Combined Parameter Training and Reduction in Tied-Mixture HMM Design

Liang Gu and *Kenneth Rose*

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

ABSTRACT

A tied-mixture HMM speech recognizer design approach is proposed, which combines parameter training with parameter reduction. 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 weight matrix, followed by parameter re-training. The additional cost in design complexity is modest. Preliminary experimental results on the E-set show that the classification error rate is reduced by over 20% compared to standard tied-mixture or semi-continuous HMM design. This improvement is obtained both inside and outside the training set.

1. INTRODUCTION

The Hidden Markov Model (HMM) is widely recognized as a useful statistical tool for automatic speech recognition. Model parameter training has long been recognized as a critical part of the system design. While the natural objective of training is accurate classification of utterances, training has traditionally been performed using the maximum likelihood (ML) criterion. The corresponding re-estimation algorithms are effective and of manageable complexity. However, ML optimization suffers from the inherent and fundamental mismatch with the natural "true" objective, namely, minimum classification error (MCE). Instead of MCE, the traditional procedure attempts to optimize ML, which is the objective of the modeling problem. Recently, a new class of methods that directly optimize MCE has been proposed [1]. 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 non-global optima (see [2] for the deterministic annealing approach to resolve this difficulty). Finally, the gains of MCE may not generalize well outside the training set. This paper is concerned with the search for an approach that offers gains over ML methods, at minimal additional design complexity cost, which generalize well outside the training set.

Optimal design of speech recognizers, based on limited training data, must take into account the fundamental tradeoff between model richness and robustness. Tied-mixture HMM (TMHMM) [3] [4] represents an important approach to optimization of this tradeoff (In this paper, we use the term TMHMM to refer generally to methods where mixtures are tied, and experiment comparison will be made with respect to the TMHMM method of [4]). With its universal set of density functions for constructing state emission mixtures, TMHMM offers the modeling capability of a large-mixture continuous HMM (CHMM), but with a substantially reduced number of free 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, two inter-related problems remain open: 1) how to optimally design the state emission density codebook; and 2) how to select and train the tying parameters. Appropriate selection of emission (typically Gaussian) density parameters is of paramount importance to the performance of continuous HMM systems [5]. Further, the choice and effective training of the tying parameters is at the heart of TMHMM, and is recognized as a challenging problem. Good selection and estimation of tying parameters can improve the model accuracy with little or no increase of model complexity and, thus, optimize the tradeoff between model complexity and robustness. Various tying techniques [6][7][8] have been developed over the last ten years, and considerable improvement has been achieved. However, these methods have mainly focused on trading performance for reduced computational complexity. The objective of the work presented here is to develop an *automatic* parameter tying approach so as to *improve* performance outside the training set.

A new approach is proposed, which is based on *combined parameter training and parameter reduction*. The Gaussian density codebook is first initialized with a large number of free parameters, and then downsized to the target codebook size using minimum-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 tying-weight matrix. The overall method is shown to significantly outperform standard TMHMM design [4]. These performance gains are achieved by automatic design without incorporating any prior phonetic knowledge as is commonly done in "manual" tying techniques [9].

2. COMBINED TRAINING AND REDUCTION

In this section we give a high-level description of, and motivation for, the combined training and reduction (CTR) approach. The approach will be applied to TMHMM design in the next section.

This work was supported in part by the National Science Foundation, 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, Intel Corp., Lernout & Hauspie Speech Products, Lockheed Martin, Lucent Technologies, Inc., Qualcomm, Inc., Panasonic Technologies, Inc., and Texas Instruments, Inc.

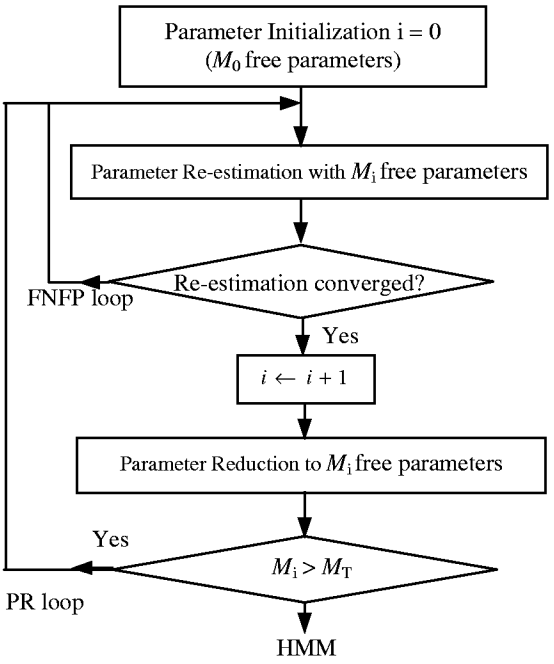

Figure 1. Combined Training and Reduction Algorithm

Let the HMM parameter set be $\lambda = (\pi, A, B)$, where π is the prior probability, A is the set of state transition probabilities, and B contains the state emission probability distributions. Let the target number of free parameters be M_T. A high-level diagram for the CTR Algorithm is given in Figure 1. 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. Here, the standard HMM training technique may be used. The other loop optimizes decisions for parameter reduction 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, continue 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 state emission parameters B. (One may consider extensions to state-tying, *i.e.*, apply reduction to A as well.) More specifically, we seek to reduce the number of codewords in a DHMM, the number of Gaussian densities per state in a CHMM model, and both the total number of Gaussian densities and the number of mixing parameters in a TMHMM model. The remainder of the paper will focus on the latter.

