Research Article

International Journal of Advanced Computer Research, Vol 6(25)
ISSN (Print): 2249-7277   ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2016.626018

A speaker model clustering method based on space position

Jing Zhang* **and Xiaomei Chen**
Cisco, School of Information, Laboratory of Language Engineering and Computing, Guangdong University of Foreign Studies, Guangzhou, China

Received: 25-June-2016; Revised: 26-July-2016; Accepted: 27-July-2016

Abstract

In the speaker recognition system with large numbers of models, the traditional computations of matching one by one can be very time-consuming. In order to solve the problem of fast recognition, this paper proposes a speaker model clustering method based on space position. The idea is: firstly, to divide the training models into multiple layers, then searches class representatives for every layer, then to cluster the training model by the gotten class representatives. This method can greatly reduce the number of models matching, and achieve the goal of fast recognition. The paper takes the Gaussian mixture model (GMM) as the speech model, the experiment shows the recognition average speed of 0.5s per person, and the correct recognition rate less than 1% loss is ensured.

Keywords

Speaker recognition, Space position, Stratification, Clustering model.

1.Introduction

With the popularity of the Internet, the online voiceprint recognition system has been greatly developed. One of the greatest features of the online voiceprint recognition system is the large number of registration in the system, but with the increase in registrations, the recognition time will be very long and may eventually lead to failure real-time recognition. The major factors impact recognition time is dimensionality feature space and the model complexity. Therefore, the research of fast recognition mostly starts from the following two aspects. Firstly, reduce the dimensionality of the space; the second is to simplify the model structure.

Apsingekar et al. [1] proposed a model pre-quantification method. The method uses the quantification criterion mentioned in the literature [2], which quantify the complex models into a single vector, and select K- means algorithm for clustering model. Thus the structure of the model got greatly simplified and the recognition speed improved. Speaker recognition method based Gaussian mixture model (GMM) of classification feature space was forwarded [3]. In the training process, the model got be quantified and clustered according to the similarity of the model, the training and clustering were done simultaneously. During the testing, the classification of the test data to be selected firstly, and then to match the test data with all members of the class likelihood score. Thereby the number of models to match and recognition time reduced. But the feature vector of speaker model has high-dimensional complexity, and then the selection of quantization criterion is a difficulty.

Sun et al. [4] proposed a stratification testing method of speaker clustering, which is based on data analysis techniques of iterative self-organization. The advantage of this method is that the classification representative selected is not the model member, but the GMM model generated from all members' data in the class for the training. The Class GMM model has a strong representation, but its building process is very long and when the scale of training model is too large the method becomes impractical. Distributed clustering method of speaker model was proposed [5]. The method clusters speaker model by using the KL divergence as a distance measurement between classes, and logarithmic likelihood distance measurement as a within class. When recognition is carried out, the selected members of the class are taken as the matching object, and then a large number of models are cut off. But it is necessary for the

*Author for correspondence

clustering algorithms to divide all members in the class, so the clustering consumes long time. A recognition method based on particle swarm optimization advance (PSOA) and kernel matching pursuit (KMP) clustering algorithm was proposed [6]. They applied particle clustering methods, for the Mel frequency ceptral coefficient (MFCC) parameters that feature a larger amount of information, of particle clustering algorithm to extract a small amount of representing the parameters characteristic of the speaker, improved the efficiency of the system, but shown the defects of slow convergence.

Matza et al. [7] used the ability of automatically selecting the order of GMM for the rival penalized EM (RPEM) algorithm, and trained the GMM with appropriate order for each speaker, which achieved better results, but in the case of large amounts of data, GMM order is difficult to stably Convergence to a specific number, so the recognition effect is not ideal. Xing et al. [8] proposed clustering algorithm which is based on Bhattacharyya distance. It is capable in quickly clustering the speakers into one class and reduced the complexity of modelling calculations, then improved the recognition efficiency of the system.

