Optimal MFCC Features Extraction by Differential Evolution Algorithm for Speaker Recognition

Mohsen Sadeghi
Electrical Engineering Department
Shahrood University of Technology
Shahrood, Semnan, Iran
Mohsen.sadeghi@shahroodut.ac.ir

Hossein Marvi
Electrical Engineering Department
Shahrood University of Technology
Shahrood, Semnan, Iran
h.marvi@shahroodut.ac.ir

***Abstract*— Speech is the most commonly and widely used form of communication and interaction between humans. The interfacing system, which is an automatic speaker recognition system, requires modeling to receive input data in the form of a feature with a minimum number and learn through this data. The purpose of this paper is to extract the optimal number of Mel-Frequency Cepstral Coefficients (MFCC) features without reducing the recognition accuracy for speaker recognition application. For this purpose, an algorithm has been proposed in which the Differential Evolution (EA) optimizer and also the probabilistic neural network (PNN) classifier are used to achieve this goal. After implementing this algorithm in MATLAB software, it was observed that the number of MFCC features, which so far had at least 13 for each frame, was reduced to 5 per frame, without any recognition accuracy being reduced.**

Keywords—MFCC; PNN; EA; Speaker Recognition; Feature Extraction

I. INTRODUCTION

The most important and most practical way to communicate and interact with each other is speech. In addition, speech is the most appropriate way to communicate and interact with human beings [1]. Speech processing involves various technologies that include speech coding, speech synthesis, speech recognition. Speech processing consists of two basic parts: extraction and classification. The most important criterion for using a speech processing system as a good speech processing system is how the feature extraction is used in this processing system. Because of feature extraction technique plays an important role in the accuracy of the speech processing system [2]. Since the signal of speech is intrinsically nonstationary; therefore, the speaker's recognition is considered as a difficult task due to gender differences, emotional state, accent, pronunciation, interpretation, speech, pitch, sound level, and speed variation in individuals. Development in speech technology stems from the fact that man tends to simulate mechanical models to create verbal communication. Speech processing allows the computer to follow and obey the various voice commands and the human languages [3]. The scope of recognition of the speaker have two main areas that are: the speaker recognition and the speaker verification. In the area of speech recognition, the goal is to the system recognizes that who is speaking in a collection of familiar sounds and speakers. For this reason, the recognition system of the speaker is also known in the limited and closed collection. But the second task of the speaker recognition system is the speaker verification. In this system, it must be verified whether the person who claims to be the person of x, is this claim valid. In fact, this system is a decision-making system, it is necessary to system determines the claim of a lie claiming to be a person of x from the correct claim [4]. In *Fig. 1*, the components of an automatic speaker system are displayed [5]. An important application of the speaker recognition system is the use of speaker recognition in forensics, which has been highly sought in recent years. Because most of the information exchanged between the two groups is done through telephone calls, these groups can be criminals. For this reason, there has been a great desire to use the automatic speaker recognition system and semi-automatic analysis methods in recent years [6-10].

The remainder of this paper is divided into the following sections: In Section 2, we will examine the method used to extract the feature of the speech signal. Section 3 provides a brief overview of the PNN structure and how it is categorized. Section 4 discusses the DE optimization algorithm, and how to derive the number of optimal MFCC features or use of this algorithm. In Section 5, the proposed algorithm is described in order to extract the optimal number of MFCC features for speaker recognition application. In Section 6, the experimental results of the proposed algorithm are presented. Finally, in Section 7, an overall conclusion is made about the proposed algorithm.

II. FEATURE EXTRACTION FROM SPEECH SIGNAL BY MFCC METHOD

The feature extraction of the speech signal means that the speech signal is converted into a series of feature vector coefficients. These features only include the information needed to identify the received speeches. Features extracted from the speech signal must meet certain criteria, the most important of these criteria are: The extracted features should be easily measurable, the extracted features should be in accordance with the time and Finally, the extracted features must be resistant against different noises and environmental conditions [11].

