TRAINING OF CONTEXT-DEPENDENT SUBSPACE DISTRIBUTION CLUSTERING HIDDEN MARKOV MODEL

Brian Mak[1] *Enrico Bocchieri*[2]

[1]Department of Computer Science, The Hong Kong University of Science & Technology, Hong Kong
[2]AT&T Labs – Research, 180 Park Ave, Florham Park, NJ 07932, USA

ABSTRACT

Training of continuous density hidden Markov models (CDHMMs) is usually time-consuming and tedious due to the large number of model parameters involved. Recently we proposed a new derivative of CDHMM, the *subspace distribution clustering hidden Markov model* (SDCHMM) which tie CDHMMs at the finer level of subspace distributions, resulting in many fewer model parameters. An SDCHMM training algorithm is also devised to train SDCHMMs directly from speech data without intermediate CDHMMs. On the ATIS task, speaker-independent context-independent (CI) SDCHMMs can be trained with as little as 8 minutes of speech with no loss in recognition accuracy — a 25-fold reduction when compared with their CDHMM counterparts [1]. In this paper, we extend our novel SDCHMM training to context-dependent (CD) modeling with the assumption of various prior knowledge. Despite the 30-fold increase of model parameters in the CD ATIS CDHMMs, their equivalent CD SDCHMMs can still be estimated with a few minutes of ATIS data.

1. INTRODUCTION

Context-dependent (CD) continuous density hidden Markov modeling (CDHMM) is commonly employed to build the acoustic models for an automatic speech recognizer of reasonably large-vocabulary because of its high accuracy. However, training CD CDHMMs usually takes a long time and requires tedious fine-tuning efforts such as state tying. The problem is mainly attributed to the large number of model parameters. For example, it is not uncommon to find laboratory recognizers with millions of model parameters [2, 3]. In general, a large model parameter space leads to large memory consumption, slow recognition, large amount of training data, and more difficult speaker/environment adaptation.

Recently we proposed a new derivative of the continuous density hidden Markov modeling which we call *subspace distribution clustering hidden Markov modeling* (SDCHMM) [4, 5] with an aim to solve the various problems through reducing the number of model parameters. A set of K-stream SDCHMMs can be derived from CDHMMs in two steps:

1. Decompose the feature space into K orthogonal subspaces or streams.
2. Cluster the subspace Gaussians from *all* states and *all* phone models in each subspace.

We refer to the tying information among subspace Gaussians of SDCHMMs together with the mappings between them and the full-space Gaussians of CDHMMs as the *subspace Gaussian tying structure* (SGTS).

We have shown in [4] and [5] that on the ATIS (Air Travel Information System) task, the acoustic space can be efficiently encoded by a much reduced number of subspace Gaussian prototypes — about two to three orders of magnitude fewer than the original number of Gaussians in the CDHMMs. The ensuing SDCHMMs work as well as the original CDHMMs they derive from, with no loss in recognition accuracy. The results thus suggest that only a few model parameters may be required to represent the whole acoustic space.

With the much reduced number of model parameters, one should be able to train these compact SDCHMMs with many fewer training data. In [1], we devised an SDCHMM training algorithm to estimate SDCHMMs directly from speech data (without intermediate CDHMMs). A novel feature of the algorithm is its implicit use of the phonetic-acoustic relationship encapsulated in the subspace Gaussian tying structure. As a result, on the ATIS task, a speaker-independent context-independent SDCHMM system trained with only 8 minutes of speech reaches the same accuracy of a similar CDHMM system trained with 105 minutes of speech. In this paper, we investigate if the same SDCHMM training algorithm can also be effective in training context-dependent SDCHMMs, which have many more model parameters than their context-independent counterparts. Moreover, various *a priori* knowledge is progressively added to the training algorithm to investigate their application to adaptation.

2. REESTIMATION OF SDCHMM PARAMETERS

SDCHMM parameters can be estimated in much the same way as the CDHMM parameters using the EM algorithm [6]. In fact, the additional constraints imposed by the subspace Gaussian tying structure (SGTS) only alter the way in which statistics are gathered from the observations in the estimation of distribution parameters. Moreover, since the SGTS concerns *all* acoustic models, the main difference between CDHMM estimation and SDCHMM estimation is that while each CDHMM may be estimated in isolation, *all* SDCHMMs have to be estimated at the same time.

