Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 61 (2013) 220 – 226

Procedia Engineering

www.elsevier.com/locate/procedia

Speaker Recognition Based on a novel hybrid algorithm

Fan-Zi Zeng , Hui Zhou*

Information Science and Engineering, Hunan University, Changsha 410081, China

Abstract

Using Probabilistic Neural Network (PNN) to recognize speaker is one of the research of branch of speaker recognition. PNN's ability of recognition is so dependent on the value of its smoothing factor that its ability of recognition is not that good. To solve this problem, we proposed a novel hybrid algorithm (DFOA-SOM-PNN) to improve PNN's ability of recognition. Firstly, it uses SOM to cluster MFCC speech characteristics parameters which can reduce storage of data and calculation, and good reflect feature of MFCC. Secondly, it uses an improved algorithm of Fruit fly Optimization Algorithm (FOA): Double group FOA (DFOA), which optimizes the smooth factor of PNN. The experimental results show that DFOA have better global convergence and fast convergence speed than FOA, and the proposed hybrid algorithm has better performance in speaker recognition.

Keywords: speaker recognition; Fruit Fly Optimization Algorithm; Double group FOA; SOM; PNN

1. Introduction

Speaker Recognition is the process of automatically verify a person's claimed identity from his voice [1].At present, speaker recognition methods mainly are Pattern Matching, Probability Statistical Model, Artificial Neural Network and so on [1]. In these methods, the mainly recognition algorithm are Dynamic Time Warping (DTW), Vector Quantization (VQ), Gaussian Mixture Models (GMM), Hidden Markov Model (HMM), Probabilistic Neural Network (PNN). As the present study hotspot, ANN is applied to the speaker recognition, still at the prospect of widely.

To some degree, ANN can imitate human cranial nerve system function. Currently, MFCC parameters are usually identified by using PNN. But its recognition performance in some cases is not so satisfied, and the data of MFCC parameters may be large, it will increase storage and calculation, and reduce recognition performance. Self-Organized Mapping (SOM) effectively solves these problems and good reflect feature of MFCC. So a kind of hybrid recognition algorithms can be proposed: SOM-PNN.

On the other hand, inappropriate PNN smooth factors value may lead to recognition accuracy rate very low, so some kind of algorithm is needed to optimize it.

Today many hybrid recognition algorithms with intelligent optimization algorithm have been widely studied. For example, Lin combined GMM with GA in speaker recognition [2]; Saeidi combined GMM with PSO in speaker recognition [3].

Fruit fly Optimization Algorithm (FOA),to compare with GA [4], PSO [5] and ACA [6], have no many function parameters , and have no complex calculation process and easy to understand, it is a very simple, robust, intelligent optimization algorithm which has fast convergence speed [7]. But it has a lot of defects, one of them is that it is easy to fall into the local extremum, so an improved algorithm of FOA : Double group FOA (DFOA) is proposed, which optimizes the smooth factor of PNN, and a mixture of supervision recognition algorithms (DFOA-SOM-PNN) can be put forward.

In this paper, Section II mainly introduces principle and algorithm procedures of DFOA. Section III introduces theoretical framework of DFOA-SOM-PNN. Section IV shows the experimental results. In the end, Section V makes a conclusion.

* Corresponding author.
E-mail address: zhouhui_1102@163.com

Selection and peer-review under responsibility of the Hunan University and National Supercomputing Center in Changsha (NSCC)
doi:10.1016/j.proeng.2013.08.007

2. The improved algorithm of FOA

2.1. Basic principle of DFOA

FOA as a kind of new evolutionary and random search optimization method which can simulate fruit flies behavior of seeking food, however, has a lot of defects: easy to fall into the local extremum; not give a general starting point; the definition of the Concentration of Taste Judgment Value(CTJV) is too simple. So an improved algorithm of FOA: DFOA is proposed.

FOA easily lead to fall into the local extremum, root cause is that the definition of the CTJV is not perfect. By designing two fruit flies groups in different direction group to seek food, and introducing direction factor and convergence factors, and determining starting point of these groups, it can quickly find the global optimal value.