The motivation for the proposed CTR is due to the following somewhat overlapping points: 1) design complexity is in the order of that of ML-based re-estimation; 2) MCE considerations are involved in the design; 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 as relatively fast but mismatched with MCE. In this paper, ML re-estimation is first performed on a large HMM parameter set, which is then downsized to the target size. The reduction procedure aims at eliminating only those parameters that offer little or no contribution to the recognition performance of the system. This may naturally be measured by MCE. Since only MCE-superfluous parameters have been removed, the system performance is roughly maintained while the 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.

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

Parameter estimation and parameter sharing have been commonly 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 for the CTR algorithm, parameter estimation and parameter sharing can be combined to achieve both complexity reduction and performance-enhancement.

Before proceeding with direct application of the approach to HMM design we introduce a further compromise to restrict design complexity. Although the MCE criterion may be used effectively for the reduction process, as explained earlier, it still involves an undesirable cost in computational complexity. In this work we chose to incorporate within the framework a minimum-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.

3. APPLICATION OF CTR TO TMHMM DESIGN

Unlike traditional CHMM, TMHMM [4] 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 $s_{m,n}$ is:

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

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

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

The proposed CTR approach is concerned with two reducible sets of parameters: 1) Universal codebook elements or Gaussian parameter vectors $\boldsymbol{v}_k$; and 2) Mixing coefficients $p_{k|m,n}$. The first set has global properties as its parameters involve all classes and states. The second set consists of state-specific parameters. Both types of reduction may be captured by operations on the *mixing weight matrix* (MWM): $\{p_{k|m,n}\}_{K\times M\times N}$, which is shown in Figure 2. In this work we restrict our attention to three possible operations:

- Row deletion - elimination of a Gaussian density from the universal codebook.
- Column sharing - distribution sharing by two states (see state clustering [8] and distribution sharing [7]).
- Column element thinning – elimination of Gaussian components from a state emission distribution

A minimum-entropy criterion has been effectively used in parameter-sharing [7]. In this paper, we apply the minimum-entropy approach to row deletion, however, in a fundamentally different way. Our focus is on the *posterior conditional entropy* as explained next.

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

$$\Pr(v_k) = \sum_{m=1}^{M}\sum_{n=1}^{N} \Pr(s_{m,n}) \cdot p_{k|m,n}$$

Consider the posterior probability

$$\Pr(s_{m,n}|v_k) = \frac{\Pr(s_{m,n})p_{k|m,n}}{\sum_{m=1}^{M}\sum_{n=1}^{N}\Pr(s_{m,n})p_{k|m,n}} \approx \frac{p_{k|m,n}}{\sum_{m=1}^{M}\sum_{n=1}^{N} p_{k|m,n}}$$

where the rightmost approximation is valid if the states are equiprobable. Thus the (posterior) conditional entropy for a Gaussian density is

$$H(v_k) = -\sum_{m=1}^{M}\sum_{n=1}^{N} \Pr(s_{m,n}|v_k)\log\{\Pr(s_{m,n}|v_k)\}$$

The minimum-entropy reduction of the universal codebook is performed by: sorting the codebook elements in increasing order of entropy $i < j \Rightarrow H(v_i) \le H(v_j)$, computing

$$\hat{L} = \arg\max_{L}\left\{\sum_{l=1}^{L} H(v_l) \le \alpha\right\},$$

where α is a pre-defined reduction rate parameter, and removing the elements $\{v_l, 1 \le l \le \hat{L}\}$.

The dynamic reduction rate defined by α can be replaced with a fixed reduction rate, where the first (constant) L pdfs are removed.

For MWM column sharing there are known techniques such [8] or [7]. Several criteria have been proposed including minimum entropy, minimum divergence, maximum likelihood, etc. In our simulations we used the minimum squared error (MSE) distance, and have found it to perform well.

