AN INVESTIGATION OF TIED-MIXTURE GMM BASED TRIPHONE STATE CLUSTERING

Guangsen WANG, Khe Chai SIM

School of Computing, National University of Singapore
COM1, 13 Computing Drive, Singapore 117417
wangguangsen@comp.nus.edu.sg, simkc@comp.nus.edu.sg

ABSTRACT

Parameter tying is a crucial scheme for robust context dependent acoustic modeling since it takes a major role in balancing the desired model complexity and the amount of data available. In this paper, a modified decision tree state clustering scheme based on tied-mixture Gaussian Mixture Model (GMM) is proposed. Instead of using a single Gaussian untied triphone system, a tied-mixture GMM triphone system is adopted as a better acoustic model for state clustering. Meanwhile, the proposed scheme allows easy incorporation of discriminative training during clustering. Experimental results show that for a varying number of state clusters, the proposed approach consistently outperforms the standard single Gaussian based state tying. The best WER performance has a 10.5% relative improvement over the conventional decision tree clustering and the proposed scheme achieves its best performance using a much smaller number of state clusters. Moreover, detailed analyses reveal that the proposed GMM clustering has a better state distribution which leads to 1) better frame-state alignments 2) better phonetic question selections. These two factors may make the proposed approach superior for clustering.

Index Terms— Tied-mixture, phonetic decision tree, state clustering, context dependent modeling

1. INTRODUCTION

Context dependent (CD) acoustic models are widely used in state-of-the-art automatic speech recognition systems to address the co-articulation phenomenon in continuous speech. However, the number of CD phonemes grows exponentially with the extent of the contexts. Each phone state is conventionally modeled by GMMs. Therefore, the total number of Gaussian parameters can be of the order of a few hundreds of thousands. Modeling all these triphones as distinct models requires a large amount of training data. Hence, a major challenge of CD modeling is how to obtain a reliable estimation for all possible triphones with limited amount of training data.

Parameter tying is a crucial scheme for robust CD modeling since it takes a major role in balancing the desired model complexity and the amount of data available. Phonetic decision tree clustering [1] is a data-driven approach to cluster all the triphone states corresponding to one base phone state. One of the major advantages is that unseen triphones can be easily synthesized. Eigentriphone [2] is proposed to extract an eigenbasis over all the "rich" triphones based on the Gaussian means. Each triphone is viewed as a point in the eigenbasis and ends up with a distinct set of parameters. However, one disadvantage of this method is that it cannot handle the unseen triphones. Yet another parameter sharing scheme is called subspace modeling, e.g., canonical state modeling [3], subspace GMMs [4]. Decision tree clustering can also be viewed as a model selection problem. Therefore, many Bayesian solutions [5, 6] are also investigated. Genone [7] is a data-driven GMM based *mixture component level* clustering scheme. The system is built by progressively untying the mixtures of a tied-mixture GMMs following an agglomerative clustering, splitting and re-estimation procedure. However, the method still cannot address the unseen triphones directly.

In this paper, we investigate a modified decision tree state clustering scheme based on tied-mixture GMMs. Single Gaussian untied triphone system is used in [1] due to the performance consideration, since the calculation of GMMs likelihood gain requires revisiting the whole training data which is intractable. In principle, better alignments can be obtained using GMMs. However, training an untied GMMs triphone system would lead to poorer alignments. Therefore, context independent (CI) GMMs are used to initialize all the triphones in the proposed approach. Meanwhile, instead of using a single Gaussian model or untied triphone GMMs, a tied-mixture GMM triphone system is adopted to perform the state clustering. By resorting to an auxiliary function, we avoid revisiting the training data during clustering and the same sufficient statistics as the standard approach can be used for GMM clustering. Moreover, discriminative training can be incorporated during clustering by training the CI GMMs with the discriminative criteria [8]. Since the new clustering scheme is still within the decision tree clustering framework, it can handle the unseen triphones easily. The proposed GMM clustering scheme has a better state distribution which leads to 1) better frame-state alignments 2) better question selections, and these two factors may benefit the state clustering procedure.