The step of DFOA is as follows:

1. The size of two fruit flies groups and the number of iterations Count are initialized. Common starting point (Xinit, Yinit) is randomly initialized, and set in the [D, D] random space (D is constant).

$$X_{init} = D \cdot RandomValue$$
$$Y_{init} = D \cdot RandomValue$$

2. [-D',D'] is random search range (D' is constant). Two groups are labeled by k (k = 1, 2). The new position coordinates of each fruit flies in two groups are as follows,

$$X_i^k = X_c^k + D' \cdot RandomValue$$
$$Y_i^k = Y_c^k + D' \cdot RandomValue$$

Where the (X_c^k, Y_c^k) is the best position coordinates.

3. Origin point is as a reference point. Distance from new coordinates to the origin is calculated,

$$Dis\tan ce_i^k = \sqrt{(X_i^k)^2 + (Y_i^k)^2}$$

4. Then CTJV is calculated,

$$S_i^k = \alpha \cdot (Dis\tan ce_i^k)^\tau$$

Where the $\alpha(\alpha \neq 0)$) is the direction factor, the t is the convergence factor.

5. The concentration of taste value is calculated,

$$SC_i^k = Function(S_i^k)$$

6. According to the concentration of taste value SC_i^k, the optimal value SC_{best}^k in two groups can be found. The best coordinates (X_{best}^k, Y_{best}^k) can be obtained,

$$X_c^k = X_{best}^k$$
$$Y_c^k = Y_{best}^k$$
$$SC_{iter}^k = SC_{best}^k$$

7. iter=iter+1,if iter≠Count, then return to step 1, or comparing each SC_{iter}^k in two groups, the global optimal value can be obtained.

3. A hybrid recognition algorithms

In recent years, many scholars and researchers try to use the hybrid neural network technology in recognition algorithm has good performance. For example, Wee Ser combined PNN with GMM in Speech emotion recognition [12]; C.Jeyalakshmi combined PNN with GMM on speech recognition for hearing-impaired people [13].

3.1. SOM theoretical framework

SOM is a competitive unsupervised neural network [16]. It is a powerful clustering algorithm. It can effectively map the input data, and can good reflect the characteristics of the input data [15]. SOM has two layers, are the input layer and competitive layer. Input layer neuron number is N, number of neurons M in competitive layer to form a one-dimensional or two-dimensional array. By self-organizing learning, input data is mapped in the competition layer one-dimensional or two-dimensional graphics, and maintain its topology unchanged.

MFCC characteristic parameters of speech samples clusters into a new characteristic parameter via SOM network. The new characteristic parameters can be a good characterization of the structural properties of the original MFCC characteristic parameters, and reduce the voice data and computation.

3.2. PNN theoretical framework

PNN [8][9] is a development of the Radial Basis Function neural network and a feedforward network, which based on the density function estimation and Bayesian decision theory, with good extension. It can be abstracted into a three-layer network [10], respectively, the input layer, pattern layer, summation layer. The number of neurons in the input is the same number of vector of the input layer, used to linear representation of the input vector. The pattern layer with the M mode node, calculating the closeness of the input vector and its function center with the kernel function of the probability density. The kernel function of the probability density is:

$$f(x)=\frac{1}{N_j\sigma^d}\sum_{i=1}^{N_j}G(\frac{\|x-x_i\|}{\sigma})$$

Where G is the kernel function, d is the dimension of the input vector, Nj is the number of samples of the j-th class category, s is the smoothing factor. Smoothing factor, to a large extent, affect the classification accuracy, so smoothing factor is very important in PNN classification.

PNN is mainly dedicated to the study of pattern classification, which is of fast training, fast convergence and robustness. So PNN can be as the final recognizer of the hybrid algorithm.

3.3. Supervised hybrid recognition algorithm principle

Speaker Recognition is a biometric verification technology, the human voice contains the characteristics of their own unique physiology and behavior [11]. It can be through some kind of computational model to simulate the human auditory system to confirm and identify. Speaker recognition is mainly subject to the sound acquisition, preprocessing, feature extraction, classification, recognition process [2] in Fig.1.

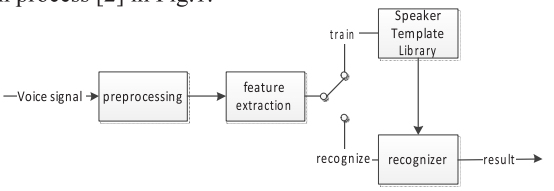

Fig. 1. General process of speaker recognition

Different components can build intelligent systems, that is, neural computation combined with artificial intelligence technology, can simulate human thought processes to a certain extent [14]. However, during training, PNN don not use the correct classification rate, so a supervised learning algorithm, DFOA-SOM-PNN, can be further proposed.

