Sub-state Tying in Tied Mixture Hidden Markov Models

Liang Gu and *Kenneth Rose*

Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106

ABSTRACT

An approach is proposed for partial tying of states of tied-mixture hidden Markov models. To facilitate tying at the sub-state level, the state emission probabilities are constructed in two stages, or equivalently, are viewed as a ‘‘mixture of mixtures of Gaussians.’’ This paradigm allows, and is complemented with, an optimization technique to seek the best complexity-accuracy tradeoff solution, which jointly exploits Gaussian density sharing and sub-state tying. Experimental results on the E-set show that the classification error rate is reduced by over 20% compared to standard Gaussian sharing and whole-state tying. The approach is then embedded within the recently developed procedure of combined parameter training and reduction technique. Experiments with the overall technique show that the error rate is further reduced by 8%.

1. INTRODUCTION

Tied-Mixture Hidden Markov Modeling (TMHMM) [1][2] has long been recognized as a useful complexity reduction technique for robust automatic speech recognition. The benefits are mainly due to its ability to maintain modeling accuracy of large-mixture probability density functions (pdf’s) at moderate complexity, by using a universal set (or codebook) of pdf’s. TMHMM may greatly reduce the model complexity by enforcing pdf sharing, but it also introduces a large number of additional mixing coefficients, which are not needed in the standard Continuous HMM (CHMM). The need to reduce the mixing coefficient set motivated further reduction approaches [2]. The premise of our work here is that additional gains may be achieved by optimizing the mixing efficiency.

Another known complexity reduction technique consists of state tying. Similarly to pdf tying, state tying addresses the fundamental conflict between accuracy of acoustic modeling and the insufficient training data. By tying some of the HMM states, training robustness is enhanced and this, in turn, makes it feasible to include a larger number of states in the HMM and, thereby, achieve higher accuracy. However, the common procedure suffers from several shortcomings. First, state tying typically involves full tying of states, or equivalently, enforcing certain states to be identical. This extreme measure yields substantial complexity reduction, but may cause serious degradation in model accuracy. Second, although efficient optimization algorithms [3][4][5] have been proposed for state tying, they are typically initialized in a greedy suboptimal fashion [4][5]. This may impact the performance, especially in the case of a large number of Markov states. Third, the optimization of state tying is normally performed separately from the optimization of pdf sharing and, hence, the overall system is suboptimally designed.

In this paper, we propose the sub-state tying (SST) approach, where Markov states are partially tied (in contrast with the traditional whole state tying). In order to develop an automatic procedure for sub-state tying, we redefine the state emission probabilities as a two-stage mixture. In other words, instead of a standard mixture of Gaussians, we view the state emission pdf as a mixture of (smaller) mixtures of Gaussians. The idea is that we create an intermediate level for tying, which is positioned between the Gaussian tying of TMHMM and whole state tying which ties the entire state mixture. Such intermediate level of tying allows one to find the right tradeoff between complexity and accuracy. Optimization of sub-state tying is automatically performed by a technique based on the Expectation-Maximization (EM). We show that with this approach, the mixing efficiency in TMHMM is improved, the need for tied-state initialization is circumvented, and a refined tradeoff yields better compromise between complexity and accuracy.

The above complexity reduction technique is complemented with a recent performance enhancement technique - combined training and reduction (CTR) of system parameters, which was proposed in [6]. The CTR procedure starts by training a system with a large universal codebook of Gaussian densities, and then alternates between codebook reduction, mixing coefficient matrix reduction, and parameter re-training. CTR has some capability of avoid poor local optima. It incorporates discriminative design in the reduction steps, at minimal cost in complexity as maximum likelihood is used in the re-estimation step.

Experimental results show significant gains of the proposed SST method over traditional state tying techniques. When complemented with CTR, the resulting automatic TMHMM design technique (denoted SST-CTR) offers performance comparable to the recent "manual" tying techniques [7], on E-set database.

