Speech Recognition Approach Based on Speech Feature Clustering and HMM

XinGuang Li
Guangdong University of Foreign Studies, Guangzhou, 510006, China
Email: lxggu@163.com

MinFeng Yao and JiaNeng Yang
Guangdong University of Foreign Studies, Guangzhou, 510006, China
Email: {hammersons, tizziyang}@gmail.com

Abstract—**The paper presents a Segment-Mean method for reducing the dimension of the speech feature parameters. K-Means function is used to group the speech feature parameters whose dimension has been reduced. And then the speech samples are classified into different clusters according to their features. It proposes a cross-group training algorithm for the speech feature parameters clustering which improves the accuracy of the clustering function. When recognizing speech, the system uses a cross-group HMM models algorithm to match patterns which reduces the calculation by more than 50% and without reducing the recognition rate of the small vocabulary speech recognition system.**

Index Terms--**HMM, Speech Feature Parameters, Segment-Mean, K-Means Clustering, Model Cross-group**

I. INTRODUCTION

Breakthrough progress has been made in studies of speech recognition techniques in recent years. And these techniques have been applied for business purposes. A regular speech recognition system can be, in general, divided into four parts, namely, speech pretreatment, feature extraction, speech recognition and semantic understanding. Speech pre-processing aims at noise elimination and endpoint detection with signal processing technology, while feature extraction is designed to extract the feature parameters of the input speech. And speech recognition refers to the process in which one explores the distance or probability between the vector sequence of the unknown speech feature and each speech sample, as well as the most analogous type, when matching the unknown speech features with different training patterns. Finally, semantic understanding means giving a grammatical and semantic analysis of the result so as to obtain a proper one which conforms to grammatical rules[1].

With the development of speech recognition techniques, Dynamic Time Warping (DTW), Vector Quantization (VQ), Hidden Markov Model (HMM) and Artificial Neural Networks (ANN) have successively been applied to the speech recognition system in a successful way, which substantially promotes its development[2]. Researchers continuously try to improve these algorithms and they have made many innovative achievements. Povey D. etc. describe an acoustic modeling approach in which all phonetic states share a common Gaussian Mixture Model structure, and the means and mixture weights vary in a subspace of the total parameter space. This style of acoustic model allows for a much more compact representation and gives better results than a conventional modeling approach, particularly with smaller amounts of training data[3]. Feng HongWei and Xue Lei introduced a new method for speech recognition which combined the hidden HMM and the algebraic neural networks. And the simulation result show the algorithm is better than the traditional algorithm in convergence speed, robustness and recognition rate improvement[4].Almost all present day large vocabulary continuous speech recognition (LVCSR) systems are based on HMM[5].Hidden Markov Model works very well in time series signal processing for its double stochastic process[6]. This statistical model has been used extensively and successfully in speech recognition systems. In this paper we attempt to study on the speech recognition system based on the integration of HMM and a new speech feature clustering model. Compared with the speech recognition system based on single HMM, the new hybrid model effectively improves the recognition speed. It provides a new reference method for the small devices to run speech recognition applications which meet the requirements of real-time system.

II. SPEECH FEATURE PARAMETERS

The system digitized analog signals according to Nyquist sampling frequency. The sampling frequency is set to 8 KHz. After pre-processing the input speech by using several algorithms, Mel-frequency cepstral coefficient (MFCC) is calculated.

A. *Pre-processing Input Speech*

Due to the pronunciation mechanism, speech signal has the characteristics of Attenuation of high frequency components. In this paper, pre-emphasis digital filter is used to enhance high frequency components. The filter flattens the spectrum of the signal and makes it possible to calculate the spectrum with the same SNR throughout the band. The filter is defined as

doi:10.4304/jcp.7.9.2269-2275

$$H(z) = 1 - az^{-1}, 0.9 \le \alpha \le 1 \quad (1)$$

Where α is the coefficient of the pre-emphasis filter, generally 0.92 or 0.94.

In this paper, we use dual-threshold comparison to detect endpoints of input speech. It combines short term energy and short term zero-crossing rate of the signal, that makes the detection become more accurate. Also the dual-threshold comparison can effectively exclude from the silent segments of the noise and enhance system performance in real-time speech signal processing.

