Substate Tying With Combined Parameter Training and Reduction in Tied-Mixture HMM Design

Liang Gu, *Member, IEEE,* and Kenneth Rose, *Senior Member, IEEE*

***Abstract*—Two approaches are proposed for the design of tied-mixture hidden Markov models (TMHMM). One approach improves parameter sharing via partial tying of TMHMM states. To facilitate tying at the substate 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 substate tying. Another approach to enhance model training is combined training and reduction of model parameters. The procedure starts by training a system with a large universal codebook of Gaussian densities. It then iteratively reduces the size of both the codebook and the mixing coefficient matrix, followed by parameter re-training. The additional cost in design complexity is modest. Experimental results on the ISOLET database and its E-set subset show that substate tying reduces the classification error rate by over 15%, compared to standard Gaussian sharing and whole-state tying. TMHMM design with combined training and reduction of parameters reduces the classification error rate by over 20% compared to conventional TMHMM design. When the two proposed approaches were integrated, 25% error rate reduction over TMHMM with whole-state tying was achieved.**

***Index Terms*—Parameter reduction, parameter training, state tying, tied-mixture HMM.**

I. Introduction

THE HIDDEN Markov Model (HMM) is widely recognized as a useful statistical tool for automatic speech recognition. Optimal design of speech recognizers on the basis of limited training data, must take into account the fundamental tradeoff between model richness and robustness. The tied-mixture HMM (TMHMM) [1], [2] represents an important approach to optimization of this tradeoff. With its universal set of density functions for constructing state emission mixtures, TMHMM offers the modeling capability of a large-mixture continuous density HMM (CHMM), but with a substantially reduced total number of Gaussian parameters to train. Thus, for the typical case of insufficient training data, TMHMM achieves significant performance gains over traditional CHMM.

In spite of TMHMM's success, several problems remain open. One of them is how to optimally design the state emission density codebook and estimate the tying parameters. While the natural design objective is accurate classification of utterances, HMM training has traditionally been performed using the maximum likelihood (ML) criterion which is, in fact, a modeling criterion. The corresponding re-estimation algorithms are effective and of manageable complexity [1], [2]. The inherent and fundamental mismatch of ML with the natural "true" objective, led to the development of a new class of methods that seek the minimum classification error (MCE) solution [3]. MCE methods offer improved performance but encounter three main difficulties. The design complexity is considerably increased, but this may not be prohibitive in practical applications where the design is typically performed off-line. MCE tends to be highly susceptible to poor nonglobal optima (see [4] for the deterministic annealing approach to combat this difficulty). Finally, the gains of MCE may not generalize well outside the training set. Several complementary measures have been proposed in recent years to determine the model complexity during HMM design. These include discriminant [5]–[7] or nondiscriminant [8], [9] measures.

Another problem in TMHMM design is how to reduce the number of free parameters using parameter sharing or reduction techniques without significant loss of recognition accuracy. Appropriate parameter sharing and reduction can reduce model complexity with minimal degradation in model accuracy and, thus, realize a better tradeoff between model complexity and robustness. One such approach is to complement distribution sharing with state tying. Similarly to pdf tying, state tying attempts to refine the design tradeoff due to the fundamental conflict between the accuracy of acoustic modeling and 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 traditional procedure suffers from several shortcomings. First, state tying typically involves full tying of states, which consists of making certain states identical. This extreme measure yields substantial complexity reduction, but may cause serious degradation in model accuracy. Second, although efficient optimization algorithms have been proposed for state tying [10]–[13], they are typically initialized in a greedy suboptimal fashion. 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, consequently, the overall system is suboptimally designed.

Manuscript received February 27, 2001; revised January 14, 2002. This work was supported in part by the National Science Foundation under Grants IIS-9978001, EIA-9986057, and EIA-0080134, the University of California MICRO Program, Conexant Systems, Inc., Dolby Laboratories, Lucent Technologies, Inc., Mindspeed Technologies, and Qualcomm, Inc. The associate editor coordinating the review of this manuscript and approving it for publication was Jerome R. Bellegarda.

The authors are with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA (e-mail: liang@scl.ece.ucsb.edu; rose@ece.ucsb.edu).

Publisher Item Identifier S 1063-6676(02)02805-5.

$$\begin{array}{c|cccccccccccccc|}
 & \multicolumn{4}{c}{\mathit{Class}\ 1} & \multicolumn{4}{c}{\mathit{Class}\ 2} & & \multicolumn{4}{c}{\mathit{Class}\ M} \\
 & s_{1,1} & s_{1,2} & \cdots & s_{1,N} & s_{2,1} & s_{2,2} & \cdots & s_{2,N} & \cdots & s_{M,1} & s_{M,2} & \cdots & s_{M,N} \\
v_1 & p_{1|1,1} & p_{1|1,2} & \cdots & p_{1|1,N} & p_{1|2,1} & p_{1|2,2} & \cdots & p_{1|2,N} & \cdots & p_{1|M,1} & p_{1|M,2} & \cdots & p_{1|M,N} \\
v_2 & p_{2|1,1} & p_{2|1,2} & \cdots & p_{2|1,N} & p_{2|2,1} & p_{2|2,2} & \cdots & p_{2|2,N} & \cdots & p_{2|M,1} & p_{2|M,2} & \cdots & p_{2|M,N} \\
v_3 & p_{3|1,1} & p_{3|1,2} & \cdots & p_{3|1,N} & p_{3|2,1} & p_{3|2,2} & \cdots & p_{3|2,N} & \cdots & p_{3|M,1} & p_{3|M,2} & \cdots & p_{3|M,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
v_K & p_{K|1,1} & p_{K|1,2} & \cdots & p_{K|1,N} & p_{K|2,1} & p_{K|2,2} & \cdots & p_{K|2,N} & \cdots & p_{K|M,1} & p_{K|M,2} & \cdots & p_{K|M,N}
\end{array}$$

Fig. 1. Mixing coefficient matrix of TMHMM.

