PROBABILISTIC CLASSIFICATION OF HMM STATES FOR LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

Xiaoqiang Luo and Frederick Jelinek

320 Barton Hall, CLSP, Dept. of ECE
The Johns Hopkins University, Baltimore, MD 21218
{xiao,jelinek}@mail.clsp.jhu.edu

ABSTRACT

In state-of-art large vocabulary continuous speech recognition (LVCSR) systems, HMM state-tying is often used to achieve good balance between the model resolution and robustness. In this paradigm, tied HMM states share a single set of parameters and are non-distinguishable. To capture the fine differences among tied HMM states, a probabilistic classification of HMM states (PCHMM) is proposed in this paper for LVCSR. In particular, a distribution from a HMM state to classes is introduced. It is shown that the state-to-class distribution can be estimated together with conventional HMM parameters within the EM [3] framework. Compared with HMM state-tying, probabilistic classification of HMM states makes more efficient use of model parameters. It also makes the acoustic model more robust against the possible mismatch or variation between training and test data. The viability of this approach is verified by the significant reduction of word error rate (WER) on the Switchboard [7] task.

1. INTRODUCTION

In state-of-art LVCSR systems, HMM state-tying is used to find a good compromise between the model resolution and robustness. Representative work includes CMU "senone" [8], SRI "genone" [4], IBM decision network [1] and HTK state clustering based on phonetic decision trees [13]. Although details differ from site to site, the essence of these methods is to find a tying point, typically HMM state, in the HMM framework, and build a relatively crude model to collect statistics based on which a classification algorithm is carried out to group HMM constructs into equivalence classes. After HMM states are clustered, the number of free parameters is reduced, and Baum-Welch[2] algorithm can be used to estimate reliably class-dependent HMM parameters.

In the tying paradigm, different tied HMM states, or equivalence classes of HMM states, are modeled with independent parameters. However, acoustic regions corresponding to tied classes are invariably overlapped. Therefore, it is inefficient to use separate parameters to model tied states. In continuous density HMM (CDHMM), the state output distribution typically takes the form of a mixture of Gaussians. Parameter independence limits the number of Gaussians that can be used since parameters can not be reliably estimated if an excessive number of Gaussians is used.

In our previous work [10], we proposed non-reciprocal data sharing (NRDS) when estimating HMM parameters. It aims at capturing the fine differences among HMM states in an equivalence class. NRDS [11] reduces WER at the cost of increasing the number of model parameters . To remedy this, we propose here to use the probabilistic classification of HMM states in LVCSR. In this method, a HMM state s is assigned to a class r based on a probability $w(r|s)$, and the class of HMM states r is modeled by a mixture of Gaussians. Since Gaussians of a class r can be used in many HMM states, this makes efficient use of model parameters so that rich distributions can be modeled without increasing the model size. At the same time, allowing a HMM state to contribute to more than one class also makes an acoustic model robust against the possible mismatch or variation between training and test data. This will be elaborated in the subsequent sections of this paper.

A recent study Nakamura [12] compares the most-likely state sequence and the forced-aligned state sequence and analyzes how often an "error" occurs. If an error is seen frequently enough, the most-likely Gaussian is augmented to the corresponding forced-aligned state. The final model structure is such that HMM states can have variable number of Gaussians and a Gaussian can be shared in more than one HMM state or equivalence class. The improvement reported in [12] suggests the importance of having a good model structure. This study is related to what we are proposing here in that a Gaussian component can be shared by many states. On the other hand, the restructuring procedure suggested in [12] is ad hoc while, as will be shown shortly, the model proposed here can be put into the EM framework and all the model parameters can be estimated simultaneously. The other differentiating point is that the model proposed here has a hierarchical structure with two levels of weights.

The rest of the paper is organized as follows. In section 2, we will derive reestimation formulas for the state to class probabilities $w(r|s)$ together with conventional HMM parameters within the EM framework. Then implementation issues will be discussed before the experimental results on the Switchboard task are presented Our experiments show that probabilistic clustering HMM scheme reduces word error rate (WER) by up to 1.7% (absolute) on the Switchboard task when compared with a state-of-art state-tied system.