2. TWO-STAGE TIED MIXTURE HIDDEN MARKOV MODELS

A. State Tying versus Sub-state Tying in TMHMM

TMHMM) [1][2] uses a universal codebook of Gaussian densities. State emission probability distributions are constructed

This work was supported in part by the National Science Foundation under grant no. IIS-9978001, the University of California MICRO Program, Cisco Systems, Inc., Conexant Systems, Inc., Dialogic Corp., Fujitsu Laboratories of America, Inc., General Electric Co., Hughes Network Systems, Lernout & Hauspie Speech Products, Lockheed Martin, Lucent Technologies, Inc., Qualcomm, Inc., and Texas Instruments, Inc.

$$
\begin{array}{c}
\\
\nu_1 \\ \nu_2 \\ \nu_3 \\ \vdots \\ \nu_K
\end{array}
\begin{array}{c}
\overbrace{\phantom{s_{1,1}\ s_{1,2}\ \cdots\ s_{1,N}}}^{Class\ 1}\ \overbrace{\phantom{s_{2,1}\ s_{2,2}\ \cdots\ s_{2,N}}}^{Class\ 2}\ \cdots\ \overbrace{\phantom{s_{M,1}\ s_{M,2}\ \cdots\ s_{M,N}}}^{Class\ M} \\
\begin{array}{cccccccccccc} s_{1,1} & s_{1,2} & \cdots & s_{1,N} & s_{2,1} & s_{2,2} & \cdots & s_{2,N} & \cdots & s_{M,1} & s_{M,2} & \cdots & s_{M,N} \end{array} \\
\begin{bmatrix}
p_{1|1,1} & p_{1|1,2} & \cdots & p_{1|1,N} & p_{1|2,1} & p_{1|2,2} & \cdots & p_{1|2,N} & \cdots & p_{1|M,1} & p_{1|M,2} & \cdots & p_{1|M,N} \\
p_{2|1,1} & p_{2|1,2} & \cdots & p_{2|1,N} & p_{2|2,1} & p_{2|2,2} & \cdots & p_{2|2,N} & \cdots & p_{2|M,1} & p_{2|M,2} & \cdots & p_{2|M,N} \\
p_{3|1,1} & p_{3|1,2} & \cdots & p_{3|1,N} & p_{3|2,1} & p_{3|2,2} & \cdots & p_{3|2,N} & \cdots & p_{3|M,1} & p_{3|M,2} & \cdots & p_{3|M,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
p_{K|1,1} & p_{K|1,2} & \cdots & p_{K|1,N} & p_{K|2,1} & p_{K|2,2} & \cdots & p_{K|2,N} & \cdots & p_{K|M,1} & p_{K|M,2} & \cdots & p_{K|M,N}
\end{bmatrix}
\end{array}
$$

Figure 1. Mixing Coefficient Matrix of TMHMM

as mixtures of densities from the codebook with appropriate mixing coefficients. Let there be *M* classes, each represented by an HMM of *N* states, and let there be a universal codebook of *K* Gaussian densities. The emission probability distribution for state is:

$$b_{m,n}(\boldsymbol{x}) = \sum_{k=1}^{K} g(\boldsymbol{x}|\boldsymbol{\nu}_k)\, p_{k|m,n} \qquad (1)$$

where $g(\cdot|\boldsymbol{\nu})$ is a Gaussian density whose mean and variance are specified in the parameter vector $\boldsymbol{\nu}$. The universal codebook may be simply represented by the set of parameter vectors $\{\boldsymbol{\nu}_k, k = 1,\ldots,K\}$. The mixing coefficients have obvious probabilistic interpretation $p_{k|m,n} = \Pr(\boldsymbol{\nu}_k \mid s_{m,n})$, and satisfy

$$\sum_{k=1}^{K} p_{k|m,n} = 1$$

State tying in TMHMM may be specified by operations on the mixing coefficient matrix: $\{\, p_{k|m,n} \,\}_{K\times M\times N}$, which is shown in *Figure* 1. The traditional whole-state tying technique combines two or more columns together and generates a tied-state which is shared across several states, that is

$$p_{k|m_1,n_1} = p_{k|m_2,n_2}, \forall 1 \le k \le K$$

Sub-state tying, which is proposed in this paper, ties parts of the columns and allows each column to be distinct from other columns, that is

$$\begin{cases} p_{k|m_1,n_1} = p_{k|m_2,n_2}, & \textit{for some } k \in [1,K] \\ p_{k|m_1,n_1} \ne p_{k|m_2,n_2}, & \textit{otherwise} \end{cases}$$

Sub-state tying enables the implementation of many intermediate levels of tying, compared to the coarse whole-state tying, and provides higher accuracy and better mixing efficiency in TMHMM. However, sub-state tying poses substantial optimization challenges, as now the tying process is carried out in a much larger space. In the next subsection we propose a simplified two-stage tying algorithm that considerably reduces the sub-state tying optimization complexity but still captures substantial gains due to partial state tying.

