Journal of Southeast University (English Edition) Vol. 20 No. 4 Dec. 2004 ISSN 1003—7985

# Speaker-independent speech recognition based on HMM state-restructuring method

Xu Xianghua Zhu Jie Guo Qiang

(Department of Electronic Engineering, Shanghai Jiaotong University, Shanghai 200030, China)

**Abstract:** Based on confusions between hidden Markov model (HMM) states, a state-restructuring method is proposed. In the method, HMM states are restructured by sharing Gaussian components with their related states, and the re-estimation to the increased-parameters, i.e., the inter-state weights, is derived under the expectation maximization (EM) framework. Experiments are performed on speaker-independent, large vocabulary, continuous Mandarin speech recognition. Experimental results show that the state-restructured systems outperform the baseline, and achieve significant improvement on recognition accuracy compared with the conventional parameter-increasing method. Such comparative results confirm that the state-restructuring method is efficient.

**Key words:** speech recognition; hidden Markov model; expectation maximization algorithm; HMM Tookit (HTK)

In recent years, most large vocabulary continuous speech recognition (LVCSR) systems are based on state-tying hidden Markov models (HMMs). During recognition, some states are commonly confused with others, i.e., state A is recognized to be state B by error, which causes a reduction in word recognition accuracy. The conventional method out of this problem is by purely increasing the number of Gaussian components[1]. However, in large vocabulary systems, where a large number of basic HMMs are used and each has only a few observations, a large number of parameters cannot be estimated robustly and are expensive to be stored.

In recent study, Luo[2] proposed the probabilistic classification of HMM states. The performance improvement reported in Ref. [2] suggested the importance of a good state structure. This study is related to what we are proposing here in that Gaussian components can be shared by different states. In the speaker-independent speech recognition system, a large number of speakers are included in the training set, which can cover the acoustic space to some extent. Therefore, in this paper, we suppose that states, which tend to be confused with other states in the training set, still tend to be confused to these states in the test set. Based on this hypothesis, a method of HMM state-restructuring is described. This method aims at increasing posterior probability to the training data. The final number of components in each state is inflated by sharing Gaussian components with the confused states. We evaluate the method by the task of speaker-independent, large vocabulary, continuous Mandarin speech recognition, and prove the method is efficient in improving system performance with limited increasing of parameters.

## 1 State-Restructuring in HMM Set

Assume an initial HMM set is generated with conventional algorithms. Fig. 1 gives the process of state-restructuring and parameter-training.

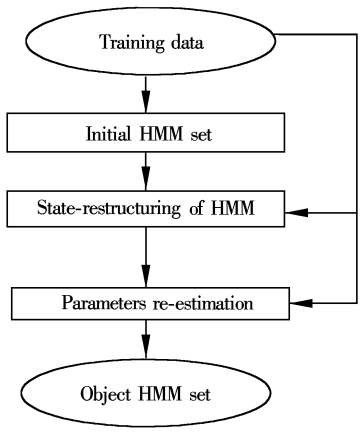


**Fig.1** Process of state-restructuring and parameters' re-estimation in HMM set

Suppose the set of states in HMM set is $\Omega$, and the training contexts are $\boldsymbol{X} = \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_i, \cdots\}$. As for $\boldsymbol{X}_i$, the corresponding observation feature vector is $\boldsymbol{O}_i = \{\boldsymbol{o}_1, \cdots, \boldsymbol{o}_t, \cdots, \boldsymbol{o}_T\}$. Based on $\boldsymbol{X}_i$, the word is

**Received** 2004-04-07.

**Foundation item:** The Science and Technology Commission Foundation of Shanghai (No. 01JC14033).

**Biographies:** Xu Xianghua (1977—), female, graduate; Zhu Jie (corresponding author), male, doctor, professor, zhujie@sjtu.edu.cn.