B. MFCC Calculation

So far, the most applicable ones of speech feature coefficients are linear predictive cepstral coefficients (LPCC) and MFCC. In this paper, we use MFCC as feature parameters. The MFCC analysis consists of four steps.

- Perform a fast Fourier transform on the input speech signal.

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{n}nk}, k = 0,1,2,..., N-1 \quad (2)$$

Where $x(n)(n = 0,1,2,..., N-1)$ is one frame discrete speech signal; N is the length of the frame; $X[k]$ is the complex sequence of N points in the frame. Then we find the mold of $X[k]$, and get the signal amplitude spectrum $|X[k]|$

- Model the frequency axis by the Mel-scale. The Mel–frequency f_{mel} can be computed from the frequency f as follows:

$$f_{mel}(f) = 2595 \cdot \lg(1 + \frac{f}{700\,Hz}) \quad (3)$$

- Calculate the amplitude spectrum of the triangle filter to the filtered signal output:

$$F(l) = \sum_{k=f_0(l)}^{f_h(l)} w_l(k)|X[k]| \qquad l = 1,2,..., L \quad (4)$$

Where

$$w_i(k) = \begin{cases} \dfrac{k - f_o(l)}{f_c(l) - f_o(l)} , & f_o(l) \le k \le f_c(l) \\ \dfrac{f_h(l) - k}{f_h(l) - f_c(l)} , & f_c(l) \le k \le f_h(l) \end{cases} \quad (5)$$

$$f_0(l) = \frac{o(l)}{\left\lfloor \frac{f_s}{N} \right\rfloor}, f_h(l) = \frac{h(l)}{\left\lfloor \frac{f_s}{N} \right\rfloor}, f_c(l) = \frac{c(l)}{\left\lfloor \frac{f_s}{N} \right\rfloor} \quad (6)$$

Here $F(l)$ denotes the filtered signal output. $w_i(k)$ is the filter coefficients of the corresponding filter. $o(l)$, $h(l)$ and $c(l)$ represent the lower frequency, higher frequency and center frequency on the actual frequency axis of the corresponding filter. f_s denotes the sampling rate; L denotes the number of the filters.

- Perform the discrete cosine transform on the logarithm of the filter-bank energies and append first order differentials. Then we obtain the expression of the MFCC

$$M(i) = \sqrt{\frac{2}{n}} \sum_{l=1}^{L} \log F(l) \cos[(1 - \frac{1}{2})\frac{i\pi}{L}], \quad i = 1,2,..., Q \quad (7)$$

Here $M(i)$ denotes the MFCC parameters[7][8]. Q denotes the order of MFCC. In this paper, Q is set to 24.

III. Speech Recognition System Based on Speech Feature Clustering and HMM

In the traditional HMM isolated word speech recognition system, the classical Viterbi algorithm which uses forward iteration of the mathematical methods solves the hidden Markov model decoding problem perfectly. However, the required computation is still very impressive while the Viterbi algorithm is used in a large vocabulary speech recognition system. Suppose a large vocabulary speech recognition system can recognize 500 words. And we establish a model for each word. Assuming they have the same number of states, we connect these models into one big model. In this way, the state number of the big model is $N \times 500$. As the Viterbi algorithm needs N^2T (T denotes the frames of input speech) orders of magnitude for the computation, compared to the computation of Viterbi algorithm in original model, the one in the big model is increased to 3 orders of magnitude. Moreover, the experiments find that the computing time is mainly used to calculate the mixture Gaussian probability distribution of the observation sequence[9]. In order to reduce the computational complexity, this paper presents a hybrid model based on speech feature clustering and HMM, which experimentally confirmed validity.

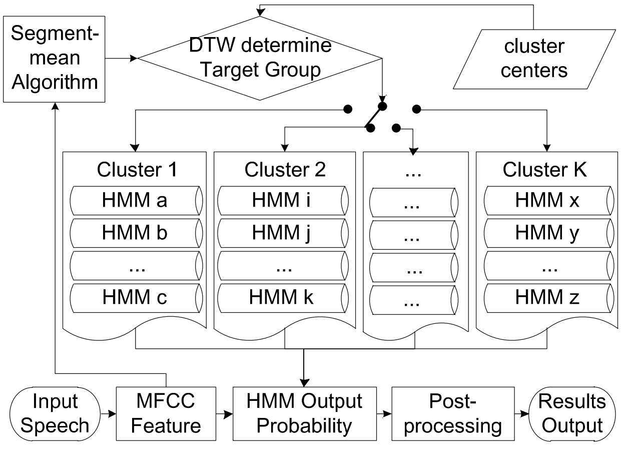


Figure 1. Structure of speech recognition system based on speech feature clustering and HMM.