A related recent class of state tying techniques is that of phonetic tied-mixture models [14]–[16], which was enhanced by probabilistic classification of HMM states [17], and is currently used in several well-know systems (such as HTK's "soft tying" [18]). In these methods, Gaussians from a particular state are allowed to be used in other mixture distributions with similar acoustics in a manner governed by a decision tree. While these approaches overcome the initialization difficulty of the state-to-class probabilities via phonetic decision trees, they are constrained by the underlying phonetic structures, namely, pre-specified or trained tying rules, and are hence suboptimal.

In this paper, we propose the *substate tying* (SST) approach [19], where Markov states are partially tied, in contrast with traditional whole state tying where tied states are made identical. In order to develop an automatic procedure for substate tying, different from the phonetic-decision-tree-based probabilistic classification or soft 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 a better tradeoff between complexity and accuracy. Optimization of substate tying is automatically performed by a technique based on the Expectation-Maximization (EM) algorithm. We show that with this approach, the mixing efficiency in TMHMM is improved, the need for tied-state initialization is circumvented, and that the refined tradeoff yields better compromise between complexity and accuracy.

To attack the model training problem, we further propose a new approach of *combined training and reduction* (CTR) of parameters [20]. The Gaussian density codebook is first initialized with a large number of free parameters, and then downsized to the target codebook size using a proposed "minimum-partial-conditional-entropy" parameter reduction techniques. The procedure simultaneously reduces the size of the density codebook, and trains the Gaussian parameters. This optimization is performed jointly with a parameter reduction procedure that dynamically reduces the mixing coefficient matrix. The overall method is shown to significantly outperform standard TMHMM design [2] when tested on the ISOLET database and its E-set subset. These performance gains are achieved by automatic design without incorporating any prior phonetic knowledge as is commonly done in "manual" tying techniques [21].

The organization of this paper is as follows. The next section introduces the substate tying (SST) approach. In Section III, combined training and reduction (CTR) of parameters is proposed. The SST and CTR algorithms are then integrated to achieve additional gains. Experimental results are summarized and discussed in Section IV.

II. Substate Tying

A. Substate Tying Versus Whole-State Tying in TMHMM

TMHMM [1], [2] uses a universal codebook of Gaussian densities. State emission probability distributions are constructed 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 $s_{m,\,n}$—state n in the HMM representing class m, is

$$\Pr(\boldsymbol{x}|s_{m,\,n}) = \sum_{k=1}^{K} g(\boldsymbol{x}|\boldsymbol{v}_k) p_{k|m,\,n} \tag{1}$$

where $g(\cdot|\boldsymbol{v})$ is a Gaussian density whose mean and variance are specified in the parameter vector $\boldsymbol{v}$. The universal codebook may be simply represented by the set of K parameter vectors $\{\boldsymbol{v}_k,\, k=1,\,\ldots,\,K\}$ corresponding to K Gaussian densities. The mixing coefficients have obvious probabilistic interpretation $p_{k|m,\,n} = \Pr(\boldsymbol{v}_k|s_{m,\,n})$, 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: $\{p_{k|m,\,n}\}_{K\times MN}$, which is shown in Fig. 1. The traditional whole-state tying technique imposes that two (or more) columns be identical and thereby ties the corresponding states

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

The proposed substate tying approach ties subsets of the column elements and hence allows the states to be distinct

$$\begin{cases} p_{k|m_1,\,n_1} = p_{k|m_2,\,n_2}, & k \in \Phi \text{ for some } \Phi \subset \{1,\,\ldots,\,K\} \\ p_{k|m_1,\,n_1} \neq p_{k|m_2,\,n_2}, & k \notin \Phi. \end{cases}$$

Substate tying enables the implementation of intermediate levels of tying, which are not achievable by whole-state tying, and provides higher accuracy and better mixing efficiency in TMHMM. However, substate tying poses a substantial optimization challenge, as now the tying process involves many more degrees of freedom. In the next subsection we propose a simplified tying algorithm that operates on two-stage mixtures and considerably reduces the tying optimization complexity but, nevertheless, captures substantial gains due to partial state tying.

B. Substate Tying With Two-Stage Mixtures

Standard TMHMM can be viewed as tying "single-stage" mixtures as specified by the matrix of Fig. 1. In this paper, two-stage mixtures are proposed as a practical way to implement substate tying. We define the state emission probability distribution as a mixture of Gaussian mixtures, or a mixture of "submixtures." We define L submixtures $\{\boldsymbol{\omega}_l,\ 1 \le l \le L\}$, each of which is a mixture of Gaussians from a codebook of K Gaussians: $\{\boldsymbol{v}_k | 1 \le k \le K\}$. The first mixing stage is at the Gaussian level, where the "Gaussian mixing coefficient matrix" (GMCM) is used to specify the submixtures in terms of the available Gaussians (see Fig. 2). The second mixing stage is at the submixture level, as given by the "submixture mixing coefficient matrix" (SMCM) of Fig. 3. The emission probability distribution for state $s_{m,n}$ is now rewritten as the mixture

$$\Pr(\boldsymbol{x}|\boldsymbol{s}_{m,n}) = \sum_{l=1}^{L}\left\{\sum_{k=1}^{K} g(\boldsymbol{x}|\boldsymbol{v}_k)q_{k|l}\right\} p_{l|m,n} \tag{2}$$

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

$$p_{l|m,n} = \Pr(\boldsymbol{\omega}_l|s_{m,n}), \qquad q_{k|l} = \Pr(\boldsymbol{v}_k|\boldsymbol{\omega}_l)$$

which naturally satisfy

$$\sum_l p_{l|m,n} = 1 \quad \text{and} \quad \sum_k q_{k|l} = 1.$$

We make the following straightforward observations about the framework of TMHMM with two-stage mixtures:

- if GMCM is diagonal (and $L = K$), the framework degenerates to standard TMHMM;
- if GMCM is diagonal and SMCM contains identical columns, we obtain the known TMHMM with whole-state tying;
- without the above constraints, GMCM and SMCM provide a generalized framework to implement both Gaussian sharing and substate sharing, where a refined tying-accuracy tradeoff is achieved, and which subsumes the standard schemes as extreme special cases.

