[illegible]

YUE Xicai YE Datian
Department of Electrical Engineering and Applied Electronic Technology,
Tsinghua University, Beijing 100084, P. R. China. E-mail: yuexc@tsinghua.edu.cn

LIU Ming
Telecommunication Institute, Air Force Engineering University, Xi’an 710077, P. R. China

Abstract: The authors combine genetic clustering algorithm with radial basis function neural network (RBFNN) for avoiding locally optimum solutions in speaker identification. The effectiveness of genetic clustering algorithm is evaluated with speech utterances by comparing with normal clustering method. Speaker identification experiments show that genetic clustering RBFNN can improve the correctness of text-independent speaker identification.

Keywords: speaker identification, vector quantization, genetic algorithm, neural networks

1. Introduction

The purpose of speaker identification is to determine speaker’s identity from his/her speech utterances. Every speaker has his/her own characteristics when speaking. These characteristics are called speaker features which can be extracted from speech utterances. Through comparing test speech with the extracted features, speaker’s identity can be recognized. The process of speaker identification consists of two steps：training and recognition. If the same text is used in both training and recognition, the mode of speaker identification is called text-dependent. Otherwise, called text-independent.

Many methods have been used in speaker identification. For text-independent speaker identification, vector quantization (VQ) method has been considered as the benchmark [1]. VQ based methods, such as the hidden Markov model (HMM) and the Gaussian mixture model (GMM) [2], are most prominent speaker identification classifiers. VQ method can also be used in RBFNN in speaker identification. The well-known technique of VQ is K-mean clustering algorithm [3], which is a very efficient algorithm. However, the clustering result varies in some extent when a different initial centroid is selected, so it is difficult to get a total optimum clustering solution by K-mean clustering algorithm. In this paper, we introduce the genetic algorithm [4], one of the total optimum methods, to VQ to get an optimized clustering result, and then combine genetic clustering algorithm with RBFNN to improve the performance of RBFNN in speaker identification. In order to evaluate the performance of genetic clustering RBFNN in text-independent speaker identification, two training methods of RBFNN, one using K-mean clustering method, and the other using genetic clustering algorithm, are compared. The results show that the genetic clustering RBFNN has better performance in speaker identification.

2. Methods

2.1 Genetic clustering algorithm

K-mean clustering algorithm is based on two necessary conditions for optimality: the centroid and the nearest neighbor conditions. Through turning clustering information into the forms of genes in chromosome and modifying nearest neighbor conditions as the fitness function, the genetic clustering method is described as follows:

Chromosome coding: Randomly taking a clustering number for each training vector as the gene in a chromosome, each chromosome can represent an initial centroid that will be optimized by genetic operation. Therefore, the length of chromosome is as long as the number of vectors in training sample, and the contain of each gene in the chromosome is the clustering number of the training vector whose sequence number is the same as the gene’s. Therefore, we can calculate clustering centroid from a chromosome:

$$C_i = \frac{1}{m_i} \sum_{x_j \in i} x_j \qquad (1)$$

where C_i and m_i are the centriod of clustering i and the total number of genes which value equals to i, and X_j is the training vector.

Fitness Function: We select the reciprocal of the Euclidean distance between training vector and the centroid as fitness function of a chromosome, listed as follow:

$$f = (\sum_{i=1}^{K} \sum_{X_j \in i} \left\| X_j - C_i \right\|^2)^{-1} \qquad (2)$$

Where K is the total number of clustering.

Selection: Using Roulette Wheel selection [4] as selection mechanism, we can select chromosomes as next generations.

Crossover: For two selected chromosomes A and B, we

Report Documentation Page

Report Date 25 Oct 2001	**Report Type** N/A	**Dates Covered (from... to)** -

Title and Subtitle Text-Independent Speaker Identification by Genetic Clustering Radial Basis Function Neural Network	**Contract Number**
	Grant Number
	Program Element Number
Author(s)	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) Department of Electrical Engineering and Applied Electronic Technology Tsinghua University Beijing 100084, P.R. China	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es) US Army Research, Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)
Distribution/Availability Statement Approved for public release, distribution unlimited	
Supplementary Notes Papers from 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 25-28, 2001, held in Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom.	
Abstract	
Subject Terms	
Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU
Number of Pages 4	

randomly select a clustering number in A and find out a gene which has the smallest fitness in this clustering by using formula (2), then we replace this gene with the gene in the same position in B. This operation is desired to change the gene that has the smallest fitness in a chromosome to reproduce a more optimized solution.