2. REESTIMATION FORMULAE

In this section we describe our model and the method to estimate the model parameters. Let $\mathcal{S}$ be the set of HMM states, and $\mathcal{R}$ the set of classes of HMM states. Apart from a normal HMM setup, a state-class distribution $w(r|s)$ is introduced, where $r \in \mathcal{R}$ is a class of HMM states, and $s \in \mathcal{S}$ is an individual HMM state. It is required that $\sum_{r\in\mathcal{R}} w(r|s) = 1$, so $w(\cdot|s)$ is a probability. State-output distributions, $q(\cdot|s)$, are defined in terms of that of classes,

$p(\cdot|r)$; That is,

$$q(\cdot|s) = \sum_{r\in\mathcal{R}} w(r|s)p(\cdot|r). \quad (1)$$

Throughout this paper, a mixture of Gaussian distributions is assumed for each class, or

$$p(\cdot|r) = \sum_{l=1}^{M_r} m_{r,l}N(\cdot|\mu_{r,l},\Sigma_{r,l}), \quad (2)$$

where M_r is the number of Gaussians used for class r, $m_{r,l}$ is the mixture weight for the l^{th} component and $\mu_{r,l}$ and $\Sigma_{r,l}$ are Gaussian mean and covariance of the l^{th} mixture component of class r. We will also use the notation $N(\cdot|r,l)$ to represent the l^{th} Gaussian of class r. Therefore, $q(\cdot|s)$ is a mixture of Gaussian mixtures.

Let $O = \{o_t\}_1^T$ be a sequence of speech feature vectors, and $S = \{S_t\}_1^T, R = \{R_t\}_1^T, L = \{L_t\}_1^T$ be sequences of HMM states, state classes and mixture labels respectively. S, R and L consist of hidden variables in this setup. Let θ be the totality of HMM parameters, θ' the value before the current iteration. With these notations, EM [3] auxiliary function $Q(\theta|\theta')$ can be expressed as

$$\begin{aligned} Q(\theta|\theta') &= E[\log P(O,S,R,L)|O] \quad (3)\\ &= E[\log P(S)|O] + E[\log w(R|S)|O] + \\ &\quad E[\log P(L|R)|O] + E[\log P(O|S,L,R)]. \quad (4)\end{aligned}$$

Similar to Juang's development in [9] (except that we have one more level of "branching" probability $w(r|s)$), let

$$\begin{aligned} \alpha_t(s) &= P(o_1,\cdots,o_t,S_t=s;\theta') \quad (5)\\ \beta_t(s) &= P(o_{t+1}^T|S_t=s;\theta') \quad (6)\end{aligned}$$

be forward and backward probabilities respectively. We also define the following posterior probabilities

$$\begin{aligned} \gamma_t(s) &= P(S_t=s|O;\theta') \quad (7)\\ \gamma_t(s,s') &= P(S_t=s,S_{t+1}=s'|O;\theta') \quad (8)\\ \gamma_t(s,r,l) &= P(S_t=s,R_t=r,L_t=l|O;\theta') \quad (9)\\ \gamma_t(s,r) &= P(S_t=s,R_t=r|O;\theta'). \quad (10)\end{aligned}$$

For simplicity of notations, the dependence on θ' is not shown on the left hand sides. The posterior probabilities can be computed efficiently by the forward-backward algorithm. In particular, since

$$\begin{aligned} & P(S_t=s,R_t=r,L_t=l,O;\theta')\\ =\ & P(S_t=s,R_t=r,L_t=l,o_1^t;\theta')P(o_{t+1}^T|S_t=s;\theta')\\ =\ & \sum_{s'}\alpha_{t-1}(s')a_{s's}w(r|s)m_{r,l}N(o_t|r,l)\beta_t(s)\\ =\ & \alpha_t(s)\beta_t(s)\frac{w(r|s)m_{r,l}N(o_t|r,l)}{\sum_{x,k}w(x|s)m_{x,k}N(o_t|x,k)}, \quad (11)\end{aligned}$$

we have

$$\begin{aligned} \gamma_t(s,r,l) &= P(S_t=s,R_t=r,L_t=l|O;\theta') \quad (12)\\ &= \frac{\alpha_t(s)\beta_t(s)}{P(O;\theta')}\frac{w(r|s)m_{r,l}N(o_t|r,l)}{\sum_{x,k}w(x|s)m_{x,k}N(o_t|x,k)}\\ &= \gamma_t(s)\frac{w(r|s)m_{r,l}N(o_t|r,l)}{\sum_{x,k}w(x|s)m_{x,k}N(o_t|x,k)} \quad (13)\end{aligned}$$

where $\gamma_t(s) = P(S_t=s|O;\theta')$ is normal state-occupancy.