B. Sub-state Tying with Two-stage Mixtures

$$
\begin{array}{c}
\\
c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_L
\end{array}
\begin{array}{c}
\begin{array}{cccc} \nu_1 & \nu_2 & \cdots & \nu_K \end{array} \\
\begin{bmatrix}
q_{1|1} & q_{1|2} & \cdots & q_{1|K} \\
q_{2|1} & q_{2|2} & \cdots & q_{2|K} \\
q_{3|1} & q_{3|2} & \cdots & q_{3|K} \\
\vdots & \vdots & \ddots & \vdots \\
q_{L|1} & q_{L|2} & \cdots & q_{L|K}
\end{bmatrix}
\end{array}
$$

Figure 2. Gaussian Tying Matrix

Standard TMHMM can be viewed as ‘‘one-stage” tying as shown in *Figure* 1. In this paper, a two-stage tying algorithm is proposed as a practical way for implementing sub-state tying. We define the state emission probability as a mixture of Gaussian mixtures, or a mixture of ‘‘sub-mixtures”. These sub-mixtures are denoted $\{\boldsymbol{\nu}_k, k = 1,\ldots,K\}$. Note that we use the same notation as for Gaussians before only to be able to use *Figure* 1 again and save space. Thus we have *K* sub-mixtures each of which is a mixtures of Gaussians from a codebook of *L* Gaussians: $\{c_l \mid 1 \le l \le L\}$. The first stage of tying is the Gaussian tying stage, where the Gaussian tying matrix (GTM) is used to specify the sub-mixtures in terms of the available Gaussians (see *Figure* 2). The second stage of tying is the sub-mixture tying stage (SMTM), which is shown in *Figure* 1, except that now the rows correspond to sub-mixtures instead of to Gaussian densities. The emission probability distribution for state $s_{m,n}$ is now rewritten as the mixture:

$$b_{m,n}(\boldsymbol{x}) = \sum_{k=1}^{K} \left\{ \sum_{l=1}^{L} g(\boldsymbol{x}|\boldsymbol{c_l}) q_{l|k} \right\} p_{k|m,n} \qquad (2)$$

where $g(\cdot|\boldsymbol{c})$ is a Gaussian density whose mean and variance are specified in parameter vector c, and

$$p_{k|m,n} = \Pr(\boldsymbol{\nu}_k \mid s_{m,n}), \quad q_{l|k} = \Pr(c_l \mid \boldsymbol{\nu}_k).$$

The two-state tying of TMHMM can be viewed as a generalization of standard TMHMM and whole-state tying:

- When GTM is diagonal (and *L*=*K*), two-stage tying TMHMM becomes standard TMHMM.

- When GTM is diagonal and several columns in SMTM are combined in their entirety, two-state tying TMHMM is equivalent to whole-state tying in TMHMM.

- When both GTM and the SMTM are full, two-stage tying TMHMM is a general framework for both Gaussian sharing and sub-state sharing, where tradeoff optimization is carried out at the sub-mixture level.

In practice, (2) can be simplified by only taking into account significant values of $p_{k|m,n}$ and $q_{l|k}$ (as is done for standard TMHMM [2]):

$$b_{m,n}(\boldsymbol{x}) = \sum_{k\in\eta(m,n)} \left\{ \sum_{l\in\zeta(m,n)} g(\boldsymbol{x}|\boldsymbol{c}_l) q_{l|k} \right\} p_{k|m,n} \qquad (3)$$

This brings about a substantial decrease in the number of free parameters without significant loss in recognition accuracy.

C. Reestimation Formulas

The reestimation of two-stage tying TMHMM is similar to that of standard TMHMM except that the reestimation process is divided into three steps. Let us denote by $\gamma_{s_{m,n}}(\vec{x}_t)$ the probability that state $s_{m,n}$ is visited at time *t*, given that the model emits $\vec{x}_t$, i.e.,

$$\gamma_{s_{m,n}}(\vec{x}_t) = \Pr(s_{m,n}|\vec{x}_t) \qquad (4)$$

(which may be calculated as in [1]).