The paper is organized as follows. The derivation of the tied-mixture GMM based decision tree state clustering is given in Section 2. Section 3 gives the recipe of how to build a tied-state CD system using the proposed state clustering scheme. Experimental evaluations are presented in Section 4. Detailed analyses and discussions of the proposed clustering scheme are given in Section 5. Section 6 summarizes the findings of the work and concludes the paper.

2. TIED-MIXTURE GMM BASED STATE CLUSTERING

Let $\mathbf{S}$ be a cluster of triphone states corresponding to one base phone state. These triphone states are modeled as a mixture of Gaussians instead of a single Gaussian in [1]. Moreover, the set of Gaussians are shared among $\mathbf{S}$. In other words, all the triphone states in $\mathbf{S}$ share the same Gaussian means $\mu(\mathbf{m})$, variances $\Sigma(\mathbf{m})$ as the base phone and cluster weights $c_m^{\mathbf{S}}$. The log likelihood of $\mathbf{S}$ is expressed as:

$$\mathcal{L}(\mathbf{S}) = \sum_{t=1}^{T} \sum_{s \in \mathbf{S}} \gamma_s(t) \log\Big(\sum_{m}^{M} c_m^{\mathbf{S}} \mathcal{N}(o_t; \mu(\mathbf{m}), \Sigma(\mathbf{m}))\Big)$$

where o_t denotes the training frame t, $s \in \mathbf{S}$ denotes an individual triphone state in cluster $\mathbf{S}$, M is the number of Gaussian components, $\gamma_s(t)$ is the expected occupancy of state s at time t, and $c_m^{\mathbf{S}}$ is the m-th component weight for the whole cluster calculated from individual weights of all the triphone states in $\mathbf{S}$. Subsequently, we refer to $c_m^{\mathbf{S}}$ as *hyper weight*.

To compute this GMM likelihood term efficiently without revisiting the training data, an auxiliary function is used to serve as the lower bound $\mathcal{L}(\mathbf{S}) \geq \mathcal{Q}(\mathbf{S})$:

$$\begin{aligned}\mathcal{Q}(\mathbf{S}) &= \sum_{t=1}^{T}\sum_{s\in\mathbf{S}}\sum_{m}^{M}\gamma_{sm}(t)\log(c_m^{\mathbf{S}}\mathcal{N}(o_t;\mu(\mathbf{m}),\Sigma(\mathbf{m})))\\ \gamma_{sm}(t) &= \gamma_s(t) * c_{sm}\end{aligned}$$

where $\gamma_{sm}(t)$ is the expected occupancy of frame o_t residing in mixture m of triphone state s and c_{sm} is the mixture weight of triphone state s for the m-th shared component. We thus have the factorized auxiliary function as:

$$\mathcal{Q}(\mathbf{S}) = \sum_{m}^{M}\log(c_m^{\mathbf{S}})\sum_{s\in\mathbf{S}}\beta_s * c_{sm} + \mathcal{K}_{\mathbf{S}}$$

where

$$\begin{aligned}\beta_s &= \sum_{t=1}^{T}\gamma_s(t)\\ \mathcal{K}_{\mathbf{S}} &= \sum_{t=1}^{T}\sum_{s\in\mathbf{S}}\sum_{m}^{M}\gamma_s(t) * c_{sm}\log\mathcal{N}(o_t;\mu(\mathbf{m}),\Sigma(\mathbf{m}))\end{aligned}$$