2.1. Reestimation of $\boldsymbol{\pi}$ and $\boldsymbol{a}$

The theory of SDCHMM modifies only the state observation pdf of the CDHMM, while the definitions of the initial-state probabilities $\boldsymbol{\pi}$ and state-transition probabilities $\boldsymbol{a}$ are kept intact. Hence, $\boldsymbol{\pi}$ and $\boldsymbol{a}$ can be reestimated in the same way as in the CDHMM.

2.2. Reestimation of b

Let the state observation pdf $b_i^\lambda(\cdot)$ of state i, $1 \le i \le N$, of a K-stream SDCHMM λ be a mixture density with M Gaussian components $b_{im}^\lambda(\cdot)$ and mixture weights c_{im}, $1 \le m \le M$, such that $b_{im}^\lambda(\cdot)$ is a product of K subspace Gaussians $b_{imk}^\lambda(\cdot)$, $1 \le k \le K$. That is,

$$\begin{aligned} b_i^\lambda(\boldsymbol{o}_t^\lambda) &= \sum_{m=1}^{M} c_{im} b_{im}^\lambda(\boldsymbol{o}_t^\lambda), \quad \sum_{m=1}^{M} c_{im} = 1 \\ &= \sum_{m=1}^{M} \left(c_{im} \prod_{k=1}^{K} b_{imk}^\lambda(\boldsymbol{o}_{tk}^\lambda) \right) \qquad (1) \end{aligned}$$

where $b_{imk}^\lambda(\cdot)$ and $\boldsymbol{o}_{tk}^\lambda$ are the projections of $b_{im}^\lambda(\cdot)$ and $\boldsymbol{o}_t^\lambda$ onto the k-th feature subspace respectively.

The reestimation formula for the mixture weights c_{im} is the same as in the case of CDHMM, since it does not depend on the functional form of the component distribution.

Now suppose there are L_k subspace Gaussian prototypes $h_{kl}(\cdot)$, $1 \le l \le L_k$, in the k-th stream of the set of K-stream SDCHMMs Λ, $1 \le k \le K$. Each subspace Gaussian, say, $b_{imk}^\lambda(\cdot)$ in stream k of the m-th Gaussian component of state i, is tied to one of the subspace Gaussian prototypes of the stream, say, $h_{kl}(\cdot)$. That is, $\forall \lambda \in \Lambda$, $\forall\, i \in [1, N]$, $\forall\, k \in [1, K]$, $\forall\, m \in [1, M]$, $\exists\, l \in [1, L_k]$ such that $b_{imk}^\lambda(\cdot) \equiv h_{kl}(\cdot)$. Then the reestimation of $b_{imk}^\lambda(\cdot)$ becomes the reestimation of $h_{kl}(\cdot)$ and may be expressed verbally as follows:

$$\begin{array}{lcl} \text{reestimation of} & & \text{reestimation of the pdf parameters as in conventional} \\ \text{the parameters of the} & = & \text{CDHMM, } \textit{but} \text{ the statistics are gathered from all frames belong-} \\ \text{pdf } h_{kl}(\cdot) & & \text{ing to } \textbf{all } b_{imk}^\lambda(\cdot) \equiv h_{kl}(\cdot) \text{ over } \textbf{all} \text{ states and } \textbf{all} \text{ models.} \end{array}$$