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])

$$\Pr(\boldsymbol{x}|\boldsymbol{s}_{m,n}) = \sum_{l\in\eta(m,n)}\left\{\sum_{k\in\zeta(m,n)} g(\boldsymbol{x}|\boldsymbol{v}_k)\tilde{q}_{k|l}\right\}\tilde{p}_{l|m,n} \tag{3}$$

where $\tilde{q}_{k|l}$ and $\tilde{p}_{l|m,n}$ are the re-normalized significant mixing coefficients. This brings about a substantial decrease in the number of free parameters without significant loss in recognition accuracy, as will be further discussed in Section III.

$$\begin{array}{c|cccc|}
 & \omega_1 & \omega_2 & \cdots & \omega_L \\
v_1 & q_{1|1} & q_{1|2} & \cdots & q_{1|L} \\
v_2 & q_{2|1} & q_{2|2} & \cdots & q_{2|L} \\
v_3 & q_{3|1} & q_{3|2} & \cdots & q_{3|L} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
v_K & q_{K|1} & q_{K|2} & \cdots & q_{K|L}
\end{array}$$

Fig. 2. Gaussian mixing coefficient matrix (GMCM) of the two-stage tied-mixture HMM (TS-TMHMM).

C. Re-Estimation

Parameter re-estimation for TMHMM with substate tying (SST-TMHMM) is similar to that of standard TMHMM except that the procedure involves three steps. Let us denote by $\gamma_{s_{m,n}}(\boldsymbol{x}_t)$ the probability that state $s_{m,n}$ is visited at time t, given that the model emits $\boldsymbol{x}_t$, i.e.,

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

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

We consider

$$\Pr(\boldsymbol{v}_k,\ \boldsymbol{\omega}_l,\ s_{m,n}|\boldsymbol{x}_t) = \gamma_{s_{m,n}}(\boldsymbol{x}_t)\,\frac{g(\boldsymbol{x}_t|\boldsymbol{v}_k)q_{k|l}\cdot p_{l|m,n}}{\Pr(\boldsymbol{x}_t|s_{m,n})} \tag{5}$$

and, by marginalization

$$\Pr(\boldsymbol{\omega}_l,\ s_{m,n}|\boldsymbol{x}_t) = \sum_{k=1}^{K}\Pr(\boldsymbol{v}_k,\ \boldsymbol{\omega}_l,\ s_{m,n}|\boldsymbol{x}_t). \tag{6}$$

Parameter re-estimation based on the EM algorithm is carried out as

1) re-estimation of Gaussian pdfs

$$\hat{\mu}_k = \frac{\sum\limits_t \sum\limits_{s_{m,n}} \sum\limits_l \Pr(\boldsymbol{v}_k,\ \boldsymbol{\omega}_l,\ s_{m,n}|\boldsymbol{x}_t)\cdot\boldsymbol{x}_t}{\sum\limits_t \sum\limits_{s_{m,n}} \sum\limits_l \Pr(\boldsymbol{\omega}_l,\ s_{m,n}|\boldsymbol{x}_t)} \tag{7}$$

$$\hat{\Sigma}_k = \frac{\sum\limits_t \sum\limits_{s_{m,n}} \sum\limits_l \Pr(\boldsymbol{v}_k,\ \boldsymbol{\omega}_l,\ s_{m,n}|\boldsymbol{x}_t)\cdot(\boldsymbol{x}_t-\hat{\mu}_k)\cdot(\boldsymbol{x}_t-\hat{\mu}_k)^T}{\sum\limits_t \sum\limits_{s_{m,n}} \sum\limits_l \Pr(\boldsymbol{\omega}_l,\ s_{m,n}|\boldsymbol{x}_t)} \tag{8}$$

where T denotes transposition;

2) re-estimation of GMCM

$$q_{k|l} = \frac{\sum\limits_t \sum\limits_{s_{m,n}} \Pr(\boldsymbol{v}_k,\ \boldsymbol{\omega}_l,\ s_{m,n}|\boldsymbol{x}_t)}{\sum\limits_t \sum\limits_{s_{m,n}} \Pr(\boldsymbol{\omega}_l,\ s_{m,n}|\boldsymbol{x}_t)}; \tag{9}$$

3) re-estimation of SMCM

$$p_{l|m,n} = \frac{\sum\limits_t \Pr(\boldsymbol{\omega}_l,\ s_{m,n}|\boldsymbol{x}_t)}{\sum\limits_t \sum\limits_{l'} \Pr(\boldsymbol{\omega}_{l'},\ s_{m,n}|\boldsymbol{x}_t)}. \tag{10}$$

$$
\begin{array}{c|ccccccccccccc|}
 & \multicolumn{4}{c}{\overbrace{\hphantom{s_{1,1}\; s_{1,2}\;\cdots\; s_{1,N}}}^{Class\;\;1}} & \multicolumn{4}{c}{\overbrace{\hphantom{s_{2,1}\; s_{2,2}\;\cdots\; s_{2,N}}}^{Class\;\;2}} & & \multicolumn{4}{c}{\overbrace{\hphantom{s_{M,1}\; s_{M,2}\;\cdots\; s_{M,N}}}^{Class\;\;M}} \\
 & s_{1,1} & s_{1,2} & \cdots & s_{1,N} & s_{2,1} & s_{2,2} & \cdots & s_{2,N} & \cdots & s_{M,1} & s_{M,2} & \cdots & s_{M,N} \\
\hline
\omega_1 & p_{1|1,1} & p_{1|1,2} & \cdots & p_{1|1,N} & p_{1|2,1} & p_{1|2,2} & \cdots & p_{1|2,N} & \cdots & p_{1|M,1} & p_{1|M,2} & \cdots & p_{1|M,N} \\
\omega_2 & p_{2|1,1} & p_{2|1,2} & \cdots & p_{2|1,N} & p_{2|2,1} & p_{2|2,2} & \cdots & p_{2|2,N} & \cdots & p_{2|M,1} & p_{2|M,2} & \cdots & p_{2|M,N} \\
\omega_3 & p_{3|1,1} & p_{3|1,2} & \cdots & p_{3|1,N} & p_{3|2,1} & p_{3|2,2} & \cdots & p_{3|2,N} & \cdots & p_{3|M,1} & p_{3|M,2} & \cdots & p_{3|M,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\omega_L & p_{L|1,1} & p_{L|1,2} & \cdots & p_{L|1,N} & p_{L|2,1} & p_{L|2,2} & \cdots & p_{L|2,N} & \cdots & p_{L|M,1} & p_{L|M,2} & \cdots & p_{L|M,N} \\
\hline
\end{array}
$$

Fig. 3. Submixture mixing coefficient matrix (SMCM) of the two-stage tied-mixture HMM (TS-TMHMM).

