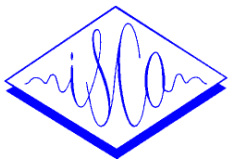

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

SEMI-CONTINUOUS SEGMENTAL PROBABILITY MODELING FOR CONTINUOUS SPEECH RECOGNITION

Jiyong Zhang, Fang Zheng, Mingxing Xu, Ditang Fang

Center of Speech Technology, State Key Laboratory of Intelligent Technology and Systems,

Department of Computer Science & Technology, Tsinghua University, Beijing, 100084, China

zhangjy@sp.cs.tsinghua.edu.cn

10.21437/ICSLP.2000-69

ABSTRACT

In this paper the design of semi-continuous segmental probability models (SCSPMs) in large vocabulary continuous speech recognition is presented. The tied Gaussian densities are trained using data from all states of all utterances while the mixture weights are estimated using data from the state being trained individually. The SCSPMs tie all the densities of all states from all Speech Recognition Units (SRUs) to form a shared *pdf* codebook, thus the number of Gaussian densities is greatly reduced. Several pruning methods are reviewed and then a new pruning criterion is proposed in order to reduce the number of tied mixture Gaussian densities while there is only a small subset of mixture Gaussian densities with larger tying weights. Our preliminary experiments show that the SCSPM incorporated with the pruning techniques can lessen the size of model storage and speed up the system with little degradation in the accuracy compared to the prior continuous model.

Keywords: semi-continuous segmental probability model (SCSPM), speech recognition unit, mixed Gaussian densities, pruning technique

1. INTRODUCTION

High accuracy large vocabulary continuous speech recognition (LVCSR) systems based on hidden Markov models (HMM)[1] have been developed in recent years. Each state of the speech recognition unit (SRU) is modeled as a mixture of elementary pdfs, for instance the Gaussians. To obtain higher accuracy, HMM-based LVCSR systems typically use continuous-density HMMs (CDHMMs), for instance, the *EasyTalk* system [2][3]

In such a system, multiple mixture Gaussian output distributions are used for each state of the SRU, and each Gaussian component must be separately evaluated in order to determine the overall likelihood without parameter tying. It will cause large scale of the acoustic model and high computation complexity. In the *EasyTalk* system, For example, the SRU (herein the syllable is used) number is 418, each SRU is divided into 6 states, and each state is described with 16 Gaussian densities, thus the total number of Gaussian densities in the system should be over 40,000. Other CDHMMs based systems have the similar embarrassment. In general, if a sufficient amount of training data is available, a large number of parameters will yield a better recognition accuracy. However, the evaluation of many thousands of elementary Gaussians during recognition time will slow down the system speed. Furthermore, if the training data are not sufficient, the accuracy of CDHMM based system will degrade greatly.

In a semi-continuous density HMMs (SCHMMs)[4], the tied Gaussians are trained using data from all states; only the mixture weights are estimated with data for the state itself, therefore far fewer data are needed to estimate a SCHMM state. SC-HMM tries all the continuous output probability densities across each individual HMM to form a shared *pdf* codebook, thus the number of mixed Gaussian densities would be greatly reduced. In a typical SCHMM based system, the adopted Gaussian density number is about 5,000.

Research on HMM distance measures has showed that the probability transition matrix contributes not so much as the observation function matrix does to the performance, so a kind of segmental probability Models (SPM) has been proposed based on the desertion of the HMM probability transition matrix with good performance, such as the mixed Gaussian continuous probability model (MGCPM).

It is interesting to incorporate the concept of semi-continuity into the SPM. Here we design the semi-continuous segmental probability model (SCSPM) for acoustic modeling in the automatic continuous speech recognition system. As some tying weights of the Gaussian densities are so small that do little contribution to the performance of the system, we ignore Gaussians with smaller weights by pruning techniques. Our experiments shows that the SCSPM with the pruning techniques can lessen the size of the model and speed up the system with little degradation in the accuracy compared to the continuous version.

In Section 2 the detail design of SCSPMs is given. First we present the outline of SCSPMs, and then show how to build a codebook with the Maximum likelihood. The formula of weight estimation is also deduced in this section. In section 3 the pruning techniques used to reduce the tying number in SCSPMs are discussed. The recognition results for SCSPMs are presented in the following section. Finally, in Section 5 the conclusion and an outlook on the further work are given.