The mathematical details of the derivation of all the reestimation formulas can be found in [7]. Here we only show the reestimation formulas for the simple case of single-mixture Gaussian density; that is, $h_{kl}(\boldsymbol{o}_{tk}) = N(\boldsymbol{o}_{tk}; \boldsymbol{\mu}_{kl}, \boldsymbol{\Sigma}_{kl})$:

$$\hat{\boldsymbol{\mu}}_{kl} = \frac{\sum_{\lambda \in \Lambda} \sum_{i\ :\ b_{ik}^\lambda \equiv h_{kl}} \sum_{t=1}^{T} \gamma_t^\lambda(i) \cdot \boldsymbol{o}_{tk}^\lambda}{\sum_{\lambda \in \Lambda} \sum_{i\ :\ b_{ik}^\lambda \equiv h_{kl}} \sum_{t=1}^{T} \gamma_t^\lambda(i)}$$

$$\hat{\boldsymbol{\Sigma}}_{kl} = \frac{\sum_{\lambda \in \Lambda} \sum_{i : b_{ik}^\lambda \equiv h_{kl}} \sum_{t=1}^{T} \gamma_t^\lambda(i) (\boldsymbol{o}_{tk}^\lambda - \hat{\boldsymbol{\mu}}_{kl})(\boldsymbol{o}_{tk}^\lambda - \hat{\boldsymbol{\mu}}_{kl})'}{\sum_{\lambda \in \Lambda} \sum_{i : b_{ik}^\lambda \equiv h_{kl}} \sum_{t=1}^{T} \gamma_t^\lambda(i)}$$

where $\gamma_t^\lambda(i)$ is the probability of being in state i of model λ at time t.

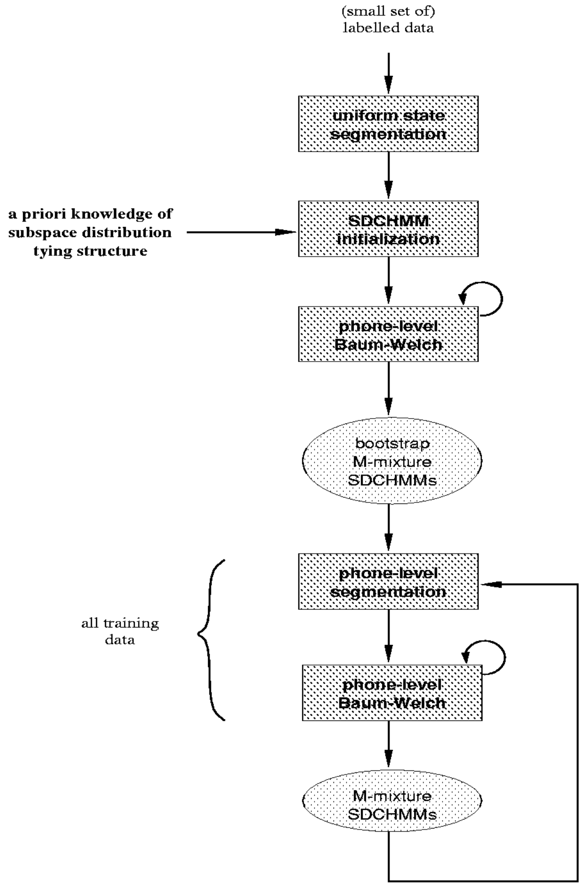

Figure 1: SDCHMM training

3. DIRECT SDCHMM TRAINING ALGORITHM

To train a set of SDCHMMs directly from a speech corpus, firstly, an SDCHMM architecture has to be defined. Part of the architecture is the subspace Gaussian tying structure which can be obtained from a set of SDCHMMs converted from CDHMMs trained on the same task. An initial set of SDCHMMs is derived from a small set of phonetically labeled data assuming uniform state segmentation.

The initial models are reestimated in an iterative procedure with all training data using a combination of Viterbi training and Baum-Welch method as shown in Figure 1. During each iteration, the old models are used in a Viterbi search to phonetically segment all training data. The phone boundaries are then fixed, and the state densities within each phone are reestimated using the Baum-Welch method. The procedure seems to be a good compromise between efficiency of Viterbi training and accuracy of Baum-Welch method.

4. EVALUATION ON ATIS

We investigate the data requirement for training context-dependent SDCHMMs with the ARPA-ATIS [8] 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 cepstral 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. Training Data Partitioning

A collection of 16,896 speech files from the ATIS-2 and ATIS-3 corpora, are divided into 16 subsets of roughly 1000 files, denoted as S1 to S16. 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 other subsets as shown in Table 1.

