Journal of mathematics and computer science 10 (2014), 61-71

Contents list available at JMCS

Journal of Mathematics and Computer Science

Journal Homepage: www.tjmcs.com

Speaker Identification by Comparison of Smart Methods

Ali Mahdavi Meimand
Department of Electrical Engineering, Sirjan Branch Islamic Azad University, Sirjan, Iran
ali.mahdavy.8@gmail.com

Amin Asadi
Department of Computer Engineering, Sirjan Branch Islamic Azad University, Sirjan, Iran
asadi.univ@yahoo.com

Majid Mohamadi
Department of Electrical Engineering, Shahid Bahonar University of Kerman
mj.Mohamadi@yahoo.com

Article history:
Received January 2014
Accepted March 2014
Available online March 2014

Abstract

Voice recognition or speaker identification is a topic in artificial intelligence and computer science that aims to identify a person based on his voice. Speaker identification is a scientific field with numerous applications in various fields including security, espionage, etc. There are various analyses to identify the speaker in which some characteristics of an audio signal are extracted and these characteristics and a classification method are used to identify the specified speaker among many other speakers. The errors in the results of these analyzes are inevitable; however, researchers have been trying to minimize the error by modifying the previous analyzes or by providing new analyzes. This study uses the modification of group delay function analysis for the first time to identify the speaker. The results obtained by this method, in comparison with the group delay function method, approve the capabilities of the proposed method.

Keywords: Speaker identification, MFCC analysis, MODGDF analysis, Auto parameters.

1. Introduction

Automatic speaker identification was introduced early 1960's as a research field in the world and researching on these systems and implementing them was maximized at 1990s. In Iran, some activities have also begun in this field since 1990's. Recently, many major companies such as IBM and Microsoft have been invested on identification systems and gained very good results. One of the cell

phone service providers in France has launched a voice portal to provide news and sports competitions results for the subscribers through the speaker identification systems. Considering the developments, it seems that in the not too distant future, speaker identification technology will be a part of our personal and professional life. It has been for a long time that various IDs are used to identify individuals. The most common IDs include national ID number, first and last name. The major drawback of these identifiers is the possibility of loss and forgery[1]. It undermines the security of identifiers and leads scientists to biometric identifiers such as fingerprints and facial and voice characteristics. In fact, the characteristics of individuals' voice are used to recognize them. Individuals' voice patterns are based on two factors, the first factor is the structure of the vocal organs, i.e. the size and the shape of throat, mouth and vocal tract characteristics; the second is the learned behavior patterns such as education, social status and the style of the speech [2,3].

To identify the speaker, the system determines whether the speaker is a particular person or among a group of persons. Speaker identification is often used in hidden systems with no known users. In this paper, after noise reduction and windowing the signal, using the mentioned analysis, the number of coefficients which depends on the number of filters is extracted.

2. Preprocessing

In the beginning of the procedure, a noise reduction step should be applied on the signal which is known as preprocessing. This is done by multiplying the signal with a first-grade filter Where the z transformation and formula in the time domain is as follows:

$$Y'(n) = y(n) - \alpha y(n-1) \quad (1)$$

Where α is considered equal to 0.9 to 0.99 [4,5].

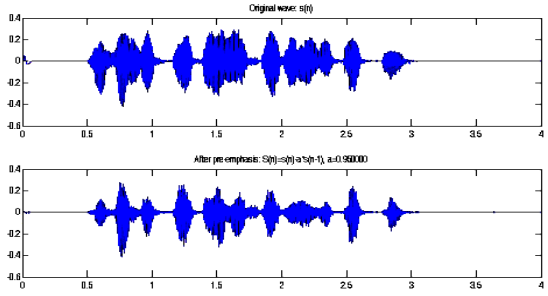

Figure 1. Preprocessing

3. Signal Windowing

The excitation function of the larynx filter for vowels is as an impulse train repeated every 2.5 ms. Therefore, we can say that the audio signals cannot be fully analyzed and to extract the characteristics of each speaker larynx filter, it must be analyzed in smaller frames and this is because the larynx filter is excited every 2.5 ms, and every 2.5 ms, the signal has specific characteristics of the filter[6].

4. MFCC Analysis

In researches conducted in the field of audio signal, the scientists found that in an audio signal, the more effective information is available at low frequencies; therefore, it can be concluded that to obtain more useful information from the signal, we should emphasize on this part of the signal. This idea leads to a method called MFCC which will be discussed at the following. The MFCC method shown in Figure 2 acts as follows: first, the size of FFT frames is calculated, and then a filter bank called Mel is used to derive the number of coefficients which depends on the number of filters. The filter bank which will be discussed, the emphasis on low frequencies is applied [7,8].

