A Characteristics Grouping Algorithm in DHMM Speech Recognition

Jing Zhang

Cisco School of Informatics, Guangdong University of Foreign Studies, Guangzhou, 510006, China

*ha_go@163.com

***Abstract*-The paper introduced a speech feature grouping algorithm for the speech recognition system based traditional Markov in accordance with the large computation of the traditional hidden Markov model and the Viterbi algorithm as well as the Gaussian mixture distribution probability. For the speech characteristic parameters, clustering was executed by K-Means algorithm on the basis of the first and second segmentation, and then obtained the grouped characteristic parameters and the parameters to be grouped, and the speech samples can be divided into different characteristic group according to these two parameters. On this basis, a grouping training algorithm was proposed by using the redundant, which improved the accuracy of grouping the speech characteristic by clustering algorithm. Compared with the traditional HMM method, the amount of calculation can be reduced more than 60% in the case of ensuring the speech recognition rate.**

Keywords- Hidden Markov Model; Speech Characteristic Grouping; Segmentation Mean; K-Means Clustering; Redundancy Factor

I. INTRODUCTION

In the traditional HMM speech recognition system, and for the isolated word, the classic Viterbi algorithm perfectly solved the decoding problem with iterative approach in the mathematics. But in the speech recognition systems, for the large vocabulary in the practical application of acoustic decoding, the amount of computation that the algorithm required is considerable. Assuming the vocabulary that a speech recognition system can recognize is 1000, then to establish a model for each word, and assuming they have the same number of states N, in order to facilitate the estimation and calculation, these models must be connected to a large model. So, the state number of the large model is 1000 times larger than that of original model. Because the orders of computation that the Viterbi algorithm required is N2T (T is the number of input speech frame), the computation amount of Viterbi algorithm of large models increased four orders than that of the original phoneme model. Moreover, according to the experiment, it was found that the computation was mainly used for calculating the multi-dimensional (mixture) Gaussian distribution probability of the obtained observation vector for each state of every frame. Therefore, the real-time requirement for the small mobile device with large vocabulary HMM speech recognition system is difficult to meet because of the limit computing capability. Contrary to the large amount of computation for the Viterbi algorithm and the Gaussian mixture distribution, the paper presented a grouping model of speech characteristics, before the matching calculation of the input speech done, the grouping judgment was executed firstly and obtained the group that the input speech belonged to, then to calculate the HMM parameters of input speech matching with that of all the speech in the same group. Then combined with the improved HMM speech recognition system, the efficiency can be improved.

II. GENERATION OF HMM GROYPING CHARACTERISTIC PARAMETERS OF SPEECH SIGNAL

An easy way to comply with the journal paper formatting requirements is to use this document as a template and simply type your text into it.

A. Generation of Grouping Characteristic Parameters

The grouping algorithm could effectively reduce the amount of computation in the recognition process, but it will also produce a certain negative impact to the recognition accuracy, and the extent is determined by the grouping accuracy. So the core problem of grouping is how to reduce the size of each grouping model library and at the same time the grouping accuracy is ensured, that is, a reasonable standard of judgment should be given. Here, a grouping characteristic parameters called *Fe* is promoted as the grouping judgments.

Assuming the MFCC characteristic parameters of W_n is the matrix $N_n * M_n$. If all lines were put to the rear of first row, then W_n can be denoted by using a row vector with dimension of $N_n * M_n$, assuming the vector is C_O vector. To cluster the n vectors corresponding to the n words using the K-means algorithm, then the entire category each word belonged to and the clustering centre of each category will be achieved. But because there are still close relevance between the clustering results and the initial cluster centres when using the K-means clustering algorithm, so the clustering centre achieved by only one time clustering can not be used as a grouping characteristic parameters. The different initial cluster centres should be used as much

DOI: 10.5963/JBAP0202001

as possible, and then to analyse the great deal of different clustering results thus the final clustering results can be gotten. To calculate the mean value of C_O vectors corresponding to all the words in each category, this value is namely the grouping characteristic parameters *Fe*.

