TRAINING OF SUBSPACE DISTRIBUTION CLUSTERING HIDDEN MARKOV MODEL

Brian Mak * *and Enrico Bocchieri*

AT&T Labs – Research
180 Park Ave, Florham Park, NJ 07932, USA
{mak, enrico}@research.att.com

ABSTRACT

In [2] and [7], we presented our novel subspace distribution clustering hidden Markov models (SDCHMMs) which can be converted from continuous density hidden Markov models (CDHMMs) by clustering subspace Gaussians in each stream over *all* models. Though such model conversion is simple and runs fast, it has two drawbacks: (1) it does not take advantage of the fewer model parameters in SDCHMMs — theoretically SDCHMMs may be trained with smaller amount of data; and, (2) it involves two separate optimization steps (first training CDHMMs, then clustering subspace Gaussians) and the resulting SDCHMMs are not guaranteed to be optimal. In this paper, we show how SDCHMMs may be trained directly from less speech data if we have a priori knowledge of their architecture. On the ATIS task, a speaker-independent, context-independent (CI) 20-stream SDCHMM system trained using our novel SDCHMM reestimation algorithm with only 8 minutes of speech performs as well as a CDHMM system trained using conventional CDHMM reestimation algorithm with 105 minutes of speech.

1. INTRODUCTION

The history of acoustic modeling is guided by the need to strike a balance between the two conflicting goals: trainability and resolution of acoustic models. That is, the acoustic models should contain enough fine acoustic details so that different models can be resolved during decoding but too much detail generally reduces the robustness of model parameters when estimated from limited amounts of training data. In the past the technique of parameter tying has been applied successfully to obtain such a balance by reducing the number of parameters in acoustic models at various granularities: phone (generalized biphones/triphones [6], context-independent phones), state (tied-state HMM [10]), observation distribution (tied-mixture/semi-continuous HMM [4]) and distribution parameters [9] have all been tied.

In the past, our subspace distribution clustering hidden Markov model (SDCHMM) was mainly presented as an approximation to the continuous density hidden Markov model (CDHMM). K-stream SDCHMMs are converted from CDHMMs by (1) decomposing the feature space into K disjoint subspaces or streams; and, (2) clustering subspaces Gaussians from *all* states and *all* phone models in each subspace. Thus one may also consider SDCHMMs as CDHMMs tied at a smaller sub-phonetic unit, namely subspace Gaussian. Since we have shown that SDCHMMs with many fewer model parameters — by one or two orders of magnitude — work as well as the CDHMMs they derive from, one should be able to train SDCHMMs directly from much less speech data. Such a direct SDCHMM training scheme will also guarantee that the trained SDCHMMs are optimal in, for example, the maximum likelihood sense.

In this paper, we present the reestimation formulas of SDCHMM parameters and show how to train SDCHMMs from speech data without intermediate CDHMMs.

*The work was accomplished while Brian Mak was a PhD candidate of Oregon Graduate Institute of Science & Technology, 20000 NW Walker Rd., Portland, OR 97006, USA.

2. REESTIMATION FORMULAS OF SDCHMM

SDCHMM parameters may be reestimated using the EM algorithm in much the same way as CDHMM parameters [8].

2.1. Basics of HMM Reestimation Using EM Algorithm

We will use the following notations in our discussion:

t	: time index
N	: number of states in a HMM
$\mathbf{o}_t$	: feature vector at time t
$\mathbf{O}$	: a time sequence of feature vectors $(\mathbf{o}_1 \mathbf{o}_2 \cdots \mathbf{o}_T)$
λ	: one current HMM
λ'	: reestimated HMM based on λ
Λ	: a set of HMMs
$\mathbf{a}$	: state transition prob. matrix $\{a_{ij}\}, 1 \le i, j \le N-1$
$\mathbf{a}_i$	: $[a_{i1}, a_{i2}, \ldots, a_{iN}]$
$\mathbf{b}$	: state observation pdf $[b_1, b_2, \ldots, b_N]$
π	: initial state probability $[\pi_1, \pi_2, \ldots, \pi_N]$
$\mathbf{q}$	: a state sequence $(q_0 q_1 \cdots q_T)$
$\gamma_t^{\lambda}(j)$	: probability of staying in state j of model λ at time t

As usual quantities in boldface are matrices or vectors, otherwise they are scalars.