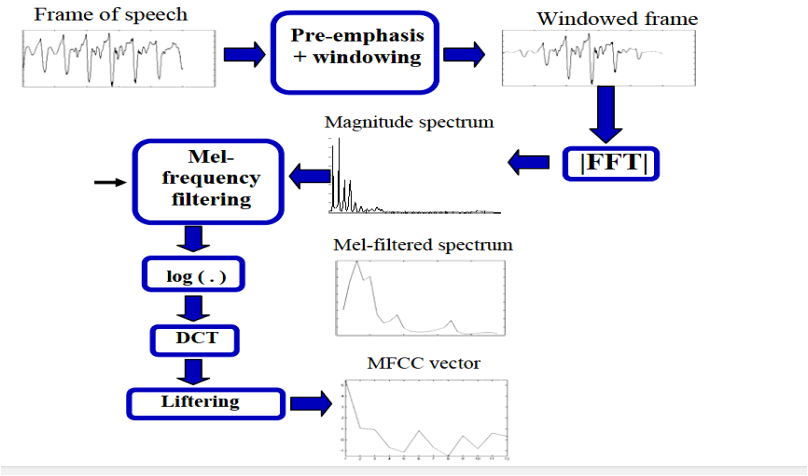

Figure 2. Diagram of MFCC Analysis

5. GDF Analysis

GDF is a negative derivative of the Fourier transform phase. Mathematically, the GDF is calculated according to the following formula:

$$GDF(w) = -\theta'(w) \tag{2}$$

The Fourier phase is correlated with the Fourier amplitude; therefore, using the following equation and formula 2, we can calculate GDF directly from the signal[9,10]:

$$GDF(w) = \frac{X_R(w).Y_R(w) + X_1(w).Y_1(w)}{|X(w)|^2} \tag{3}$$

5.1. Superiority of Group Delay Function

This function has a very important property which makes it superior to other analyzes and it is a very high resolution.

5.2. High Resolution

Group delay function has a high capability for accurate decomposition. To demonstrate this property, a tri-polar filter as shown in the Figure is considered as a hypothetical larynx filter whose poles are very close together. Then, according to the vowels formation mechanism, an impulse function is applied to the input and the output signal is considered as an audio signal [11,12].

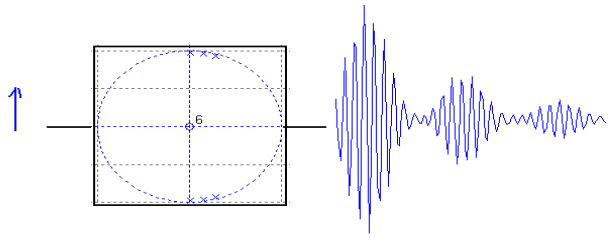

Figure 3. Tri-polar filter whose poles are very close together

6. Size Reduction using DCT

In this method, which is used in most update studies which are using MFCC and GDF analyses, at first, the DCT of the frame is calculated using the following relation[13]:

$$c(n) = \sum_{k=0}^{k=Nf} F(k)\cos\left(n(2k+1)\,\pi/_{Nf}\right) \qquad \text{n= [0, NF]} \qquad (4)$$

In the above formula, f(k) is the component of the frame, N_f is the frame length, k represents the k^{th} component of the frame. Then, 18 first coefficients are selected as the representative of the entire frame.

7. Calculation of Auto1 Parameter

To calculate Auto1 using the following one-dimensional correlation function, twenty correlation coefficients between a frames with the next frame must be derived:

$$RF(a) = \sum_{-\infty}^{+\infty} F_i[x].F_{I+1}[x+a]' \qquad (5)$$

In addition, to calculate this parameter for the last frame we should use the first frame since there is no other frame.

7.1. Calculation of Auto2 Parameter

To calculate Auto2, at first we form a matrix including the first and the next 16 frames. Then, using the following correlation formula while considering b=0 and changing from 0 to 17, 18 correlation coefficients are derive from the matrix and is called Auto2.

$$R(a,b)=\sum_{-\infty}^{+\infty}\sum_{-\infty}^{+\infty}F(x,y).F(x+a,y+b) \tag{6}$$

After this step, frame number 2 and 16 next frames are considered in a matrix and 18 coefficients are derived as before. This step is repeated for all frames and 18 coefficients are derived for each frame.