With the help of (7-10), (4) can be expressed as

$$\begin{aligned} Q(\theta|\theta') &= \sum_t\sum_{s',s}\gamma_t(s,s')\log P(s'|s) +\\ &\quad \sum_t\sum_{s,r,l}\gamma_t(s,r,l)\log w(r|s) +\\ &\quad \sum_t\sum_{s,r,l}\gamma_t(s,r,l)\log m_{r,l} +\\ &\quad \sum_t\sum_{s,r,l}\gamma_t(s,r,l)\log P(o_t|r,l). \quad (14)\end{aligned}$$

Maximizing (14) will yield the reestimation formulae we are after. Particularly, maximizing the first term will give us the update formula for state transition probabilities while maximizing the second to fourth term will give us update formulae for state-class probabilities $w(r|s)$, mixture weights $m_{r,l}$ and Gaussian means and covariances respectively. The first three terms of (14) have the form

$$\sum_i a_i\log x_i, \quad (15)$$

where $a_i \geq 0, \sum_i a_i > 0$ and $x_i \geq 0, \sum_i x_i = 1$. It is easy to show that $x_i = \frac{a_i}{\sum_j a_j}$ maximizes (15). So update formulae for $P(s'|s)$, $w(r|s)$ and $m_{r,l}$ are

$$\begin{aligned} \hat{P}(s'|s) &= \frac{\sum_t\gamma_t(s,s')}{\sum_t\sum_{s''}\gamma_t(s,s'')} \quad (16)\\ \hat{w}(r|s) &= \frac{\sum_t\sum_{l=1}^{M_r}\gamma_t(s,r,l)}{\sum_t\sum_x\sum_{k=1}^{M_x}\gamma_t(s,x,k)} \quad (17)\\ \hat{m}_{r,l} &= \frac{\sum_t\sum_s\gamma_t(s,r,l)}{\sum_t\sum_s\sum_k\gamma_t(s,r,k)} \quad (18)\end{aligned}$$

The last term of (14) can be written as

$$\begin{aligned} & \sum_t\sum_{s,r,l}\gamma_t(s,r,l)\log P(o_t|r,l)\\ =\ & \sum_t\sum_{s,r,l}\gamma_t(s,r,l)\big(-\frac{n}{2}\log(2\pi)-\frac{1}{2}\log|\Sigma_{r,l}|\\ & -\frac{1}{2}(o_t-\mu_{r,l})'\Sigma_{r,l}^{-1}(o_t-\mu_{r,l})\big). \quad (19)\end{aligned}$$

Maximizing this function is similar to the derivation of maximum likelihood estimate (MLE) for multi-variate Gaussian distributions [10], and final reestimation formulae are

$$\begin{aligned} \hat{\mu}_{r,l} &= \frac{\sum_t\sum_s\gamma_t(s,r,l)o_t}{\sum_t\sum_s\gamma_t(s,r,l)} \quad (20)\\ \hat{\Sigma}_{r,l} &= \frac{\sum_t\sum_s\gamma_t(s,r,l)(o_t-\hat{\mu}_{r,l})(o_t-\hat{\mu}_{r,l})'}{\sum_t\sum_s\gamma_t(s,r,l)}. \quad (21)\end{aligned}$$

This completes the derivation of reestimation formulae for the probabilistic classification model.

It is easy to verify that reestimation formulae (16-18) and (20-21) are consistent with those of a deterministic classification of HMM states. Note that if a classification of HMM states is deterministic, or

$$w(r|s) = \begin{cases} 1, & \text{if } r = r_s\\ 0, & \text{otherwise}\end{cases}, \quad (22)$$

where r_s is the class of state s, then

$$\gamma_t(s, r_s, l) = \gamma_t(s) \frac{m_{r_s,l} N(o_t|r_s, l)}{\sum_k m_{r_s,k} N(o_t|r_s, k)} \quad (23)$$

and $\gamma_t(s, r, l) = 0$ for all $r \neq r_s$. Plug (23) into (16-18) and (20-21) and we get normal reestimation formulae.