Fig. 1 shows that when recognizing, the system calculates the feature of input speech and determines its cluster group k at first. While Viterbi decoding is under progress, it only calculates the HMM parameters in the k-group. In the case of appropriate cluster group the system will save a considerable amount of computation.

A. Speech Feature Dimension Reduction and Cluster Cross-grouping Model

Speech feature parameters must be structured before the clustering. This paper proposes a segment-mean algorithm to reduce the dimension of the speech feature parameters, so that they can keep the same orders and frame length. After the dimension reduction, the proposed cluster cross-grouping model effectively improves the accuracy of the speech feature parameters clustering function.

1) Segment-mean Algorithm

When studying on clustering algorithm in the field of speech recognition, most of the literatures take it as a means of pattern classification[10][11]. In this paper, the improved speech feature clustering model will have the words with similar acoustic characteristics clustered into the same group. When recognizing, it only calculate the HMM parameters in the selected group. After a number of experiments we find that the group accuracy results of directly using traditional K-means clustering algorithm to cluster the speech feature are not optimistic.

The segment-mean algorithm fragments the speech feature parameters into segments with the same dimension. Define the speech feature parameters as $s(K,J)$. Where K denotes the orders of the MFCC parameters; J denotes the number of fragmented frames. Assumes T is the number of frames before fragmented. Then fragment the speech feature parameters into N segments can be:

$$M(i) = s(K,J), \quad J = [\frac{T}{N}(i-1)+1],..., [\frac{T}{N} i] \tag{8}$$

$M(i)$ represents the i-th segment of the fragmented speech feature parameters. The value of N is set to the statue number of the HMM.

After fragmenting the speech feature parameters into average segments, we continue fragment $M(i)$ into M average segments (The value of M is set to the observation sequence number of the HMM). The calculations of child segments see (8). The mean of each child segments is given by $\overline{M(i)_k}$, $k = 1,2,..., M$. Merge all the mean of the child segments into a matrix. The matrix denotes the speech feature parameters output after dimensionality reduction. It is defined as $\overline{s(K,T)}$. The size of $\overline{s(K,T)}$ is $MN \times K$.

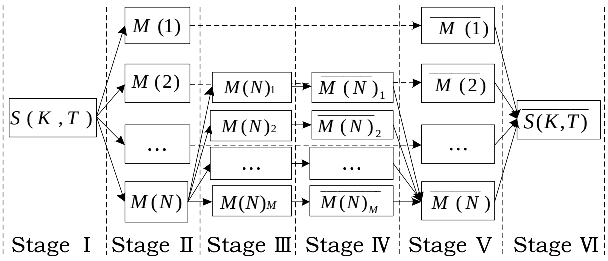


Figure 2. Schematic diagram of the segment-mean algorithm.

The total numbers of parameters in Fig. 2 are shown in Table Ⅰ. The segment-mean algorithm turns the size of feature parameters matrix from $T \times K$ to $MN \times K$. That is to say the algorithm successfully removes the frame length T from the matrix. This means, the matrix (dimensionality reduction) keeps the same size after the segment-mean calculation. And the size of feature parameters matrix is determined for K (the orders of the speech feature parameters), N (size of the segment) and M (size of the child segment).

TABLE I.
NUMBER OF PARAMETERS IN EACH STAGE

Stage	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	Ⅵ
Matrix size	TK	$\frac{T}{N}K$	$\frac{T}{NM}K$	$\frac{T}{MN}\frac{MN}{T}K$	MK	MNK
Number	TK	TK	TK	KMN	KMN	KMN

2) Cluster Cross-grouping Model

This paper presents a new cluster cross-grouping model to enhance the performance in the field of speech feature clustering.

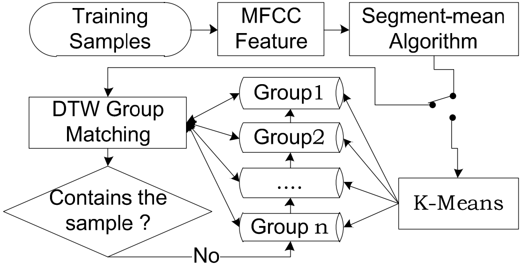


Figure 3. Schematic diagram of cluster cross-grouping.