first expanded into its phonemic representation, and each phoneme is mapped into a sequence of states: $\Xi = \{s_1, \cdots, s_t, \cdots, s_T\}$, where $s_t$ is the state assigned to frame $t$ by the Viterbi alignment method[3]. We define $\Xi$ the actual state sequence. When $\boldsymbol{X}_i$ is decoded to the sequence of states: $\Psi = \{r_1, \cdots, r_t, \cdots, r_T\}$, where $r_t$ is the state assigned to frame $t$ by the Viterbi decoding algorithm[3]. Similarly, define $\Psi$ the recognized state sequence. Compare the two sequences, we get two states: $s_t$ and $r_t$ for feature vector $\boldsymbol{o}_t$. If $s_t \neq r_t$, define $r_t$ the related state of $s_t$ and the confusion $C_{s_t|r_t}$ is

$$C_{s_t|r_t} = \frac{P(\boldsymbol{o}_t \mid r_t)}{P(\boldsymbol{o}_t \mid s_t)} \tag{1}$$

Since state $s_t$ is recognized by $r_t$ by error, when $s_t \neq r_t$, $P(\boldsymbol{o}_t \mid r_t) > P(\boldsymbol{o}_t \mid s_t)$, i.e., $C_{s_t|r_t} > 1$. From definition (1), we can see the larger $C_{s_t|r_t}$ is, the more possible it is that $s_t$ is recognized by $r_t$. Thus, if $r_t$ is shared with $s_t$, i.e., share the weighted components in $r_t$ with $s_t$ and accordingly change the structure of $s_t$, the posterior probability $P(\boldsymbol{o}_t \mid s_t)$ will increase, therefore, the error reduction will be achieved.

For arbitrary state $s \in \Omega$, define its related state set $R^s$. Each state in $R^s$ is a related state of $s$. Restructure $s$ with $r\ (r \in R^s)$, the final Gaussian component function is

$$b(\cdot \mid s) = \sum_{r \in \boldsymbol{R}^s} w_{s|r} P(\cdot \mid r) + w_0 P(\cdot \mid s) = \sum_{r \in R^{s'}} w_{s|r} P(\cdot \mid r) \tag{2}$$

where $R^{s'} = R^s \cup \{s\}$ and $w_{s|r} = w_0$. Initialize $w_0$ to $1 - D$, where $D$ is a hand-set constant. The initial weights $w_{s|r}$ and the probability function $P(\cdot \mid r)$ are

$$w_{s|r} = \begin{cases} D \dfrac{C_{s|r}}{\sum_{r \in R^s} C_{s|r}} & r \neq s \\ 1 - D & r = s \end{cases} \tag{3}$$

$$P(\cdot \mid r) = \sum_{l=1}^{L} m_{r,l} N(\cdot \mid \boldsymbol{\mu}_{r,l}, \boldsymbol{\Sigma}_{r,l}) \tag{4}$$

where $L$ is the number of Gaussian mixtures; $N(\cdot \mid \boldsymbol{\mu}_{r,l}, \boldsymbol{\Sigma}_{r,l})$ represents multi-Gaussian function; $m_{r,l}$, $\boldsymbol{\mu}_{r,l}$ and $\boldsymbol{\Sigma}_{r,l}$ are the mixture weight, mean vector and dialogue covariance matrix of the $l$-th component in state $r$. Accordingly, the restructured-state $s$ has two level weights: the inter-state weight $w_{s|r}$ and the intra-state weight $m_{r,l}$. They satisfy:

- Intra-state weight

$$\sum_{l=1}^{L} m_{r,l} = 1 \qquad 0 \leqslant m_{r,l} \leqslant 1$$

- Inter-state weight

$$\sum_{r \in R^{s'}} w_{s|r} = 1 \qquad 0 \leqslant w_{s|r} \leqslant 1$$

After state-restructuring, the inter-state weights are re-estimated with the maximum likelihood criterion. The objective function is given by the log-likelihood function as

$$L(\boldsymbol{O}_s) = \sum_{\boldsymbol{o} \in \boldsymbol{O}_s} \log(p(\boldsymbol{o} \mid s)) \tag{5}$$

where $\boldsymbol{O}_s$ represents the set of observation feature vectors that belongs to state $s$ in training contexts. Our goal here is to find weight $w_{s|r}$ which maximizes the upper log-likelihood function. However, since there is no direct solution for the maximization, an EM algorithm[4] is used. The EM algorithm requires an auxiliary function as

$$Q(w_{s|r}, \bar{w}_{s|r}) = E[\log P(\boldsymbol{O}_s, s \mid \bar{w}_{s|r}) \mid \boldsymbol{O}_s, w_{s|r}] \tag{6}$$

We can update $w_{s|r}$ by equating the gradient of Eq.(6) with respect to $\bar{w}_{s|r}$ to zero.