2. DESIGN OF SCSPMs

We first give an overview of a prototypical design of SCSPMs in this section. We show the principle of SCSPM, and then two main steps adopted in the SCSPMs training procedure are enumerated in detail respectively.

2.1 Principle of SCSPMs

The SCSPM can be regarded as the combination of the mixed Gaussian density scheme and the MGCPM; here let's take an overview on it.

SCHMM (or tied mixture HMM [5]) systems use a mixture of – generally Gaussian *pdfs* – to model a state. The observation likelihood of state s for a frame vector $\vec{X}$ is given by

$$F_s(\vec{X}) = \sum_{i=1}^{N} w_{s_i} \times g_i(\vec{X}) \qquad (1)$$

where N is the size of the Gaussian set, the codebook of Gaussian *pdfs*, w_{s_i} is the weight for Gaussian i in State s and the likelihood of Gaussian i is denoted by $w_i(\vec{X})$.

MGCPMs adopt a left-to-right non skipping topology, the state transition is controlled by the high robust Nonlinear Partition (NLP) algorithm which is based on the equal feature variance sum (EFVS)[6] criterion in the training procedure while the EFVS based search or modified Viterbi algorithm [7] in the recognition procedure. The mixed Gaussian densities (MGDs) are used to describe the intra-state feature space.

MGCPMs can achieve a satisfying recognition accuracy in the continuous speech recognition, but the acoustic model is large and the problem of lacking of training data exists. Here we try to tie all the MGDs into a codebook, thus we get the SCSPMs.

A standard procedure of SCSPMs training can be illustrated in Fig. 1.

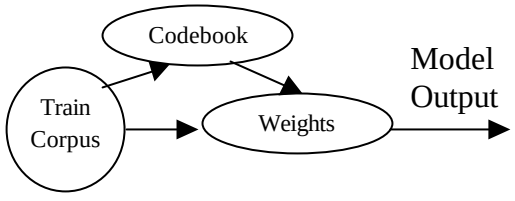

Fig. 1. The structure of SCSPMs

2.2 Codebook Generation

Each *pdf* in the codebook of the SCSPMs contains the mean vector and the covariance matrix. Two algorithms can be used to obtain the codebook of SCSPMs. One is the classification and the other is the mergence. Here we give these two algorithms respectively.

Let *N* be the expected codebook size, the classification algorithm can be described as follows.

1. Initialization. The mean vector and the variance vector are calculated from all vectors in the training corpus. The current codebook size *n* is 1.
2. The LBG algorithm [8] is used to divided size *N* codebook into a size *n**2 codebook. The current codebook size is changed to *n**2.
3. The MGD clustering is used to adjust the codebook. The object function is Equation (1) and we expect it reach its maximum.
4. If $n = N$, exit, otherwise go to step 2.

The classification algorithm ensures that the obtained codebook can reach the maximum likelihood, but the calculation complexity is very large. For example, if *N*=4,096, the last MGD clustering should adjust all the 4096 Gaussian at the same time. Another algorithm for obtaining the codebook is the mergence algorithm. The MGCPMs are consisting of Gaussian densities. We can use them as the initial codebook, and then merge them to a fixed number of codebook. The mergence algorithm can be described as:

1. Choose the nearest pair of Gaussian densities among the codebook. Here we define a simple distance between two Gaussian densities as follows:

   $$D(g_1, g_2) = \| \vec{m}_1 - \vec{m}_2 \|, \qquad (2)$$

   where g_t is the *t*-th Gaussian, and $\vec{m}_t$ the mean vector of g_t.
2. Let g_i, g_j be the two Gaussians selected by the last step, remove both g_i and g_j from the codebook, and add a new Gaussian g into the codebook with the mean vector $\vec{m}$ and the diagonal covariance vector $\vec{s}$ are such that

   $$\vec{m} = (\vec{m}_i + \vec{m}_j)/2, \quad \vec{s} = (\vec{s}_i + \vec{s}_j)/2 \qquad (3)$$

3. Let n be the current codebook size, if $n = N$, exit, otherwise go to step 1.

2.3 Weight Estimation

Having obtained the codebook, we can estimate the weights of Gaussian densities for each state of SRUs. The Maximum Likelihood Estimation (MLE) method can be adopted to make

the model reach its maximum of the likelihood of corresponding training data. We define the object function as follows:

$$h(q) = \ln(p(Z|q)) = \ln(\prod_{j=1}^{J} p(z_j|q)) = \sum_{j=1}^{J} \ln p(z_j|q), \quad (4)$$

Where Z is the output feature vector set of an acoustic model q, and $Z = \{z_j : j = 1,2,...,J\}$.

