American Journal of Computational Mathematics, **2013, 3, 37-42**
http://dx.doi.org/10.4236/ajcm.2013.31005 Published Online March 2013 (http://www.scirp.org/journal/ajcm)

English Sentence Recognition Based on HMM and Clustering

Xinguang Li[1], Jiahua Chen[1], Zhenjiang Li[2]
[1]Cisco School of Informatics, Guangdong University of Foreign Studies, Guangzhou, China
[2]School of Business Adminstrtion, South China University of Technology, Guangzhou, China
Email: lxggu@163.com

Received December 2, 2012; revised January 7, 2013; accepted January 18, 2013

ABSTRACT

For English sentences with a large amount of feature data and complex pronunciation changes contrast to words, there are more problems existing in Hidden Markov Model (HMM), such as the computational complexity of the Viterbi algorithm and mixed Gaussian distribution probability. This article explores the segment-mean algorithm for dimensionality reduction of speech feature parameters, the clustering cross-grouping algorithm and the HMM grouping algorithm, which are proposed for the implementation of the speaker-independent English sentence recognition system based on HMM and clustering. The experimental result shows that, compared with the single HMM, it improves not only the recognition rate but also the recognition speed of the system.

Keywords: English Sentence Recognition; HMM; Clustering

1. Introduction

The pauses between English words can simplify speech recognition. Because the endpoint detection of a word (*i.e.* detecting the starting point and the end point of the word) is relatively easy, and the Coarticulation effect between words can be reduced to the minimum. In addition, generally the word pronunciation is more serious, because there must have pauses between words which make less fluent reading. In view of the above reasons, many techniques can be used for the English word speech recognition system [1].

Compared with English word, more feature data and more complex changes in pronunciation make the English sentence speech recognition more difficult. Firstly, English sentence has a larger vocabulary and no obvious pause between words with pronunciation. That is to say, there is no clear boundary between sub-words. Secondly, every word pronunciation in English sentence is usually more natural, and associated language pronunciation is more casual than isolated word pronunciation, thus the coarticulation effect is more serious. Furthermore, affected by the context, in the process of English pronunciation, rhythm, intonation, stress and speed in English sentence may be different, even the same speaker at different times or in different environment, the prosodic features are different.

As a mainstream technology for large-vocabulary speaker-independent continuous speech recognition system, the Hidden Markov Model (HMM) [2-5] has achieved considerable success. Analyzing the short-term energy of speech signal and extracting the speech feature with a frame length, this paper takes Markov modeling on the whole sentence [6,7]. Model training uses a training set recorded by many speakers and the statistical theory is used to resolve the differences between the individual and the whole, so as to make the speaker independent single sentence Markov modeling robust. When recognizing speech, the system uses Viterbi algorithm to decode and find out the correct recognition result. Using Markov modeling on single sentence can describe the correlation of the words within each sentence. Under the condition of sufficient training speech, the speaker independent small statement English sentences modeling can be achieved with a high accuracy. However, HMM needs prior statistical knowledge of speech signal and has weak classification decision ability and other problems, including the computational complexity of the Viterbi algorithm and mixed Gaussian distribution probability. These shortcomings make it difficult to further improve the recognition performance of the single HMM [8].

Most of the literatures [9-14] in the field of speech recognition improve clustering algorithm within HMM and take them as the method of pattern classification, to optimize the model parameters estimation, but the effect for sentence recognition was not ideal. For English sentences with a large amount of data and complex pronunciation changes, the shortage of HMM is more apparent,

making recognition time longer. In order to effectively improve the recognition efficiency, this paper, on the basis of the single HMM, attempts to integrate clustering algorithm with HMM and apply to the English sentence recognition. According to the characteristics of English sentences and the similarity between them, the English sentences data set is divided into several groups, each of which consists of some sentences with similar phonetic feature. So when recognize an English sentence, there is no need for all the sentences on Viterbi decoding, just to calculate the HMM parameters within the selected group which the input speech belongs to. In the case of appropriate clustering groups, the system will save a considerable mount of calculation, and the recognition performance can be greatly improved. This is not only to provide a new reference method for speech recognition in small device applications which meet the requirement of real-time, but also to lay the foundation of speech recognition for a new English sentence evaluation system.