III. Combined Parameter Training and Reduction

A. Basic Approach

In this subsection we give a high-level description of, and motivation for, the combined training and reduction (CTR) approach, which can be applied to general HMM design, and is specialized here to two-stage mixture TMHMM design.

Let the parameter set be $\lambda = (\pi, A, B, C)$, where π is the prior probability, A is the set of state transition probabilities, B contains the Gaussian codebook parameters, and C contains all (Gaussian and submixture) mixing coefficients. Let the target number of free parameters be M_T. A high-level diagram for the CTR Algorithm is given in Fig. 4. The training process consists of two iterative optimization loops: the inner loop optimizes the system for a fixed number of free parameters (FNFP), and is hence referred to as the FNFP loop. Here, a standard HMM training technique may be used. The outer 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 performance 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.

The reduction procedure targets a subset of the HMM parameters. We will restrict our treatment to the Gaussian parameters B and mixing coefficients C. In the general derivation of CTR, the target parameters depend on the type of HMM. We seek to reduce the number of codewords in a DHMM, the number of Gaussian densities per state in a CHMM, and both the total number of Gaussian densities and the number of mixing parameters in a standard TMHMM or two-stage mixture TMHMM. The remainder of the paper will focus on the latter.

The motivation for our particular choice of CTR implementation is due to the following somewhat overlapping points: 1) design complexity is in the order of that of ML-based re-estimation; 2) cross-model considerations are involved in the design; and 3) parameter training and parameter reduction are combined.

Most of the computation performed during CTR design is in the form of ML re-design of HMM systems in the FNFP loop. ML-based re-estimation formulas are known to be relatively fast, but it ignores cross-model considerations. In the proposed approach, however, ML re-estimation is first performed on a large HMM parameter set, which is then downsized to the target size. The reduction procedure attempts to eliminate only those parameters that offer little or no contribution to the recognition performance of the system. This may be measured naturally by MCE [3], or implicitly by Maximum Mutual Information (MMI) [22], such as the approach proposed in [7]. Since only the most superfluous parameters have been removed, the system performance is roughly maintained while the total number of parameters is reduced. Once the PR loop is completed, the parametric structure of the system has been changed, and it is no longer expected to be at a local optimum. A new round of re-estimation may therefore be carried out based on the now improved initial values, and so on.

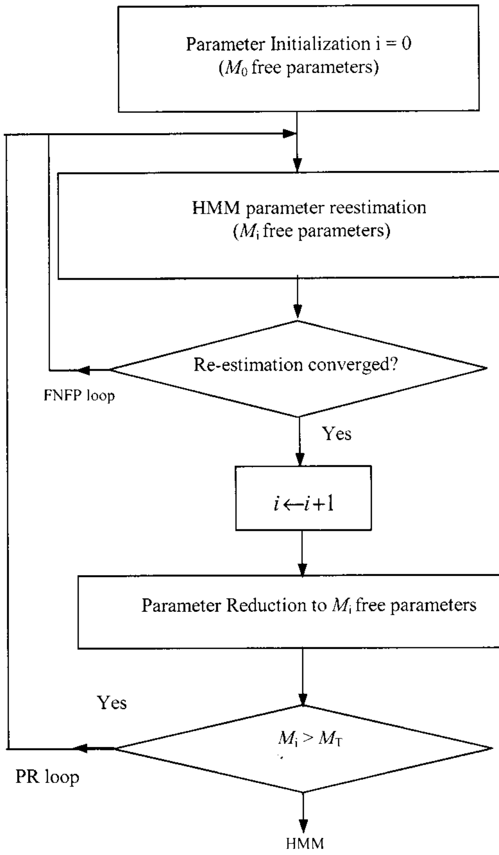

Fig. 4. CTR algorithm for HMM design.

By using ML re-estimation for the FNFP loop and alternating it with an MCE or MMI based PR loop we achieve the desired properties enumerated above. The design complexity is largely determined by the ML re-estimation procedure, and is therefore only moderate. However, cross-model information is not ignored, as the PR loop takes into account inter-class relationships to adjust the design for better discrimination.

Parameter estimation and parameter sharing have typically been considered separately in the literature. Parameter estimation is viewed as a performance-enhancing procedure. Parameter sharing techniques are mainly used for complexity-reduction at the cost of reduced recognition accuracy. However, this is not necessarily always the case. In fact, as will be shown by the CTR algorithm, parameter estimation and parameter sharing can be combined to achieve both complexity reduction and performance-enhancement.

Before proceeding with direct implementation of the approach to TMHMM design we introduce a further compromise to reduce the design complexity. Although the MCE or MMI criteria may be used effectively for the reduction process, as explained earlier, they still involve an undesirable and substantial cost in computational complexity. In this work we chose to incorporate within the framework a "minimum partial conditional entropy" parameter reduction algorithm, which substantially reduces the computational burden and, yet, achieves considerable gains. The evaluation of the merits of a high complexity approach that optimizes combined parameter estimation and reduction solely with respect to the MCE criterion is currently under investigation.

B. Parameter Reduction

The proposed parameter reduction approach in two-stage mixture TMHMM is concerned with three reducible sets of parameters: 1) Universal codebook elements or Gaussian parameter vectors v_k; 2) GMCM mixing coefficients $q_{k|l}$; and 3) SMCM mixing coefficients $p_{l|m,n}$. The reduction may be performed by operations on the matrices GMCM or SMCM, and will be explained while referring to Figs. 2 and 3. We restrict our attention to the following operations:

- row deletion in GMCM—elimination of a Gaussian density from the universal codebook;
- column deletion in GMCM, and corresponding row deletion in SMCM—elimination of a sub-Gaussian mixture;
- column element thinning in GMCM—elimination of Gaussian components from a submixture;
- column element thinning in SMCM—elimination of sub-Gaussian mixtures from a state emission distribution.