Given $\lambda'(\cdot) = \lambda'(\mathbf{a}, \mathbf{b}, \pi)$, the auxiliary Q function used in the EM algorithm is:

$$Q(\lambda, \lambda') = \sum_{q} P(\mathbf{O}, \mathbf{q} \mid \lambda) \cdot \log P(\mathbf{O}, \mathbf{q} \mid \lambda'). \quad (1)$$

Since

$$\log P(\mathbf{O}, \mathbf{q} \mid \lambda') = \log(\pi_{q_0}) + \sum_{t=1}^{T} \log(a_{q_{t-1} q_t}) + \sum_{t=1}^{T} \log(b_{q_t}(\mathbf{o}_t)),$$

we may separate Eqn.(1) into a sum of three independent Q functions:

$$Q(\lambda, \lambda') = Q_{\pi}(\lambda, \pi) + \sum_{i=1}^{N} Q_{\mathbf{a}_i}(\lambda, \mathbf{a}_i) + \sum_{i=1}^{N} Q_{b_i}(\lambda, b_i) \quad (2)$$

0-7803-4428-6/98 $10.00 © 1998 IEEE

where,

$$Q_\pi(\lambda,\pi)=\sum_{i=1}^{N} P(\mathbf{O},q_0=i\mid\lambda)\ \log(\pi_i) \tag{3}$$

$$Q_{\mathbf{a}_i}(\lambda,\mathbf{a}_i)=\sum_{j=1}^{N}\sum_{t=1}^{T} P(\mathbf{O},q_{t-1}=i,q_t=j\mid\lambda)\ \log(a_{ij}) \tag{4}$$

$$Q_{b_i}(\lambda,b_i)=\sum_{t=1}^{T} P(\mathbf{O},q_t=i\mid\lambda)\ \log(b_i(\mathbf{o}_t)). \tag{5}$$

Maximization of $Q(\lambda,\lambda')$ can be done by maximizing the three independent Q functions separately, since each involves a different set of optimization variables.

2.2. Reestimation of π and a in SDCHMM

As we only manipulate the state observation pdf $b_j(\mathbf{o}_t)$, the reestimation formulas for the initial state probabilities (π) and the state transition probabilities (a) will remain the same as those of conventional HMM.

2.3. Reestimation of b in SDCHMM

The state observation pdf for each state, j, is assumed to be a mixture density with M components, b_{jm}, and mixture weights, c_m, $1\le m\le M$. Then by the definition of SDCHMM [2, 7] with K *locally* independent streams, we have

$$b_j(\mathbf{o}_t)=\sum_{m=1}^{M} c_m \prod_{k=1}^{K} b_{jmk}(\mathbf{o}_{tk}) \tag{6}$$

where b_{jmk} and $\mathbf{o}_{tk}$ are the projections of b_{jm} and $\mathbf{o}_t$ onto the k-th feature subspace (or stream).

Since an HMM state with an M-mixture density is equivalent to a multi-state with single-mixture densities [1], for simplicity reason, let us consider without loss of generality only the case that there is 1 mixture in each state observation pdf. Thus

$$b_j(\mathbf{o}_t)=\prod_{k=1}^{K} b_{jk}(\mathbf{o}_{tk}). \tag{7}$$

Substituting Eqn.(7) into Eqn.(5), we have

$$\begin{aligned} Q_{b_j}(\lambda,b_j) &= \sum_{t=1}^{T} P(\mathbf{O},q_t=j\mid\lambda)\left(\sum_{k=1}^{K}\log(b_{jk}(\mathbf{o}_{tk}))\right) \\ &= \sum_{k=1}^{K}\left(\sum_{t=1}^{T} P(\mathbf{O},q_t=j\mid\lambda)\log(b_{jk}(\mathbf{o}_{tk}))\right) \\ &\equiv \sum_{k=1}^{K} Q_{b_{jk}}(\lambda,b_{jk}). \end{aligned} \tag{8}$$

As the streams or subspaces are assumed independent in the *local* acoustic space (**not** global acoustic space as in tied-mixture HMM), each $Q_{b_{jk}}(\lambda,b_{jk})$ can be maximized independently.