Feature vectors of speech signals typically use spectral analysis techniques, which are: MFCC, LPC, wavelet

transformations, etc. In this paper, the MFCC feature extraction method is used to extract the feature.

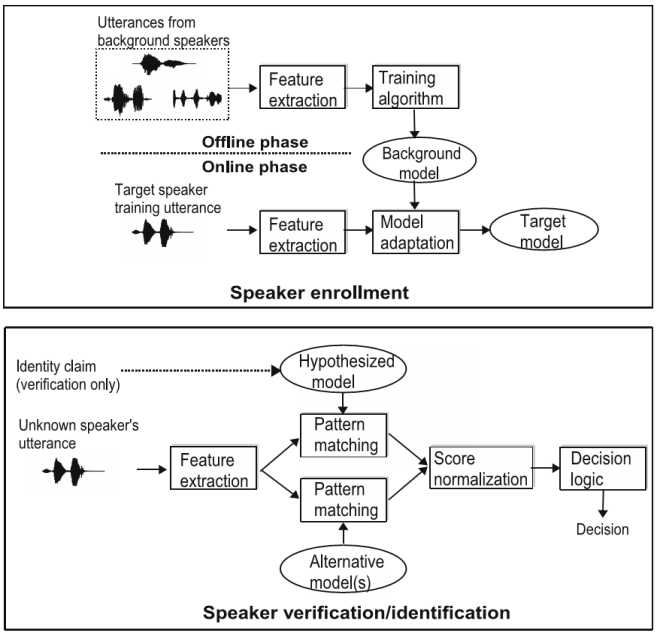

Fig. 1. The components of an automatic speaker recognition system [5]

The human sound features can be extracted by the MFCC method and the features extracted by this method show the short-time power spectrum of human sound. Frequency bands in this method are equal on the Mel scale. Because it is an exact approximation of human voice. For the conversion of the normal frequency f to the Mel scale, the relation (1) is used [2].

$$m = 2595 \log_{10}(1 + \frac{f}{100}) \quad (1)$$

In this method, Delta-Cepstrum is used to achieve changes between different frames. Delta's advantage is to have time-varying (energy + MFCC) as new features which represents the velocity and acceleration of (energy + MFCC) [2].

$$\Delta C_m(t) = \frac{[\sum \pi = -M^M C_m(t+\pi)\pi]}{[\sum \pi = -M^M \pi^2]} \quad (2)$$

The next step is to apply DCT on the energy logic (E_k) that are obtained through triangular band-pass filters in order to obtain the L cepstral coefficient of the Mel scale. The DCT formula is (3) [2]:

$$C_m = \sum_k 1^N \cos\left[m \times (k - 0 \cdot 5) \times \frac{\pi}{N}\right] \times E_k \quad (3)$$

$$m = 1.2 \ldots . \mathrm{L}$$

In the above relation, N is the number of triangular band-pass filters, L is the number of cepstral coefficients of the Mel scale. The value of M is usually considered to be 2. The steps for MFCC feature extraction algorithm are shown in *Fig. 2* [2].

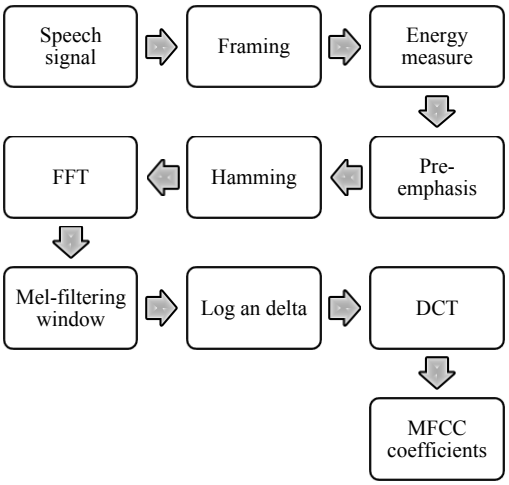

Fig. 2. MFCC feature extraction algorithm steps [2]

III. PROBABILISTIC NEURAL NETWORK (PNN)

Probabilistic neural network was introduced by Specht in 1990. PNN was defined as an implementation of Kernel discriminate analysis (one kind of statistical), in which case operations are organized as a multilayer direct network with 4 layers. These layers are described as input layer, pattern layer, summation layer and output layer. One of the basic architectures of PNN is shown in *Fig. 3* [13].