Then how to analyse these clustering results achieved from different initial cluster centres? Because the words belonged to the same class have the similar number of category and should keep the accordant changes of their categories, so the analysis used in this paper was mentioned as the following method: after m times clustering, write down the category number after each word being clustered and denote it with a row vector V_n. The V_n denoted the category the word belonged to. The number of average values was expressed by E_n, the mean value of V_n, and the changes of category was represented by σ_n, the standard deviation of V_n, and then P_n, the multiplication of E_n and V_n can be used to characterize the grouping that the word belongs to. In this way, the grouping of a word can be represented by the value P_n. At this point, the representation way for a word W_n has been converted from a row vector with dimensional of $N_n * M_n$ to a multiplication P_n, then the clustering problem for n words was transformed into a problem for an n-dimensional row vector. Many clustering methods are appropriate for such row vectors with simple and small amount of data. For convenience, k-means algorithm is used to cluster the row vector in the experiment and was tested feasible and appropriate.

B. The First and the Second Sub-segments for MFCC Parameters of Speech Signal

The MFCC parameters was selected as the speech characteristic parameters as described in this papers. Based MFCC parameters, each speech generates a new parameter, that is C_o, the characteristic parameters to be divided. Then generate the grouping characteristic parameters *Fe* according to all the speech characteristics parameters to be divided C_o. Finally, all the speeches in library were grouped according to characteristic parameters. In order to get the better group effect, some following grouping training was necessary. The parameters need to be defined include the following:

To define a parameter D_o as the first sub-segments number for the MFCC parameters of speech signal;

To define a parameter D_t as the second sub-segments number for the MFCC parameters of speech signal;

To define a parameter C_O as one-dimensional array, this is used to characterize a speech signal;

To define a parameter C_n as the elements number of C_O, that is $C_n = D_o * D_t * 24$;

To define a parameter S_n as the grouping number of speech signal;

To define a parameter H_n as the number of words in the speech database;

To define a parameter F_e as the matrix of $S_n * C_n$, of which each row is grouped characteristic parameters.

Assume that the speech signal MFCC parameters has 12 frames, and D_o, the number of first segment dividing is 3 and for the second, D_t is 12, that is $C_n = 3*2*24 = 144$. The average segmentation was taken for every time of segments dividing. Fig. 1 shows the first and second segments dividing process for the MFCC parameters of speech signal.

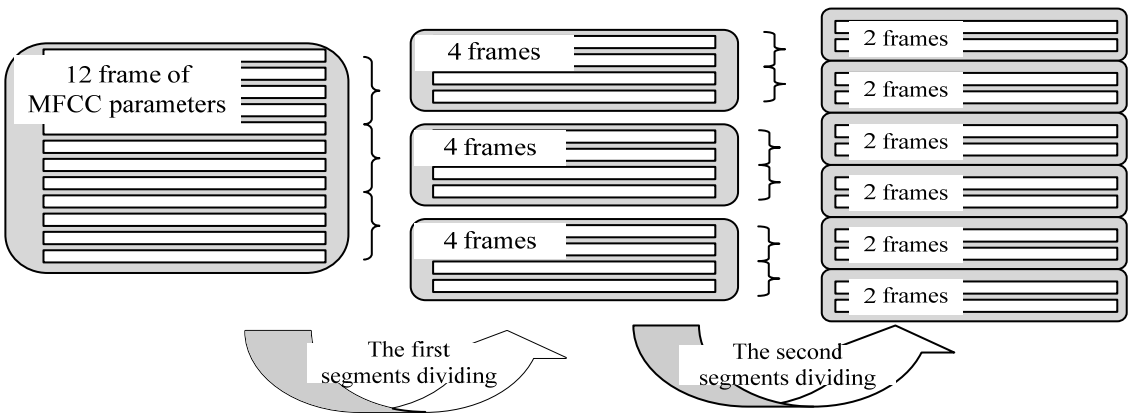

Fig. 1 Progress of first and second segments dividing for the speech MFCC parameters

DOI: 10.5963/JBAP0202001

C. The Generation of C_o Parameter

After the second segments dividing, cluster all the MFCC parameters in each short segment into one category by traditional K-means algorithm, finally, integrate the average of the MFCC parameters in each segment to a one-dimensional array, i.e. the parameters C_o. Take the hypothesis in Function 2.2 as an example, the generation process of C_o parameters as shown in Fig. 2.