But the generation of Bhattacharyya kernel function resulting large number matrix operations, which increased the computational complexity. Shan et al. [9] proposed universal background model (UBM) order reduction algorithm of background model to improve the computing speed, but at the same time the recognition rate would be affected with order reduced.

Huaqiao et al. [10] proposed clustering method of speaker model based on approximate Kullback–Leibler (K-L) distance and vector quantization techniques. The method uses a variable search strategy, coordinated the relationship between the recognition rate and recognition time, but the computational complexity increased. Based on the traditional clustering methods, Patnaik et al [11] proposed a method that fragmenting the speech segments with the same nature in a long sentence. It greatly reduced the recognition time, but achieved the low stability of the clustering and grouping.

This paper proposed a speaker clustering algorithm based on the spatial position that is spatial location-speaker clustering model algorithm (SL-SCMA). The advantage is reflected in three aspects: 1) the selected similarity measurement technology for the class choosing does not rely on vector quantization (VQ), which avoids the loss of model information caused by quantification; 2) to find the class representative layer by layer according to the spatial location of the training model, time of the algorithm is short. 3) The clustering process features self-organization, capable of self-appointing the class representatives and cancelling it.

2.A speaker clustering algorithm based on spatial location (SL-SCMA)

The SL-SCMA idea is: suppose the speaker model generated in training process can be represented by a point in high-dimensional space, then some of these points can be considered close while other far. Based on this assumption, to select the class representatives of the layer that belong to from the outermost layers individually until no class represents meet the requirements appear. Then to cluster all the models according to class represents obtained from each layer, and to calculate the log-likelihood score from each model to all classes represents, and classify each model to the class achieved the highest score. Finally, cancel the classes represent that number of class members is not met the maximum and minimum requirements, and allocate the members of another class according to the principle of proximity, and ultimately achieve the clustering of speaker model. The generation of class represents in each layer as the algorithm described was shown in *Figure 1*.

2.1Similarity criterion

Similarity criterion not only needs an accurate measure of similarity degree between the models, but also a small amount of computation. Since the GMM model is a complex structure with multi-parameters, which belongs to the distribution model, the similarity measure between models is to complete the distance measure between the two types of distribution. K-L divergence commonly used to measure the distance between the two types of distribution. As (1) described.

$$\lambda_k^{cr} = \arg\min_{1\le s\le S} d_1(\lambda_s, r_k), 1 \le n \le K \qquad (1)$$

But currently there is no analytical using K-L divergence to find the distance between two GMM distributions. So the K-L divergence cannot be directly used to find the distance. So K-L divergence approximation methods of calculating the difference of a logarithmic likelihood between the feature vectors and two GMM models were proposed [12].

Papers referenced this method and used (2) as a measure of similarity between the GMM.

$$D(\lambda_1,\lambda_2)=|\frac{1}{N}\sum_{i=1}^{N}\log p(\frac{X_1^{tr}}{\lambda_1})-\frac{1}{N}\sum_{i=1}^{N}\log p(\frac{X_1^{tr}}{\lambda_2})| \quad (2)$$

Where, $D(\lambda_1,\lambda_2)$ means the distance from model λ_1 to λ_2; and X_1^{tr} is training feature vector of model λ_1, when use (2), it needs to be processed as (3) described.

$$D_{(\lambda_1,\lambda_2)}=\frac{1}{2}(D(\lambda_1,\lambda_2)+D(\lambda_2,\lambda_1)) \quad (3)$$