For column element thinning, instead of the fixed number reduction which was applied in [1], we propose to use a probability-based dynamic reduction. For each state $s_{m,n}$, the thinning is performed by: sorting the mixing weights in ascending order $i \le j \Rightarrow p_{i|m,n} \le p_{j|m,n}$, computing

$$\hat{L} = \arg\max_{L}\left\{\sum_{l=1}^{L} p_{l|m,n} \le \beta\right\}$$

(where β is a predefined reduction rate parameter), and thinning the state's mixture by setting to zero the first $\hat{L}$ mixing

$$\begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_K \end{array}
\begin{array}{c}
\overbrace{\phantom{s_{1,1}\ s_{1,2}\ \cdots\ s_{1,N}}}^{Class\ 1}\ \overbrace{\phantom{s_{2,1}\ s_{2,2}\ \cdots\ s_{2,N}}}^{Class\ 2}\ \cdots\ \overbrace{\phantom{s_{M,1}\ s_{M,2}\ \cdots\ s_{M,N}}}^{Class\ M} \\
s_{1,1}\ \ s_{1,2}\ \cdots\ s_{1,N}\ \ s_{2,1}\ \ s_{2,2}\ \cdots\ s_{2,N}\ \cdots\ s_{M,1}\ \ s_{M,2}\ \cdots\ s_{M,N} \\
\begin{bmatrix}
p_{1|1,1} & p_{1|1,2} & \cdots & p_{1|1,N} & p_{1|2,1} & p_{1|2,2} & \cdots & p_{1|2,N} & \cdots & p_{1|M,1} & p_{1|M,2} & \cdots & p_{1|M,N} \\
p_{2|1,1} & p_{2|1,2} & \cdots & p_{2|1,N} & p_{2|2,1} & p_{2|2,2} & \cdots & p_{2|2,N} & \cdots & p_{2|M,1} & p_{2|M,2} & \cdots & p_{2|M,N} \\
p_{3|1,1} & p_{3|1,2} & \cdots & p_{3|1,N} & p_{3|2,1} & p_{3|2,2} & \cdots & p_{3|2,N} & \cdots & p_{3|M,1} & p_{3|M,2} & \cdots & p_{3|M,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
p_{K|1,1} & p_{K|1,2} & \cdots & p_{K|1,N} & p_{K|2,1} & p_{K|2,2} & \cdots & p_{K|2,N} & \cdots & p_{K|M,1} & p_{K|M,2} & \cdots & p_{K|M,N}
\end{bmatrix}
\end{array}$$

Figure 2. Mixing Weight Matrix of TMHMM

methods	No. states per HMM	No. of distinct Gaussian pdfs	Train Set Error Rate	Test Set Error Rate
CHMM	13	234	7.8 %	17.7 %
Standard TMHMM	13	234	7.5 %	15.2 %
TMHMM - CTR	13	234	5.6 %	11.3 %

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

Methods	No. states per HMM	No. of free parameters	Train Set Error Rate	Test Set Error Rate
CHMM	22	28512	5.6 %	11.0 %
Standard TMHMM	19	28044	4.8 %	10.8 %
TMHMM - CTR	19	27936	3.9 %	8.1 %

Table 2. Performance comparison of HMM design methods with similar number of free parameters

coefficients: $p_{l|m,n} = 0, l = 1, \ldots, \hat{L}$.

4. EXPERIMENT RESULTS

To test the performance of CTR on TMHMM design, experiments were carried out on the E-set speech database obtained from OGI [10]. 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 is 30ms, the analysis frame step is 10ms, and a Hamming Window is used. Two HMM models were used 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 various HMM design methods 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 [4], to TMHMM-CTR. Note that TMHMM in this case has more free parameters because of the mixing weights. In Table 2, further comparison is given between CHMM, standard TMHMM and TMHMM-CTR at similar number of free HMM parameters. TMHMM-CTR offers the best performance and achieves reduction in error of more than 20% relative to standard TMHMM. Alternatively, TMHMM-CTR can achieve similar recognition accuracy but with a greatly reduced set of Gaussian components, and with enhanced robustness (this alternative version of the results is not shown for lack of space)

5. CONCLUSION

Selection and training of the Gaussian density codebook and the tying parameter set are two critical issues for TMHMM. The combined training and reduction (CTR) algorithm proposed in this paper maintains complexity similar to that of ML-based training while providing improved training results. Experiments 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 in other set-ups as well. Future work will focus on incorporation of powerful optimization tools within the CTR framework to achieve further improvement.

6. REFERENCES

[1] B. H. Juang and S. Katagiri, "Discriminative learning for minimum classification", *IEEE Transactions on Acoustics, Speech, and Processing*, vol. 40, no. 12, pp. 3043-3054, Dec. 1992.

[2] A. Rao, K. Rose and A. Gersho, "Deterministically annealed design of speech recognizers and its performance on isolated letters", *IEEE ICASSP'98*, May 1998.

[3] 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.

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

[5] Y. Zhang; R. Togneri, and M. Alder, "Phoneme-based vector quantization in a discrete HMM speech recognizer", *IEEE Transactions on Speech and Audio Processing*, vol.5, no.1, Jan. 1997.

[6] 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.

[7] 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.

[8] 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.

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

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