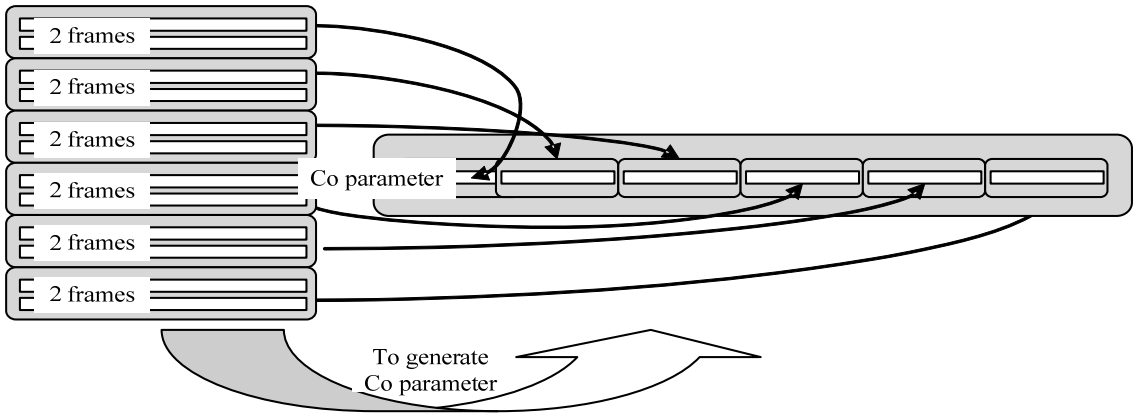

Fig. 2 Generation process of C_o parameters

D. The Generation of F_e Parameters

The characterization of the speech achieved by C_o parameters shown in Figure 2 is defined as characteristic parameters of words to be divided. In speech database, all the speech characteristic parameters formed a matrix, in which, each row is a characteristic parameters to be divided of a word.

The forming of grouping characteristic parameters: firstly, for the characteristic parameters of all words in the speech database, 300 times clustering and grouping was done by K-means algorithm. For the *i*-th word, to record the grouping number that each clustering and grouping belonged to with a cell unit $temp(i)$, which embodied the changes of grouping that the characteristic parameters that the *i*-th word belongs to, so, $temp(i)$ can be called second-order parameters to be divided of the *i*-th word. Then to cluster and group the second-order parameters to be divided of all the words by using K-means algorithm, and obtain the words that each group contained. Finally, to average the parameters to be divided of all words in each group, that was the grouping characteristic parameters of that group.

Assuming $Y_m(1 \le m \le 300)$ is a congregation, the congregation recorded the group that each speech belongs to when the C_O parameters was clustered by K-means algorithm in the *m*-th. That is the Equation (1).

$$Y_m = index(Kmeans_m(CO, S_n), Y_m = \{Y_{m1}, Y_{m2}, \cdots Y_{mHn}\},\ (1 \le m \le 300) \tag{1}$$

Assume

$$Temp_i = \{y_{1i}, y_{2i}, \cdots y_{300i}\}\ \ C''_O = \{Temp_1, Temp_2, \cdots Temp_{Hn}\}$$

where, C''_O is the second order parameter to be divided.

Assume $Z_p = find(index(Kmeans(C''_O, Sn)) == p)$, $1 \le p \le Sn$, then got Equations (2) and (3).

$$Temp'_p = Kmeans(C_o(Z_p), 1) \tag{2}$$

$$F_e = \{Temp'_1, Temp'_2, \cdots, Temp'_{sn} \tag{3}$$

where, F_e is the grouping characteristic parameters.

Assume H_n, the number of words in the current speech database is 12, and S_n, the grouping number of speech signal is 4. Fig. 3 roughly showed the formation process of F_e parameters.