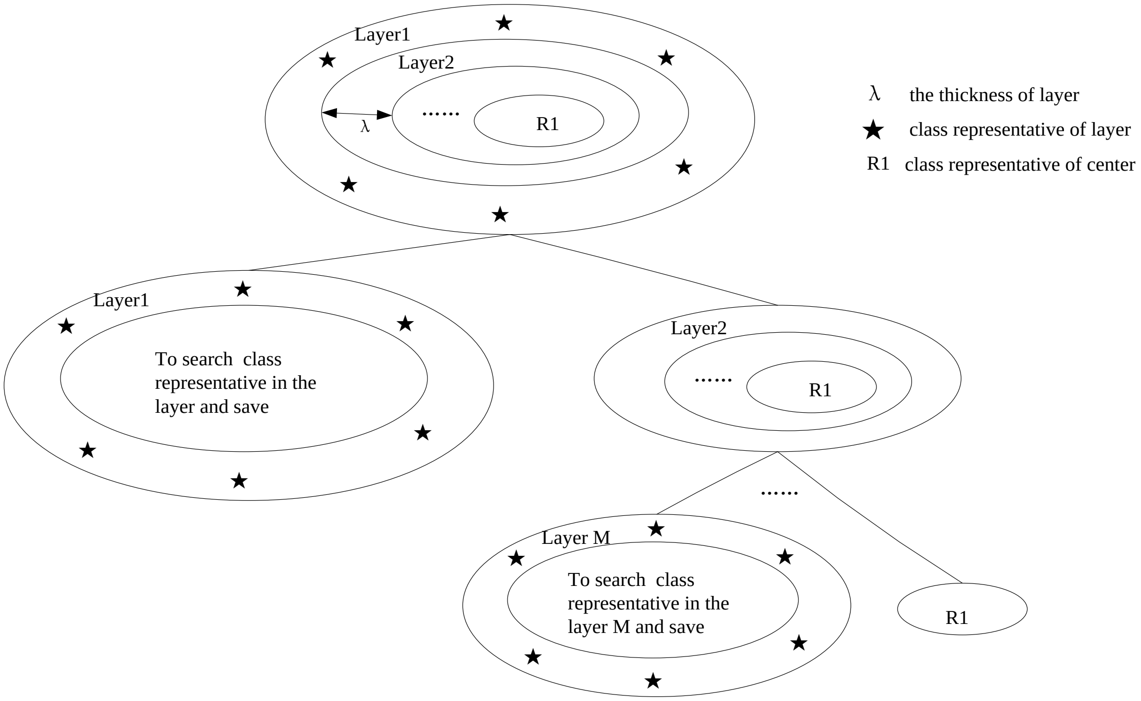
Layer1
Layer2
......
R1
λ
λ the thickness of layer
★ class representative of layer
R1 class representative of center
Layer1
To search class representative in the layer and save
Layer2
......
R1
......
Layer M
To search class representative in the layer M and save
R1

Figure 1 SL-SCMA algorithm

Since the experiment proved $D(\lambda_1,\lambda_2)\neq D(\lambda_2,\lambda_1)$, in order to make the distance between λ_1 and λ_2 equal the distance between λ_2 and λ_1, the paper used the value processed according to (3) to measure the distance the two models. The easiest way to format your manuscript is to simply download the template, and replace the content with your own material.This template provides authors with most of the formatting specifications needed for preparing the electronic versions of their papers.

2.2The uniform grouping policy based on density polyethylene con

In order to quickly recognize, the class number M gotten by clustering method could neither too small nor too large. According to the calculation analysis of the amount of the testing process, the traditional recognition method need N (N is the model library size) times likelihood calculations of GMM model while model clustering method requires only M + E times, E is the total number of contained models after class selection. If M is too small, will inevitably lead to the larger E, then the M+E could not significantly reduce compared with M; if M is too large, although E would be reduced, but the sum of both would still very large and the recognition time could not be significantly reduced. Visible, the cluster size M should be appropriate, and more importantly, the number of members of each class should be relatively uniform. If the number of members of any class is too large, it means that the more speaker models contained in the class and more recognition time needed, which will result, excessive recognition time for part of speaker and the system does not have real-time. Therefore, the paper evenly divided model, the number of members in each class were set as between 5 to 15. In order to achieve uniform division, the paper proposed the concept of density, as defined in the (4)

$$F_{\lambda i}=\sum_{j=1}^{s}D_{(\lambda_i,\lambda_j)} \quad (4)$$

In (4), the size of $F_{\lambda i}$ can reflect the intensity of each model around model λ_i, the smaller $F_{\lambda i}$ means more dense, and the bigger $F_{\lambda i}$ means more sparse. s is the number of class members. Intensive can be used to achieve uniform division of model, the specific steps are:

Step one: According to the principle of minimum distance to cluster the model library. As (5):

$$S_i^w = \arg \min_{\substack{1 \le i \le N \\ 1 \le w \le M}} D_{(\lambda_i, R_w)} \quad (5)$$

In (5), S_i^w indicates that the i-th model is divided into a class where the w-th class represents in, M is the number of class representatives.

Step two: To judge the number of Count(w) of the class w whether meets the maximum or minimum member requirements or not. If Count (w) ＜Lmin, to cancel the class representative; If the Count (w) <Lmax, to cancel the class representatives and carry class growth calculation, the like growth algorithm is as follows***:***

1) To calculate the intensity F of all members in the class, and press F value from small to large order. To appoint the member that corresponding to the minimum F as the firstly grow out class representative.
2) To verify the class representative qualification of other members by pressing F from small to large order, followed by validation of. If the formula (5-5) is meeting, then the member is appointed as the new class representative. Until all possible class representatives were found out.

After class growth stops, return to step one, to clustering all models again. Until the number of members contained in all class meet L_{min}≤Count (w) ≤L_{max}, (1≤w≤M), then exit the loop, otherwise return step(1) to continue to look for new class representatives.

3.SL-SCMA algorithm

3.1Generation of class representative

First, randomly selected a model R_1 from the model library as the central class represents and then set a logarithmic likelihood threshold L_0 as well as layer thickness Δ .To calculate all log-likelihood scores that from models to center class represents (except the central class represents itself), if it were larger than L_0 then the model will be assigned to the center class, whose class representative is R_1 , else, assign model to candidate set C_d .The logarithmic likelihood $P(x)$ is calculated as shown in (6).

$$P(x) = \sum_{j=1}^{C} \log(\sum_{i=1}^{C} (N(x;i) * A)) \quad (6)$$

Where C is the order of GMM, in this experiment the order is eight. A is weight, and $N(x;i)$ is the density function speech feature vectors x to the Gaussian model of No. i, and the equation is expressed as (7).

$$N(x,i) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_i|}} \exp[\frac{1}{2}(x-\mu_i)\Sigma_i^T(x-\mu_i)] \quad (7)$$

In formula (7), μ is GMM model mean, and Σ is model covariance, and D is the dimensions number of x . Next, to check the qualifications of all members of C_d . The check rules is: calculate the sum logarithm likelihood S of each member to other members, S is expressed by (8) below.

$$S(i) = \sum_{i=1}^{n} \sum_{j=1, j \ne i}^{n} D(\lambda_j, \lambda_i) \quad (8)$$

Where, n is the members' number of current layer in C_d , and $D(\lambda_j, \lambda_i)$ is the logarithm likelihood scores from model i to j , then to sort the data in S by descending order. To appoint the member that corresponding to the maximum likelihood as the first class representatives in the current layer and search the next class representative in S by descending order. If the logarithm likelihood of next members of former representative were less than L_0 , then to appoint the member as second class representative in this layer, else continue the searching, until the validation completion of the 4/5 top-ranking members. Since the remaining 1/5 of the members are on the edge of the layer, so directly to cancel the qualification of the class representative. After this validation, return to the step of generation of candidate class represents and start the validation for the lower layer, at this time $L_0 = L_0 + \Delta$, the center class representative is unchanged and still R_1 . So the cycle continues, until the validation of class represents for all layers end.