2. Whole Design Process

As shown in **Figure 1**, first to pretreat the input speech signal, including pre-emphasis, frame processing, window adding and endpoint detection. Then extract the speech feature parameters MFCC and reduce the dimensionality of MFCC by segment-mean algorithm. The dynamic time warping (DTW) algorithm is followed to determine the speech feature clustering group *K*. Then calculate the HMM parameters within Group *K* and finally output the recognition results with post-processing.

3. Core Algorithm

3.1. Segment-Mean Algorithm

As K-means clustering algorithm has the iterative characteristics with randomly selected sample point, coupled with the higher dimensionality of speech feature parameters, so the stability of clustering results is relatively poor. For this reason, this article explores the segment-mean algorithm for dimensionality reduction of speech feature parameters, as shown in **Figure 2**.

Fragmenting the speech feature parameters into segments with the same dimension, the Segment-Mean algorithm consists of four steps:

1) 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:

$$M(i)=S(K,J),J=\left[\frac{T}{N}(i-1)+1\right],\cdots,\left[\frac{T}{N}i\right] \quad (1)$$

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

2) 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 the above formula.

3) The mean of each child segments is given by $\overline{M(i)}_k$,

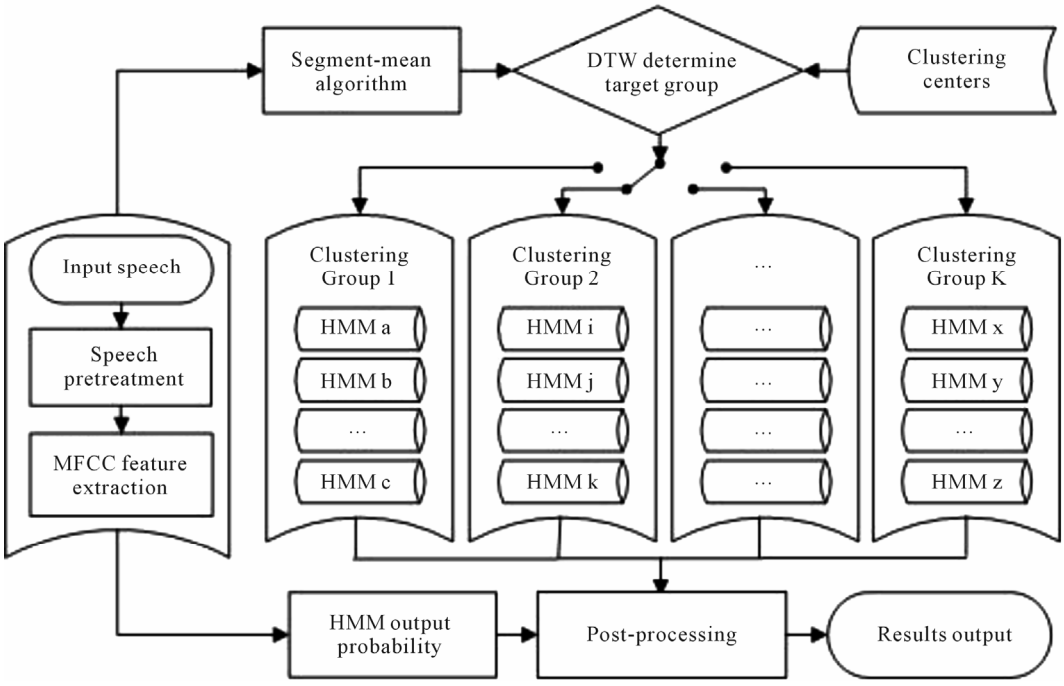

Figure 1. The frame diagram of speech recognition based on HMM and clustering.