DOI: 10.5963/JBAP0202001

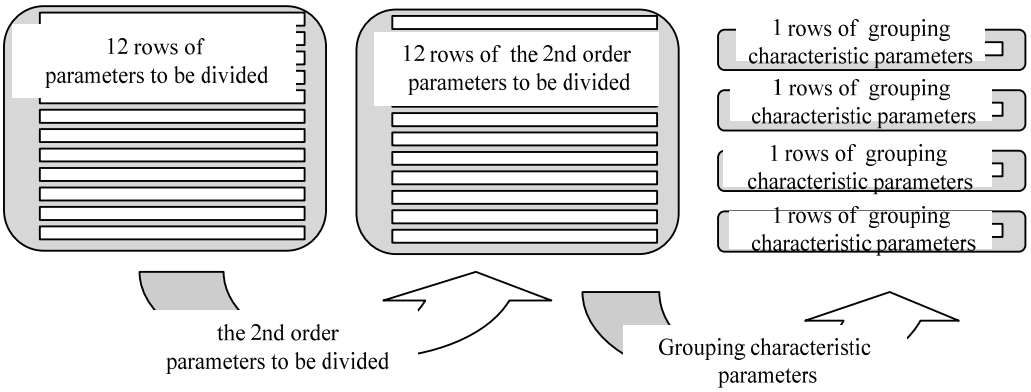

Fig. 3 Generation progress of F_e parameters

III. USE REDUNDANCY TO ENSURE THE GROUPING ACCURACY

Since the grouping accuracy directly impacts the speech recognition accuracy, grouping accuracy must be high enough. The method as following mentioned was used to add a redundant word in the appropriate category: input the collected samples into the system, and determine their category through the grouping characteristic parameters. If the judgment is correct, then input the next sample; otherwise add the word corresponding to current sample in the class.

To match each sample parameter with the distance between the parameter and the sample centre successively with the DTW algorithm, and then set the minimum distance as the target group and test whether the target group contains the words expressed by the input characteristic parameters, if it is, then the classification is correct, else then add the word to target group.

After the redundancy was added, the number of words contained in each category will be increased. Assume the number of words a category that contained is K_p, and the number of all words is *n*, the grouping number of words is *m*. The Cross-grouping coefficient $\alpha = \sum_{p=1}^{m} K_p$. The smaller α, the less cost in identification process. When $\alpha \to 1$, then it means each group approximately contained all the words, this phenomenon is called false grouping, which should be avoided.

During the experiment, the grouping characteristic parameters was used for initial grouping, and then redundancy was added by imputing the sample, where the α is about 0.39, while the grouping accuracy reached at 99% or more.

The specific process is shown in Fig. 4.

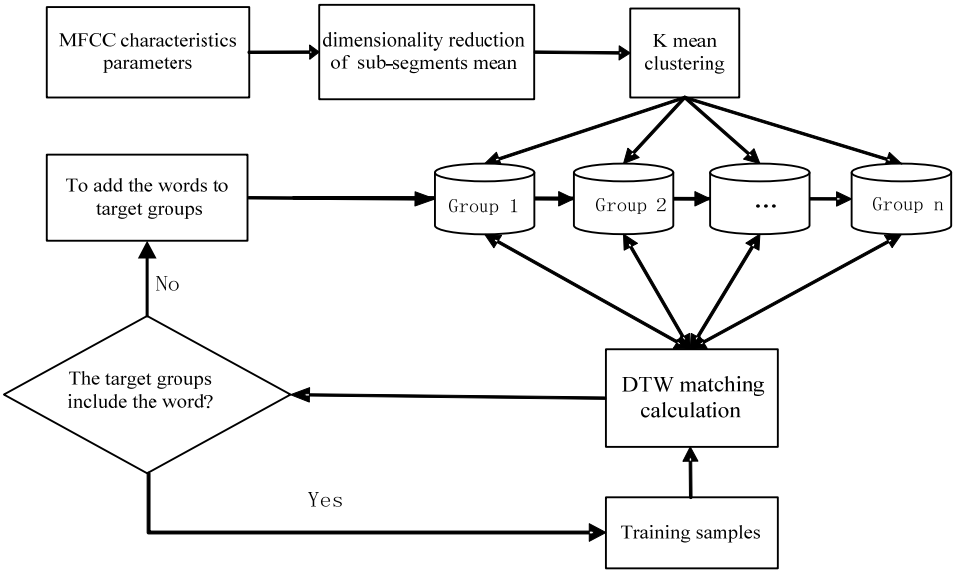

Fig. 4 Training Algorithm of Using Redundancy

After the first grouping, then $\sum_{p=1}^{m} K_P = n$ so for the first grouping, $a = \frac{1}{m}$. After several grouping training, all the number of

DOI: 10.5963/JBAP0202001

words contained in each group K_P increased, now, $K_P \to n$, that a is closed to 1.