Then, to assign all models except the class representative to the corresponding class according to the principle of the highest log-likelihood score, as (9) shown.

$$C_g^s = \arg \max_{1 \le s \le N} D(R_s, \lambda_g) \quad (9)$$

Where, N is the total clustering number, and C_g^s means to assign the g-th train model to the class that the class representative R_s belongs to. Finally, to cancel the class representative that does not meet the requirements of minimum or maximum members, and to assign member to the other class according to the principle of proximity, and complete the clustering of the speaker model finally.

3.2Selection of algorithm parameter selection of threshold

During the clustering, the log-likelihood threshold L_0 would be used by both the layer division and the represents validation inside layer, so the choice of L_0 directly impact on the performance of the clustering results. The method that the papers selected is: firstly to recognize the unknown speaker using traditional recognition methods, to choose 300 correctly recognized speaker for the recognition results, and get the log-likelihood score, and store which in set A. To calculate the mean and variance of the data in A, the initial value of L_0 is selected as the sum of $\overline{\omega}$ and S then multiplied by the activity factor σ ($0 < \sigma < 1$), the formula is expressed as (10) below.

$$L_0 = \sigma(\overline{\omega} + S) \tag{10}$$

Then, to appropriately adjust the size of σ according to the member number obtained by experimental stratification, for example: if the member inner layer is too small, to reduce σ, thereby increasing the L_0 (L_0 is negative); else, to increase σ, thus L_0 reduced. If the members in layer are too many, then it will spend a lot of time for the validation of class representatives, then it must be avoided.

Selection of Layer Thickness Δ: For the Training model with unknown spatial distribution, which shared the layers will directly affect the algorithm time and the final clustering result. If the layers were too many, then class representatives obtained are not representative; else layers were too few, the calculated amount will increase significantly. In this experiment, the selection method is: Select the absolute value that k times one percent L_0, and k is a positive integer. As expressed in (11).

$$\Delta = k\square\frac{|L_0|}{100} \tag{11}$$

The layer thickness will be different due to the different values of k. We hope layers are not too many, and the final number of clusters is not too little. In order to achieve a good clustering effect, a suitable k value should be found combined with the empirical method.

The selection of θ: In the process of generating the initial class representatives and class growth, it is desirable to produce a new class representative of the model closely with the surrounding and between the existing class represents the distance θ is large enough. If θ is small, the new class representatives get more numbers, may lead to the eventual number of clusters is too large; less new class represents a larger number if θ is obtained, may not meet the requirements of uniform division. In this system, $\theta \approx \mu \bullet \Delta$ ($0 < \mu \leq 1$), to be combined with the actual situation to adjust μ, get the ideal cluster.

4.Analysis of experimental results

The experiments used TIMIT speech recognition database, which contains 630 speakers, and a total of about 6300 English sentences with 2s length. The performance testing of spatial location clustering was carried out by choosing the speaker models of 600 speakers that can be recognized correctly among 630 people.

The first seven sentences of each speaker were chosen for training and generated GMM model, while the last three for testing. The experimental platform is MATLAB2010b based Windows 7, and the host is configured with dual-core 2.3GHz CPU as well as 2G. GMM memory and degree of mixing is 8.

Since any clustering algorithm is to reduce the recognition rate for the recognition time, misclassification is unavoidable. Thus, in the testing phase, according to the descending order of log-likelihood scores obtained by the test data for each class represent, select w classes to construct a matching subset, then to match the test data and all models in matching subset, the speaker obtained the highest log-likelihood score is determined as target speaker.

4.1The influence of matching subset *w* to the average recognition time and recognition rate

Figure 2 reflected the relationship of w and average recognition time, and *Figure 3* reflected the relationship of w and recognition accuracy. As *Figure 2* shown, the average recognition time increased with the greater matching subset w value. In order to achieve rapid identification purposes, the smaller w is better. But it can be seen from *Figure 3* that, (total clustering is 25) when w = 4, the correct rate is only 93.6%, and the recognition rate of loss is too large to be desirable; when *w* = 6, the correct rate reach 99.18%, which meet the requirement that the loss of recognition rate less than 1%, in this case the average recognition time is about 0.5s.