Now suppose in each subspace k, subspace pdf's are clustered into L pdf codewords $h_{kl}(\cdot)$ where $1\le l\le L$. That is, $\forall\lambda'\in\Lambda, \exists l\in[1,L]$ such that $b_{jk}(\cdot)=h_{kl}(\cdot)$.

Hence, the reestimation of $b_{jk}(\cdot)$ becomes the reestimation of $h_{kl}(\cdot)$ and may be expressed verbally as follows:

reestimation of parameters of pdf h_{kl} = reestimation of parameters of pdf in convention CDHMM, *but* the statistics are gathered from all frames belonging to **all** $b_{jk}(\cdot)\equiv h_{kl}(\cdot)$ over **all** states and **all** models

In particular if the pdf's are Gaussians, that is,

$$h_{kl}(\cdot)\equiv N(\mathbf{O}_k;\mu_{kl},\Sigma_{kl})$$

then

$$\mu_{kl}' = \frac{\sum_{\lambda\in\Lambda}\sum_{j\,:\,b_{jk}\equiv h_{kl}}\sum_{t=1}^{T}\gamma_t^{\lambda}(j)\cdot\mathbf{o}_t}{\sum_{\lambda\in\Lambda}\sum_{j\,:\,b_{jk}\equiv h_{kl}}\sum_{t=1}^{T}\gamma_t^{\lambda}(j)}$$

$$\Sigma_{kl}' = \frac{\sum_{\lambda\in\Lambda}\sum_{j\,:\,b_{jk}\equiv h_{kl}}\sum_{t=1}^{T}\gamma_t^{\lambda}(j)(\mathbf{o}_t-\mu_{kl})(\mathbf{o}_t-\mu_{kl})^t}{\sum_{\lambda\in\Lambda}\sum_{j\,:\,b_{jk}\equiv h_{kl}}\sum_{t=1}^{T}\gamma_t^{\lambda}(j)}$$

The extension to state observation pdf with M-mixture density is straightforward. Note that though the reestimation formulas look like those of tied-mixture HMM, they are only equivalent for state pdf with 1-mixture density and represent different updates in the general case when the state pdf is an M-mixture density function.

3. SDCHMM TRAINING

If there is a priori knowledge of the tying structure of the subspace Gaussians in the SDCHMMs, they can be trained directly from a speech corpus without going through CDHMM training in a general scheme shown in Figure 1. The tying structure may be obtained from SDCHMMs converted from CDHMMs trained on the same task, on a speech corpus recorded under similar environment (channel, SNR, speaking style) or from a general phone recognizer.

4. EVALUATION ON ATIS

We test the hypothesis that SDCHMMs should require less data to train than CDHMMs with the ARPA-ATIS [3] recognition task. ATIS (Airline Travel Information Service) is a medium-vocabulary task containing spontaneous goal-directed speech for air travel information queries.

4.1. Signal Processing

At every 10ms, 12 MFCCs (after mean subtraction) and power, their first and second order time derivatives are extracted from a 20ms frame of speech producing a 39-dimensional feature vector.

4.2. Recognition

All CDHMMs and SDCHMMs trained in this paper are evaluated on the 1994 official test set of 981 utterances (91 mins.) using a vocabulary size of 1532 words, a word-class bigram language model with a perplexity of about 20 and a one-pass beam search with a fixed pruning threshold.

