FUZZY PARAMETER CLUSTERING METHOD IN SPEECH RECOGNITION

Xianghua Xu and Jie Zhu

Department of Electronic Engineering of Shanghai Jiaotong University, Shanghai, 200030, P.R. China

ABSTRACT

In large vocabulary continuous speech recognition system, to efficiently decrease parameter size and improve the robustness of parameter training, a parameter clustering method by fuzzy clustering analysis is proposed. Based on the structure of phonetic decision tree, leaf nodes are used for Gaussian fuzzy clustering and root node or shallow leaf nodes are used for covariance fuzzy sharing. Experimental results show when the number of Gaussians is reduced by 50%, recognition accuracy only decreases by 0.55% compared to the baseline. By combining covariance fuzzy sharing, a significant performance increasing is achieved over the conventional system with approximately the same parameter size.

1. INTRODUCTION

In large vocabulary continuous speech recognition (LVCSR) system, the use of Gaussian components HMM is increasingly popular. On one hand, good enough Gaussian components can improve acoustic resolution by modeling a fine structure of the underlying distribution; on the other hand, large number of components will not only increase computational complexity in training and testing, but require large memory space. In most state-of-the-art systems, the way out of this problem is through parameter sharing [1-2]. The main idea of parameter sharing is to merge acoustic similar models or components by clustering. Therefore, clustering can be carried out at three levels: phoneme, state and density. In this paper, we decide to concentrate on the third level, which includes Gaussian clustering and covariance sharing.

The algorithm of fuzzy clustering analysis can be used to determine the sample classification [3]. Because of its good effect, this method has been adopted in the field of speech recognition [4]. For the first time, we propose to use Fuzzy Clustering Method Based on Perturbation (FCM BP) [5] for parameter clustering. Based on the hierarchy of the phonetic decision tree, leaf nodes are used for Gaussians clustering and root nodes or shallow leaf nodes are used for covariance sharing. Compared with other data-driven based clustering method (agglomerative or divisive hierarchical methods), FCM BP method needs no updating to the distance matrix, so it is more accurate and has an advantage of computational simplicity. Experimental results on large vocabulary Mandarin speech recognition show the fuzzy clustering method can give good performance with small parameter size.

This paper is organized as follows: Section 2 briefly reviews FCM BP fuzzy clustering analysis; Section 3 explains how this method is used in Gaussian fuzzy clustering and covariance fuzzy sharing; Section 4 describes experimental results; Section 5 summarizes this paper and offers a consideration of future work.

2. FUZZY CLUSTERING ANALYSIS

Fuzzy clustering analysis as it is known to us transforms a fuzzy similarity matrix R into a fuzzy equivalence matrix R^*, the finally clustering is then made by the λ-cut matrix R_λ^* [3] of R. Definitions of the three matrices are given as follows: for matrix $R=(r_{ij})_{n\times n}$ with $0\le r_{ij}\le 1$, if it satisfies:

1) reflexivity: $r_{ii}=1,(i=1,2,...,n)$;

2) symmetry : $r_{ij}=r_{ji},(i,j=1,2,...,n)$;

3) transitivity: $R\circ R\subset R$

where "$\circ$" represents the composite operation in fuzzy mathematics [3], then R is a fuzzy equivalent matrix. Else, if it only satisfies the upper two terms, R is a fuzzy similarity matrix R^*. The λ-cutting matrix $R_\lambda^*=(r_{ij}^\lambda)_{n\times n}$ of R^* satisfies:

$$r_{ij}^\lambda=\begin{cases}1, & r_{ij}^\lambda\ge\lambda\\ 0, & r_{ij}^\lambda<\lambda\end{cases} \tag{1}$$

Here, λ（$0\le\lambda\le1$）is a hand-determined threshold. Fuzzy clustering analysis clusters the elements corresponding to 1 in each row of matrix R_λ^*. Thus, by changing λ, we can get different clustering results. Suppose the number of clusters is m ($1\le m\le n$) after fuzzy clustering, obviously, m is in inverse ratio with λ. Two special cases are m=1 with λ=0 and m=n with λ=1.

Among the methods of transforming a fuzzy similarity matrix into a fuzzy equivalent matrix, the

objective-function-based method is very popular for its simple designing [6]. FCM BP method [5] is widely used objective-function-based method, and He *et al*. [5] have proved the fuzzy equivalent matrix gained by the FCM BP method is the closest one to the given fuzzy similarity matrix. For a fuzzy similarity matrix $S=(s_{ij})_{n\times n}$, the objective function of FCM BP is:

$$F(R)=\frac{1}{2}\|R-S\|_F^2=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(r_{ij}-s_{ij}\right)^2 \qquad (2)$$

The optimal fuzzy equivalent matrix R^* under the Frobenius norm $\|\cdot\|_F$ is sought using FCM BP scheme [5]. The scheme is the iteration through necessary condition of the optimality Eq. (2).