A minimum-entropy criterion has been previously proposed and used for distribution-sharing [10]. In this paper, we apply a minimum-entropy approach to row deletion within GMCM and SMCM, but, in a fundamentally different way. Our focus is on the *partial conditional entropy* (PCE) as explained next.

The marginal probability of a universal codebook element $\boldsymbol{v}_k$ is

$$\Pr(\boldsymbol{v}_k) = \sum_{l=1}^{L} \Pr(\boldsymbol{\omega}_l) \cdot q_{k|l} \tag{11}$$

where

$$\Pr(\boldsymbol{\omega}_l) = \sum_{m=1}^{M} \sum_{n=1}^{N} \Pr(s_{m,n}) \cdot p_{l|m,n}. \tag{12}$$

Consider the posterior probability

$$\Pr(\boldsymbol{\omega}_l|\boldsymbol{v}_k) = \frac{\Pr(\boldsymbol{\omega}_l) q_{k|l}}{\Pr(\boldsymbol{v}_k)} \approx \frac{\sum_{m=1}^{M} \sum_{n=1}^{N} p_{l|m,n} \cdot q_{k|l}}{\sum_{l'=1}^{L} \sum_{m=1}^{M} \sum_{n=1}^{N} p_{l'|m,n} \cdot q_{k|l'}} \tag{13}$$

where the last approximation is valid if the states are roughly equiprobable. Note that in general the criterion will be calculated without the approximation, but complexity can be saved when it is valid.

Let $H(\boldsymbol{\Omega}|\boldsymbol{v}_k)$ be the posterior submixture entropy conditional on Gaussian density $\boldsymbol{v}_k$

$$H(\boldsymbol{\Omega}|\boldsymbol{v}_k) = -\sum_{l=1}^{L} \Pr(\boldsymbol{\omega}_l|\boldsymbol{v}_k) \log\{\Pr(\boldsymbol{\omega}_l|\boldsymbol{v}_k)\} \tag{14}$$

whose computation may employ the approximation of (13). Note that a high entropy value of $H(\boldsymbol{\Omega}|\nu_k)$ may indicate less discriminatory information for the specific Gaussian pdf (i.e., all bins are uniformly distributed), and, hence, limited importance to the overall system performance, compared with other low-entropy pdfs. Moreover, let us define the partial conditional entropy $H_k(\boldsymbol{\Omega}|\boldsymbol{V})$ which measures the contribution of Gaussian density $\boldsymbol{v}_k$ to the overall conditional entropy $H(\boldsymbol{\Omega}|\boldsymbol{V})$

$$H_k(\boldsymbol{\Omega}|\boldsymbol{V}) = \Pr(\boldsymbol{v}_k) \cdot H(\boldsymbol{\Omega}|\boldsymbol{v}_k), \tag{15}$$

and

$$H(\boldsymbol{\Omega}|\boldsymbol{V}) = \sum_{k} H_k(\boldsymbol{\Omega}|\boldsymbol{V}).$$

One may view $H_k(\boldsymbol{\Omega}|\boldsymbol{V})$ as measure of the contribution of Gaussian $\boldsymbol{v}_k$ to the overall uncertainty given the selected pdf from the Gaussian pool. Hence, a higher value of $H_k(\boldsymbol{\Omega}|\boldsymbol{V})$ corresponds to less discriminatory information.

The Minimum-PCE approach to reduce the universal codebook consists of removing codebook elements with high PCE value. The Gaussian pdfs to be eliminated are selected as

$$\{\boldsymbol{v}_k\colon H_k(\boldsymbol{\Omega}|\boldsymbol{V}) \geq \alpha,\ \forall 1 \leq k \leq K\} \tag{16}$$

where α is a pre-defined entropy threshold. In this way, the number of free parameters is reduced, while the recognition accuracy is roughly maintained, which leads to a more efficient use of model parameters.

A similar definition for partial conditional entropy of states given submixture is

$$H_l(S|\boldsymbol{\Omega}) = \Pr(\boldsymbol{\omega}_l) \cdot H(S|\boldsymbol{\omega}_l) \tag{17}$$

where $\Pr(\boldsymbol{\omega}_l)$ is defined in (12) and

$$\begin{aligned} H(S|\boldsymbol{\omega}_l) &= -\sum_{m=1}^{M} \sum_{n=1}^{N} \Pr(s_{m,n}|\boldsymbol{\omega}_l) \log\{\Pr(s_{m,n}|\boldsymbol{\omega}_l)\} \\ &\approx -\sum_{m=1}^{M} \sum_{n=1}^{N} p_{l|m,n} \log(p_{l|m,n}). \end{aligned} \tag{18}$$

Here too, the last approximation assumes the equiprobability of states similarly to (13). The minimum-PCE reduction of submixtures is performed by removing the high entropy submixtures, i.e., remove

$$\{\boldsymbol{\omega}_l\colon H_l(S|\boldsymbol{\Omega}) \geq \beta,\ \forall 1 \leq l \leq L\}. \tag{19}$$

For column element thinning in SMCM, we performed a probabilistic dynamic reduction as is commonly done in traditional TMHMM design. For each state $s_{m,n}$, the thinning is performed by: sorting the mixing weights in ascending order $i \leq j \Rightarrow p_{i|m,n} \leq p_{j|m,n}$, computing

$$\hat{K} = \arg\max_{K} \left\{ \sum_{k=1}^{K} p_{k|m,n} \leq \gamma_{SMCM} \right\} \tag{20}$$

(where γ_{SMTM} is a predefined reduction rate parameter), and thinning the state's mixture by setting to zero the first $\hat{K}$ mixing coefficients: $p_{k|m,n} = 0,\ k = 1, \ldots, \hat{K}$.