8. Modeling using Multi-Layer Perceptron Neural Networks

The objective of this study is to compare several speaker identification methods in the same condition, and this is preferably done by a multi-layer back propagation neural network [14].

8.1. Neural Networks in Speaker Identification

When using a neural network, several parameters have to be determined as the following:

1. Number of layers

If a network has three layers, it will be possible to solve any problem with any degree of complexity.

2. Neurons in each layer

It is possible any number of neurons to be available at the input and the hidden layers and they are selected using different criteria. Large numbers of neurons in these layers increases the computational size and the few numbers of neurons in this layer lowers the accuracy of the network. At first, the number of neurons in the hidden layer is considered as a fraction of the number of inputs, and then the problem is simulated. If you did not achieve a good coverage and generalization power, the number of neurons in the hidden layer will be increased by 1 and the simulation is repeated again. This must be continued until an appropriate convergence and generalization power is achieved.

In this project we consider it equal to 15.

The number of neurons in the first layer is considered as 5 using trial and error approaches.

The number of neurons in the outer layer should be equal to the number of speakers who must be identified (18 layers).

3. Number of inputs

The number of inputs must be equal to the size of the feature vector.

4. The function used in each layer

Usually, the function of neurons in hidden and the first layers is a *tansig* function and in the last layer is *logsig*. The *logsig* function is used for the last layer because we want the outputs to be between 0

and 1 during the test to attribute a probability between 0 and 1 to each row which indicates the probability of each speaker.

8.2. Network training

Network training includes two steps:

1. To create a carrier matrix of feature vectors

   The matrix consists of *n* rows and *d* columns, each column containing one observation or in other words one input frame and each row containing different aspects of the input.

2. To create the desired output matrix (t)

   This matrix contains the desired outcomes of the network and the number of the columns is equal to the carrier matrix of the feature vector. The data in each column indicates that to which speaker the corresponding feature vector belongs.

   Meanwhile, the number of rows equals to the number of outputs (speakers). If for example a column corresponds to the first speaker, the first row is equal to 1 and the others rows are 0.

   Similarly, if a column corresponds to the second speaker, the second row is equal to 1 and other are. Continuing this, t is formed. In this case, the network learns that when the input corresponds to a certain speaker, the related row will be equal to 1.

8.3. Testing the Network

To test the network, first, a sample of the voice of a certain speaker must be tested is divided to frames and each frame is separately used to derive the specifications. Then, the feature vectors corresponding to each speaker are applied to the input of the network. The output of the network is a probability between zero and one for each speaker.

This is repeated for all frames and the number of derived probabilities will be equal to the number of frames in the signal. Then the averages of the obtained probabilities are calculated and the maximum average indicates that the voice belongs to that speaker.

9. Data Base Specifications

This study uses TIMIT database which contains 10 terms for each speaker two first of which is identical for each speaker and other terms vary for the speakers. For simulation, we used 18 speakers and we have 10 terms for each speaker 70% of which is used for training and 30% for network testing[15].

10. Text-Independent Simulation Approach

In this study, text-independent approach was used in which the network is trained by a set of words and is tested by other series that is not related to training data. We used this approach due to the database utilized in this study. The data base contains10 terms for each speaker and different terms have no connection with each other. These two methods are obviously different in identification percentages and text-dependent has identification percentages much higher than text-independent.

11. Simulation for Comparing MODGDF and MFCC

First, all training data are calculated as the following:

A) Noise reduction using a first-grade filter is applied on all vocal samples from each speaker.

B) The signal are divided to frames with length=20ms and a frame shift=10ms.

C) The FFT of each frame and its size are calculated.

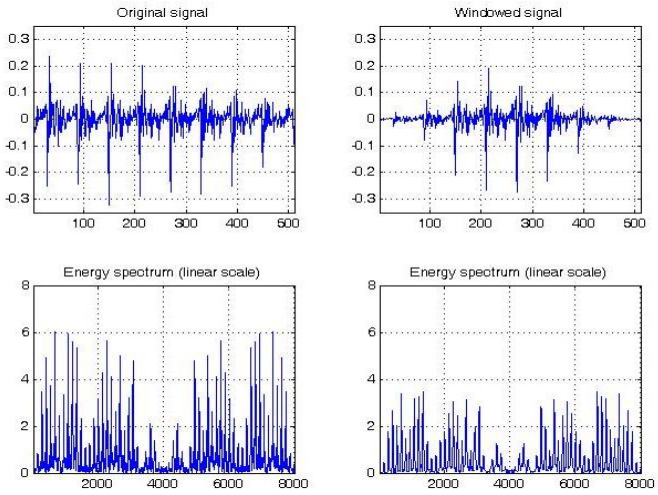

Figure 4. The FFT of Desired frame

D) The filter bank is constructed using 43 filters.