3. FUZZY PARAMETER CLUSTERING

3.1. Gaussian fuzzy clustering

Once HMM states with single Gaussian component have been tied through phonetic decision tree, the number of Gaussian components for each tied-state iteratively increases until the performance doesn't improve with the increasing of Gaussian components. Then, we use FCM BP fuzzy clustering to reduce the number of Gaussian components within each leaf node as illustrated in Figure 1. The detail process is as follows.

Firstly, construct Gaussian fuzzy similarity matrix. When diagonal covariance matrices are assumed, distance $d(p,q)$ between Gaussians $G(\mu_p, \Sigma_p)$ and $G(\mu_q, \Sigma_q)$ is defined as the sum of the Kullback-Leibler divergence [2]:

$$d(p,q)=\sum_{i=1}^{V}\left[\frac{\delta_p^2(i)-\delta_q^2(i)+\left(\mu_p(i)-\mu_q(i)\right)^2}{\delta_q^2(i)}+\frac{\delta_q^2(i)-\delta_p^2(i)+\left(\mu_q(i)-\mu_p(i)\right)^2}{\delta_p^2(i)}\right] \qquad (3)$$

where V is the dimensionality of the speech feature vector, $\mu_p(i)$ is the i-th element of the mean vector and $\delta_p^2(i)$ is the i-th diagonal element of the covariance matrix Σ_p. When $i\neq j$, define:

$$l(i,j)=1/d(i,j) \qquad (4)$$

Then, set $Avg=\frac{n(n-1)}{2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}l(i,j)$ and construct matrix $S=(s_{ij})_{n\times n}$ as:

$$s_{ij}=\begin{cases}1 & i=j\\ l(i,j)\Big/\left(\max_{p\neq q}\left(l(p,q)\right)+Avg\right) & i\neq j\end{cases} \qquad (5)$$

Obviously, S satisfies reflexivity and symmetry, besides, each element in S reflects the similarity between Gaussians, thus, S can be used as a fuzzy similarity matrix for the following Gaussian fuzzy clustering.

Secondly, from the upper fuzzy similarity matrix S, gain its fuzzy equivalent matrix R^* through FCM BP method. Given threshold λ, get the cutting matrix R_λ^* of R^*. Then Gaussian components within each tied-state are clustered through R_λ^*.

Thirdly, give the renewed parameters definition for Gaussian centers after Gaussian fuzzy clustering. Suppose the number of Gaussian centers is M and that of Gaussian components before clustering is N. Let P_m denote the m-th set of Gaussian codebook: $P_m=\{G(c_{m,k}, \mu_{m,k}, \delta^2_{m,k})|k=1,2,\ldots,K_m\}$. Here, $c_{m,k}$ is the weight of $G(\mu_{m,k}, \delta^2_{m,k})$ within a tied-state, and K_m is the size of the Gaussian codebook. Assume the size of data samples from each Gaussian is equal [7], replace P_m by one Gaussian $G(w_m, \mu_m, \delta^2_m)$ as an approximation to minimize the loss in likelihood:

$$\begin{aligned}\mu_m(i)&=\frac{1}{K_m}\sum_{k=1}^{K_m}\mu_{m,k}(i)\\ \delta_m^2&=\frac{1}{K_m}\left(\sum_{k=1}^{K_m}c_{m,k}\left(\mu_{m,k}^2(i)+\delta_{m,k}^2(i)\right)\right)-\left(\mu_m(i)\right)^2\\ w_m&=\frac{1}{K_m}\sum_{k=1}^{K_m}c_{m,k}\end{aligned} \qquad (6)$$

Then mixture weights in each state are normalized from the upper renewed weights. The final model parameters are re-estimated by Baum-Welch algorithm.

3.2. Covariance fuzzy sharing

As covariance has less distinctiveness than mean and weight within Gaussian components, smoothing covariance by covariance sharing can not only decrease the number of parameters, but improve the robustness of covariance training [7]. In this section, we will discuss covariance fuzzy sharing based on FCM BP method.