Similarly, column element thinning in GMCM is implemented by sorting the mixing weights in ascending order $i \leq j \Rightarrow q_{i|k} \leq q_{j|k}$ and setting to zero the first $\hat{L}$ mixing coefficients: $p_{l|k} = 0,\ l = 1, \ldots, \hat{L}$, where

$$\hat{L} = \arg\max_{L} \left\{ \sum_{l=1}^{L} p_{l|k} \leq \gamma_{GMCM} \right\}. \tag{21}$$

IV. Experiments

To test the performance of TMHMM design with substate tying and combined parameter training and reduction, experiments were carried out on the speaker-independent ISOLET database and its E-set subset (obtained from OGI [23]). The E-set database's recognition task is to distinguish between nine confusable English letters, {b, c, d, e, g, p, t, v, z}, while the complete ISOLET involves all the 26 English letters. Both databases were generated by 150 speakers (75 male and 75 female) and include 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 was repeated 300 times with random re-partition into training and test sets, and the average performance over all trials was recorded. All reported results are in terms of test set performance.

In our experiments, 26-dimensional speech features were used, with 12 mel-frequency cepstral coefficients (MFCC) and 12 delta cepstrums, complemented with log energy and delta log energy. The analysis frame width is of 30 ms and the frame step is 10 ms. A Hamming Window was used. Two whole-word HMM models were included for each letter, to allow for variation between male and female speakers. For simplicity, but without loss of generality, we only tested TMHMM with diagonal covariance matrices. A similar design approach can be applied to TMHMM with full covariance matrices, albeit at higher complexity.

More specifically, N Markov states were assigned to each of the M whole-word HMM's. Hence, before parameter reduction, the total number of HMM states is $M \times N$. The Gaussian codebook consists of K pdfs. If the HMM's are initialized with one Gaussian pdf per state, as in all of our experiments, K equals the total number of HMM states, i.e., $K = M \times N$. Moreover, L submixtures were adopted in the proposed two-stage TMHMM (TS-TMHMM) design, where L was initialized as $M \times N$ in our experiments, to enable initialization by the conventional TMHMM training technique.

The experiments were organized as follows. We first compared the proposed substate tying within TS-TMHMM with standard CHMM, TMHMM, as well as TMHMM with whole-state tying. Secondly, we evaluated the performance of parameter reduction based on minimum-PCE and dynamic thinning, incorporated it within the framework of combined training and reduction (CTR) of parameters, and compared it to different types of HMM model design. Finally, we optimized TS-TMHMM design with the CTR algorithm, and thereby integrated both new approaches, namely, substate tying and CTR.

We recognize that the database adopted here, namely, ISOLET with the English alphabet and its E-set, represents a highly confusable task, where many letters show similar or identical vowels. As such, it offers a design challenge for testing our proposed techniques. However, there is no guarantee that the gain level achieved in this specific recognition task will be sustained in the case of large vocabulary continuous-density HMM systems.

A. Substate Tying in Tied Mixture HMM

As explained in Section II, substate tying is an intermediate level between pdf sharing (TMHMM) and whole-state tying. EM-based re-estimation is used to optimize the tying process. In our experiment, we compared substate tying with standard TMHMM [2] and TMHMM with whole-state tying [10], while CHMM is included as a base-line model. For CHMM, one single Gaussian pdf was assigned and trained for each Markov state. These Gaussian pdfs were pooled together as the initial Gaussian codebook for standard TMHMM design. During the design of whole state tying, the Markov states were clustered via the minimum entropy criterion defined in [10]. In TMHMM with substate tying, the Gaussian submixtures were initialized with Gaussian mixtures for each Markov state as in traditional TMHMM design, where GMCM was equivalent to the mixing coefficients matrix in conventional TMHMM, and SMCM was set as a diagonal matrix. For a concrete example, if the number of states for each HMM is 7 (as shown in the first row of Table I), there will be 7 distinct Gaussian pdfs in each CHMM (one for each state), a distribution pool of 7 single Gaussian pdfs in each TMHMM, and a submixture pool of 7 Gaussian submixtures, consisting of 7 distinct single Gaussian pdfs, in each TS-TMHMM. While the number of single Gaussian pdfs and Gaussian submixtures were fixed here, the constraint can be relaxed by the parameter reduction technique shown in next subsection.

The experimental results for the E-set and the ISOLET database are shown in Tables I and II, respectively. They demonstrate that, at the same number of HMM states (i.e., the same number of single Gaussian pdfs), substate tying

TABLE I
ERROR RATE (%) OF VARIOUS PARAMETER-SHARING TECHNIQUES ON THE E-SET SPEECH DATABASE ($N = 7 \sim 21, M = 18, K = 18N, L = 18N$)

No. of states per HMM	CHMM	TMHMM	TMHMM with whole-state tying	TMHMM with sub-state tying
7	16.8	15.4	14.8	13.0
9	15.3	14.0	13.0	11.3
11	14.1	12.7	11.5	9.8
13	13.2	12.0	10.8	9.1
15	12.8	11.6	10.3	8.6
17	12.6	11.2	9.9	8.1
19	12.5	10.8	9.5	7.6
21	12.4	10.7	9.4	7.4

TABLE II
ERROR RATE (%) OF VARIOUS PARAMETER-SHARING TECHNIQUES ON THE ISOLET DATABASE ($N = 7 \sim 21, M = 52, K = 52N, L = 52N$)

No. of states per HMM	CHMM	TMHMM	TMHMM with whole-state tying	TMHMM with sub-state tying
7	8.5	7.8	7.5	6.8
9	7.8	7.3	6.9	5.9
11	7.3	6.8	6.1	5.3
13	7.0	6.4	5.7	4.9
15	6.8	6.1	5.5	4.6
17	6.7	5.9	5.3	4.5
19	6.6	5.7	5.2	4.4
21	6.6	5.6	5.1	4.4

offer consistent performance gains over CHMM, standard TMHMM, and TMHMM-based whole-state tying. Although the number of mixing coefficients are increased from TMHMM to TS-TMHMM, this additional complexity is minimal after probabilistic dynamic reduction is applied to both GMCM and SMCM, as has been described in Section III-B.

B. Parameter Reduction

Parameter reduction in TS-TMHMM can be performed over various sets of parameters, including Gaussian pdfs and mixing coefficients in the GMCM and SMCM of Figs. 2 and 3. As described in Section III-B, Gaussian pdfs and submixtures are reduced based on the minimum-PCE criterion, while the column elements of GMCM and SMCM are dynamically thinned. We performed several experiments to investigate the feasibility and merits of the reduction procedure. α, β, γ_{GMCM} and γ_{SMCM} are chosen dynamically so that a fixed percentage (10% in our experiments) of parameters are eliminated in each parameter set during each parameter reduction iteration, with a total of 5 iterations for each experiment. The TS-TMHMM parameters were initialized as described in the previous subsection.