In conclusion, state-restructuring is implemented as follows.

**Step 1** Align the initial state set and speech data by Viterbi decoding and get the most likely recognized state-sequence $\{\eta\}_i$ for context $\boldsymbol{X}_i$.

**Step 2** As for context $\boldsymbol{X}_i$, construct a network: $\mathrm{LM}_i$. Based on the network $\mathrm{LM}_i$, the Viterbi alignment method is used to give the actual state-sequence $\{\gamma\}_i$.

**Step 3** Do step 1 and step 2 for all $\boldsymbol{X}$.

**Step 4** Compare the two sequences of $\{\eta\}$ and $\{\gamma\}$, and from $\{\eta\}$, select the related state set $R^s$ for $s$ $(s \in \Omega)$ and compute the confusion $C_{s|r}\ (r \in R^s)$. For simplicity, select the top hand-determined $T$ candidate states of $R^s$ and remark the new state set as $R_s$.

**Step 5** Based on Eq.(3), compute the inter-state weight $w_{s|r}$. Share Gaussian components in state $r$ with state $s$ (as shown in Eq.(2)).

**Step 6** Do step 4 and step 5 for all states.

## 2 Experimental Results

The proposed state-restructuring method is evaluated on the LVCSR Mandarin dictation task. The databases Er-Wai[5] from Microsoft Research Asia are used for training. Er-Wai contains 19 688 utterances from 100 male students. The corresponding test set is MSR[5], which contains 500 utterances from 25 male speakers. Note that the two sets have no speaker overlap at all, i.e., all experiments are strictly speaker-independent. Acoustic features are 12 MFCCs and log

energy plus first and second derivatives (total 39 dimensions). Word-internal continuous triphone HMMs corresponding to 184 total initial/final set and a silence model are trained by HMM Tookit (HTK)[1]. To illustrate the effectiveness of the proposed algorithm, only the acoustic performances are given. The baseline system B-8 with 8 Gaussian components per state achieves the recognition accuracy of 52.27%. B-8 consists of 2 392 tied-states after decision-tree based state tying[6], with $2\,392 \times 8 \times 79 = 1\,511\,744$ parameters. The digit 79 denotes 39 mean, 39 dialogue covariance, and 1 intra-state weight per Gaussian component.

Let S-$T$ ($T$ = 1, 2, 3, 4, 5) represents the state-restructured system. Fig. 2 plots the recognition accuracy and the number of inter-state weights against different thresholds $T$. The first observation is that S-$T$ outperforms B-8, which indicates the shared Gaussian components within states of S-$T$ enhance the posterior probability. The second observation is that the number of inter-state weights in S-$T$ is no more than $2\,392 \times T$. The reason is that the size of some related-state sets is less than $T$. The third observation is that the performance of S-$T$ increases as threshold $T$ increases, while the trend of the increasing speed descends, implying that, although some states have many related states, using only the top 2 to 3 ones can achieve an improved performance.