As illustrated in Figure 1, the hierarchy of the phonetic decision tree is used to guide the following covariance fuzzy sharing. Here, shallow leaf nodes are defined as leaf nodes of a lower decision tree, which is created by increasing thresholds during conventional phonetic decision tree based state tying. C_I method denotes all leaf nodes in a decision tree share the same covariance codebook and C_II method denotes different shallow leaf nodes share different covariance codebooks. By setting means to zero in Eq. (3), the distance measure $\Sigma(p,q)$ between covariances Σ_p and Σ_q is defined as:

$$\Sigma(p,q) = \sum_{i=1}^{V}\left[\frac{\delta_p^2(i)-\delta_q^2(i)}{\delta_q^2(i)} + \frac{\delta_q^2(i)-\delta_p^2(i)}{\delta_p^2(i)}\right] = \sum_{i=1}^{V}\left[\frac{\delta_p^2(i)}{\delta_q^2(i)} + \frac{\delta_q^2(i)}{\delta_p^2(i)} - 2\right] \quad (7)$$

As covariance fuzzy sharing is similar to Gaussian fuzzy clustering in Section 3.1, the rest detail process is omitted in this section. Similarly, still assume the number of data samples from each covariance is equal, then, the initial parameters of the shared covariance may be expressed as the average of all the covariances belong to the same classification. Finally, shared covariances are re-estimated by Baum-Welch algorithm.

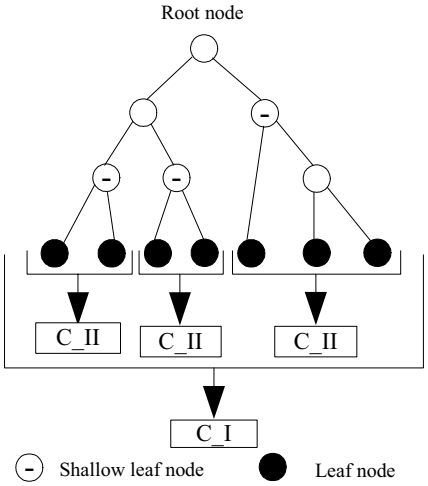

Figure 1 FCM BP parameter clustering based on the phonetic decision tree

4. EXPERIMENTS

The proposed clustering method is evaluated on a LVCSR Mandarin dictation task. Database Er-Wai [8] from Microsoft Research Asia is used for training. Er-Wai contains 19,688 utterances from 100 male students. The corresponding test set is MSR [8], which contains 500 utterances from another 25 male speakers. The acoustic feature vector has 39 elements, consisting of 12 MFCC coefficients and the normalized energy plus their first and second time derivative. The system uses context dependent tri-phone units for modeling Mandarin tonal syllables, which use the set of 185 phones proposed in [9]. After decision-tree based state tying, a total of 2893 tied-states are used in following acoustic modeling. Due to insufficient training data for estimating parameters in each individual Gaussian component, word (i.e., tonal syllable) accuracy (WA) doesn't greatly improve after the number of Gaussian components increases to 16. Therefore, the model with 2893×16 Gaussians is used as the first baseline system. Hereinafter, let 2893×k denote the standard state-tied HMM system with k Gaussian components per state.

Firstly, FCM BP method for Gaussian fuzzy clustering is evaluated in the following experiment. By adjusting λ in Gaussian λ-cutting matrix R_λ^*, we get different Gaussian-clustered systems. Table 1 shows word accuracy reduction (WAR) in different systems compared to the baseline. In Table 1, conventional agglomerative hierarchical clustering method (AHC) [10] for Gaussian clustering uses the same similarity measures as FCM BP.

Table 1 Recognition results for Gaussian clustering

system	#.G	WA (%)	WAR (%)
2893×16	46288	54.19	baseline
2893×12	34716	53.75	0.81
2893×8	23144	52.43	3.25
2893×4	11572	50.66	6.51
FCM BP	34716	54.33	–0.26
	23144	53.89	0.55
	11572	52.24	+3.60
AHC	34716	54.22	–0.06
	23144	53.05	2.10
	11572	50.98	5.92

#.G: number of Gaussian components

In Table 1, the front 4 rows are regularly trained state-tied HMM systems. Comparative results demonstrate Gaussian-clustered systems through FCM BP and AHC actually perform better than the regular system with approximately the same parameter size. For example, when the number of Gaussian is reduced by 25%, the two Gaussian-clustered systems even slightly outperform the baseline. Such results prove Gaussian clustering can effectively improve the robustness of parameters in some extent. Compared with AHC, FCM BP needs no updating to distance matrix, which simplifies the computation and avoids some error introduction, thus FCM BP is more accurate than AHC. For example, compared with the baseline, FCM BP system with 23144 Gaussian centers decreases WA by 0.55%, while AHC with 23144 Gaussian centers decreases WA by 2.10%.