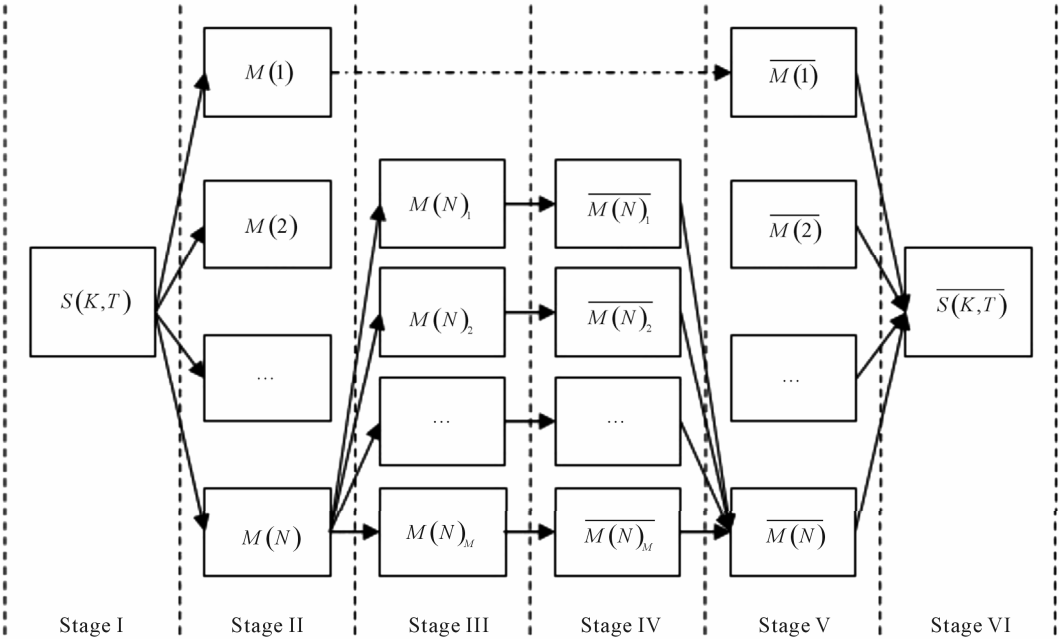

Figure 2. Segment-mean algorithm for dimensionality reduction of voice feature coefficients.

$k = 1, 2, \cdots, M$.

4) 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 $K \times M \times N$.

The total numbers of parameters in **Figure 2** are shown in **Table 1**. The segment-mean algorithm turns the size of feature parameters matrix from $T \times K$ to $K \times M \times N$. 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).This makes speech with different length can be structured as a matrix of the same size, which largely facilitates the implementation of speech feature clustering algorithm.

3.2. Clustering Cross-Grouping Algorithm

In order to further enhance the performance in the field of speech feature clustering, this paper presents a new secondary training method-clustering cross-grouping algorithm.

As shown in **Figure 3**, the clustering cross-grouping algorithm consists of three steps:

1) Cluster the features of the training speech samples using K-means clustering algorithm.

2) Calculate 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.

3) Check whether the target group contains the training sample. If included, the classification is correct; else the sentence will be added to the target group.

3.3. HMM Grouping Algorithm

In the recognition system based on single HMM, 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 sentence, each output probability is calculated by Viterbi algorithm within n HMMs respectively. Because each isolated sentence has a unique HMM parameter with corresponding. We are able to have the sentences in the feature clustering groups mapped to the corresponding HMM parameters. Therefore we achieve the clustering grouping HMM model as **Figure 4** shown.

As the clustering cross-grouping algorithm is good in grouping performance, the number of the HMM parameters in the clustering 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 integrate the feature clustering model and HMM to form a hybrid model—English sentence recognition system based on clustering and HMM (as **Figure 1** shown).

4. Experimental Results and Analysis

In order to verify the validity of the proposed model, the recognition rate and time on the single HMM and the

Table 1. The parameter table of voice feature coefficients processing segment-mean algorithm.

Stage	I	II	III	IV	V	VI
Size of matrix	$T \times K$	$\left(\frac{T}{N}\right) \times K$	$\left(\frac{T}{NM}\right) \times K$	$\left(\frac{T}{NM} \times \frac{1}{\frac{T}{NM}}\right) \times K$	$M \times K$	$(M \times N) \times K$
Number of parameters	$T \times K$	$T \times K$	$T \times K$	$K \times M \times N$	$K \times M \times N$	$K \times M \times N$