Note the evaluation of $\mathcal{K}_{\mathbf{S}}$ still needs revisiting the training data. However, since tied-mixture is used, $\mathcal{K}_{\mathbf{S}}$ stays the same before and after a split. Let $\mathbf{S}_1$ and $\mathbf{S}_2$ be the two new clusters after splitting $\mathbf{S}$, the equation

$$\mathcal{K}_{\mathbf{S}} = \mathcal{K}_{\mathbf{S}_1} + \mathcal{K}_{\mathbf{S}_2} \quad (1)$$

holds. Thus $\mathcal{K}_{\mathbf{S}}$ does not contribute to the change in likelihood. In other words, the likelihood change only depends on the weight terms, $c_m^{\mathbf{S}}$ and c_{sm}. In this way, we avoid evaluating $\mathcal{K}_{\mathbf{S}}$ thus revisiting the training data during decision tree building. β_s is the expected occupancy of triphone state s computed on all the training data and can be obtained from the Baum-Welch estimation. This is also the same sufficient statistics used for standard decision tree clustering [1]. The *hyper weight* $c_m^{\mathbf{S}}$ can be computed as:

$$\begin{aligned}c_m^{\mathbf{S}} &= \frac{\sum_{s\in\mathbf{S}}\sum_{t=1}^{T}\gamma_{sm}(t)}{\sum_{s\in\mathbf{S}}\sum_{m}^{M}\sum_{t=1}^{T}\gamma_{sm}(t)}\\ &= \frac{\sum_{s\in\mathbf{S}}c_{sm}\beta_s}{\sum_{s\in\mathbf{S}}\beta_s}\end{aligned}$$

With these sufficient statistics, the auxiliary function can be reformulated as:

$$\mathcal{Q}(\mathbf{S}) = \sum_{m}^{M}c_m^{\mathbf{S}} * \log(c_m^{\mathbf{S}})\sum_{s\in\mathbf{S}}\beta_s + \mathcal{K}_{\mathbf{S}}$$

The change of $\mathcal{Q}$ values after a split can be expressed as:

$$\Delta\mathcal{Q}(\mathbf{S}) = \mathcal{Q}(\mathbf{S}_1) + \mathcal{Q}(\mathbf{S}_2) - \mathcal{Q}(\mathbf{S})$$

Note that since equation 1 holds, $\Delta\mathcal{Q}(\mathbf{S})$ depends only on the *hyper weight* $c_m^{\mathbf{S}}$. Therefore, for each node splitting, the question which leads to the maximal $\Delta\mathcal{Q}(\mathbf{S})$ is chosen and used to split $\mathbf{S}$. The procedure is repeated until the maximal $\Delta\mathcal{Q}(\mathbf{S})$ falls below a threshold.

3. SYSTEM BUILDING RECIPE

The procedure of building a tied-mixture GMM based CD tied-state system is basically the same as [1]. The only difference is that the model before the clustering is a well trained CD tied-mixture GMM system. The following procedure is performed for each base phone state q_j over its corresponding triphone states.

Step 1: Monophone GMM/HMM system is firstly built. The likelihood of a base phone state q_j is denoted as:

$$\mathcal{L}_{q_j}(o_t) = \sum_{m=1}^{M}c_{jm}\mathcal{N}(o_t;\mu_{jm},\Sigma_{jm})$$

Step 2: The monophone GMM/HMM of base phone state q_j is cloned to initialize all its corresponding triphone states.

Step 3: Re-estimate only the weights of the tied-mixture system while keeping Gaussian means and variances fixed as q_j. The likelihood of a triphone state q_j^s thus can be expressed as:

$$\mathcal{L}_{q_j^s}(o_t) = \sum_{m=1}^{M}c_{jm}^{s}\mathcal{N}(o_t;\mu_{jm},\Sigma_{jm})$$

where c_{jm}^s is the m-th shared component weight of triphone state q_j^s.

Step 4: Perform GMM tied-mixture based decision tree clustering using the re-estimated tied-mixture GMMs for each base phone state and the unseen triphone states are synthesized using the corresponding decision tree.

Step 5: For each physical state $\mathbf{S}$, untie all the mixture components of the clustered model so that each physical state has its own Gaussian means, variances and weights. These cluster specific parameters are then retrained.