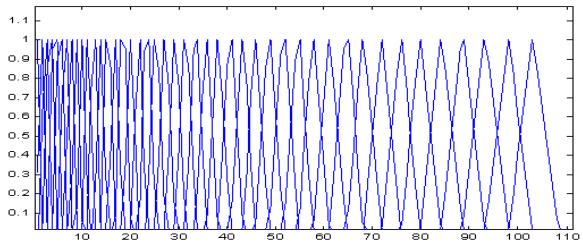

Figure 5. Mell filter bank

E) Each frame is multiplied by each filter of the filter bank and then the average energy is calculated.

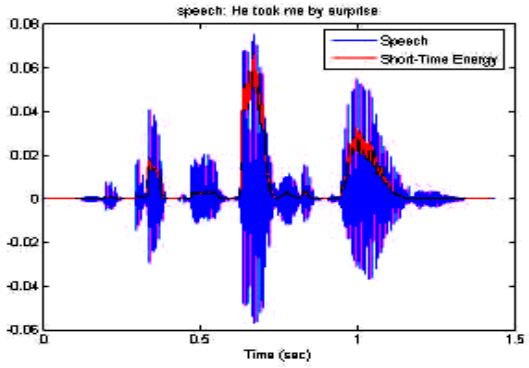

Figure 5. Average energy

F) 43 coefficients equal to 43 filter banks are extracted from each frame.

G) The logarithm of the obtained coefficients is calculated.

H) DCT of the obtained coefficients is calculated and the first 18 coefficients are derived.

Then, using these data, a back propagation neural network with a size of [18, 15, 5] is trained. After this step, using the same steps described above, MFCC is calculated for the test data. Then each feature vector obtained from test data which belongs to a certain frame is applied to the neural network input and the output is a probability for each frame. Finally, the probabilities obtained for each frame of test data are averaged, and the test data is attributed to the speaker with the maximum average probability.

Finally, the result of the simulation was equal to 78.45% .

12. Then the training data is calculated using MODGDF method

First, all vocal samples from each speaker are de-noised using a first-degree filter.

Then the signal is divided to frames with a frame length = 20ms and a frame shift = 10ms.

The FFT of the windowed signal x[n] is calculated and called X (k). The FFT of nX [n] is also calculated and called Y[k].

The spectrum S (ω) is calculated using Cepstrum technique and considering Lifterw = 5.

Then MODGDF is formed.

DCT of the obtained coefficients is calculated and the first 18 coefficients are derived.

Then, using these data, a back propagation neural network with a size of [18, 15, 5] is trained. After this step, using the same steps described above, MODGDF is calculated for the test data. Then each feature vector obtained from test data which belongs to a certain frame is applied to the neural network input and the output is a probability for each frame. Finally, the probabilities obtained for each frame of test data are averaged, and the test data is attributed to the speaker with the maximum average probability.

Finally, the result of the simulation was equal to 89.56% .

"Table1 . ComparingMODGDF to MFCC "

Type of Analysis	Type Size of Feature Vector	Type of Neural Network	The size of the Neural Network	Learning Algorithm	Of pattern recognition
MODGDF	12	Feed forward back propagation	[5,15,18]	LM	89.56%
MFCC	12	Feed forward back propagation	[5,15,18]	LM	78.45%

13. MODGDF Simulation using Auto Parameters

In the previous section, it was proved that MODGDF works much better than MFCC. After this step, we intend to compare *Auto* parameter, which was proposed in this study, with other parameters using MODGDF analysis.

The parameter MODGDF is calculated as discussed in the previous section (with no size reduction). In this case, 18 coefficients are obtained for each frame. Then the neural network is trained and tested as discussed in the previous section.

In this case, the simulation result was 89.56% .

In the next step, Auto1 is calculated using the analyzed signal of various frames as discussed previously and 18 coefficients was derived. Then the neural network is trained and tested as discussed in the previous section.

In this case, the simulation result was equal to 75.27%, which not only had no improvement, but the result was even worse. In the next step, Auto2 is calculated using the analyzed signal of various frames as discussed previously and 18 coefficients was derived. Then the neural network is trained and tested as discussed in the previous section.