As shown in Fig. 3, Cluster Cross-grouping consists of three steps:

- Cluster the features of the training speech samples using K-means clustering algorithm.
- Calculating the distances between the training speech samples and the cluster centers using dynamic time warping (DTW) algorithm. For each sample, the minimum distance determines its target group.
- Check whether the target group contains the training sample. If included, the classification is correct; else the word will be added to the target group.

Set the cluster group number to K, the number of vocabulary to N. The number of words in the k-th group is $S_k, k = (1,2,..., K)$. After the first time clustering, we have $\sum_{k=1}^{K} S_k = N$. Define the clustering coefficient of cluster cross-grouping model as

$$\varphi = \frac{\sum_{k=1}^{K} S_k}{KN} \tag{9}$$

Easy to see that after the first time clustering, we get $\varphi = \frac{1}{K}$. A question must be asked is, after training the cluster group for the second time using cross-grouping algorithm, whether $\sum_{k=1}^{K} S_k$ will tend to or equal to $K \times N$, or φ tends to 1. If φ tends to 1, that means the feature clusters are degraded to ungrouped.

B. System Design and Implementation

The system improvable ratio is key indicator of the improving analysis. Experimental Analysis of system improvable ratio which bases on the pattern matching stage shows that the proposed method is reasonable and effective.

1) System Improvable Analysis and HMM Parameters Grouping

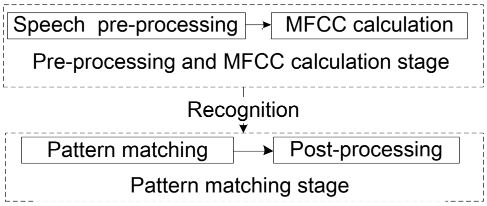


Figure 4. Speech recognition system based on single HMM.

There are two main stages when the system recognizing words, as shown in Fig. 4: Pre-processing and MFCC calculation stage; Pattern matching stage. The pattern matching stage takes up most of the recognizing time. Table Ⅱ shows the recognizing time of ten different samples 'apple' in the system based on single HMM. According to the statistics given in Table Ⅱ, pattern matching time account for average 96.4% of the total recognition time.

TABLE II.
RECOGNIZING TIME OF TEN DIFFERENT SAMPLES 'APPLE'

id	*Pre-processing Time (sec)*	*Pattern Matching Time (sec)*	*Recognition Time (sec)*	*Pattern Matching Time /Recognition Time*
1	0.03	0.93	0.96	96.4%
2	0.03	0.77	0.80	96.4%
3	0.03	0.76	0.79	96.5%
4	0.03	0.92	0.95	96.6%
5	0.03	0.86	0.89	96.6%
6	0.03	0.75	0.78	96.4%
7	0.03	0.74	0.77	96.5%
8	0.03	0.71	0.74	96.4%
9	0.03	0.69	0.72	96.4%
10	0.02	0.46	0.48	95.8%

The improved speech feature clustering model has the words with similar acoustic characteristics clustered into the same group. Combining this with the HMM Parameters Grouping model, we achieve the speech recognition system based on speech feature clustering and HMM which greatly improves the system efficiency.

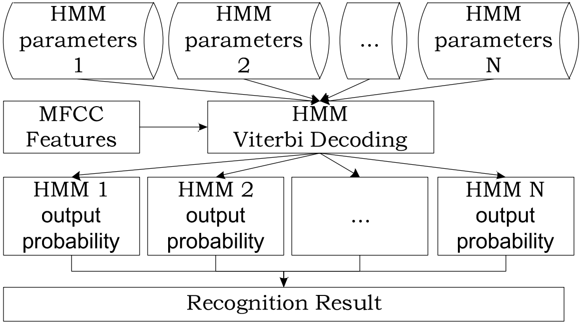


Figure 5. Speech recognition system based on single HMM.

As Fig. 5 shown, when using Viterbi algorithm to do decoding operations, all the model parameters must be involved in the computation. Assume the number of system vocabulary is n. Then the number of HMM parameters is n. When recognizing a word, each output probability is calculated by Viterbi algorithm and involved the total n HMM parameters. Because each isolated word has a unique HMM parameter with corresponding. We are able to have the words in the feature cluster groups Mapped to the corresponding HMM parameters. Therefore we achieve the HMM parameters grouping model as Fig. 6 shown.