Tables III–VI show the E-set performance before and after each type of parameter reduction, at a typical number of states per HMM. Table III shows that about 30% of the Gaussian pdfs can be removed without significant loss of accuracy. Similarly in Table IV, 40% of the submixtures can be eliminated with minimal impact on system performance. Tables V and VI show that, on the average, about 80% of the elements can be discarded from GMCM and SMCM with little or no increase in recognition error. The results demonstrate that, at least for the E-set database, various free parameter sets can be substantially reduced in TS-TMHMM at the cost of minimal decline in performance. (TMHMM is a special case of TS-TMHMM and exhibits similar behavior.)

TABLE III
THE IMPACT OF MINIMUM-PCE GAUSSIAN pdf REDUCTION ON E-SET PERFORMANCE (BEFORE REDUCTION, $N = \{7, 13, 19\}, M = 18, K = 18N$)

No. of states per HMM	Before Reduction		After Reduction	
	No. of Gaussian pdfs	Error rate (%)	No. of Gaussian pdfs	Error rate (%)
7	126	15.0	98	15.2
13	234	12.0	174	12.1
19	342	8.0	245	8.0

TABLE IV
THE IMPACT OF MINIMUM-PCE SUBMIXTURE REDUCTION ON E-SET PERFORMANCE (BEFORE REDUCTION, $N = \{7, 13, 19\}, M = 18, K = 18N, L = 18N$)

No. of states per HMM	Before Reduction		After Reduction	
	No. of sub-mixtures	Error rate (%)	No. of sub-mixtures	Error rate (%)
7	126	15.4	110	15.8
13	234	12.0	200	12.0
19	342	8.0	270	7.9

TABLE V
THE IMPACT OF DYNAMIC GMCM COLUMN THINNING ON E-SET PERFORMANCE (BEFORE REDUCTION, $N = \{7, 13, 19\}, M = 18, K = 18N, L = 18N$, THE DIMENSION OF GMCM IS $K \times L$)

No. of states per HMM	Before Reduction		After Reduction	
	No. of elements in GMCM	Error rate (%)	No. of elements in GMCM	Error rate (%)
7	15900	15.0	330	15.4
13	54800	11.7	680	12.0
19	117000	7.8	1120	8.0

TABLE VI
THE IMPACT OF DYNAMIC SMCM COLUMN THINNING ON E-SET PERFORMANCE (BEFORE REDUCTION, $N = \{7, 13, 19\}, M = 18, K = 18N, L = 18N$, THE DIMENSION OF SMCM IS $K \times L$)

No. of states per HMM	Before Reduction		After Reduction	
	No. of elements in SMCM	Error rate (%)	No. of elements in SMCM	Error rate (%)
7	15900	15.8	130	15.4
13	54800	12.4	280	12.0
19	117000	8.3	460	8.0

C. CTR Experiments

The combined training and reduction algorithm has been evaluated on the E-set database. In the below experiments, all TMHMM parameters (including Gaussian pdfs) are initialized by two iterations of CHMM re-estimation. All HMM models use the same fixed number of states for each utterance. For

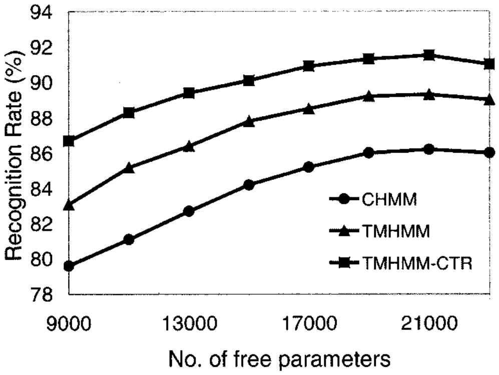

Fig. 5. E-set performance comparison of TMHMM-CTR with standard CHMM and TMHMM, shown versus model complexity.

TABLE VII
ERROR RATE (%) OF VARIOUS PARAMETER TRAINING AND PARAMETER SHARING TECHNIQUES ON THE E-SET SPEECH DATABASE

No. of free parameters	TMHMM	TMHMM with whole-state tying	TS-TMHMM	TS-TMHMM with CTR
7,000	15.2	14.6	13.3	12.0
13,000	11.6	10.8	9.5	8.5
22,000	12.4	9.4	7.4	6.8

TABLE VIII
ERROR RATE (%) OF VARIOUS PARAMETER TRAINING AND PARAMETER SHARING TECHNIQUES ON THE ISOLET DATABASE

No. of free parameters	TMHMM	TMHMM with whole-state tying	TS-TMHMM	TS-TMHMM with CTR
60,000	7.8	7.5	6.8	6.0
110,000	6.4	5.8	5.1	4.6
180,000	5.6	5.2	4.4	4.1

combined training and reduction, TMHMM is initialized with a larger number of HMM states (about 30% more than the target number of free parameters), and is gradually downsized to the target parameter size via three PR loop iterations (see Fig. 4). The results are shown in Fig. 5. For the E-set database, CTR gained in performance relative to standard CHMM and TMHMM at the same number of parameters. Notice that as the training set is limited, the HMM model becomes over-trained when the number of parameters approaches 21 000. CTR-trained TMHMM achieved higher accuracy at this level of saturation, and demonstrated its superiority over both CHMM and TMHMM. Similar improvement had also been achieved for the ISOLET database.

D. TS-TMHMM Design With CTR

The CTR algorithm can be applied to the design of TS-TMHMM to realize the combined benefits of two new approaches, as is seen from Tables VII and VIII. The results demonstrate that, under equivalent complexity (i.e., same number of free parameters) and for both E-set and the ISOLET database, TS-TMHMM designed by CTR achieved over 20% error rate reduction over TMHMM-based whole state tying, and more than 25% error rate reduction over standard TMHMM.