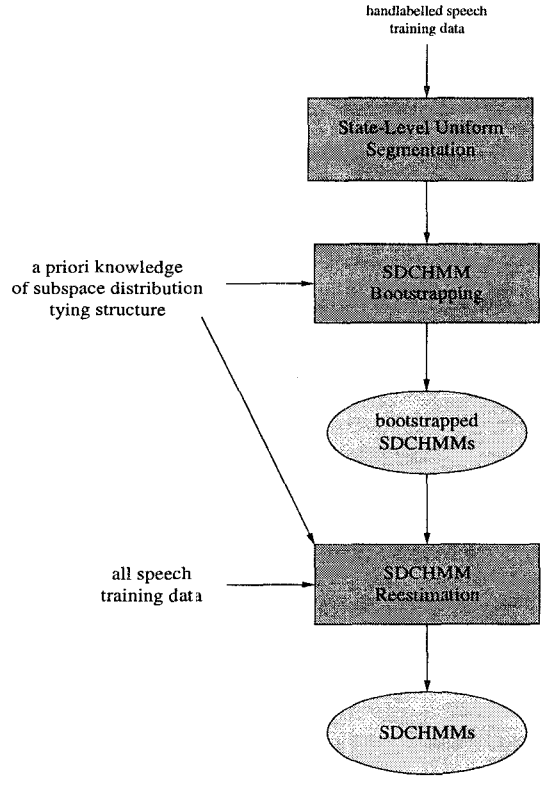

Figure 1: True SDCHMM training

Table 1: Training data subsets of ATIS

SUBSET	#FRAME	TIME(min)	DESCRIPTION
S1-16	8883240	1480	16897 files
S1-4	2140470	357	4226 files
S1-2	1080650	180	2114 files
S1	527599	88	1055 files
S0	249565	42	500 files in S1
A	101309	17	100 files in S16
B	49616	8.3	50 files in A
C	27811	4.6	25 files in B
D	12421	2.1	12 files in C

4.3. Training Data Partitioning

A collection of 16896 speech files from the ATIS-2 and ATIS-3 corpora, which were collected over five sites: BBN, CMU, MIT, NIST and SRI are employed in the experiments. They are divided into 16 subsets of roughly 1000 files, denoted as S1 to S16, so that data from the five sites are spread out into each subset as equally as possible. The 100 longest utterances from S16 are selected for bootstrapping HMMs and is denoted as subset A. Four other smaller subsets denoted as S0, B, C and D are derived from these 17 subsets as shown in Table 1. Subsets S5 to S16 are not used in this paper.

Table 2: Number of Gaussians in CDHMMs trained with various numbers of mixtures and data subsets

#MIXTURES	TRAINING DATA SUBSETS				
	A	S0	S1	S1-2	S1-4
1	142	142	142	142	142
2	257	273	280	283	283
4	452	516	535	559	563
8	735	927	1022	1077	1117
16	1019	1563	1863	2050	2143

4.4. Training Procedure

Training is done with datasets: A only, B only, C only, D only, S0 only, S1 only, S1-2, and S1-4 (meaning S1, S2, S3 and S4). Hand-labeled data are simulated by segmenting the bootstrapping data, subset A, at the phone level with our context-independent (CI) ATIS CDHMMs trained over 12,000 files. There are altogether 48 monophones. Bootstrapped CDHMMs or SDCHMMs are initialized from the segmented subset A data by assuming uniform state-level segmentation followed by 20 iterations of EM algorithm. The bootstrapped CDHMMs or SDCHMMs are used to perform state-level segmentation for the set of training data under experiment. The segmented data are then used to reestimate HMM parameters in the segmental k-means training(SKM) procedure [5]. The SKM procedure is repeated at most twice to obtain the final CDHMMs or SDCHMMs. However, for SDCHMM training with subsets A, B, C and D, only SDCHMM bootstrapping is done to obtain the final SDCHMMs as more SKM iterations do not improve results.

In all SDCHMM training, the subspace Gaussian tying structure is obtained from the 20-stream SDCHMMs (WER = 9.5%) which is converted from the best 16-mixture, 2143-Gaussians CDHMMs trained with S1-4 (WER = 9.0%), and they have 128 subspace Gaussian codewords per stream.

In all CDHMM training, the number of mixtures is varied from 1 to 16 for each training dataset.

5. RESULTS

Figure 2 compares recognition word error rates (WER) between SDCHMMs trained with SDCHMM training algorithm and CDHMMs trained with the CDHMM training algorithm when the amount of training data is progressively reduced.

As the amount of training data decreases, in general, we are limited to less complicated HMMs (with fewer mixtures and Gaussian components) and the resulting HMMs may not generalize well to test data. However, as seen in Figure 2, 20-stream 128-codeword SDCHMMs trained with the SDCHMM training algorithm allows us to reduce the amount of training data to 17 mins. of speech (subset A) before performance degrades. On the other hand, CDHMM training requires much more training data; thus we only can train CDHMMs with the subsets A, and S0 up to S1-4. The number of Gaussians in the CDHMMs in each training condition is summarized in Table 2.