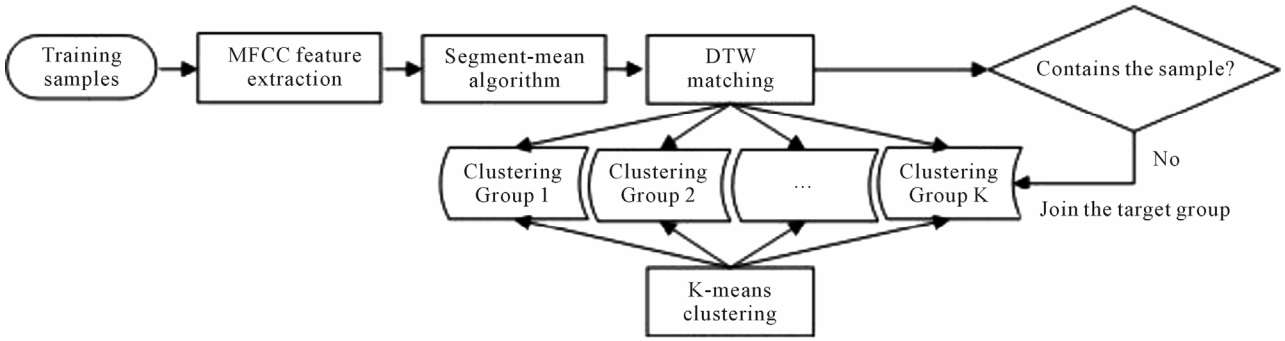

Figure 3. Clustering cross-grouping algorithm.

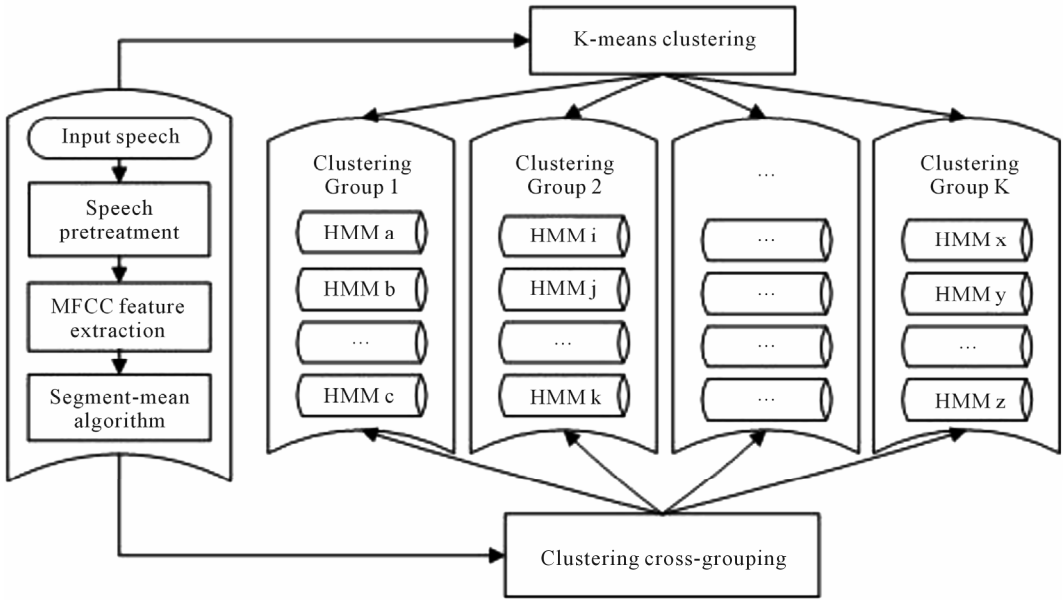

Figure 4. HMM grouping algorithm.

hybrid model based on HMM and clustering were compared in speaker-independent English sentence recognition systems. The number of system vocabulary is 30. This experiment selects 30 different English sentences as standard sentences, 900 English sentences recorded by 30 individuals as training samples and 450 English sentences recorded by 15 individuals as test samples.

Take Sentence 1 "Can I have breakfast served in my room?" as an example to show the recognition rate and time in different recognition methods.

For example, comparing the sentences from Student 1, the system gives recognition results as shown in **Figure 5**.

No matter whether of the single model or the hybrid model the recognition results are correct, but the former total recognition time is 1.85 seconds, the latter total recognition time was 1.41 seconds, only 76.22% of the former. That is to say, the recognition time decreases, and the system efficiency is improved.

Compare the sentences from all the students (student 1 to 15), the results are show as **Table 2**. The experiments show that the recognition rate of the single HMM and the

Figure 5. The recognition result of sentence 1 from student 1.

Table 2. The recognition time table of sentence 1 (15 samples) in different recognition methods.

Recognition time (s) \ No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Single model	1.44	1.52	1.86	1.64	1.44	1.45	1.41	1.37	1.94	1.79	1.62	1.64	1.40	1.59	1.52
Hybrid model	1.23	1.25	1.30	1.25	1.26	1.24	1.26	1.26	1.31	1.29	1.25	1.24	1.22	1.29	1.23

Table 3. The overall recognition performance table in different methods.