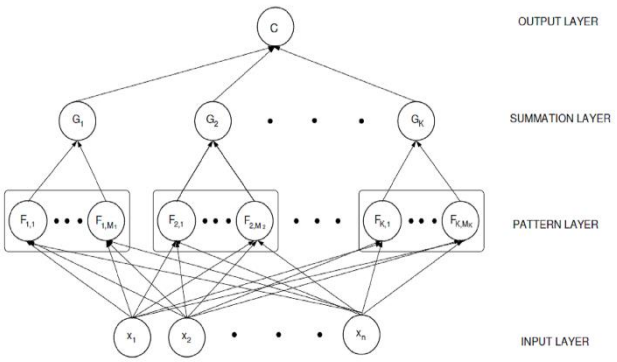

Fig. 3. Base Architecture of PNN [13]

Probability neural networks have successfully used for various pattern recognition applications. Some of these applications can be mentioned as images recognition, texture recognition, signal processing, financial and medical applications [14], [15]. PNN is a classifier pattern, in which very applicable Bayes strategy of decision making with Parson nonparametric estimator [16] has been used for estimating of probability. In contrary to other neural networks structures, PNN can be implemented and interpreted by designing [17]. One of the vital benefits of PNN is its fast learning process, which is intrinsically a parallel structure and training samples can be added or subtracted without requiring extensive retraining. According to these facts and benefits, it can be decisively expressed that PNN can operate impressively as a neural network with monitoring over recognition of speaker and speech [13].

Finally, In PNN, first layer is the location of extractive features as the input of neural network. The second layer, which is the pattern layer, is contained of Gaussian functions, in which given set of points are used as centers. By assuming that X and W_i are normalized by one length unit, the output of pattern unit can be represented as equation 3 [13].

$$f(X.W_i) = \exp[-\frac{(X-W_i)^T(X-W_i)}{2\sigma^2}] \quad (3)$$

In abovementioned equation, i, σ, W, X respectively indicates number of patterns, parameter smoothing, weight array and input array. In this structure, every set has a summation unit related to itself and every summation unit is only connected to the pattern units of its set. Summation units, sum up the output of Kernel functions of pattern units which are linked to itself. Output layer is actually a decision maker layer that selects the biggest amount according to the amounts of output of previous layer, which is binary, and then determines the sticker of the respective class.

IV. Differential Evolution Algorithm(DE)

Evolution algorithms are generally considered as algorithms that can be effective in all of optimization problems. These algorithms are able to reach close answers in real problems and mathematics [18]. The reason of the high acceptability of evolution optimization methods compared to classic and analytical methods is that the classical and analytical methods need more computational time for optimizing in many of problems. Evolution algorithms have various types. One of these algorithms, which has been represented recently, is deferential evolution algorithm. The main reason of representing this algorithm can be known as overcoming the main flaw of genetic algorithm which is lack of local search. Order of applying mutation operators, and recombining, and also the way that selection operator acts, can be known as the main differences between genetic algorithm and DE algorithm. The process of implementation of this algorithm is as shown in *Fig. 4*.

DE algorithm benefits from using one differential operator for producing new population. This operator yields in information exchange between members of population. One of the advantages of this algorithm is having memory that preserves the information of appropriate answers in current population. In this algorithm, all of the members of the population have the equal chance for being selected as a parent. In the way that the generation of new-borns are compared to the generation of parent by competency, which is evaluated with fitness function. Then best members enter the next step as subsequent generation. The most essential feature of DE algorithm is high velocity, simplicity and high potency. This method starts only with adjusting three parameters. These three parameters are NP, F and Cr, which are representing the size of population, mutation weight and the probability of performing crossover, respectively, that are multiplied in the difference of two vectors. Usually, the amount of parameter F is considered between 0 and 2, and the amount of parameter Cr is considered between 0 and 1 [18].