When the vocabulary amount contained in each group is similar to that of speech database, according to the grouping principle, it is known that, at this time, the calculation of system and the times of matching for the input speech to be carried out is not significantly reduced, and then the efficiency and real-time of the system improved little. Therefore, the variation of crossing grouping coefficient a needs to be noted when grouping training was being carried. When a was large, the training should be stopped.

IV. EXPERIMENTAL RESULTS AND ALGORITHM EVALUATION

A. Experimental Results

The interface of grouping statistics and the recognition results is as shown in Fig. 5 and Fig. 6. Fig. 5 shows the proportion of average recognition rate, average grouping accuracy and the average recognition computation when the speech database was divided according to different grouping numbers. Fig. 6 shows the variety of charts generated and various types of parameters and recognition results of the speech files in the current path.

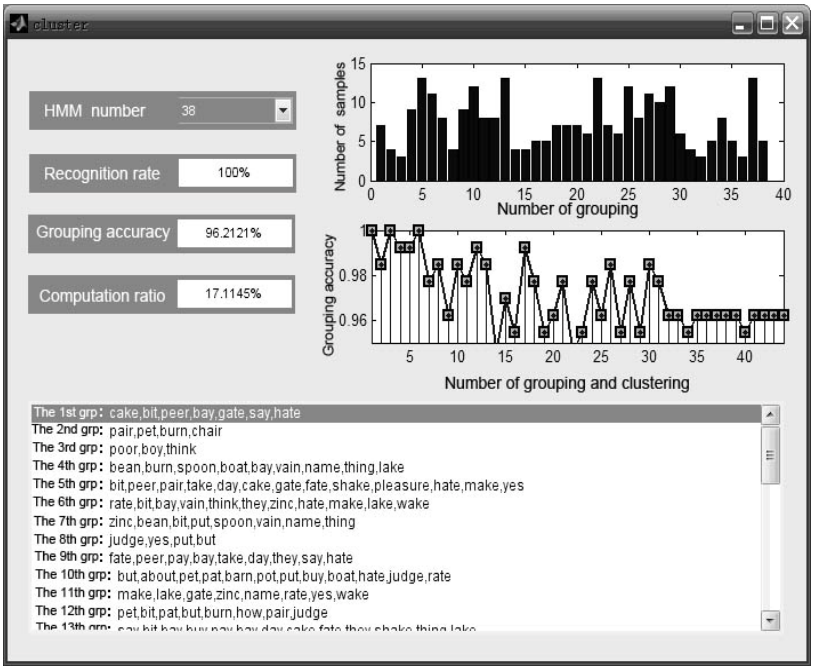

Fig. 5 Grouping statistics interface

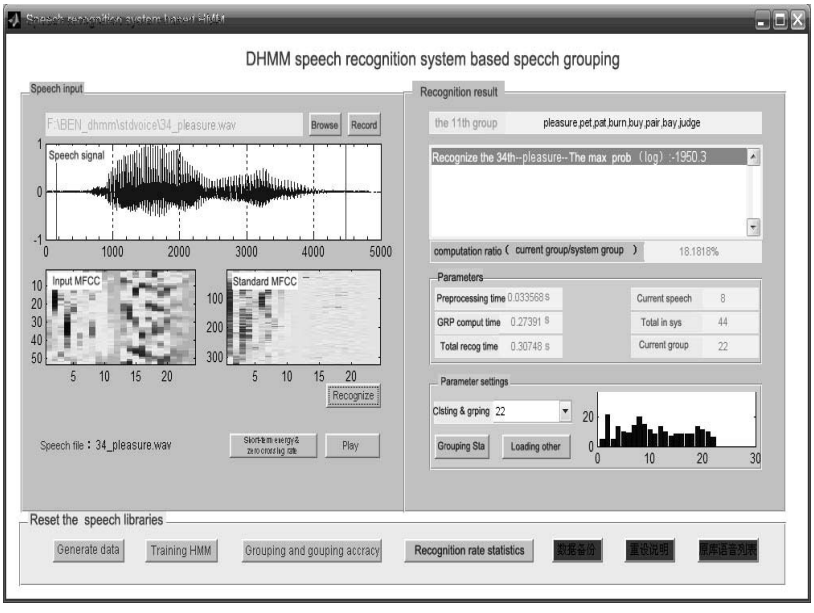

Fig. 6 Speech recognition results

DOI: 10.5963/JBAP0202001

B. The Algorithm Evaluation

In this experiment, the number of words to be identified is 44, and the number of clustering and grouping is 3, and the total test samples are seven groups. Table 1 shows the initial grouping accuracy rate (due to space constraints, here only given 6 word groupings), and Table 2 shows the grouping after adding redundancy, while Table 3 shows the system recognition efficiency.

TABLE I THE INITIAL GROUPING ACCURACY RATE

order	word	Sample1	Sample 2	Sample 3	Sample 4	Sample 5
1	bean	1	1	1	1	1
2	bit	1	1	2	2	2
3	pet	3	3	3	3	3
4	pat	3	3	3	3	3
5	barn	3	3	3	3	3
6	pot	3	3	3	3	3
7	born	1	3	3	3	3
8	put	3	3	3	3	3
9	spoon	1	1	3	1	1
10	but	3	3	3	3	3
11	burn	1	3	3	3	3
12	about	3	3	3	3	3
13	bay	1	1	1	1	1
14	buy	3	3	3	3	3
