have a minimal effect on the distance estimate. If the smoothing parameter $\sigma$ is large, the estimated density is forced to be smooth and the limit becomes a multivariate Gaussian with covariance $\sigma^2 I$. On the other hand, a smaller value of $\sigma$ allows the estimated density to assume non-Gaussian shapes, but with the drawback that wild points may have too great an effect on the estimate (Specht, 1991).

4. Experimental analysis and results

4.1. Experimental set-up

The speaker identification system consists of a combination of speech signal feature extraction and voice classification using an

---

**Fig. 2.** Structure of generalized regression neural network.

**Fig. 3.** Experimental setup of the speaker identification system.
artificial neural network. In the experimental work, a desirable database plays an important role in verifying a classification methodology. It was recognized the speaker database is limited in its usefulness for improving and comparing several classification algorithms. Fig. 3 shows the experiment setup of the proposed speaker identification system. In this section, we first present the database. The recording environment is a laboratory environment. There are 36 individual speakers with 18 individual male and 18 individual female speaking five Chinese short sentences. As shown in Table 1, each speaker repeated the allocated sentences 50 times. The recording equipment included a microphone (PCB 130D20) and a data acquisition system (NI-6024/CB-68LP). The sampling rate was 16 kHz. In the stage of feature extraction, the EMD method divided the crude signal into many IMF components. Accordingly, the energy calculation displayed the feature distributions in different speaker voice states. In the voice classification procedure, the identification system used the generalized regression neural network (GRNN) to train the features and classify the different speakers, and then compared the performance with the traditional back-propagation neural network (BPNN).

4.2. Speech signal analysis

This experimental work recorded all the crude signals of the speaker’s voice. The complete analysis flowchart is shown in Fig. 4. It includes two parts: feature extraction and classifier. To simplify the data of the EMD for speech signal, the equation and feature component are used. Fig. 5 demonstrates a male and a female speaker’s speech signal diagrams in the time domain. However, using only this signal makes it almost impossible to recognize the speakers. To further analyze the crude signals, the EMD method decomposed these signals into many components. This method shows the varied modes effectively. Fig. 6 shows the results of decomposition in a female speaker condition at instance 3. In the figure, the imf1 to imf6 represent the different frequencies from high to low. The residue displays useless information. Accordingly, an energy equation is used to sum the total energies of every IMF component as one of features. This method can decrease the dimensions of the features. This equation is designated as (Zhang & Chen, 2008)

\[
\text{energy} = \sum_{i=1}^{n} \log (s_i^2)
\]

where \(s_i\) means the unit of the IMF component. \(n\) represents the total samples.

4.3. Classification and results

After feature extraction, both the BPNN and the GRNN are implemented and compared in the proposed experimental system. In this experiment, each condition employs 50 data for training the networks and 30 data for testing the network in this classification work. The identification rate is defined as

\[
\text{identification rate} = \frac{\text{number of correct classified samples}}{\text{number of total testing samples}} \times 100\%
\]

<table>
<thead>
<tr>
<th>Instance</th>
<th>Chinese phonation</th>
<th>English equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Wo Shih Wang Siao Ming</td>
<td>I’m Wang Siao Ming</td>
</tr>
<tr>
<td>2</td>
<td>Cing Kai Men</td>
<td>Open the door</td>
</tr>
<tr>
<td>3</td>
<td>San Loui Sih Ba</td>
<td>3 6 4 8</td>
</tr>
<tr>
<td>4</td>
<td>Fa Dong Yin Cing</td>
<td>Start the engine</td>
</tr>
<tr>
<td>5</td>
<td>Fang Dao Ci Dong</td>
<td>Enable security system</td>
</tr>
</tbody>
</table>

Table 1
Speech data of experiment.

**Fig. 4.** Block diagram of speaker identification system.

**Fig. 5.** Speech signal diagrams in time domain: (a) male and (b) female.
The performance of the identification rate with the number of training data using the BPNN and the GRNN are summarized in Table 2. It shows the classification accuracy of GRNN (89%) is better than the performance of BPNN (78%). The GRNN rapidly performs the training procedure, as mentioned in Section 3.2. The smoothing parameter is an important parameter and is decisive in whether the GRNN is able to smoothly estimate the data. Such an illumination is only directed at one situation where the smoothing parameter is determined. On the weight values of GRNN are regularly, they are determined by the vector of training data. However, the random weight values in the BPNN system, the results of BPNN are not exactly the same each time. Accordingly, the GRNN does not require much time in the training procedure. The experimental results show the EMD method with GRNN can be effectively used in the speaker identification system.

### Table 2

Performance of identification rate with number of training data using BPNN and GRNN.

<table>
<thead>
<tr>
<th>Instance</th>
<th>Number of training data</th>
<th>BPNN</th>
<th>GRNN</th>
<th>BPNN</th>
<th>GRNN</th>
<th>BPNN</th>
<th>GRNN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>71.51</td>
<td>81.25</td>
<td>73.29</td>
<td>87.1</td>
<td>78.38</td>
<td>89.89</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>71.51</td>
<td>83.79</td>
<td>75.37</td>
<td>85.71</td>
<td>77.13</td>
<td>87.47</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>72.14</td>
<td>85.59</td>
<td>76.15</td>
<td>87.76</td>
<td>77.65</td>
<td>88.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>73.02</td>
<td>83.34</td>
<td>75.74</td>
<td>86.03</td>
<td>79.37</td>
<td>87.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>74.82</td>
<td>85.91</td>
<td>87.59</td>
<td>87.25</td>
<td>90.39</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average (%)</td>
<td>72.6</td>
<td>83.98</td>
<td>75.01</td>
<td>86.84</td>
<td>78.16</td>
<td>88.82</td>
</tr>
</tbody>
</table>

The performance of the identification rate with the number of training data using the BPNN and the GRNN are summarized in Table 2. It shows the classification accuracy of GRNN (89%) is better than the performance of BPNN (78%). The GRNN rapidly performs the training procedure, as mentioned in Section 3.2. The smoothing parameter is an important parameter and is decisive in whether the GRNN is able to smoothly estimate the data. Such an illumination is only directed at one situation where the smoothing parameter is determined. On the weight values of GRNN are regularly, they are determined by the vector of training data. However, the random weight values in the BPNN system, the results of BPNN are not exactly the same each time. Accordingly, the GRNN does not require much time in the training procedure. The experimental results show the EMD method with GRNN can be effectively used in the speaker identification system.

### 5. Conclusions

In this paper, a speaker identification system based EMD for feature extraction and classification using artificial neural network has been developed. The experimental results indicated the proposed system is an effective recognition method for speaker identification. The EMD can effectively sift the riding wave from every complex signal of the time series. The sifted IMF components represented important information in the entire signal set. The energy correlates closely with every IMF component. The proposed procedure of feature extraction reduced the dimension of feature more effectively. In the identification system, BPNN and GRNN were used as the classifiers. As a comparison, GRNN is much faster, also, GRNN performed well in the recognition rate. The experimental results revealed the proposed EMD method with GRNN can accomplish speaker identification in a short time and achieve a satisfactory recognition rate.

### Acknowledgement

The study was supported by the National Science Council of Taiwan, Republic of China, under Project No. NSC-97-2221-E-018-008.

### References