Total number of samples is $N=\sum_{j=1}^{M}N_j$, the number of correctly identified in input training samples is n, so recognition accuracy is $f=\frac{n}{N}$, f is determined by σ, that is,

$$f=F(\sigma)$$

But σ is determined by X_i^k and Y_i^k ($i=1,2,\cdots,U$; $k=1,2,\cdots,V$), U is size of group, V is size of iteration. That is,

$$\sigma=G[(X_1^k,X_2^k,\cdots X_U^k),(Y_1^k,Y_2^k,\cdots Y_U^k)]$$

So f is determined by X_i^k and Y_i^k, That is,

$$f=\max_{k=1,2,\cdots V}(F\{G[(X_1^k,X_2^k,\cdots X_U^k),(Y_1^k,Y_2^k,\cdots Y_U^k)]\})$$

Because of DFOA's random search capabilities, the best σ value can be easy obtained. Recognition process is as shown in Figure 2.

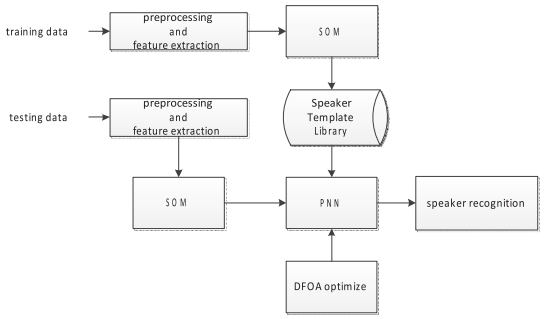

Fig. 2. The general process of the hybrid recognition algorithm

4. Experimental Analysis

4.1. DFOA global optimization experiments

The two functions, (1) and (2), are set as follows:

$$y = \sin\left(\frac{(6x^2-5)(x+6)}{(x+2)^2+3}\right) \tag{1}$$

$$y = 8-(x+100)^2 \tag{2}$$

A contrast experimental research is to find the global minimum value of (1) which is carried out on the FOA and DFOA. Another research is to find the global maximum value of (2) in three typical t value, and to analyze its convergence. The function (1) in the definition domain has very many minimum value and maximum value, and they are very close to each other. The global minimum value is approximately -0.3332. The graph of function (2) is relatively simple, its global maximum value is 8.

For function (1), FOA initial population starting point is set to random domain [-1, 1], and iterative random field is set to [-1, 1], and the number of populations of individuals is set to 20. Two groups' initial population starting point in DFOA is set to random domain [-1, 1], and iterative random field is also set to [-1,1], and the number of populations of individuals is also set to 20, and convergence factor is set to -1, and orientation factor is respectively set to 1 and -1. From Fig.3, FOA only converge to about -0.1163, but DFOA convergence to -0.3327 in Fig.4, which is very close to a given exact value. DFOA's global convergence of the complex function is fully reflected in the above comparative experiments.

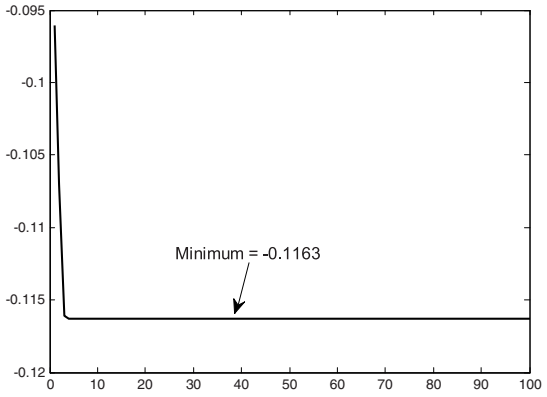

Fig. 3. Iterative process curve of FOA optimization process

As is shown in Fig.5, when the convergence factor is -1, the function (2) can't converge to the global maximum value. When the convergence factor is 1, the function (2) correctly converge to the global maximum value. But when the onvergence factor is 2, the convergence speed is significantly increased. The experiments show that the introduction of the

convergence factor, for CTJV, is an effective promotion.