15	boy	1	1	3	3	3
16	boat	1	3	3	3	3
17	how	3	3	3	3	3
18	peer	1	1	2	1	1
19	pair	1	1	3	3	3
20	poor	3	3	3	3	3
21	pay	1	1	1	1	1
22	bay	1	1	1	1	1
23	take	1	1	1	1	1
24	day	1	1	1	1	1
25	cake	1	1	1	1	1
26	gate	1	1	1	1	1
27	fate	1	1	1	1	1
28	vain	1	1	1	1	1
29	think	1	1	1	1	3
30	they	1	1	1	1	1
31	say	3	1	1	1	3
32	zinc	1	1	1	1	1
33	shake	1	1	1	1	1
34	pleasure	1	3	1	1	1
35	hate	1	1	1	1	1
36	chair	1	1	1	2	1
37	judge	3	2	3	2	2
38	make	1	1	1	1	1
39	name	1	1	1	1	1
40	thing	1	1	1	1	3
41	rate	1	1	1	1	1
42	lake	1	1	1	1	1
43	yes	2	2	2	2	2
44	wake	1	1	1	1	1
grouping accuracy				0.8796		

TABLE II THE GROUPING AFTER ADDING REDUNDANCY

group	the word contained in clustering and grouping
group 1	Pet Pat Spoon Peer Pair Poor Cake Gate Say Shake Pleasure Hate Chair Yes Wake make
group 2	Bean Bit Born Burn Boy Boat Pay Bay Gate Say Vain Think They Zinc Name Lake Wake
group 3	Barn Pot Put But About Bay Buy How Hate Take Day Fate Udge Make Thing Rate Wake Shake

DOI: 10.5963/JBAP0202001

TABLE III THE RECOGNITION TIME WITH DIFFERENT NUMBER OF MODEL LIBRARY

Item / Model groups	Recognition time(s)	Recognition rate	The time savings
2	0.587432793	99.50%	17.54%
3	0.471864087	99.00%	33.76%
4	0.393995398	99.50%	44.69%
6	0.30373318	99.00%	57.36%
7	0.284968684	97.50%	60.00%
8	0.230544148	96.50%	67.64%
10	0.192162851	99.00%	73.03%
11	0.184630581	97.50%	74.08%
12	0.184794962	99.00%	74.06%
13	0.147031782	98.00%	79.36%
14	0.144813982	97.50%	79.67%
16	0.117887713	98.00%	83.45%
18	0.104763784	98.00%	85.29%

It can be seen from Classification results shown in Table I, due to the introduction of segmentation mean and grouping clustering algorithm, the clustering results had better stability, of which the grouping for 10 words had no error. During crossing-group training the 10 words would not be involved in the re-grouping, which will help reduce the $\sum_{p=1}^{m} K_P$ and achieve good grouping results. The TableII showed the grouping with increased redundancy.