From Figure 2, it appears that SDCHMMs require almost an order of magnitude less of training data to achieve comparable accuracies of the CDHMMs. In particular, SDCHMMs trained on datasets D, C, B and A give approximatively the same accuracies

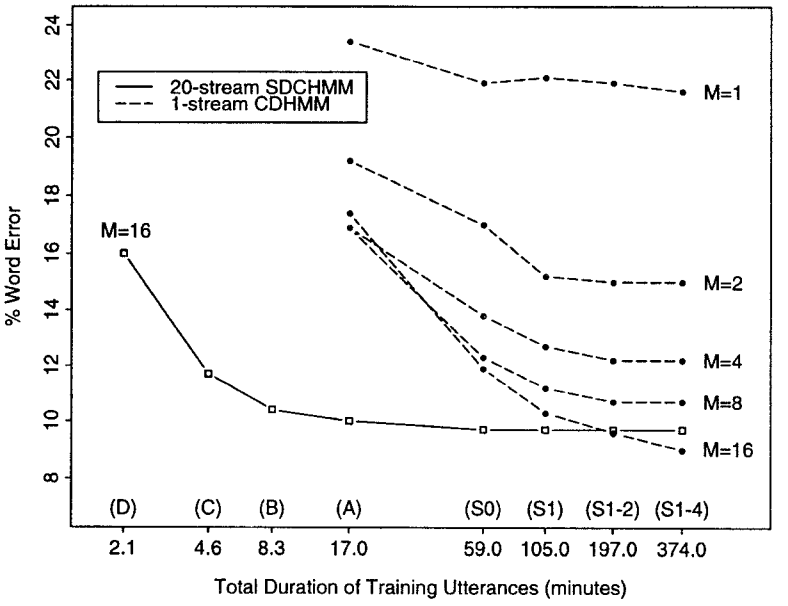

Figure 2: Comparison between the amount of training data required in CDHMM training and SDCHMM training: Its effect on ATIS recognition (where M = #mixtures, and the 20-stream SDCHMMs have 128 subspace Gaussian codewords per stream)

as the CDHMMs trained on subsets A, S0, S1, and S1-2 respectively.

One common approach to improving the CDHMMs accuracy is by training fewer Gaussian mixture components when only few training data are available. Our result shows that the use of SDCHMMs is much more effective than this strategy.

6. CONCLUSION

From the result of our experiment, SDCHMM training allows estimating SDCHMM parameters with roughly 10 times less data than CDHMM training when the amount of data is relatively small. Although SDCHMM training requires a priori knowledge of the subspace Gaussian tying structure, and in our experiment, the tying structure is derived from an existing recognizer on the same task, our results are still significant and may have the following application: instead of doing conventional speaker or environment adaptation, one may estimate speaker- or environment-specific SDCHMMs with few enrollment utterances using a tying structure derived from a speaker- or environment-independent system for the same task. Further, for our future work, we will investigate whether the subspace Gaussian tying structure is speaker or environment independent, or even task independent.

7. REFERENCES

[1] S.E. Levinson B.H. Juang 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.

[2] E. Bocchieri and B. Mak. "Subspace Distribution Clustering for Continuous Observation Density Hidden Markov Models". In *Proceedings of Eurospeech*, volume 1, pages 107–110, 1997.

[3] D. Dahl et al. "Expanding the Scope of the ATIS Task: The ATIS-3 Corpus". *Proceedings of ARPA Human Language Technology Workshop*, March 1994.

[4] X. Huang and M.A. Jack. "Semi-continuous Hidden Markov Models for Speech Signals". *Journal of Computer Speech and Language*, 3:239–251, 1989.

[5] B.H. Juang and L.R. Rabiner. "A Segmental K-means Algorithm for Estimating Parameters of Hidden Markov Models". *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(9):1639–1641, September 1990.

[6] K.F. Lee. "Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition". *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(4):599–609, April 1990.

[7] B. Mak, E. Bocchieri, and E. Barnard. "Stream Derivation and Clustering Schemes for Subspace Distribution Clustering HMM". In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 1997.

[8] L. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[9] S. Takahashi and S. Sagayama. "Four-Level Tied-Structure for Efficient Representation of Acoustic Modeling". In *Proceedings of ICASSP*, volume I, pages 520–523, 1995.

[10] S.J. Young and P.C. Woodland. "The Use of State Tying in Continuous Speech Recogniser". In *Proceedings of Eurospeech*, pages 2203–2206, 1993.