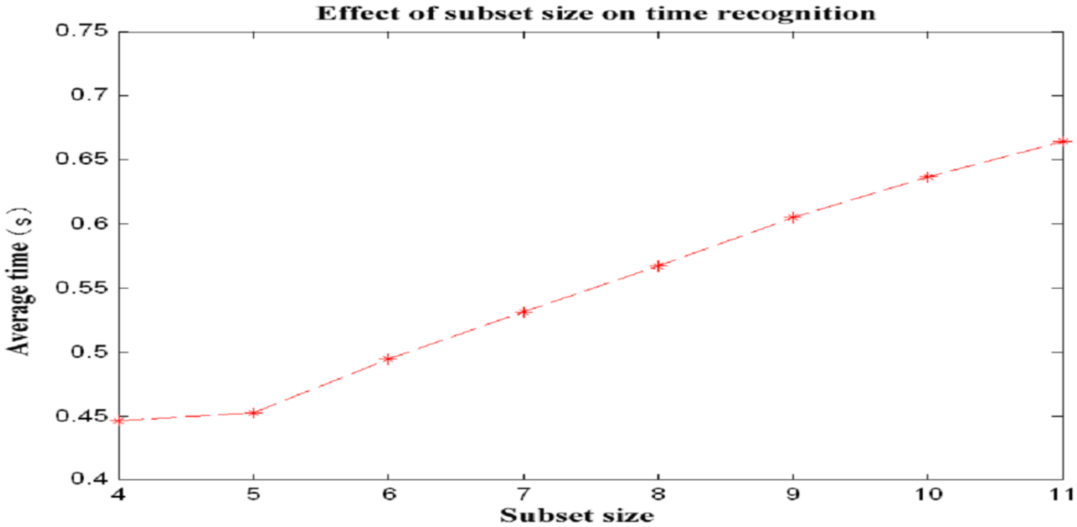

Figure 2 Relationship of *w* and average recognition time

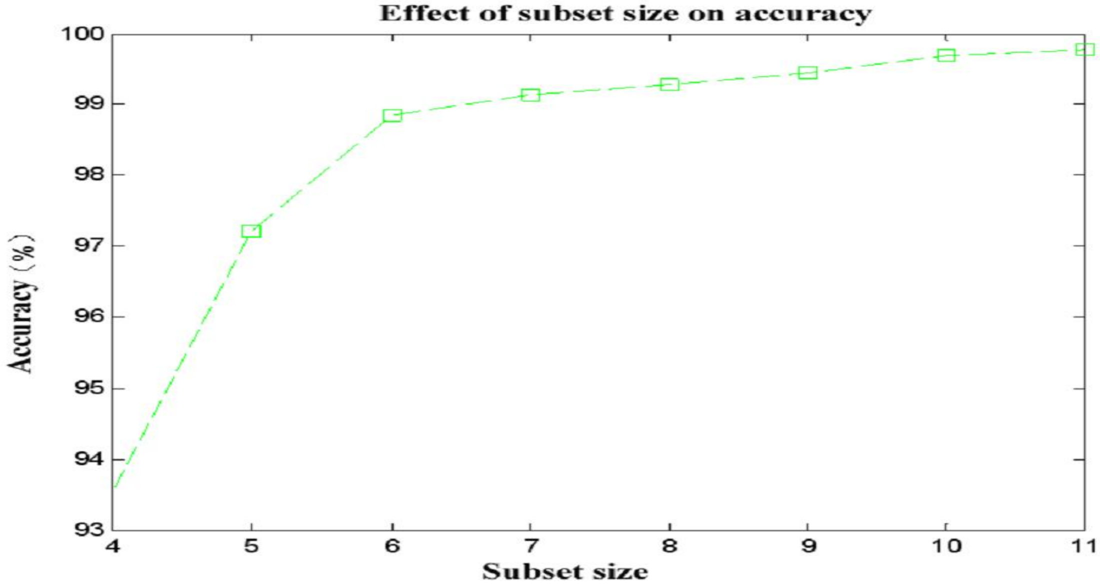

Figure 3 Relationship of *w* and recognition accuracy

4.2The recognition effect comparison of SL-SCMA method to traditional methods

Successively select 120,250,380,630 models for experiments from TIMIT database, and to compare the recognition effect of recognition method based SL-SCMA to traditional methods as the paper proposed. *Figure 4* is the comparison result for mean time of two methods; *Table 1* is error loss by using SL-SCMA Recognition in different size of the model library.

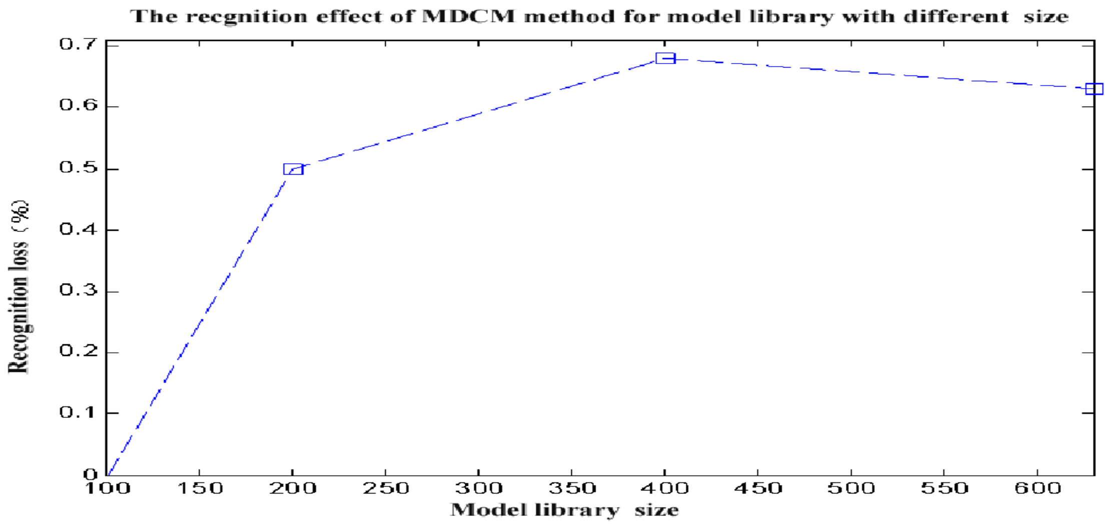

Figure 4 The average recognition time of two methods

It can be known from *Figure 4* that for the traditional identification methods, with the increase in model library, the recognition time-consuming trend was increasing rapidly. But for SL-SCMA recognition method, although the average recognition time is also increased with the increase of the number of models, but it is significant that the increase is much more than the traditional method, and the average recognition time is maintained at a low level. For the model size is 100, the SL-SCMA recognition speed is 1.7 times than that of conventional methods; when the model size is 630, the recognition rate reached 3.65 times than that of traditional methods. And with the increase in the size of the model, the multiple of the increase was a significant uptrend.

The recognition performance under different model library scale according to SL-SCMA recognition method as shown in *Table 1*

Table1 Recognition results using SL-SCMA recognition method

Scale of Model Library	180	250	380	630
Mean recognition time (s)	0.82	1.49	1.35	2.11
Loss of recognition rate (%)	0.38	0.55	0.66	0.62

Table 1 shows that with the increase in the size of the model library, the recognition rate of loss has increased, but remained at 0.7%, when the model library size is 630, the recognition rate of loss of 0.62%. To exchange for 3.65 times of recognition speed by 0.62% recognition rate loss, the effect is more significant.