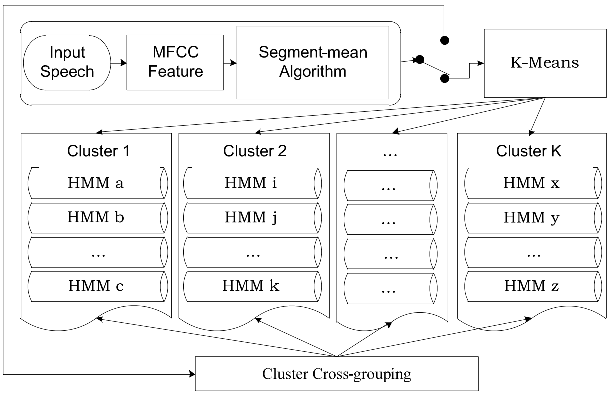
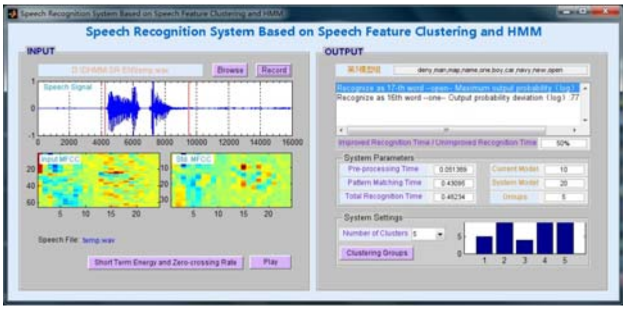


Figure 6. HMM parameters grouping model.

As the feature clustering algorithm is good in grouping performance, the number of the HMM parameters in the cluster group is always less than or equal to the number of system vocabulary. Also, the improved speech feature clustering model ensures a high grouping accuracy rate. Hence, this paper proposes to combine the feature clustering model and HMM to form a hybrid model--speech recognition system based on speech feature clustering and HMM (as Fig. 1 shown).

2) Implementation

According to the hybrid model, we complete the system on the Matlab platform. The main interface of the system shows in Fig.7. The state number of HMM is set to 8[12].



Speech Recognition System Based on Speech Feature Clustering and HMM

INPUT

Browse Record

Speech File: temp.wav

Short Term Energy and Zero-crossing Rate Play

OUTPUT

deny,man,map,name,one,boy,car,navy,new,open

Recognize as 17-th word --open-- Maximum output probability (log)

Recognize as 16th word --one-- Output probability deviation (log) :77

Improved Recognition Time / Unimproved Recognition Time 50%

System Parameters

Pre-processing Time 0.051369

Pattern Matching Time 0.43095

Total Recognition Time 0.48234

Current Model 10

System Model 20

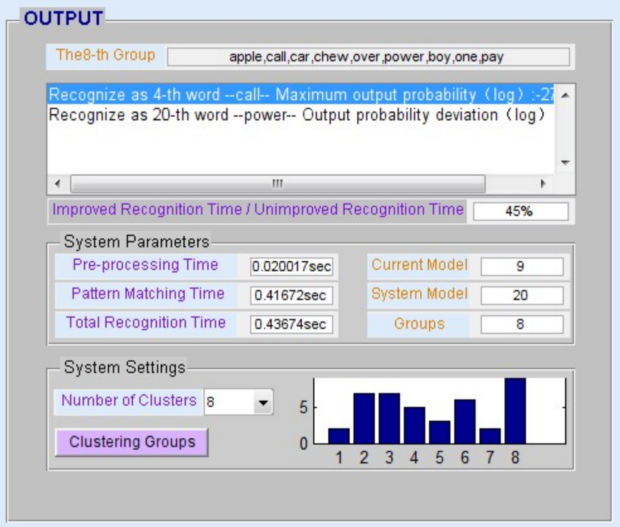
Groups 5

System Settings

Number of Clusters 5

Clustering Groups

Figure 7. Main interface of the system.