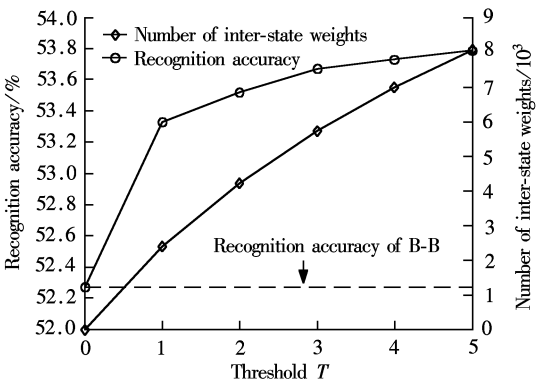


**Fig.2** Recognition accuracy and number of inter-state weights against threshold $T$

To compare the conventional parameter-increasing method by splitting Gaussian mixtures[1] with the state-restructuring method, Tab. 1 lists the comparative results yielded by the two methods, where WRA denotes word recognition accuracy, and B-$n$ ($n$ = 10, 12, 14, 16) represents the system created by the parameter-increasing method with $n$ Gaussian components per state. As we can see, with an increase in $n$, the number of parameters is germinated, whereas the increase in performance is limited. However, S-$T$ achieves better performance with only a slight increasing of parameters. For example, in S-3, the number of inter-state weights is 5 729, corresponding to 0.38% increase in parameters, and the word recognition accuracy is increased by 2.68%. We can attribute such better performance to the increase of related components per state. For example, in S-2, the number of inter-state weights is 4 222, and the number of average Gaussian components within each state is 22. This can explain why S-2 is superior to B-16. Similarly, S-1, with the average Gaussian components of 16, has similar performance to B-16.

**Tab.1** Comparative results in different systems %

| System | WRA | Increase of WRA | Increase of parameters |
|---|---|---|---|
| B-8 | 52.27 | baseline | baseline |
| B-10 | 52.34 | 0.13 | 25.00 |
| B-12 | 52.50 | 0.44 | 50.00 |
| B-14 | 53.05 | 1.49 | 75.00 |
| B-16 | 53.42 | 2.20 | 100.00 |
| S-1 | 53.33 | 2.03 | 0.16 |
| S-2 | 53.52 | 2.39 | 0.28 |
| S-3 | 53.67 | 2.68 | 0.38 |
| S-4 | 53.73 | 2.79 | 0.46 |
| **S-5** | **53.79** | **2.91** | **0.53** |

## 3 Conclusion

In order to increase the accuracy of HMM set and improve the system performance, a state-restructuring method is proposed. Experimental results show the method can improve the word recognition accuracy with a limited increase in the number of parameters. Since the state-restructuring method is based on the training data and obtains significantly good recognition results, we can expect to gain better performance by generalizing the method to use speaker adaptation data. This theory will be executed in the later work.

## References

[1] Young S, Jansen J, Odell J, et al. The HTK book [EB/OL]. http://htk.eng.cam.ac.uk/. 2003-10-03/2004-02-16.

[2] Luo X O, Jelinek F. Probabilistic classification of HMM states for large vocabulary continuous speech recognition [A]. In: *Proc of ICASSP* [C]. **Phoenix, Arizona, 1999, 1**: 353－356.

[3] Rabiner L, Juang B H. *Fundamentals of speech recognition* [M]. New Jersey: Prentice Hall, 1993. 339－342.

[4] Moon T K. The expectation-maximization algorithm [J]. *IEEE Signal Processing Magazine*, **1996, 13**(1): 47－60.

[5] Chang E, Shi Y, Zhou J L, et al. Speech lab in a box: a Mandarin speech toolbox to jumpstart speech related

research [A]. In: *Proc of Eurospeech* [C]. Aalborg, Denmark, 2001, **3**: 2779－2782.
[6] Reichl W, Chou W. Robust decision tree state tying for continuous speech recognition [J]. *IEEE Trans Speech and Audio Processing*, 2000, **8**(5): 555－566.

# 基于 HMM 状态结构调整的非特定人语音识别

徐向华　朱　杰　郭　强

(上海交通大学电子工程系, 上海 200030)

**摘要**: 利用 HMM 模型状态间的混淆度, 提出了一种新的状态结构调整算法, 使不同的状态可以共享相同的高斯混合函数, 并在 EM 算法的框架下推导出对状态结构调整后的增加参数, 即状态间权值的重估公式. 并对非特定人进行大词汇量汉语连续语音识别实验, 实验结果表明状态结构调整后的系统不仅优于基线系统, 还获得了比传统的参数增加方法更高的识别率, 由此证明了状态结构调整方法的有效性.

**关键词**: 语音识别; HMM; EM 算法; HTK
**中图分类号**: TN912.34; TP391.42