Table 1: ATIS: Training datasets

DATASET	#FRAMES	TIME (min.)	DESCRIPTION
Test	545,642	91	981 files
X	13,000,205	2,167	baseline
S1–16	8,883,240	1,480	16,896 files
S1–4	2,140,470	357	4,226 files
S1–2	1,080,650	180	2,114 files
S1	527,599	88	1,055 files
S0	249,565	42	500 files from S1
A	101,309	17	100 files from S16
B	49,616	8.3	50 files from A
C	27,811	4.6	25 files from B
D	12,421	2.1	12 files from C

4.3. Training Experiments

We start with a baseline context-dependent CDHMM system consisting of 3,916 tied states and 76,154 39-dimensional Gaussians. It has a word error rate (WER) of 5.2% on the official test set. It is converted to a set of 20-stream SDCHMMs with 64 subspace Gaussian prototypes per stream. The latter has a WER of 5.0%. The subspace Gaussian tying structure is extracted from the latter and is used in all SDCHMM training experiments. To save computation, all training data is phonetically labeled with the baseline CDHMM system; they will not be re-labeled during training.

All CDHMMs and SDCHMMs trained in this paper are evaluated on the 1994 official test set of 981 utterances (91 minutes) using a vocabulary size of 1536 words, a word-class bigram language model with a perplexity of about 20, and a one-pass Viterbi beam search with a fixed pruning threshold.

(I) With Prior Knowledge of SGTS

For datasets no smaller than S0, SDCHMMs are initialized with the SGTS obtained above using the phonetically labeled dataset A. Then it is found that one BW iteration is enough to get the bootstrap context-dependent SDCHMMs. The bootstrap models are reestimated by running the BW training algorithm on the training data under study. Again one BW iteration is enough for the models to converge. On the other hand, for the smaller datasets (i.e. A only, B only, C only, and D only), the bootstrapping — uniform state segmentation followed by SDCHMM initialization and one BW iteration — alone is found to be sufficient.

The first two curves from the top of Figure 2 show the number of unseen triphones in each training dataset and the recognition accuracy of the context-dependent SDCHMMs trained on the dataset. It can be seen that even when only 5–30% of the triphones are covered in subset D (2 minutes) to S0 (59 minutes), reasonable word recognition accuracies of about 7% are obtained. The low coverage seems to have caused the irregular performance of the context-dependent SDCHMMs trained in these datasets. However, the asymptotic performance does not meet the baseline performance (WERs of 5.5% vs. 5.0%), and it requires more than 735 minutes of training speech (dataset S1–8). One possible explanation is the insufficient triphone coverage: Even with all the data from S1–16, about 8% of the triphones are unrepresented. This is because the baseline system models all triphones appearing even once in all the ATIS data. To do that, it is not only trained with more ATIS data, but also with 8000 additional utterances from the Wall Street Journal corpus to increase the coverage for the rare triphones.

(II) With Prior Knowledge of SGTS and Mixture Weights

Analysis of Experiment I shows that

- even with the smallest dataset D (12,421 frames of speech), although the triphone coverage is small, the 64 subspace Gaussian prototypes of each stream are well represented.
- the main effect of insufficient training data is that the mixture weights are not learnt.

Hence, to confirm our conjecture that the poorer performance of the context-dependent SDCHMMs trained above (as compared with the baseline converted SDCHMMs) is due to the poor triphone coverage in the given training data, we repeat Experiment I by borrowing the mixture weights from the baseline context-dependent SDCHMMs, and by fixing them during SDCHMM training. For the small datasets A, B, C, and D, two to five BW iterations are now required, whereas only one BW iteration after bootstrapping is still adequate for larger datasets. The result is presented in the third curve from the top in Figure 2. By incorporating additional *a priori* knowledge of the mixture weight (in additional to the SGTS), context-dependent SDCHMMs can now be trained from as little as 8.3 minutes of speech (dataset B) with no degradation in performance when compared with the baseline context-dependent CDHMMs, even when only 14% of the triphones are observed in the training data.