OUTPUT

The8-th Group apple,call,car,chew,over,power,boy,one,pay

Recognize as 4-th word --call-- Maximum output probability（log）:-2

Recognize as 20-th word --power-- Output probability deviation（log）

Improved Recognition Time / Unimproved Recognition Time 45%

System Parameters

Pre-processing Time 0.020017sec

Pattern Matching Time 0.41672sec

Total Recognition Time 0.43674sec

Current Model 9

System Model 20

Groups 8

System Settings

Number of Clusters 8

Clustering Groups

Figure 8. Recognition and data analysis module.

As Fig. 8 shown, in the recognition and data analysis module we can get the recognition results and the clustering items. Users can customize the number of cluster groups and compare different recognition cases.

IV. Experiments

In order to verify the validity of the new model, we analyze the segment-mean algorithm, cluster cross-grouping model and the hybrid model (speech feature clustering and HMM) through several experiments.

The test samples are composed of 1,000 speech samples recorded by 10 individuals. The number of system vocabulary is 20.

A. The Effect of Segment-mean Algorithm

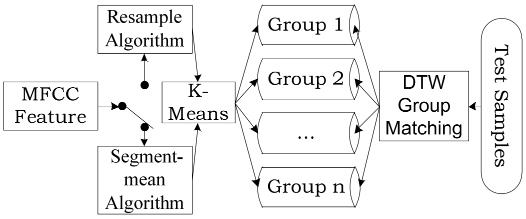


Figure 9. Schematic diagram of speech feature parameters clustering experiment.

According to the procedures in Fig. 9, for the speech feature parameters whose dimension are not reduced, we choose the resample algorithm to wrap their length into the same size. As the length of the frames is generally 30 to 80 when using 8 KHz sampling frequency, we set the resample length to 50. DTW algorithm is used to calculate the Euclidean distance between the test speech samples and the cluster centers. And the target group is the one who has the minimum Euclidean distance. Check whether the target group contains the input sample. If included, the classification is correct.

TABLE III.
CLUSTERING RESULTS

Number of clusters	*Direct MFCC Clustering Accuracy*	*Segment-mean clustering Accuracy*
1	1.0000	1.0000
2	0.7000	0.8400
3	0.4000	0.8900
4	0.3800	0.9400
5	0.3000	0.8200
6	0.2000	0.8450
7	0.3500	0.9150
8	0.2000	0.8850
9	0.1500	0.8750
10	0.1000	0.9000
11	0.1000	0.9050
12	0.1500	0.8800
13	0.1000	0.8900
14	0.1000	0.9000
15	0.1000	0.8850
16	0.1000	0.9100
17	0.0500	0.9050
18	0.0500	0.9200
19	0.1000	0.9100
20	0.0500	0.9050

The experimental data shows in Table Ⅲ, the average clustering accuracy of the untreated MFCC feature is 23.40%. Such a low accuracy cause that we cannot use this kind of clusters in the system. Experimental data also shows that, the average clustering accuracy which uses the Segment-mean Algorithm to the MFCC feature is much better with a value of 89.60%. Although the segment-mean clustering accuracy relatively in a higher level, the cluster still cannot be used in the system. Because that its classification performance will still reduce the recognition rate of the system in a certain degree. So we should continue try others methods to improve the clustering accuracy of the speech feature clustering.

B. Analysis of Cluster Cross-grouping Model

Take experiment according to the schematic diagram of cluster cross-grouping as illustrated in Fig. 3. The experiment result is shown in Table Ⅳ. In the experiment, the number of the clusters is set to 3. There are 10 groups test samples which contain the speech recorded by 10 individuals, 200 words/times. The words in each cluster are given in Table Ⅴ. And the clustering accuracy is 89.00%.

TABLE IV.
CLUSTERING RESULTS

id ***word***	1	2	3	4	5	6	7	8	9	10
Apple	2	2	2	2	2	2	2	2	2	2
Baby	1	1	1	1	1	1	1	1	1	1
Boy	3	3	3	3	3	3	3	3	3	3
Call	2	2	3	2	1	2	1	2	2	2