4. EXPERIMENTS

4.1. Experimental Setup

This section presents the experimental results comparing the tied-mixture GMM based and the baseline conventional single Gaussian based decision tree state clustering using the WSJCAM0 [9] corpus. There are 18.3 hours of training data, comprising 9889 utterances. Testing set "si_dt5a" which consists of 0.73 hours of speech for the 5K WSJ0 task is used for performance evaluation. The phone set has 41 monophones including one silence model and one short pause model. Each triphone is modeled as a left-to-right 3-state HMM and each state has a Gaussian density of $M = 16$. The features are the standard 39-dimensional MFCCs consisting of 13 static coefficients (12 MFCC plus one C0 energy term) and the first and second derivatives. HTK[1] is adopted for decoding. Word recognition is performed using a bigram full decoding followed by a trigram rescoring.

4.2. Tied-mixture GMM Based State Clustering

The word recognition performance under the baseline state clustering [1] (Single Gaussian Clustering) and proposed tied-mixture GMM based state clustering scheme is shown in Figure 1 in terms of word error rate (WER%). Consistent performance improvement

[1] Hidden Markov Model Toolkit, http://htk.eng.cam.ac.uk

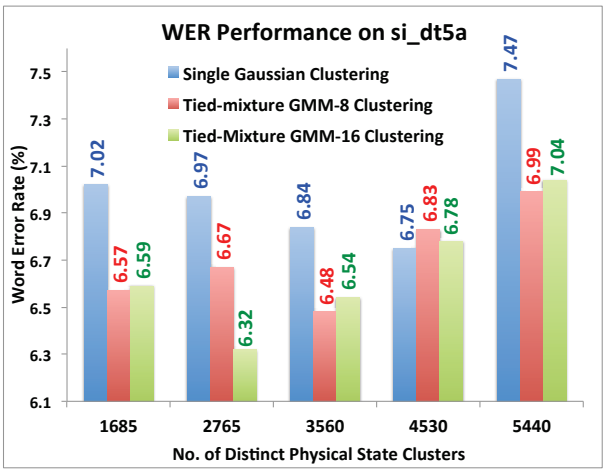

Fig. 1. *WER comparison of three clustering schemes. Numbers after GMM denote the Gaussian density of the tied-mixture system for GMM clustering. All clustering schemes are evaluated using their resulting tied-state triphone models with 16 components per state.*

is obtained with various number of state clusters using both tied-mixture GMM-8 and GMM-16 clustering. The best performance obtained from GMM-16 clustering (6.32%) has a 6.4% relative (0.43% absolute) improvement over the baseline (6.75%). More interestingly, the best performance of GMM-16 clustering uses only 2765 clusters, whereas for the baseline, 4530 clusters are needed. This illustrates an advantage of the proposed clustering scheme: a much better performance can be obtained with a much smaller parameter size. The notable performance declines after 4530 may be because there is not enough training data for a reliable estimation of this many clusters.

4.3. Incorporation of Discriminative Training

We further investigate whether the incorporation of discriminative training would further boost the clustering performance of the proposed tied-mixture GMM-16 clustering. One way of achieving this is training the CI GMMs in step 1 (see section 3) using discriminative training criteria, e.g., MMI, MPE [8]. Note only CI GMMs are discriminatively trained. After clustering, the resulting tied-state triphone system is trained with maximum likelihood (ML) and then used for decoding. Compared to the GMM-16 clustering initialized with ML trained CI GMM-16, a further performance boost (6.04 vs. 6.32) is obtained with an even smaller number of state clusters (1685 vs. 2765) using MMI criterion. This also translates to a 10.5% relative improvement over the baseline single Gaussian based clustering. Significance test using SCTK[2] reveals that the improvement over the baseline clustering is significant with $p = 0.029$. Therefore, GMM based state clustering retains its clustering performance even with the incorporation of discriminative training. Recall all the triphone states given a base phone state share the same set of Gaussians. The further reduction of cluster number may be because the discrimination among all the triphone states initialized by the discriminatively trained GMM-16 CI models before clustering is better modeled, thus less clusters are needed to distinguish them.