We consider:

$$\eta_{l,k,s_{m,n}}(t) = \gamma_{s_{m,n}}(\vec{x}_t) \frac{g(\vec{x}_t|c_l) q_{l|k} \cdot p_{k|m,n}}{\Pr(\vec{x}_t|s_{m,n})} \qquad (5)$$

$$\xi_{k,s_{m,n}}(t) = \gamma_{s_{m,n}}(\vec{x}_t) \frac{p_{k|m,n}\left[\sum_{l=1}^{L} g(\vec{x}_t|c_l)\cdot q_{l|k}\right]}{\Pr(\vec{x}_t|s_{m,n})} \qquad (6)$$

Parameter reestimation based on the EM algorithm is carried out as:

1) Reestimation of Gaussian pdfs

$$\hat{\mu}_l = \frac{\sum_t \sum_{s_{m,n}} \sum_k \eta_{l,k,s_{m,n}}(t)\cdot \vec{x}_t}{\sum_t \sum_{s_{m,n}} \sum_k \eta_{l,k,s_{m,n}}(t)} \qquad (7)$$

$$\hat{\Sigma}_l = \frac{\sum_t \sum_{s_{m,n}} \sum_k \eta_{l,k,s_{m,n}}(t)\cdot (\vec{x}_t - \hat{\mu}_l)\cdot(\vec{x}_t - \hat{\mu}_l)^T}{\sum_t \sum_{s_{m,n}} \sum_k \eta_{l,k,s_{m,n}}(t)} \qquad (8)$$

where ^{T} denotes matrix transpose.

2) Reestimation of GTM:

$$q_{l|k} = \frac{\sum_t \eta_{l,k,s_{m,n}}(t)}{\sum_t \sum_l \eta_{l,k,s_{m,n}}(t)} \qquad (9)$$

3) Reestimation of SMTM

$$p_{k|m,n} = \frac{\sum_t \xi_{k,s_{m,n}}(t)}{\sum_t \sum_k \xi_{k,s_{m,n}}(t)} \qquad (10)$$

3. SUB-STATE TYING WITH COMBINED TRAINING AND REDUCTION

A high-level diagram for the application of sub-state tying with CTR algorithm in TMHMM is given in *Figure* 3. The training process builds on two iterative optimization loops: one loop optimizes the system for a fixed number of free parameters (FNFP), and is referred to as the FNFP loop. In this loop, the two-stage tying TMHMM is trained. The other loop optimizes decisions for parameter reduction (PR) and is called the PR loop. The initial number of free parameters is M_0, and either a fixed or a variable parameter reduction rate may be employed. A group of parameters is identified and eliminated in each iteration. The decision is based on a minimum-entropy-increase criterion derived from the previous FNFP loop. The overall process, of parameter estimation and reduction, continues until the target number of free parameters has been reached. (For detailed information on the CTR framework see [6].)

The inner FNFP loop is susceptible to local minima due not only

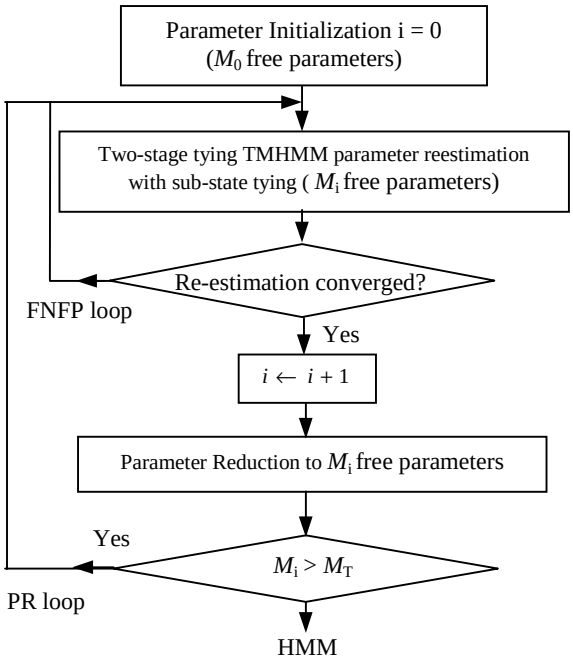

Figure 3. Sub-state tying in TMHMM with CTR Algorithm

methods	No. states per HMM	No. of distinct Gaussian pdfs	Train Set Error Rate	Test Set Error Rate
CHMM	13	234	7.0 %	17.3 %
TMHMM	13	234	6.6 %	12.0 %
Two-stage tying TMHMM	13	234	6.2 %	9.1 %

Table 1. Performance comparison of HMM design methods at the same number of states and Gaussian densities

Methods	No. distinct Gaussian pdfs	No. of free parameters	Train Set Error Rate	Test Set Error Rate
Whole-state tying	342	34200	5.0 %	9.5 %
Sub-state tying	330	32600	4.3 %	7.6 %
Sub-state tying – CTR	330	33000	3.9 %	7.0 %

Table 2. Performance comparison of tying methods in TMHMM with similar number of free parameters

to the suboptimality of EM, but also to the fact that HMMs are separately designed for ML optimality. However, in the outside PR loop, the number of free parameters is downsized based on a global performance criterion and this often helps to break through the barriers of poor FNFP local minima.