5.Conclusion

A clustering method of speaker model based spatial position was proposed. It started from the space distribution of the model, using search methods layer by layer, has the effect of self-organizing clustering model and can search out the most representative members for training models as the representative of each clustering. In addition, the method also has such advantage as short training time and high class selection accuracy. Therefore, it is of significance for the clustering method of speaker model based spatial location in achieving rapid speaker recognition.

Acknowledgment

This work is partially supported by Guangdong Provincial Science and technology projects#2013B040401015.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Apsingekar VR, De Leon PL. Speaker model clustering for efficient speaker identification in large population applications. IEEE Transactions on Audio, Speech, and Language Processing. 2009; 17(4):848-53.

[2] De Leon PL, Apsingekar V. Reducing speaker model search space in speaker identification. In biometrics symposium 2007 (pp. 1-6). IEEE.

[3] Xiao WW, Zheng J, Hua J, Zhan E. Speaker identification based on classification sub-space

Gaussian mixture model. In international conference on image analysis and signal processing 2011 (pp. 607-11). IEEE.

[4] Sun B, Liu W, Zhong Q. Hierarchical speaker identification using speaker clustering. In proceedings of international conference on natural language processing and knowledge engineering 2003 (pp. 299-304). IEEE.

[5] Than K, Ho TB, Nguyen DK. An effective framework for supervised dimension reduction. Neurocomputing. 2014; 139:397-407.

[6] Dong A, Chao-qun R, Dan Y, Jiao W. Speaker recognition method based on PSOA clustering and KMP algorithm. Chinese Journal of Scientific Instrument. 2013; 6: 015.

[7] Matza A, Bistritz Y. Speaker recognition with rival penalized EM training. In IEEE international workshop on machine learning for signal processing 2011(pp. 1-6). IEEE.

[8] Xing Y, Tan P. A novel SVM Kernel with GMM super-vector based on bhattacharyya distance clustering plus within class covariance normalization. In international conference on natural computation (ICNC) 2015 (pp. 47-51). IEEE.

[9] Shan ZY, Yang YC. Universal background model reduction based efficient speaker recognition. Journal of Zhejiang University (Engineering Science). 2009; 6: 003.

[10] Huaqiao X, Jianbin Z, Enqi Z, Yang W, Jia H. Speaker recognition based on speaker model clustering. Computer Engineering and Applications. 2014, 50(2):133-6.

[11] Patnaik M, Mathew A, Gill MS, Pradhan D. FastRec: a fast and robust text independent speaker recognition system for radio networks. In international conference on recent advances and innovations in engineering (ICRAIE) 2014 (pp. 1-7). IEEE.

[12] Wang HL, Han JQ, Zheng GB. K-L divergence based model clustering method for fast speaker identification. Pattern Recognition and Artificial Intelligence. 2010; 23(6): 856-61.

Jing Zhang was born in November 24; 1977.Zhang graduated from Shenyang University of Technology in July of 2000 and received a bachelor of engineering degree. Then entered Guangdong University of Technology and received a master of engineering degree in July of 2003 and received a doctor of engineering degree in 2012. The major field of study is embedded system and pattern classification. Her job is teaching and works in Guangdong University of Foreign Studies. She has presided over the education of Humanities and Social Science Fund Project (10YJCZH220), Guangdong Provincial Science and Technology Projects (2013B040401015). She published a number of papers, which are included by EI OR SCI.
Email: ha_go@163.com

Xiaomei Chen was born in January 22, 1973. She graduated from South China Normal University in July of 1995 and received a Bachelor of Science degree. Then in June of 2002 from Jiangxi University of Finace and Economics received a master of Management degree. The major field of study is algorithms classification. Her job is teaching and works in Guangdong University of Foreign Studies.