[2]NIST Speech Recognition Scoring Toolkit, http://www.itl.nist.gov/iad/mig//tools

Table 1. WER performance comparison of two baseline systems

#state clusters	1685	2765	3560	4530	5440
baseline	7.02	6.97	6.84	**6.75**	7.47
+re-estimation	6.75	6.68	**6.59**	6.96	7.64
GMM-16 clustering	6.59	**6.32**	6.54	6.78	7.04

5. ANALYSES AND DISCUSSIONS

Given the performance gain of the proposed tied-mixture GMM-16 clustering scheme, we further investigate two possible factors which make GMM based clustering superior.

5.1. Alignment of Training Data and Base Unit Modeling

Decision tree based state clustering has an assumption that the initial frame-state alignments do not change during the tree building procedure. Otherwise, a re-alignment using all the training data would be required for each possible partition which is intractable. However, a single Gaussian may provide poor alignments for clustering since it is inadequate to represent the variability in the data. On the other hand, better alignments may be obtained by tied-mixture GMMs in the proposed approach. To verify this, the same tied-mixture triphone GMM-16 after step 3 is used as the alignment model to perform a two-model re-estimation [10] on the single Gaussian untied triphone system before state clustering. Thus, both clustering schemes now have the same alignment model to generate the frame-state alignments. After the re-estimation, the conventional decision tree based clustering procedure is performed to get the tied-state triphone system for decoding. The WER performance is given in Table 1. As expected, a small performance boost (6.59 vs. 6.75) is observed after the re-estimation. However, the performance of the baseline clustering and the one after re-estimation is very close although there are some performance declines which are also reported in [11]. Therefore, the fact that the proposed approach outperforms the baseline system may be due to that GMM is used so that the variability in the data can be better modeled. This benefits the following clustering procedure in the sense that the separability among the triphone states before clustering is enlarged, which makes them easy to distinguish. However, the performance of the baseline clustering after re-estimation is still worse than the proposed approach. Therefore, other factors may exist that the proposed scheme can further benefit from.

5.2. Investigation of Phonetic Questions

We further study the phonetic questions used by the two systems during decision tree building since the questions determine the state partitions and eventually the state clusters. The results given here are based on the configuration that both schemes lead to 4530 state clusters. 7089 questions are used by the tied-mixture GMM-16 clustering scheme, which is 1205 more than the baseline clustering. This means more partitions are considered, which results in a larger search space and may lead to a better clustering eventually. Figure 2 shows the count of the top 10 most important questions used by each system. The count for each question is calculated as: 1) for each decision tree, all the questions used are sorted according to the likelihood gain in descending order and the top 10 questions are considered most important for this decision tree 2) if one question appears in the top 10 most important questions of any of the 117 decision trees, its count is increased by one. After counting, the

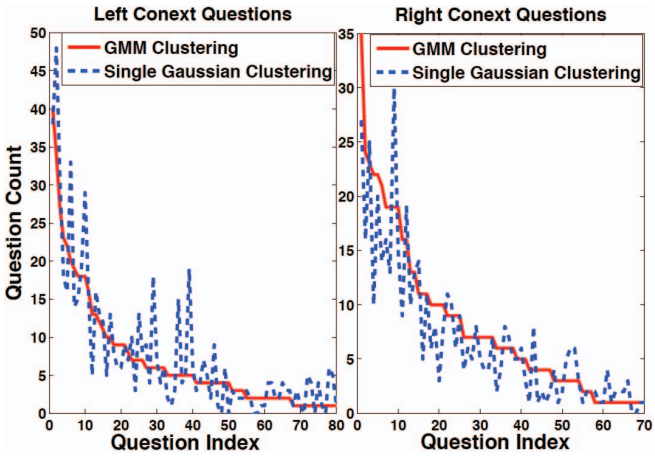

Fig. 2. *Question counts under two clustering schemes based on their importance.*

questions of the GMM based clustering are sorted based on their count in descending order and then indexed. According to the question indices, the corresponding question count of the single Gaussian based state clustering is drawn. This procedure is done on questions concerning left contexts (left subfigure) and right contexts (right subfigure) respectively.