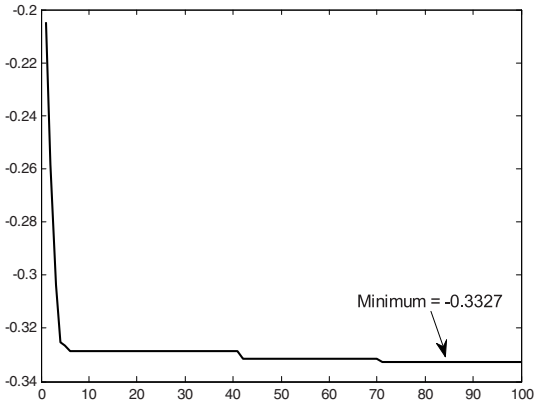

Fig. 4. Iterative process curve of DFOA optimization process

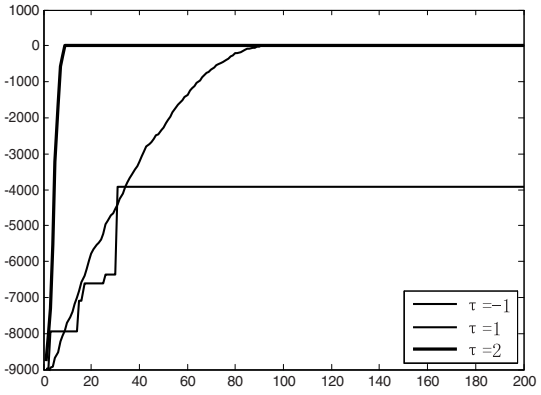

Fig.5. Different convergence curve in different convergence factors

4.2. Speaker recognition experiments

Comparative experiments of speaker recognition will be shown in this section: (a). PNN for Speaker Recognition; (b). SOM-PNN for Speaker Recognition; (c). DFOA-SOM-PNN for Speaker Recognition. Voice data in experiments is from 20 people which are a mixture of men and women. The voice text is a two-syllable Chinese word. Voice text is recorded 30 times per person, which are total 600 samples. Taking into account the short period of time may have fixed mode of pronunciation, so the first 15 times and second 15 times voice samples are recorded separately by 4 weeks. Voice samples of each person are divided into two parts, the first 18 times are used for training, the rest are used for testing. Recording environment is in the laboratory of a little noise. All voice samples' sampling rate is 8KHz, length is 4 seconds. After preprocessing, we take MFCC parameters and its 1, 2-order differential coefficient as a new characteristic parameter.

According to the number of training sample size, experiment is divided into five groups. For example, the first group takes the first 2 times voice samples of each speaker as the training sample. To obtain the optimal smoothing factor, algorithm (c) adopt cross-training. After a number of experiments, we can know that when smoothing parameter is set to in [0.01 1] can make speaker recognition to take the maximum recognition accuracy, therefore, the smoothing parameter in the above algorithm (a) and algorithms (b) take the empirical value 0.1. The network topology value of the SOM is set to 3.

It can be seen from Fig.6, recognition accuracy rate of DFOA-SOM-PNN algorithm is significantly higher than the other algorithms, and recognition accuracy rate is getting higher and higher with the increase of training samples. This is because DFOA provides a supervised training program, the more training samples the more smoothing factor value is accurate. Obviously, table I shows that SOM-PNN is better than PNN, and DFOA-SOM-PNN is the best of all.

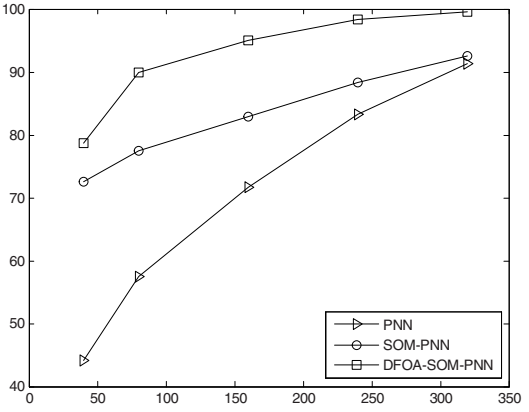

Fig. 6. Comparison of recognition accuracy in different algorithm

Table 1. Comparison of recognition accuracy in different algorithm

Number of training samples	Recognition accuracy		
	PNN	SOM-PNN	DFOA-SOM-PNN
40	44.17%	72.50%	78.75%
80	57.50%	77.50%	90.00%
160	71.67%	82.92%	95.00%
240	83.33%	88.33%	98.33%
320	91.25%	92.50%	99.57%

Experiments show that speaker recognition accuracy rate would be increased after DFOA optimization. It is worth mentioning that SOM-PNN can reach more than 70% recognition accuracy in the case of low sampling rate (8KHz) and the minimal training sample.