4. EXPERIMENT RESULTS

To test the performance of sub-state tying in TMHMM design, experiments were carried out on the E-set speech database obtained from OGI [8]. The recognition task is to distinguish between nine confusable English letters {b, c, d, e, g, p, t, v, z}. The database was generated by 150 speakers (75 male and 75 female) and includes one utterance per speaker. Of the 150 speakers, 60 male and 60 female speakers were selected at random for training, and the remaining 30 speakers were set aside for the test set. The experiment of random selection followed by design was repeated 300 times and the average performance over all trials was recorded.

In our experiment, 36-dimension LPCC parameters were used as the speech features, with 18 LPC-derived cepstrums plus 18 delta-cepstrums. The analysis frame width was 30ms, the analysis frame step was 10ms, and a Hamming Window was used. Two HMM models were included for each utterance, to allow for variation between male and female speakers. The experiment results are shown in *Table* 1 and *Table* 2. *Table* 1 summarizes the performance of the competing HMM model structures and design techniques, which are compared at the same number of states and number of Gaussian components. The results demonstrate that the performance is monotonically improving from CHMM, through standard TMHMM [2], to two-stage tying TMHMM. Note that TMHMM and two-stage tying TMHMM in this case have more free parameters because of the mixing coefficients. In *Table* 2, another comparison is given between whole-state tying [4], sub-state tying, and sub-state tying with CTR at similar number of free HMM parameters, all based on TMHMM. Sub-state tying obtained a 20% error rate reduction over whole-state tying method, while sub-state tying with CTR offers the best performance and achieves 8% further reduction in error relative to plain sub-state tying. This combined approach provided 93% recognition rate, which is comparable to the performance of state-of-art "manual" tying techniques on E-set database [7], while the results presented here are achieved automatically.

5. CONCLUSION

Gaussian sharing and state tying are two approaches for complexity reduction in HMM design. Basic TMHMM shares Gaussians across states and classes, while state tying shares the mixing coefficients among selected subsets of states. The proposed sub-state tying (SST) method implements partial state tying that builds on redefining state emission probabilities as two-stage mixtures and results in a refined tradeoff between complexity and accuracy. The method jointly optimizes Gaussian sharing and sub-state tying by EM-based re-estimation. In simulations over the E-set, SST reduced the recognition error rate by 20% compared to existing TMHMM with whole-state tying. SST was then embedded within the combined training and reduction (CTR) framework to provide further reduction of 8% in error rate. Future work will focus on the application of powerful optimization tools to the joint optimization of all components of the general tying framework.

6. REFERENCES

[1] J. R. Bellegarda and D. Nahamoo, "Tied mixture continuous parameter modeling for speech recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 2033-2045, vol. 38, Dec. 1990.

[2] X. D. Huang, "Phoneme classification using semicontinuous hidden Markov models", *IEEE Trans. Signal Processing*, pp. 1062-1067, vol. 40, May 1992

[3] O. Cappe, C. E. Mokbel, D. Jouvet and E. Moulines, "An algorithm for maximum likelihood estimation of hidden Markov models with unknown state-tying", *IEEE Trans. Speech Audio Processing*, pp. 61-70, vol. 6, Jan. 1998.

[4] M. Y. Hwang and X. Huang, "Shared-distribution hidden Markov models for speech recognition", *IEEE Trans. Speech Audio Processing*, pp. 414-420, vol. 1, Oct. 1993.

[5] S. J. Young and P. C. Woodland, "State clustering in hidden Markov model-based continuous speech recognition", *Computer Speech and Language*, pp. 369-383, vol. 8, Oct. 1994.

[6] L. Gu and K. Rose, "Combined parameter training and reduction in tied-mixture HMM design", Proc. IEEE ASRU, Keystone, CO, Dec. 1999.

[7] P. C. Loizou and A. S. Spanias, "High-performance alphabet recognition", *IEEE Trans. Speech and Audio Processing*, pp.430-445, vol. 4, Nov. 1996.

[8] R. Cole, Y. Muthusamy, and M. Fanty, "The ISOLET spoken letter database", *Tech. Rep. 90-004, Oregon Graduate Inst.*, 1990.