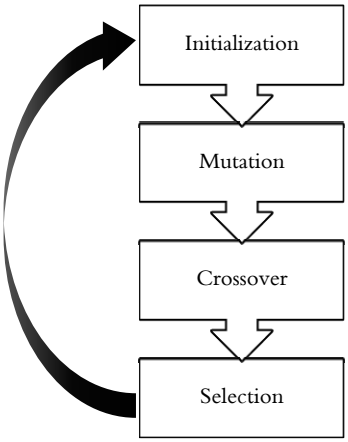

Fig. 4. Implementation process of DE algorithm [18]

V. The proposed algorithm for extracting optimal number of MFCC feature

For automatic speaker recognition system, the first step is to extract features from databases, which are the voices of different people. However, there are various methods for extracting these features which are described in the following. In this paper MFCC technique is used for feature extracting. The instruction of this technique is illustrated in *Fig. 2*. This technique was selected due to its acceptability and its vast application in feature extracting compared to other features. The reasons why this method of feature extracting is one of the most applicable methods can be probed over four factors: creating a good distinction, low affinity between coefficients, as it is not based on linear features, it resembles the system of hearing perception of human; finally, required phonetic features can be collected [12].

The process of the algorithm that used to extract the optimal number of MFCC features without damaging the accuracy of the speaker recognition system is shown in *Fig. 5*. Scholars in the area of signal processing and speech in recognition systems are always looking for a feature that can have the most accurate recognition with minimum length. In other words, with minimum length of feature, desired signal, which can be wave, sound, speech or even image, is able to be modelled. For this purpose, in this paper, we attempt to find out minimum and optimal number of MFCC features for speaker recognition without decrease in recognition accuracy. Minimum number of features considered for MFCC feature in various references and different papers is number of 13 per each frame. In this paper, we make effort to find out a less number for the number of MFCC features by using DE algorithm as an optimizer and PNN as a classifier without having decrease in recognition accuracy.

As shown in *Fig. 5*, the proposed algorithm process in this paper is aimed at achieving at least the optimal number of features of the MFCC in two directions, namely, two direct paths. One is to train the network and extract the minimum number of features and paths, and the other to test the number of extracted feature. The function of this algorithm is that, first, the data of the training which are here speakers is entered into the system, Then, after an initial population, the number of MFCC

features is considered, the number of these attributes is placed in the Local Solution block. Then the number of features defined in the previous block is extracted from the input signal for each frame of the MFCC feature. Then these features are classified into the PNN classifier. Then, depending on the objective function proposed for differential evolution optimization algorithm, the optimization and extraction process starts with the optimal number of features and this trend continues until the end of the number of repetitions defined for the optimization algorithm. The next step is the test phase. Test data is entered, and by the number of features, the number of these features is obtained for each frame in training step, it is extracted from the test signals. And then, by the PNN classifier, the accuracy of the recognition is obtained which indicates whether the number of features obtained during the optimization phase is also optimized. The objective function that is considered in this algorithm for the differential evolution optimizer is shown in (4).

$$\text{Fitness} = (100 - \text{Accuracy}) * \text{M}^2 \quad (4)$$

In the above relation M is the number of MFCC features Given the relationship that the goal is to minimize it, this objective function consists of two parameters: One of these parameters is the recognition accuracy and the other is the number of MFCC features. Researchers always seek to maximize the accuracy of recognition and minimize the number of extracted features from each frame. That is why two contradictory parameters are targeted at this function, one desirability is to be maximized and the other is desirable to be minimized. Since we seek to minimize the objective function. For this reason, using the (100-Accuracy), we set both the parameters involved in this objective function on the one hand, which is the same as minimization. By doing this, after optimization, a number of optimal features are obtained that have maximum recognition accuracy.

VI. Experimantal Results

In this paper, in order to achieve the optimal number of MFCC features, the following algorithm that it shown in *Fig. 6*, has been used. The combination of two probabilistic neural network algorithms and differential evolution optimization algorithm have been used. In the proposed algorithm, the optimizer DE is used as an optimizer and PNN as a classifier. The algorithm written for the intended purpose is written and implemented in MATLAB software. Table (1) shows the simulation results of this algorithm.