V. 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 substate 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, under the new structure of two-stage tied-mixture HMM (TS-TMHMM) jointly optimizes Gaussian sharing and substate tying by EM-based re-estimation. In simulations over the ISOLET database and its E-set subset, SST reduced the recognition error rate by 15% compared to the conventional techniques of TMHMM with whole-state tying.

Model training is another critical problem in HMM design. For TMHMM design, this procedure includes selection and estimation of the Gaussian density codebook and the mixing coefficients. The combined training and reduction (CTR) algorithm proposed in this paper maintains complexity similar to that of ML-based training, but employs the minimum-partial-conditional-entropy criterion to provide improved training results. Experiments on the E-set and ISOLET database demonstrate that CTR can reduce the recognition error rate by over 20% compared with the benchmark TMHMM model. The basic CTR algorithm is not restricted to TMHMM, and is expected to improve HMM training performance significantly with other structures.

TS-TMHMM for substate tying can be further embedded within the combined training and reduction framework to provide additional improvement. In our experiment on the E-set and ISOLET databases, TS-TMHMM designed by CTR achieved more than 25% error rate reduction over conventional TMHMM with whole-state tying. Future work will focus on incorporation of powerful optimization tools within the CTR framework to further exploit the potential of the framework.

REFERENCES

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

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

[3] B. H. Juang and S. Katagiri, "Discriminative learning for minimum classification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 40, no. 12, pp. 3043–3054, Dec. 1992.

[4] A. Rao and K. Rose, "Deterministically annealed design of hidden Markov model speech recognizers," *IEEE Trans. Speech Audio Processing*, vol. 9, Feb. 2001.

[5] L. R. Bahl and M. Padmanabhan, "A discriminant measure for model complexity estimation," in *Proc. ICASSP*, 1998, pp. 453–455.

[6] Y. Normandin, "Optimal splitting of HMM Gaussian mixture components with MMIE training," in *Proc. ICASSP*, 1995, pp. 449–452.

[7] M. Padmanabhan and L. R. Bahl, "Model complexity adaptation using a discriminant measure," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 205–208, Mar. 2000.

[8] L. R. Bahl *et al.*, "Performance of the IBM large vocabulary continuous speech recognizer on the ARPA Wall Street Journal task," in *Proc. ICASSP*, 1995, pp. 41–44.

[9] S. S. Chen and P. S. Gopalakrishnan, "Clustering via the Bayesian information criterion with applications in speech recognition," in *Proc. ICASSP*, 1995, pp. 645–648.

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

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

[12] C. Dugast, P. Beyerlein, and R. Haeb-Umbach, "Application of clustering techniques to mixture density modeling for continuous-speech recognition," in *Proc. ICASSP*, 1995, pp. 524–527.

[13] 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*, vol. 6, pp. 61–70, Jan. 1998.

[14] G. Zavaliagkos *et al.*, "The BBN BYBLOS 1997 large vocabulary conversational speech recognition system," in *Proc. ICASSP*, 1998, pp. 905–908.

[15] A. Sankar, "A new look at HMM parameter tying for large vocabulary speech recognition," in *Proc. ICSLP*, 1998, pp. 2219–2222.

[16] A. Lee *et al.*, "A new phonetic tied-mixture model for efficient decoding," in *Proc. ICASSP*, 2000, pp. 1269–1272.

[17] X. Luo and F. Jelinek, "Improved clustering techniques for class-based statistical language modeling," in *Proc. ICASSP*, Phoenix, AZ, 1999, pp. 2044–2047.

[18] D. Povey and P. C. Woodland, "Frame discrimination training of HMM's for large vocabulary speech recognition," in *Proc. ICASSP*, Phoenix, AZ, 1999, pp. 333–336.

[19] L. Gu and K. Rose, "Sub-state tying in tied mixture hidden Markov models," in *Proc. ICASSP'2000*, Istanbul, Turkey, Jun. 2000, pp. 1013–1016.

[20] ——, "Combined parameter training and reduction in tied-mixture HMM design," in *Proc. IEEE ASRU*, Dec. 1999.

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

[22] V. Valtchev, J. J. Odell, P. C. Woodland, and S. J. Young, "MMIE training of large vocabulary recognition systems," *Speech Commun.*, vol. 22, no. 4, pp. 303–314, Sept. 1997.

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

Liang Gu (S'96–M'02) received the B.S. and M.S. degrees in electronic engineering from Tsinghua University, Beijing, China, in 1993 and 1996, respectively, and the Ph.D. degree in electrical and computer engineering from University of California, Santa Barbara, in 2001.

From January 1998 to February 2002, he was with the Signal Compression Laboratory, University of California, Santa Barbara, as a Graduate Student Researcher and later as a Postdoctoral Researcher. He is currently a Researcher with the IBM T. J. Watson Research Center, Yorktown Heights, NY. His research interests lie in digital signal processing including automatic speech recognition, speech-to-speech translation, pattern recognition, multimedia indexing, and real-time embedded system design. His current research involves robust machine-translation and speech recognition in multilingual speech-to-speech translation.

Dr. Gu received the ISCA Best Student Paper Award addressing original speech communications at Eurospeech 2001.

Kenneth Rose (S'85–M'85–SM'01), received the B. Sc. and M.Sc. degrees in electrical engineering from Tel-Aviv University, Tel-Aviv, Israel, in 1983 and 1987, respectively. He received the Ph.D. degree from the California Institute of Technology, Pasadena, in 1991. >

From July 1983 to July 1988 he was with Tadiran Ltd., Israel, where he carried out research in the areas of image coding, image transmission through noisy channels, and general image processing. In January 1991, he joined the Department of Electrical and Computer Engineering, University of California, Santa Barbara, where he is currently a Professor. His main research activities are in information theory, source and channel coding, image coding and processing, speech and general pattern recognition, and nonconvex optimization in general. He is also particularly interested in the relations between information theory and statistical physics, and their potential impact on fundamental and practical problems in diverse disciplines.

Dr. Rose currently serves as Editor for Source-Channel Coding for the IEEE TRANSACTIONS ON COMMUNICATIONS. He co-chaired the technical program committee of the 2001 IEEE Workshop on Multimedia Signal Processing. In 1990, he received the William R. Bennett Prize Paper Award (with A. Heiman) from the IEEE Communications Society.