We use the following probability density functions to describe the distribution of feature vector z_j

$$p(z_j|q) = \sum_{n=1}^{N} \{w_n p(z_j|C_n)\} \quad (5)$$

where C_n is the *n*-th *pdf* in the codebook and g_n is the weight of the *n*-th *pdf*. The w_n's are such that

$$\sum_{n=1}^{N} w_n = 1 \quad (6)$$

Equation (5) is a conditional extremum , the Lagrange function is:

$$f = h(q) + l(\sum_{n=1}^{N} w_n - 1) \quad (7)$$

For each $t = 1,2,..N$, we can get the following differential equations:

$$\sum_{j=1}^{J} p(z_j|q)^{-1} \nabla_{w_t} (\sum_{n=1}^{N} w_n p(z_j|C_n)) + l = 0 \quad (8)$$

Then we can get

$$\ddot{e} \cdot w_t = -\sum_{j=1}^{J} p(z_j|q)^{-1} w_t p(z_j|C_t) \quad (9)$$

By summing the upper equation over *t* and using Equation (6), we get

$$\ddot{e} = -\sum_{t=1}^{N}\sum_{j=1}^{J} p(z_j|q)^{-1} w_t p(z_j|C_t) = -J \quad (10)$$

From (9)(10), we can get the iterative equation of w_n:

$$\hat{w}_t = J^{-1} \cdot \sum_{j=1}^{J} p(z_j|q)^{-1} w_t p(z_j|C_t) \quad (11)$$

This is the estimation equation of tying weights of each *pdf* for the acoustic model q.

Each acoustic model has *N* tying weights after the above weights estimate process. We find that some tying weights of the Gaussian densities are so small that they do little contribution to the performance of the system. In the next section we adopt some pruning techniques to reduce the number of tying weights.

3. Pruning Techniques

There are three pruning methods for SCHMMs that can be used to reduce the number of tying weights [9]:

1. Pruning by a Bigger-Weight criterion: if the weights of a given state do not exceed a certain threshold, the corresponding Gaussians are omitted.
2. Pruning by a Fixed-Number criterion: The Gaussians are sorted in the descending order according to the tying weights, only a fixed number of Gaussians with highest weights are kept for each state.
3. Pruning by a Fixed-Probability-Percentage criterion: the Gaussians with the highest weights are selected up to the point where the sum of these weights reaches a predefined percentage threshold.

Neither of the above three criteria is perfect. The first one may conflict with Equation (6). The second and third one may omit some Gaussians that have bigger weights.

As far as the training procedure of SCSPMs is concerned, the weights of each state are obtained through an iterative process. We can get the following new pruning method.

4. Pruning by a Bigger-Weight criterion only during the weight estimation process: during each iteration step, if the weights in a given state do not exceed a certain threshold, the corresponding Gaussians are omitted.

The difference between Criteria 1 and 4 is that the later one is iteratively applied, the threshold can be set to be small enough, and the sum of the weights satisfy Equation (6) at the end of the weight estimation.

4. EXPERIMENTAL RESULTS

A continuous Chinese speech corpus from 863 materials is used in the following experiments. The corpus contains 13 speakers' data and there are 520 utterances available for each speaker. All the recorded materials are obtained in a low noise environment through a close-talk noise-canceling microphone. Ten speakers' data are used as the training database for SCSPMs, the remaining part is used for testing. They are digitized at a sampling frequency of 16KHZ. A 32ms Hamming window is applied to each frame of speech. And then the cepstral coefficients derived from 16-order LPC are extracted every 16ms. We choose the Chinese syllables as the SRUs and each syllable is divided into 6 states by the NLP algorithm.

Our experiments have three parts: firstly we will do the experiment with different codebook size of SCSPMs. Secondly the different pruning methods are adopted during the SCSPMs training. At last a contrast experiment of SCSPMs and MGCPMs is done to illustrate the performance of SCSPMs.

4.1 Experiment of codebook size

The codebook size is an important factor for SCSPMs, here we have codebooks with 1,024, 2,048, and 4,096 Gaussians respectively. The fourth pruning criterion is chosen in this experiment.