Item \ Model	Single model	Hybrid model
Average recognition rate	96.89%	99.78%
Average recognition time (s)	1.7687	1.2248

proposed model are both 100%; but the former average recognition time is 1.5753 seconds, the latter average recognition time of 1.2587 seconds, only 79.90% of the former, so as to improve the recognition efficiency.

Table 3 is the overall recognition performance comparison in different recognition methods. The experimental results show that, compared with the English sentence recognition system based on single HMM, the average recognition rate of the English sentence recognition system based on HMM and clustering (the proposed model) increased by 2.89%, while the average recognition time accounted for only 69.25% of the former, improving the system efficiency.

5. Conclusion

On the basis of the English sentence recognition method and the traditional HMM speech recognition technology, an improved algorithm based on HMM and clustering is proposed for the implementation of the English sentence recognition system. The experimental results show that the improved system in accordance with the method proposed in this paper, not only improve the recognition rate of the system, but also reduce the amount of computation of the system (*i.e.*, the recognition time), to achieve the goal of improving system performance. But how to determine the clustering groups to further improve the recognition efficiency and applied to more large-scale English sentence recognition is in need of further research.

6. Acknowledgements

This work was financially supported by the Guangdong Science and Technology Foundation (2011B031400003).

REFERENCES

[1] M. Zhu, X. Wen, J. Huang and L. Zhou, "Computer Speech Technology," Revised Edition, Beijing University of Aeronautics and Astronautics Press, Beijing, 2002.

[2] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceeding of the IEEE*, Vol. 77, No. 2, 1989, pp. 257-286. doi:10.1109/5.18626

[3] L. R. Rabiner and B. H. Juang, "Fundamentals of Speech Recognition," 1st Edition, Prentice Hall, Upper Saddle River, 1993.

[4] Q. He and Y. He, "An Extension of MATLAB Programming," 1st Edition, Tsinghua University Press, Beijing, 2002.

[5] J. Han, L. Zhang and T. Zheng, "Speech Signal Processing," 1st Edition, Tsinghua University Press, Beijing, 2004.

[6] L. Lippmann, E. Martin and D. Paul, "Multi-Style Training for Robust Isolated-Word Speech Recognition," *Proceedings of the* 1987 *IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, 6-9 April 1987, pp. 705-708.

[7] L. R. Rabiner, J. G. Wilpon and F. K. Soong, "High Performance Connected Digit Recognition Using Hidden Markov Model," *Proceedings of the* 1988 *IEEE International Conference on Acoustics, Speech and Signal Processing*, New York, 11-14 April 1988, pp. 119-122.

[8] Y. Bao, J. Zheng and X. Wu, "Speech Recognition Based on a Hybrid Model of Hidden Markov Models and the Genetic Algorithm Neural Network," *Computer Engineering & Science*, Vol. 33, No. 4, 2011, pp. 139-144.

[9] S. K. Bhatia, "Adaptive K-Means Clustering," *Proceed-*

ings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, Miami, 12-14 May 2004, pp. 695-699.

[10] A. Likas, N. Vlassis and J. Verbeek, "The Global K-Means Clustering Algorithm," *Pattern Recognition*, Vol. 36, No. 2, 2003, pp. 451-461. doi:10.1016/S0031-3203(02)00060-2

[11] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman and A. Y. Wu, "An Efficient K-Means Clustering Algorithms Analysis and Implementation," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 7, 2002, pp. 881-892. doi:10.1109/TPAMI.2002.1017616

[12] X. Ma, Y. Fu and J. Lu, "The Segmental Fuzzy c-Means Algorithm for Estimating Parameters of Continuous Density Hidden Markov Models," *Acta Acustica*, Vol. 22, No. 6, 1997, pp. 550-554.

[13] L. Zhao, C. Zou and Z. Wu, "The Segmental Fuzzy Clustering Algorithm for Estimating Parameters of the VQ-HMM," *Journal of Circuits and Systems*, Vol. 7, No. 3, 2002, pp. 66-69.

[14] H. Wang, L. Zhao and J. Pei, "Equilibrium Modified K-Means Clustering Method," *Journal of Jilin University* (*Information Science Edition*), Vol. 24, No. 2, 2006, pp. 172-176.