Where the α is about 0.39.

Table III showed the recognition time and recognition rate for the improved system by selecting the different number of models group, when the model group number was 1, the system degenerated to the unmodified state, then the model grouping did not affect the recognition rate. In the case of ensuring the recognition rate of 99.50% of the system, the number of model group was taken as 5 and the recognition time is 45.44% to the unimproved system, which achieved the goal of improving the system efficiency.

V. CONCLUSIONS

The core problem of grouping speech recognition is how to reduce the size of each grouping model library when the accuracy rate of grouping is guaranteed at the same time. A good grouping characteristic parameters can get better initial clustering results, which makes the α be smaller after adding redundancy so as to achieve a better group result and to improve the real-time ability of speech recognition.

ACKNOWLEDGMENT

This work is supported by the ministry of education of humanities and social science project #10YJCZH220.

REFERENCES

[1] Jiang Guanxing, Wang Jianying. An improved method of testing speech end point, Journal of Micro-computer information, 22(2006)138-139.

[2] Yeqing Yun, Jiang Jia. The improved algorithm based speech MFCC feature, Journal of Wuhan University of Technology, 29(2007) 150-152.

[3] Feng Yun, Jing Xinxing, Ye Mao. Improving the MFCC Features for Speech Recognition, Journal of COMPUTER ENGINEERING & SCIENCE, 31(2009)146-148.

[4] Li Dongdong, Wu Zhaohui, Yang Yingchun. The recognition method of emotional speech clustering for the Speaker based base-band, Journal of Pattern Recognition and Artificial Intelligence, 22(2009)139:140.

[5] Zhang Jie, Huang same, Xiao-Lan Wang. Speech recognition, hidden Markov model state number of the selection principle, Journal of Computer Engineering and Applications,1(2009)67 -69.

[6] Seide F. Peng Yu, Chengyuan Ma, Chang E. Vocabulary-independent search in spontaneous speech. Proceeding of ICASSP, 2004.

DOI: 10.5963/JBAP0202001

[7] FengQin Yang,Changhai Zhang, GeBai. A novel Genetic Algorithm Based on Tabu Search for HMM Optimization. Proceeding of Fourth International Conference on Natural Computation, (2008)57-61.

[8] Zhao Hui, Gu Ya-qiang, Tang Chao-jing. Speech Recognition Method of Dual-mode Based Multiplication HMM, Journal of Computer Engineering, 36(2010)7-9.

[9] Yu Mei-juan, Ma Xi-rong. An Improvement of dynamic gesture recognition based HMM. Journal of Computer Science, 38(2011) 251-252.

[10] Zhang Jian-ping, Li Ming, Suo Hong-bin. The Application of Long speech features in the speaker recognition, Journal of Acoustics, 35(2010)267-269.

Jing Zhang was born in Dalian City, Liaoning Province, 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. Now Zhang is studying for a doctor's degree in Guangdong University of Technology. The major field of study is digital signal procession and artificial intelligence.

Her current job is TEACHING and works in Guangdong University of Foreign Studies. Previous publications are: The HMM Speech Recognition Algorithm Based Speech Feature Clustering (Information-An International Interdisciplinary Journal, 2012), Speech Material Recognition Technology on an Objective Evaluation System for the Rhythm of English Sentences (Advances in computer science, Intelligent system and environment, 2011), Objective Evaluation System of Speech Quality Based HMM (Applied Mechanics Research, 2011), A Speech Recognition System Based Improved Algorithm of Dual-template HMM (Procedia Engineering, 2011), A Speech Recognition Method Based Clustering Neural Network Integration (ICECE, 2011), a design of embedded multimedia player based on WINCE (Procedia Engineering, 2011), Software copyright (DTW model for speech recognition system Based on the optimization template, 2011), Software copyright (HMM Speech Recognition System Based Characteristics Grouping, 2011), Software copyright (Speech recognition system Based on java platform, 2012).

DOI: 10.5963/JBAP0202001