Table 1. Accuracy Rate with different codebook size

Codebook		Accuracy Rate (%)		
Size	**Criterion**	**Top 1**	**Top 5**	**Top 10**
1,024	Classification	60.39	85.74	91.56
	Mergence	61.33	86.08	91.64
2,048	Classification	69.50	88.79	92.76
	Mergence	70.21	88.93	92.94
4,096	Classification	-	-	-
	Mergence	75.49	90.48	94.42

The codebook with 4,096 Gaussians through classification the method is not built due to the computational complexity. Table 1 shows that the error rate can be reduced by about 36.6% when increasing the codebook size from 1,024 to 4,096. It also shows that the mergence method has better performance than the classification method.

4.2 Pruning techniques

In this experiment, we select the codebook of size 2,048 and use the mergence method to generate codebook.

Table 2. Accuracy rate when using different pruning criterion

Pruning Criterion	Accuracy Rate (%)		
	Top 1	**Top 5**	**Top 10**
1. (th=0.005)	69.52	88..76	92.69
2. (th=50)	68.74	87.89	91.45
3. (th= 90%)	69.11	88.24	92.17
4. (th=0.00001)	70.21	88.93	92.94

Being adopted the pruning criterion on the model training, the size of SCSPMs can be only about 1/4 of the model without pruning criterion. And the above table shows that when using Criterion 4 the performance of SCSPMs is the best.

4.3 SCSPMs vs. MGCPMs

We build MGCPMs with 4 and 8 mixed Gaussians for each state. The SCSPMs has 4,096 Gaussians in the codebook, and pruning Criterion 4 is adopted.

Table 3. Accuracy rate of MGCPMs and SCSPMs

Model		Accuracy Rate (%)		
Name	**Gaussian Num**	**Top 1**	**Top 5**	**Top 10**
MGCPMs (4)	9,669	74.35	89.15	93.78
MGCPMs (8)	18,685	75.87	90.62	94.55
SCSPMs	4,096	75.49	90.48	94.42

From the above table we can see that the recognition accuracy rate of SCSPM is higher than MGCPMs of 4 mixtures but lower than that of 8 mixtures. And the SCSPMs have far fewer Gaussians than MGCPMs have, thus the computational complexity can be greatly reduced for SCSPMs.

5. SUMMARY

In this paper the SCSPMs for continuous speech recognition is proposed and studied. The experimental results show that the SCSPMs can reduce the computational complexity during the recognition process with only a little degradation in accuracy compared to the MGCPMs. The codebook size is an important factor for SCSPMs. We can conclude that the accuracy rate increases with the large codebook size, so does the computational complexity. To reduce the number of tied Gaussians in each state, we study four pruning criteria respectively and the experimental result shows that the criterion of pruning small tying weights during the training outperforms other criteria. In our future study, we will do more experiments on the relationship between the accuracy rate and the codebook size and apply the SCSPMs to the systems with other SRUs such as initial/final or phoneme.

REFERENCES

1. L.R.Rabiner, "A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition", *Proc. IEEE*, 72(2):257-286, Feburary 1989
2. Zheng F., Mou X.-L., Xu M.-X., Wu J.,Song Z.-J., "Studies and Implementation of the Techniques for Chinese Dictation Machines," J.Software, 10(4):436-444, April 1999
3. Zheng, F., Song, Z.-X, Xu, M.-X, "EASYTALK: A large-vocabulary speaker-independent Chinese dictation machine", EuroSpeech'99, Vol.2, pp.819-822, Budapest, Hungary, Sept.1999
4. Huang X.- D., M.A.Jack, "Semi-continuous Hidden Markov Models for Speech Signals," *Computer Speech and Language*, 3:239-251, 1989
5. Bellegarda, J.R., Nahamoo, D., "Tied Mixture Continuous Parameter Modeling for Speech Recognition," *IEEE Trans. on ASSP*, vol.ASSP-38, No.12, pp.2033-2045, Nov. 1990
6. Xu, M.-X,. Zheng, F., Wu, W.-H, A fast and effective state decoding algorithm, EuroSpeech'99, Vol. 1, pp.187-190, Budapest, Hungary, Sept. 1999

7. Viterbi, A.J., “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm,” *IEEE Trans. on IT*, 13(2), Apr. 1967

8. Sadaoki, F., “Digital Speech Processing, Synthesis, and Recognition”, Tokai University Press, 1985

9. Jacques D., Kris D., Dirk V.C., “Fast and Accurate acoustic Modeling with Semi-Continuous HMMs”, Speech Communication 24 (1998) 5-17