Mutation: To explore new genetic information that bas not been generated in the population during the initialization, we select a chromosome, and randomly change the value of a gene in the range of 1 to K in this chromosome.

Crossover rate is a possibility factor used to control the balance among the rate of exploration, the new recombined building block and the rate of disruption of good individuals, while the mutation rate is a possibility factor for controlling the balance between random search and the rate of disruption of good individuals. For there is no appropriate method to determine these parameters [5], we determine them by experiments.

Repeating selection, crossover and mutation, until the fitness of total chromosomes does not increase. The chromosome of the biggest fitness is selected as the clustering result.

2.2 Genetic clustering RBFNN

Neural networks are popular in pattern recognition. Typical ones of them used in speaker identification are forward networks, such as back propagation (BP) and RBFNN. RBFNN is a two-layer neural network whose output nodes form a linear combination of the basis (or kernel) functions computed by the hidden layer nodes. The architecture of RBFNN is showed in figure 1. The most commonly used basis in RBFNN is a Gaussian function of form:

$$\mu_{1,j} = \exp\left[-\frac{(x-w_{1,j})^T(x-w_{1,j})}{2\sigma_j^2}\right] \qquad j=1,2,\ldots,N_1 \tag{3}$$

Where $\mu_{1,j}$ is the output of the jth node in the hidden layer, x is the input pattern. N_1 is the number of nodes in the hidden layer. Parameters $w_{1,j}$ and σ_j can be estimated in many ways. The output of network forms a liner combination of the output of the hidden layer:

$$y_j = w_{2,j}^T \mu_1 \qquad j=1,2,\ldots N_2 \tag{4}$$

The first reason for using RBFNN is the successful use of GMM in speaker identification, for RBFNN has the same underlying structure as GMM when Gaussian function is selected as the type of basis function. The second and more important reason is that it is easy to introduce the genetic clustering algorithm to RBFNN. In RBFNN, the hidden layer and the output layer can be trained separately, and the hidden layer can be trained by clustering algorithm [6]. If we optimized clustering centriods by genetic clustering algorithm, the result of total RBFNN should be improved. Therefore, we combine the genetic clustering algorithm in RBFNN and named it as genetic clustering RBFNN whose training algorithm is described as follows:

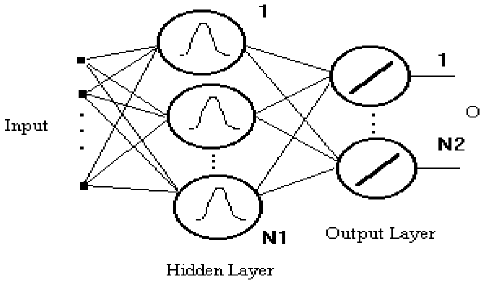

Fig. 1 Architecture of RBFNN

1. Parameters in the hidden layer can be estimated by:

$$W_{1,j} = \frac{1}{m_j}\sum_{x_i \in j} x_i$$
$$\sigma_j^2 = \frac{1}{m_j}\sum_{x_i \in j}(x_i - w_{1,j})^T(x_i - w_{1,j}) \tag{5}$$

Where j is the node number, which corresponds to number j th centriod calculated by genetic clustering.

2. Weights in the output layer can be calculated by least mean squares (LMS) algorithm [6].

3. Speaker Identification experiments

3.1 Database

The database consists of speech utterances read from Chinese newspapers randomly by 20 male postgraduate students in a noisy environment. The speech signal was sampled at 8KHz with 16 bits A-D. The word intervals in the speech are detected and eliminated by the zero cross rate and energy threshold.

3.2 Feature extraction

Although the exact features in speech signal for speaker characteristics are not exactly known, it is a fact that many speaker identification systems rely on spectral-based features [1, 2]. One of most commonly used is cepstral coefficients, which can be obtained from linear predictive (LP) coefficients. The principle of LP is that speech signal can be approximated as a linear combination of past samples. This can be expressed as follows:

$$\hat{s}_n = \sum_{k=1}^{N} a_k s_{n-k} \tag{6}$$

where s_n are speech samples, n is time index, and the a_k are the LP coefficients. After the LP coefficients $a_k(1 \le k \le N)$ are calculated, the cepstral coefficients c_k are obtained from a_k by the recursive relationship[7]:

$$\begin{aligned} c_1 &= a_1 \\ c_n &= \sum_{k=1}^{n-1}(1-\frac{k}{n})a_k c_{n-k} + a_n \\ &1 < n < P \end{aligned} \tag{7}$$

If the dimension of the cepstral vector P is great than N, then a_n is set to zero for $P<n\le N$. The P-dimensional vectors are used as the input to the classifier. We shall use linear predictive coefficients cepstrum (LPCC) as feature vectors of a speaker in this paper.