In this case, the simulation result was equal to 92.34 % which indicates a performance better than the previous ones.

"Table 2. MODGDF Simulation using Auto Parameters "

Type of Analysis	Type Size of Feature Vector	Type of Neural Network	The size of the Neural Network	Learning Algorithm	Of pattern recognition
MODGDF	20	Feed forward back propagation	[18,5,15,18]	LM	89.56%
Auto1	20	Feed forward back propagation	[18,5,15,18]	LM	75.27%
Auto1	20	Feed forward back propagation	[18,5,15,18]	LM	92.34%

14. Conclusions

Unlike the analyses that have been used to identify the speaker, GDF analysis uses the angle of Fourier transform rather than the size and according to the modifications applied on GDF, it is known as MODGD analysis. Through the modifications applied on the group delay function analysis, a new better approach was developed for speaker identification in comparison with group delay function. MFCC analysis emphasizes on low frequencies and when comparing MFCC and MODGDF methods, as observed, MODGDF has a performance much better than MFCC. Then MODGDF analysis was compared to Auto1 and Auto2 (according to the previous comparison) and the results indicate that in comparison with Auto1, it not only did not improve the results, but also the results were worse; but better results were obtained in comparison with Auto2.

REFRENCE

[1] Richard Duncan, Mississippi State University, A Description And Comparison Of The Feature Sets Used In Speech Processing Ph (601) 325-3149 - Fax (601) 325-3149.

[2] Tomi Kinnunen "Spectral Features for Automatic Text-IndependentSpeaker Recognition".LICENTIATE'STHESIS University of Joensuu Department of Computer Science P.O. Box 111, FIN-80101 Joensuu, Finland. December 21,2003.

[3] Rangsit Campus & Klongluang & Pathum-thani "Voice Articulator for Thai Speaker

Recognition" Thammasat Int. J. Sc. Tech., Vol.6, No.3,September-December2001.

[4] Antanas LIPEIKA, Joana LIPEIKIEN ÿ E, Laimutis TELKSNYS."Development of Isolated Word Speech Recognition System".September 2001.

[5] Rangsit Campus & Klongluang & Pathum-thani "Voice Articulator for Thai Speaker Recognition" Thammasat Int. J. Sc. Tech., Vol.6, No.3, September-December 2001.

[6] Tomi Kinnunen a,*, Haizhou Li b 'An overview of text-independent speaker recognition: From features to supervectors' Speech Communication 52 (2010) 12–40.

[7] Richard Petersens Plads. "Mel Frequency Cepstral Coefficients: An Evaluation of Robustness of MP3 Encoded Music". Informatics and Mathematical Modeling Technical University of Denmark Richard Petersens Plads - Building 321 DK-2800 Kgs. Lyngby - Denmark2002.

[8] Hat Yai, Songkhla " MODIFIED MEL-FREQUENCY CEPSTRUM COEFFICIENT". Department of Computer Engineering Faculty of Engineering Prince of Songkhla University Hat Yai, Songkhla Thailand, 90112.

[9] Ramya & Rajesh M Hegde & Hema A Murthy. "Significance of Group Delay based Acoustic Features in the Linguistic Search Space for Robust Speech Recognition" Indian Institute of Technology Madras, Chennai, India. Department of Electrical Engineering, Indian Institute of Technology Kanpur, Kanpur, India 2008.

[10] Rajesh M. Hegde, Hema "Significance of the Modified Group Delay Feature in Speech Recognition". 2007.

[11] Rajesh M. Hegde, Hema "Significance of the Modified Group Delay Feature in Speech Recognition"2007.

[12] C.F. Chen, L.S. Shieh, A Novel Approach to Linear Model Simplification, International Journal of Control. 8 (1968) 561 – 570.

[13] G. Parmer, R. Prasad, S. Mukherjee, Order Reduction of Linear Dynamic Systems using Stability Equation Method and GA, World Academy of Science, Engineering and Technology. 26 (2007) 72 - 78.

[14] Adjoudj Réda & Boukelif Aoued. "Artificial Neural Network & Mel-Frequency Cepstrum Coefficients-Based Speaker Recognition". Evolutionary Engineering and Distributed Information Systems Laboratory,EEDIS, Computer Science Department, University of Sidi Bel-Abbès, Algeria March 27-31, 2005.

[15] Julien Neel."Cluster analysis methods for speech recognition" Department of Speech,

Music and Hearing Royal Institute of Technology S-100 44 Stockholm.2005-02-18.