3. IMPLEMENTATION ISSUES

We address some implementation issues in this section, including how to determine classes of HMM states, how to initialize the state-to-class probabilities and how to train efficiently the entire set of parameters.

In the derivation in Section 2, it is assumed that the HMM classes $\mathcal{R}$ are known and only state-to-class probabilities $w(r|s)$ and other HMM parameters remain to be estimated. In practice, $\mathcal{R}$ is unknown as well. In this work, we use HTK-based phonetic decision trees [13] to determine HMM classes. This step is described in [13]. So leaf nodes of these decision trees will represent HMM classes. At the end of tree constructions, there will be an underlying Gaussian distribution for each leaf node. To obtain an initial state to class assignment, we compute the distance between the two underlying Gaussian densities for any leaf node-pair so that for each leaf node s, we can order the other leaf nodes based on the distances. The closest few nodes are chosen as the candidate HMM classes for HMM states belonging to the leaf node s. Or in other words, s will have zero probabilities going to all other HMM classes.

After this alternation of the acoustic model, a few EM iterations are run to obtain an initial state-to-class probabilities $\{w(r|s)\}$, starting from the model with single Gaussians. Then the state-to-class probabilities $\{w(r|s)\}$ and normal mixture weights $\{m_{r.l}\}$ are collapsed so that full-scale training can be carried out using HTK. The other advantage of collapsing the two types of weights is that the decoder can be kept intact.

4. EXPERIMENTS

The test data used in this study is WS97 [5] dev-test set which consists of 2427 utterances and about 18 thousand words. The baseline was built in LVCSR WS97 and is a state-tied cross-word triphone system. The baseline system has about seven thousand equivalence classes of HMM states and there are 12 Gaussians per class.

The system implementing the probabilistic clustering of HMM states is built based on the baseline system. First of all, classes of HMM states are chosen as leaf nodes in the decision trees of the baseline system. Initial state-to-class assignments are obtained by examining the distances between underlying Gaussians, as described in Section 3.

	no-GMLLR	GMLLR
12G(Baseline)	39.4	36.6
12G-PC3	**38.0**	**35.6**
12G-PC5	38.2	35.8
24G-PC3	**37.7**	**35.5**

Table 1: Comparison of WER of the PCHMM vs. the baseline system using trigram LM

The new model is used to rescore lattices generated by the baseline model. Results are tabulated in Table 1 when a trigram language model is used. The numbers in the second and third columns are WERs without and with global MLLR [6] speaker adaptation respectively. The first two digits in the first column denote the number of Gaussians per state and the last digit stands for how many classes a state can belong to with positive probability. Therefore, "12G-PC3" means there are 12 Gaussians per class and a state can belong to 3 classes. As seen from the table, up to 1.7% WER reduction is obtained compared with the baseline system. Note that "12G-PC?" systems have roughly the same number of parameters as the baseline. In other words, 1.4% WER reduction can be obtained without increasing the number of model parameters.

A question one may ask is whether the same amount of WER reduction is achievable by simply increasing the number of Gaussians per state in a tied-state system. To answer this question, we conducted a series of experiments by changing the number of Gaussians per state and the results are tabulated in Table 2. Note that a bigram LM is used to speed up the experiments. The second through fifth row are results of the state-tied system with various numbers of Gaussians per state. For instance, "18-G" means that each state has 18 Gaussians. The last three rows correspond to results of the systems with probabilistic clustering of HMM states. The first observation is that the probabilistic clustering systems always outperform the state-tied systems. The second observation is that simply increasing the number of Gaussians per state does not help without speaker adaptation. This is not very surprising since some of the Gaussians will be poorly estimated as the number of mixture components increases. The third observation is that a larger model helps if combined with speaker adaptation. This is probably because a larger model is likely to match test data better after adaptation than a smaller model. The best model reduces WER by 1.7% compared with the state-tied baseline system (i.e, the 12-G system), and 0.7% with the best tied system (the 24-G system). The best 12-G system reduces WER by 1.4% compared with the 12-G baseline system.

	no-GMLLR	GMLLR
12-G(Baseline)	41.8	39.1
18-G	**41.6**	38.6
24-G	41.8	**38.1**
36-G	42.5	38.8
12-G-PC3	**40.3**	37.9
12-G-PC5	40.6	37.7
24-G-PC3	40.5	**37.4**

Table 2: Comparison of WER of the PCHMM vs. state-tied systems with various number of Gaussian components per state

Before closing this section, we'd like to discuss advantages of adopting a structural model such as probabilistic clustering. First of all, the probabilistic clustering scheme makes use of parameters in a more economic fashion to achieve the same acoustic resolution as a state-tied system. To be specific, assume that each class uses 12 Gaussians, and we allow a state to contribute to 3 classes, then there are 36 Gaussians that have contributions to each state. We would need 36 Gaussians for each class in a state-tied system to match this. More Gaussians per state will make it possible to model in detail distributions that may otherwise be ignored.