TABLE I. Results of Simulation of the Proposed Algorithm

MFCC Count	3	4	5	8	13	17
Fitness	562.5	400	312.5	800	2112.5	3612.5
Accuracy	37.5	75	87.5	87.5	87.5	87.5
Processing Time	0.3746	0.1395	0.1300	0.1289	0.1342	0.1363

According to Table (1) and *Fig. 6*, it is observed that the best and lowest value for the target function is 312.5, this objective

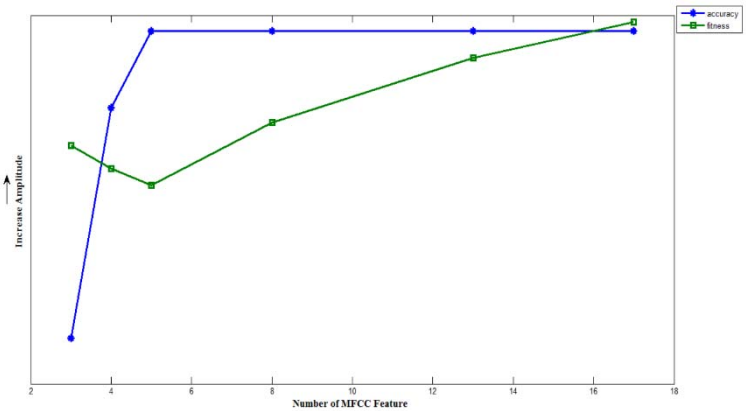

Fig. 6. Recognition Accuracy & Fitness Value Against Changing MFCC Features Count

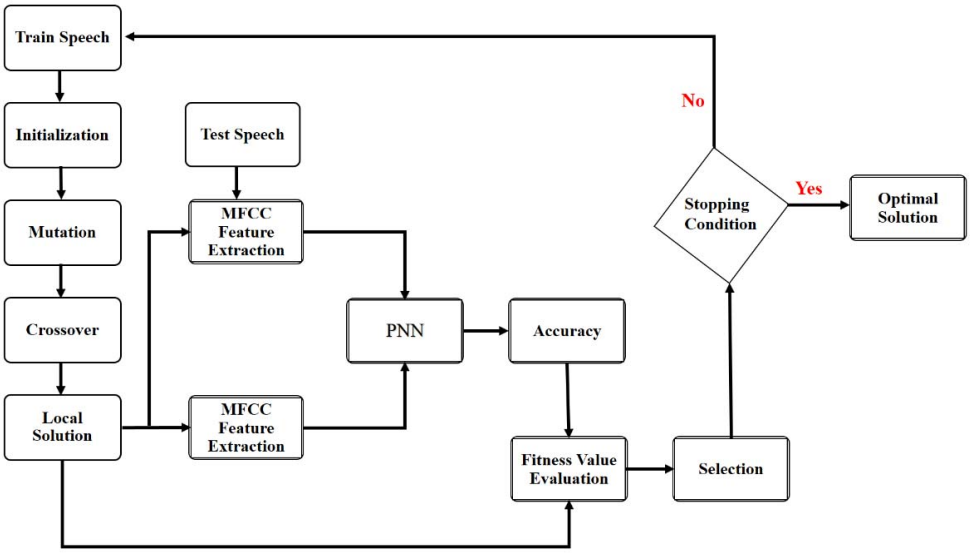

Fig. 5. The Proposed Algorithm for Extracting the Optimal Number of MFCC Features

function is obtained for the number of attributes 5 for each frame. The recognition accuracy for this number is 87.5% and it can be seen that the recognition accuracy for the 13 MFCC features is also the same. So, in this paper we were able to reduce the number of MFCC properties to 5 to reach the maximum of the accuracy of the speaker recognition which was already 13. The databases used in this article are at least 135 frames per second, the number of extracted features is 1755, but using the proposed algorithm in this paper, we were able to achieve the number of attributes of 5 numbers per frame. Now, if you want to query the optimal number of features obtained from the feature input from database, the total number of features will be 675. This means a reduction of about 1080 features, by which we were able to model the system with an accuracy of 87.5% with fewer features.