Speaker features are obtained as follows: The digital speech data is preemphasized with $1\text{-}0.95Z^{-1}$, then segmented into frames of 256 sample points with the overlap of 128 sample points. 16-order LPCC of each frame is adopted as speaker's features. A part of features are used for training, and the others used for test.

3.3 Results

For 10 seconds speech of 10 speakers, K-mean clustering and genetic clustering algorithms are used for clustering respectively. The number of clustering is 32 in both methods. Some parameters in genetic clustering are selected as follows: population size is 64, and the crossover rate is 0.9, while the mutation rate is 0.01. The distortion distance to the clustering centroids of a speaker, which is used as performance criterion for both of the K-mean clustering and genetic clustering, is calculated by:

$$d = \frac{1}{N}\sum_{i=1}^{N}\min_{j=1}^{K}\|X_i - C_j\|^2 \tag{8}$$

where N is the total number of vectors in test sample. For 10 seconds test speech, the distortion distances from the clustering centroids of a typical speaker to all speakers are calculated by two clustering methods respectively, showing in table 1. From the table, we can see that the average distortion distance of imposer has increased approximately 6% when using genetic clustering algorithm, while verification distortion distance has decreased approximately 1%. These changes are both beneficial to separate this typical speaker from others, so it is easier to verify a speaker when using genetic clustering method. It is worth to note that the distortion distances do not decrease for all imposers in table 1, but from the whole, we still think that genetic clustering algorithm could optimized the clustering results, for genetic algorithm is a totally optimum algorithm.

Table 1 only shows the data of a typical speaker. For all of 10 speakers, the average distortion distance of imposer increases 5.7% when using genetic clustering algorithm, and the average verification distortion distance decreases 1.8%. This result indicates that it is helpful to use genetic clustering algorithm in speaker identification.

Table 1. The distortion distance for VQ and GA

	VQ (K-mean)	GA	(GA-VQ)/VQ
Imposer 1	0.1934	0.2183	12.8%
Imposer 2	0.1458	0.1451	-0.48%
Imposer 3	0.1148	0.1125	-2.00%
Imposer 4	0.0892	0.0907	1.68%
Imposer 5	0.1187	0.1193	0.51%
Imposer 6	0.1192	0.1125	-5.62%
Imposer 7	0.2262	0.2293	1.37%
Imposer 8	0.1912	0.2146	12.24%
Imposer 9	0.2635	0.3077	16.77%
Imposer's average	0.1624	0.1722	6.03%
Verification	0.0957	0.0946	-1.15%

From table 1, we notice that the distortion distance for imposer 4 is smaller than the distortion distance for verification in both of VQ and genetic clustering algorithm, although there is 1.7% improvement in genetic clustering algorithm. This result indicates that training with 10 seconds speech and test with 10 seconds speech are not enough for speaker verification. Furthermore, it implies that using VQ or genetic clustering algorithm alone is not enough to identify these speakers, and some forms of weighted distance measure are needed. RBFNN can be regarded as a weighted distance measure especially when hidden layer is trained by clustering algorithm. This is the key for us to use genetic clustering RBFNN in speaker identification.

Two training methods of RBFNN, the normal RBFNN using K-mean clustering method and the genetic RBFNN using genetic clustering algorithm, are trained respectively. The numbers of nodes in RBFNN for identification 20 speakers are 16 in the input layer and 20 in the output layer. Determining the number of nodes in hidden layer is the first problem we should solved in RBFNN's training. We select number of nodes in hidden layer as 16, 32, 64, and 128 to train RBFNN respectively. Training with 30 seconds speech, and tested 10 times for each speaker with 15 seconds speech respectively, the identification results for two kinds of RBFNN are showed in Fig. 2. In the Fig, we can see that the correctness curve of the genetic RBFNN is above that of the normal RBFNN, showing that the speaker identification performance of genetic RBFNN is better than that of normal RBFNN. The correctness of speaker identification increase with increasing of the number of nodes in hidden layer for both two types of RBFNN respectively, and the larger the number of nodes is, the higher the correctness the network can yield. The correctness improvement becomes smaller and smaller when the number of nodes in hidden layer increases for both types of RBFNN respectively. For genetic RBFNN, after the number of nodes is bigger than 64, the correctness improvement becomes very small. Consideration both of the improvement of speaker identification performance and the complex of computation, we select 64 as the number of nodes in the hidden layer.