Then covariance fuzzy sharing is evaluated through C_I (listed in Table 2). In FCM BP method, the FCM BP system with 34716 Gaussian centers is taken as the baseline. In AHC method, the AHC system with 34716 Gaussian centers is taken as the baseline. In the two systems, the corresponding number of covariance is 34716. From Table 2, we notice with 67.52% degradation in the number of covariances, FCM BP method only reduces recognition accuracy by 0.74% while AHC method reduces that by 1.75%. This demonstrates FCM

BP method is more efficient than AHC for both Gaussian and covariance clustering. Furthermore, by comparing FCM BP based covariance-shared system with 11275 covariance centers to the 2893×8 system in Table 1, we notice when more Gaussian means than covariances are allocated, the performance is improved slightly (2.86%) over the model which has about the same parameter size but has equal number of Gaussian means and covarances.

Thirdly, the hierarchy of the phonetic decision tree for covariance fuzzy clustering is evaluated through C_I and C_II (list in Table 3). In the experiment, the FCM BP system with 23144 Gaussian centers is taken as the baseline. Compared with C_I, although C_II increases the number of covariance codebooks, C_II can effectively decrease the size of each covariance codebook. Thus, C_II can achieve higher performance with fewer covariance centers. For example, when the number of shared covariance is 5600, C_I decrease performance by 3.47%, while C_II only decrease that by 1.97%.

Table 2 Recognition results for covariance sharing with FCM BP and AHC

Method	#.C	WA(%)	WAR(%)
FCM BP	34716	54.33	-
	22995	54.12	0.39
	11275	53.93	0.74
AHC	34716	54.22	-
	22995	53.86	0.66
	11275	53.27	1.75

#.C: number of covariances

Table 3 Recognition results for covariance fuzzy sharing through C_I and C_II

Method	#.C	WA (%)	WAR (%)
Baseline	23144	53.89	-
C_I	17360	53.66	0.43
	11424	53.10	1.47
	5600	52.02	3.47
C_II	17360	53.73	0.30
	11424	53.26	1.17
	5600	52.83	1.97

#.C: number of covariances

5. CONCLUSIONS

In this paper, a parameter clustering method based on FCM BP is proposed. The method uses a well-trained, large-sized state-tied HMM system as a baseline, in which the distance between Gaussians within each tied-state is used to build Gaussian fuzzy similarity matrix. Based on the Gaussian fuzzy clustering model, covariance in similar acoustic contexts is further shared guided by the hierarchy of the phonetic decision tree. Experimental results prove the effectiveness of the FCM BP method for parameter fuzzy clustering.

FCM BP method is not restricted to Gaussian clustering and covariance sharing, and is expected to significantly improve HMM training performance in other set-ups as well. Future work will focus on incorporation of powerful optimization tools within the FCM BP framework to achieve further improvement in LVCSR systems.

REFERENCES

[1] V. Digalakis, P. Monaco, and H. Murveit, "Genones: generalized mixture tying in continuous hidden Markov model-based speech recognitzers," *IEEE Trans. on Speech and Audio Processing*, vol.4, no.4, pp.281-289, 1996.

[2] K. Shinoda, and K.-I. Iso, "Efficient reduction of Gaussian components using MDL criterion for HMM-based speech recognition," *Proc. of ICASSP2002*, vol.1, pp.869-872, 2002.

[3] G. Bojadziey, and M. Bojadziev, *Fuzzy Sets, Fuzzy Logic, Applications*, World Scientific, New Jersey, 1995.

[4] M.-P. Mills, and J. Bowles, "Fuzzy logic enhanced symmetric dynamic programming for speech recognition," *Proc. of FUZZ-IEEE96*, pp.2013-2019, 1996.

[5] Q. He, H.-X. Li, Z.-Z. Shi, and E.-S. Lee. "Fuzzy clustering method based on perturbation," *Computers and Mathematics with Applications*, vol.46, no.5-6, pp.929-946, 2003.

[6] X.-B. Gai, and X. Xie, "Advances in theory and applications of fuzzy clustering," *Chinese Science Bulletin*, vol.45, no.11, pp.961-969, 2000.

[7] M.-Y. Hwang and X.-D. Huang, "Dynamically configurable acoustic models for speech recognition," *Proc. of ICASSP98*, vol.1, pp.669-672, 1998.

[8] E. Chang, Y. Shi, J.-L. Zhou, and C. Huang, "Speech lab in a box: a Mandarin speech toolbox to jumpstart speech related research," *Proc. of Eurospeech2001*, pp.2779-2782, 2001.

[9] E. Chang, J.-L. Zhou, S. Di, C. Huang, and K.-F. Lee, "Large vocabulary Mandarin speech recognition with different approaches in modeling tones," *Proc. of ICSLP2000*, pp.983-986, 2000.

[10] C.-F. Olson, "Parallel algorithms for hierarchical clustering," *Parallel Computing*, vol.21, no.8, pp.1313-1325, 1995.