VII. Conclusion

In this paper, a new method for extracting the optimum number of MFFC feature has been proposed for modeling the speech signal for speaker recognition application. We proposed an algorithm that uses two probabilistic neural network algorithms as a classifier and differential evolution optimization algorithm as an optimizer. Experimental results suggest that the optimal number of extracted features by the MFCC method, which used to be 13 already per frame, is decreased to 5 per frame, while the accuracy of the speaker recognition system is the same. Using this algorithm, we will be able to model speech signals with fewer MFCC features.

References

[1] Yousra F., Al-Irhaim Enaam Ghanem Saeed, "Arabic word recognition using wavelet neural network", Scientific *Conference in Information Technology,* November 2010.

[2] Hibare, Rekha, and Anup Vibhute. "Feature Extraction Techniques in Speech Processing: A Survey." International Journal of Computer Applications 107.5 (2014).

[3] Bhabad, Sanjivani S., and Gajanan K. Kharate. "An Overview of Technical Progress in Speech Recognition." International Journal of advanced research in computer science and software Engineering 3.3 (2013).

[4] Reynolds, Douglas A. "An overview of automatic speaker recognition technology." Acoustics, speech, and signal processing (ICASSP), 2002 IEEE international conference on. Vol. 4. IEEE, 2002.

[5] Kinnunen, Tomi, and Haizhou Li. "An overview of text-independent speaker recognition: From features to super vectors." Speech communication 52.1 (2010): 12-40.

[6] Alexander, Anil, et al. "The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications." Forensic science international 146 (2004): S95-S99.

[7] Gonzalez-Rodriguez, Joaquin, et al. "Robust likelihood ratio estimation in Bayesian forensic speaker recognition." Eighth European Conference on Speech Communication and Technology. 2003.

[8] Niemi-Laitinen, Tuija, et al. "Applying MFCC-based automatic speaker recognition to GSM and forensic data." Proc. Second Baltic Conf. on Human Language Technologies (HLT'2005), Tallinn, Estonia. 2005.

[9] Pfister, Beat, and René Beutler. "Estimating the weight of evidence in forensic speaker verification." Eighth European Conference on Speech Communication and Technology. 2003.

[10] Thiruvaran, Tharmarajah, Eliathamby Ambikairajah, and Julien Epps. "FM features for automatic forensic speaker recognition." Interspeech. 2008.

[11] Radha, V., and C. Vimala. "A review on speech recognition challenges and approaches." doaj. org 2.1 (2012): 1-7.

[12] Madan, Akansha, and Divya Gupta. "Speech Feature Extraction and Classification: A Comparative Review." International Journal of computer applications 90.9 (2014).

[13] Li, Xin Guang, et al. "The Application of Probabilistic Neural Network in Speech Recognition Based on Partition Clustering." Applied Mechanics and Materials. Vol. 263. Trans Tech Publications, 2013.

[14] Chtioui, Younes, et al. "Comparison of multilayer perceptron and probabilistic neural networks in artificial vision. Application to the discrimination of seeds." Journal of chemometrics 11.2 (1997): 111-129.

[15] Wang, Yue, et al. "Quantification and segmentation of brain tissues from MR images: a probabilistic neural network approach." IEEE transactions on image processing 7.8 (1998): 1165-1181.

[16] Parzen, Emanuel. "On estimation of a probability density function and mode." The annals of mathematical statistics 33.3 (1962): 1065-1076.

[17] Specht, Donald F. "Probabilistic neural networks." Neural networks 3.1 (1990): 109-118.

[18] Mansouri, R., and H. Torabi. "Application of Differential Evolution (DE) Algorithm for Optimizing Water Distribution Networks (Case Study: Ismail Abad Pressurized Irrigation Network)." (2015): 81-95.