Many spikes of the baseline clustering curve are observed for both left and right context question figures. This means that the two systems perform quite differently on question selections and node partitions during the decision tree building procedure. Moreover, there are more outstanding spikes on the left context questions than the right contexts. This may imply that the single Gaussian based clustering puts more emphasis on the left context. We further examine these 1170 (117×10) questions to verify this hypothesis. It turns out, for the baseline clustering, there are 618 questions concerning the left contexts and 496 about the right contexts. While for the proposed scheme, 564 left contexts and 563 right contexts are used, which is much more balanced than the baseline. The reason why these numbers do not sum to 1170 is for some decision trees, the total number of questions used is smaller than 10. The imbalance of left and right context questions for the single Gaussian based clustering will probably lead to many biased partitions which may hurt the performance.

Moreover, further examination of these questions reveals another interesting fact: the single Gaussian based state clustering tends to choose more specific questions. We define "*specific questions*" as: 1) L/R_Silence 2) L/R_Nasal 3) L/R_phone (e.g., L_ae, R_iy). The total number of specific questions for the baseline is 328 while only 276 for the proposed clustering. In general, more specific questions tend to give more state clusters thus may require more training data for a robust estimation. This insight also explains why the proposed clustering scheme can achieve its best performance with a smaller number of clusters than the baseline system.

6. CONCLUSIONS

In this paper, we have investigated a tied-mixture GMM based state clustering scheme as an alternative to the conventional single Gaussian based decision tree state clustering scheme. Decision tree clustering is performed on a tied-mixture GMM system, instead of a single Gaussian system used in the conventional approach. Experimental results show that for a varying number of state clusters, the proposed approach consistently outperforms the standard single Gaussian based clustering. The best WER performance of the proposed approach has a 10.5% relative improvement over the conventional approach. In addition, the proposed scheme achieves its best performance using a much smaller number of clusters. Detailed analyses reveal that the proposed GMM clustering has a better state distribution which leads to 1) better frame-state alignments 2) better phonetic question selections. These two factors may contribute to the performance improvement over the conventional decision tree clustering scheme. Future work includes the investigation of discriminative splitting criteria using the tied-mixture GMM based state clustering scheme. Applying the proposed clustering scheme to larger tasks and databases, e.g., spontaneous speech, is also necessary to see whether the clustering performance can be retained.

7. ACKNOWLEDGEMENT

This research is done for CSIDM Project No. CSIDM-200806 partially funded by a grant from the National Research Foundation (NRF) administered by the Media Development Authority (MDA) of Singapore.

8. REFERENCES

[1] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *HLT*, 1994, pp. 307–312.

[2] T. Ko and B. Mak, "A fully automated derivation of state-based eigentriphones for triphone modeling with no tied states using regularization," in *Interspeech*, 2011, pp. 781–784.

[3] M. J. F. Gales and K. Yu, "Canonical state models for automatic speech recognition," in *Interspeech*, 2010, pp. 58–61.

[4] D. Povey and L. Burget et al., "The subspace gaussian mixture model-a structured model for speech recognition," *Comput. Speech Lang.*, vol. 25, pp. 404–439, 2011.

[5] W. Chou and W. Reichl, "Decision tree state tying based on penalized Bayesian information criterion," in *IEEE., ICASSP*, 1999, pp. 345–348.

[6] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Variational Bayesian estimation and clustering for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 12, pp. 365–381, 2004.

[7] V. Digalakis, P. Monaco, and H. Murveit, "Genones: Generalized mixture tying in continuous hidden markov model-based speech recognizers," *IEEE Trans. on Speech Audio Processing*, vol. 4, pp. 281–289, 1996.

[8] D. Povey, *Discriminative training for large vocabulary speech recognition*, PhD thesis, Cambridge University, 2004.

[9] T. Robinson and J. Fransen et.al, "WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition," in *IEEE., ICASSP*, 1995, pp. 81–84.

[10] S. J. Young et al., *The HTK Book, version 3.4*, Cambridge University Engineering Department, Cambridge, UK, 2009.

[11] H.J. Nock, M.J.F. Gales, and S.J. Young, "A comparative study of methods for phonetic decision-tree state clustering," in *European Conf. on Speech Commun. & Tech.*, 1997, pp. 111–114.