Varying length of the test speech, two types of RBFNN are tested 10 times with 5 seconds and 10 seconds respectively, and the results are showed in table 2. In the table, we can find that correctness of genetic RBFNN is higher than that of the normal RBFNN for both 5, 10 and 15 seconds test speech, and even tested with 10 seconds speech, genetic RBFNN is slightly better than normal RBFNN tested with 15 seconds speech. This result shows that genetic clustering RBFNN is better than normal RBFNN in speaker identification. From the tables, we also see that there are 6% and 3% improvements for 5 seconds and 15 seconds test speech, respectively. Therefore, the genetic RBFNN is better for shorter speech situation. The shorter the test speech is, the higher improvement the genetic RBFNN should be. This is helpful for the situation where speech is difficult to acquire.

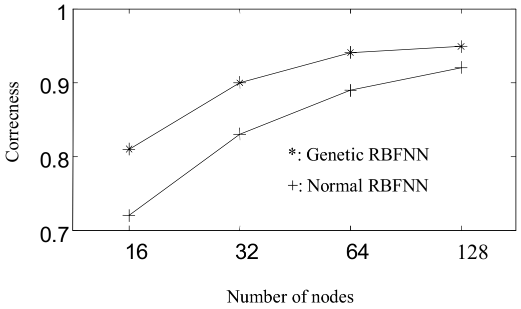

Fig. 2. Speaker identification performance of two types of RBFNN

Table 2. Speaker identification performance with different length of test speech

Length of test Speech (s)	Normal (N)	Genetic (G)	G-N
5	84%	90%	6%
10	89%	93%	4%
15	92%	95%	3%

4. Conclusions and discussions

The genetic clustering algorithm uses a large amount of population to search the total optimized solution of a clustering problem. With the genetic operations, such as crossover and mutation, the role of each population is just like an isolated K-mean clustering process. From this view, there are many "K-mean clustering processes" in the same time in genetic clustering. Therefore, the optimized centriod can be selected from all of the population by fitness function. Using genetic clustering algorithm, we can get a better clustering result, which has the property of a bigger distortion distance for intra-classes and a smaller distortion distance for inter-classes. Therefore, the genetic clustering algorithm is more suitable for pattern recognition. The speaker identification performance improvement of genetic RBFNN firstly benefit from the optimized clustering centriods. Furthermore, after combination with neural network method, genetic RBFNN is more powerful in pattern recognition, for a weighted distance measurement is more efficient than a simple distance measurement when using genetic clustering algorithm alone. Our text-independent speaker identification experiments show that genetic RBFNN can improve correctness of speaker identification. The experiment results also show that genetic clustering RBFNN is more effective for cases where data is difficult to get. Although genetic clustering has many advantages, its calculation cost is a little too big, for the amount of population that should be handled with in the same time is big.

Acknowledgment

This work is partly supported by the National Natural Science Foundation of China under the contract number 39870212.

References

1. Farral K R, Mammone R J, Assaleh K T. "Speaker recognition using neural networks and conventional classifiers". IEEE trans on speech and audio processing, 1994, 2(1): 194-205
2. Reynolds D, Rose R. "Robust text-independent speaker identification using Gaussian mixture speaker models". IEEE trans on speech and audio processing. 1995, 3(1): 72-83
3. Yuan Z X, Xu B L, Yu C Z. "Binary quantization of feature vectors for robust text-independent speaker identification", IEEE trans on speech and audio processing, 1999, 7(1): 70-78
4. Man K F, Tang K S, Kwong S. Genetic algorithm, Springer-Verlag, London, 1999
5. Grefenstette J J. "Optimization of control parameters for genetic algorithm". IEEE trans on system, man and cybernetics, 1986, 16(1): 122-128
6. Hush D R, Horne B G. "Progress in supervised neural networks". IEEE signal processing mag, 1993, 10(1): 8-39
7. Atal B S. "Effectiveness of linear prediction characteristics of speech wave for automatic speaker identification and verification". J. Am. Soc. Acou, 1974, 55: 1034-1312