5. Conclusion

In this paper, SOM neural network optimize and cluster MFCC characteristic parameters, which not only effectively reduces the amount of computation and storage of data, but also better reflects the speaker's voice characteristics. The experimental results show that recognition accuracy rate of SOM-PNN is higher than using only PNN. Because of DFOA's global random search capability, PNN's smoothing factor can be optimized by DFOA, and the experimental results show that DFOA-FOA-PNN hybrid algorithm has the highest recognition accuracy. The experiment also shows DFOA effectively compensate for the defects of the FOA. In summary, DFOA-SOM-PNN hybrid algorithm is an effective method of speaker recognition.

References

[1] Campbell J.P. and Jr., "Speaker Recognition: A Tutorial," Proceedings of the IEEE, Vol. 85, pp. 1437-1462, 1997.
[2] Lin Lin and Shuxun Wang, "A New Genetically Optimized GMM for Speaker Recognition," Intelligent Control and Automation, 2006. WCICA 2006. The Sixth World Congress on, Vol.2, pp. 10235-10239, 2006.
[3] R. Saeidi, H.R.S. Mohammadi, T. Ganchev and R.D. Rodman, "Particle Swarm Optimization for Sorted Adapted Gaussian Mixture Models," Audio, Speech, and Language Processing, IEEE Transactions on, Vol.17, pp. 344-353, 2009.
[4] J.H.Holland, "Adaptation in Natural and Artificial System," The Univ, Michigan Press, 1975.
[5] J Kennedy and R Eberhart, "Particle swarm optimization," Proc IEEE Int Conf on Neural Networks, Perth, pp. 1942-1948, 1995.
[6] A Colorni, M Dorigo, V Maniezzo, D Elettronica P Milano, "Distributed optimization by ant colonies," Proc of European Conf on Artificial Life, Paris, pp.134-142, 1991.
[7] Pan Wen-Tsao, "A new fruit fly optimization algorithm: Taking the financial distress model as an example," Knowledge-Based Systems, Vol.26, pp. 69-74, 2011.

[8] D.F. Specht, “Probabilistic Neural Networks for Classification, Mapping, or Associative Memory,” Proceedings, IEEE International Conference on Neural Networks, Vol.1, pp. 525-532, 1988.
[9] D.F. Specht, “Probabilistic Neural Networks,” Neural Networks, Vol.3, pp. 109-118, 1990.
[10] T. Song, M Jamshidi, R.R. Lee and M. Huang, “A Novel Weighted Probabilistic Neural Network for MR Image Segmentation,” Systems, Man and Cybernetics, 2005 IEEE International Conference on, Vol.3, pp.2501-2506, 2005.
[11] Fang Ye and Zhou Yabin, “Pnn-Based Algorithm for the Recognition Of Speakers,” Electronic Measurement & Instruments, 2009. ICEMI '09. 9th International Conference on, pp. 1104-1107, 2009.
[12] Wee Ser, Ling Cen and Zhu Liang Yu, “A Hybrid PNN-GMM Classification Scheme for Speech Emotion Recognition,” Pattern Recognition, 2008. ICPR2008. 19th International Conference on, pp. 1-4, 2008.
[13] C.Jeyalakshmi, Dr.Krishnamurthi.V and Dr.A.Revathi, “Speech Recognition of Deaf and Hard of Hearing People Using Hybrid Neural Network,” Mechanical and Eletronics Engineering (ICMEE), 2010 2th International Conference on, pp. 83-87, 2010.
[14] M. Minsky, “Logical Versus Analogical or Symbolic Versus Connectionist or Neat Versus Scruffy,” AI Magazine, Vol.65, pp. 34-51, 1991.
[15] Tian Dan and Fan Linan, “A Brain MR Images Segmentation Method Based on SOM Neural Network,” Bioinformatics and Biomedical Engineering, 2007. ICBBE 2007. The 1st International Conference on, pp.686-689, 2006.
[16] T. Kohonen, E. Oja, , O. Simula, A. Visa and J. Kangas, “Engineering applications of the self-organizing map,” Proceedings of the IEEE, Vol.84, pp.1358-1384, 2002.