Car	2	2	2	2	2	2	2	2	2	2
Chew	2	1	2	2	1	1	1	2	2	1
Deny	2	1	2	2	1	1	2	2	2	2
Dress	2	2	2	2	2	1	2	2	2	2
Man	1	1	1	1	1	1	1	2	2	2
Many	1	1	1	1	1	1	1	1	1	1
Map	2	2	1	2	2	2	2	2	2	2
Movie	1	1	1	1	1	1	1	1	1	1
Name	1	1	1	1	1	1	2	2	1	1
Navy	1	1	1	1	1	1	1	1	1	1
New	1	1	1	1	1	1	1	1	1	1
One	3	3	3	3	1	1	3	3	3	3
Open	3	3	3	3	3	3	3	3	3	3
Over	3	3	3	3	3	3	3	3	3	3
Pay	1	1	1	1	1	1	1	2	2	1
Power	2	2	2	2	2	2	2	2	2	2

TABLE V.
GROUPS OF THE SEGMENT-MEAN MFCC CLUSTERING (3 GROUPS)

	The Words in the Cluster Groups
Group1	Baby Chew Man Many Movie Name Navy New Pay
Group2	Apple Call Car Deny Dress Map Power
Group3	Boy One Open over

The clustering results are in good stability after using the segment-mean algorithm. We can see from Table Ⅳ, there are 10 words in the test that do not occur any grouping error.

Therefore, these 10 words will not be re-grouping while using the cluster cross-grouping algorithm. This will help to reduce the value of $\sum_{k=1}^{K} S_k$ .And the cross-group will achieve better results. From the result in Table Ⅵ, after the train using the cluster cross-grouping algorithm, the total number of words in each group is $\sum_{k=1}^{K} S_k = 36$. And $K \times N = 3 \times 20 = 60$. So the clustering coefficient of cluster cross-grouping model is $\varphi = \frac{\sum_{k=1}^{K} S_k}{KN} = 0.6$. Hence, cluster cross-grouping model is good in performance.

TABLE VI.
TRAIN RESULT OF USING THE CLUSTER CROSS-GROUPING ALGORITHM

	The Words in the Cluster Groups
Group1	Baby Chew Man Many Movie Name Navy New Pay Apple Boy Call Deny Dress Map One
Group2	Apple Call Car Deny Dress Map Power Boy Chew Man Name New One over Pay
Group3	Boy One Open over Call

After using the cluster cross-grouping algorithm to retrain the speech feature clusters, the average clustering accuracy raises to 98.75%. The accuracy is 99.50% when the number of clusters is set to 3 which greatly improved compared to the previous 89.00%. Fig. 10 shows the accuracy in different number of the clusters.

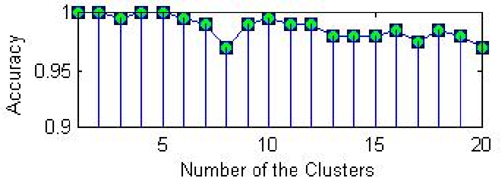


Figure 10. Accuracy of cross-group training.

Thus, after using the segment-mean algorithm and the cluster cross-grouping algorithm, the accuracy of speech feature clustering is raises to above 98%.

C. Improved Effectiveness Analysis of the Speech Recognition System

TABLE VII.
RECOGNITION TIME AND RECOGNITION RATE UNDER THE NUMBER OF DIFFERENT HMM PARAMETERS CLUSTERS

the number of HMM parameters clusters	***Recognition Time (sec)***	***recognition rate***	***improved recognition time / unimproved recognition time***
1	0.712402641	99.50%	100.00%
2	0.587432793	99.50%	82.46%
3	0.471864087	99.00%	66.24%
4	0.393995398	99.50%	55.31%
5	0.323681682	99.50%	45.44%
6	0.30373318	99.00%	42.64%
7	0.284968684	97.50%	40.00%
8	0.230544148	96.50%	32.36%
9	0.256837057	97.00%	36.05%
10	0.192162851	99.00%	26.97%
11	0.184630581	97.50%	25.92%
12	0.184794962	99.00%	25.94%
13	0.147031782	98.00%	20.64%
14	0.144813982	97.50%	20.33%
15	0.135196623	98.00%	18.98%
16	0.117887713	98.00%	16.55%
17	0.108788616	97.00%	15.27%
18	0.104763784	98.00%	14.71%
19	0.104496738	97.50%	14.67%
20	0.092861028	97.00%	13.03%

As Table Ⅶ shown, when the number of HMM parameters clusters is set to 1, the system is degraded to unimproved, so that the recognition rate is not affected by the cluster cross-grouping algorithm at all. In the case of ensuring that the recognition rate is above 99.50%, and when the number of HMM parameters clusters is set to 5, the improved recognition time account for 45.44% of the unimproved recognition time.