Second, allowing a state to contribute to more than one class also means that a model will be more robust against possible mismatch or variation between the test and training data than a state-tying model would be. To illustrate this point, let's consider the following example. Say state s is assigned to class A in a state-tied system; and in probabilistic clustering, s has positive probability of going to both class A and B. Then if test data happens to fall in to the acoustic regions described by class B, then the state-tied system is likely to fall apart while the soft-clustering will be able to handle the variation gracefully. We expect that this kind of a situation is more likely to be encountered in conversational speech such as Switchboard than in read speech.

5. CONCLUSIONS

In this paper we have proposed the probabilistic classification of HMM states, in which scheme a HMM state is assigned to a class with a probability, and not by fixed rules as in state-tying. The reestimation formulas for the state-to-class probabilities together with the Gaussian parameters are derived using the EM algorithm. A two-step procedure is used to train parameters for probabilistically clustered HMMs. Significantly better recognition results (up to 1.7% absolute WER reduction) on the Switchboard task are obtained by using the proposed method.

6. ACKNOWLEDGMENTS

We would like to thank Bill Byrne for allowing us to access his numerous improvements on top of the standard HTK and to use his state-of-art baseline system. The system provides us with a good starting point for the new models and also serves as a rigorous test-bed for the proposed technique. Without his significant contribution, this study could not be finished so timely, to say the least. We also want to thank Ciprian Chelba and Asela Gunawardana for many useful discussions.

7. REFERENCES

[1] L.R. Bahl, P.V. de Souza, P.S Gopalakrishnan, D. Nahamoo, and M. A Picheny. Robust methods for using context-dependent features and models in a continuous speech recognizer. In *Proc. of ICASSP-94*, volume I, pages 533–536, 1994.

[2] L. E. Baum. An inequality and associated maximization techniques in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, 3:1–8, 1972.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[4] Vassilios V. Digalakis, Peter Monaco, and Hy Murveit. Genones: Generalized mixture tying in continuous hidden Markov model-based speech recognition. *IEEE Transactions on Speech And Audio Processing*, 4(4):281–289, July 1996.

[5] Frederick Jelinek ed. Research notes 14: 1997 lvcsr summer research workshop technical reports. Technical report, CLSP, The Johns Hopkins University, 1998.

[6] M.J Gales and P.C Woodland. Mean and variance adaptation within the MLLR framework. *Computer Speech and Language*, 10:249–264, October 1996.

[7] John J. Godfrey, Edward C. Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Proc. of ICASSP*, volume I, pages 517–520, 1992.

[8] Mei-Yuh Hwang, Xuedong Huang, and Fileno A. Alleva. Predicting unseen triphones with senones. *IEEE Transactions on Speech and Audio Processing*, 4(6):412–419, November 1996.

[9] B. H. Juang, Stephen E. Levinson, and M. M. Sondhi. Maximum likelihood estimation for multivariate mixture observations of Markov chains. *IEEE Transactions on Information Theory*, IT-32(2):307–309, March 1986.

[10] Xiaoqiang Luo and Frederick Jelinek. Nonreciprocal data sharing in estimating HMM parameters. Technical Report 32, CLSP, The Johns Hopkins University, 1998.

[11] Xiaoqiang Luo and Frederick Jelinek. Nonreciprocal data sharing in estimating HMM parameters. In *Proc. ICSLP*, 1998.

[12] A. Nakamura. Restructuring gaussian mixture density functions in speaker-independent acoustic models. In *Proc. ICASSP*, volume II, pages 649–652, Seattle, May 1998.

[13] S.J Young, J.J Odell, and P.C Woodland. Tree-based state tying for high accuracy acoustic modelling. In *Proceedings of ARPA Workshop on Human Language Technology*, pages 307–312, March 1994.