V. SUMMARY

Through the analysis of the speech recognition system based on single HMM, the problem that small devices cannot meet the requirements of real-time system for the enormous computation is pointed out. To solve this problem, a new speech recognition system based on speech feature clustering and HMM is proposed. The main techniques of the system are as follows:

- A new segment-mean algorithm for reducing the dimension of the speech feature parameters;
- The cluster cross-grouping model is proposed;
- HMM parameters clustering.

The results of experiment indicate that, after improving the system by using the method which this paper proposed, the improved recognition time account for less than 45.44% of the unimproved recognition time. Therefore, the purpose of improving the system efficiency is achieved.

ACKNOWLEDGMENT

This work is supported by GDSF（2011B031400003）, GDNSF (915104200100017) and GDSF (2008B080701007).

REFERENCES

[1] Wang Xianbao1, Chen Yong, Tang Liping, “Speech recognition research based on MFCC analysis and biomimetic pattern recognition”, Computer Engineering and Applications, vol.47 No.12, 2011, pp.20-22.

[2] Mehryar Mohria, Fernando Pereirab and Michael Rileya, “Weighted finite-state transducers in speech recognition”, Computer Speech & Language, vol.16, No.1, Jan. 2002, pp.69-88.

[3] Povey D., Burget L., Agarwal M., “Subspace Gaussian Mixture Models for speech recognition”, 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Mar. 2010, pp. 4330-4333.

[4] Feng HongWei, Xue Lei,” Application of speech recognition system based on algebra algorithm and HMM”, Computer Engineering and Design, Vol. 31, No.24, Dec. 2010, pp.5324-5327.

[5] Mark Gales, Steve Young, ” The application of hidden Markov models in speech recognition”, Foundations and Trends in Signal Processing, vol. 1 No.3, Jan. 2007, pp.195–304.

[6] Rabiner L.R., “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” Proceedings of the IEEE, vol. 77, No.22, Feb. 1989, pp. 257-286.

[7] Ye Qingyun and Jiang Jia, “Improved Extraction Algorithm for MFCC Feature,” Journal of WuhanUniversity of Technology, vol. 29(5), May. 2007, pp. 150-152.

[8] Feng Yun, Jing Xinxing and Ye Mao, “Improving the MFCC Features for Speech Recognition,” COMPUTER ENGINEERING & SCIENCE, vol. 31(12), 2009, pp. 146-168,doi: 10. 3969/ j. issn. 10072130X. 2009. 12. 042.

[9] Yuan Jun, “The Viterbi Algorithm Optimization and Application on Continuous Speech Recognition Base on HMM,” Electronic Technology, vol. 2, 2001, pp. 48-51.

[10] Yu Xiangdong, Suo Xiuyun and Zhai Jianren, “Speech Recognition Based on Fuzzy Clustering,” Fuzzy Systems and Mathematics, vol. 16(01), Mar. 2002, pp. 75-79.

[11] Li Dongdong, Wu Zhaohui and Yang Yingchun, “Speaker Recognition Based on Pitch-Dependent Affective Speech Clustering,” Pattern Recognition and Artificial Intelligence, vol. 22(01), Feb. 2009, pp. 139-140.

[12] Zhang Jie, Huang Zhitong and Wang Xiaolan, “Principle of Selection of States Number of HMM in Speech Recognition and its Analysis,” Computer Engineering and Applications, vol. 36(01), 2009, pp. 67-69.

XinGuang Li Hunan Province, China. Birthdate: Jan, 1963. Circuit and System Ph.D., graduated from School of Electronics and Informatics, South China University of Technology; with research interests in artificial intelligence.

He is a professor of Cisco School of Informatics Guangdong University of Foreign Studies.

MinFeng Yao Henan Province, China. Birthdate: Nov, 1977. Graduated from Leeds University with a Master Degree in Computer Science; research interests in artificial intelligence.

He is a lecturer of Cisco School of Informatics, Guangdong University of Foreign Studies.

JiaNeng Yang Guangdong Province, China. Birthdate: May, 1987. bachelor's degree in computer science and technology. And research interests in artificial intelligence and machine learning.

He is a candidate for Master of business management in School of management Guangdong University of Foreign Studies.