(III) With Prior Knowledge of SGTS, Mixture Weights, and Gaussian Variances

To further alleviate the effect of unseen triphones, Experiment II is repeated by borrowing the Gaussian variances from the converted SDCHMMs. The result is shown in the bottom curve in Figure 2. Now even with only 2.1 minute of speech, the performance is almost the same as that of the baseline CDHMMs (WERs of 5.3% vs. 5.2%) and it reaches that of the baseline SDCHMMs with 59 minutes of training data.

5. DISCUSSION AND CONCLUSION

Context-dependent modeling usually requires substantial training data, and tedious fine-tuning efforts. Experiment I shows that the amount of training data may be dras-

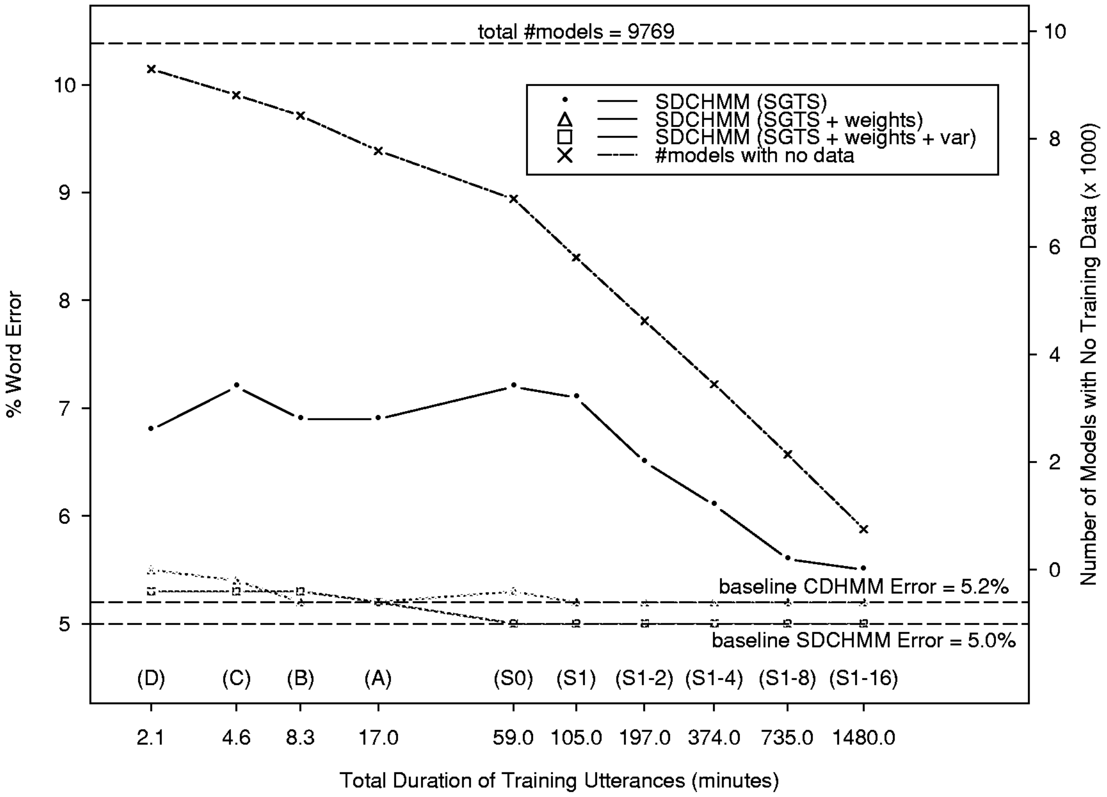

Figure 2: ATIS: Data requirement for context-dependent SDCHMM training

tically reduced with acceptable accuracy if the phonetic-acoustic relationship is made use of as in SDCHMM training using the subspace Gaussian tying structure. However, with few training data comes the triphone coverage problem. Experiment II shows that if we have *a priori* knowledge of the mixture weights, speaker-independent context-dependent SDCHMMs can be trained in a few minutes of speech. Even fewer training data is required if the Gaussian variances are also known. With the progressive addition of knowledge from the baseline system, the boundary between adaptation and training becomes blurred.

Although in our experiments, training context-dependent SDCHMM requires some prior knowledge of the original CDHMM system, 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 (and perhaps with the addition of mixture weights and/or Gaussian variances) derived from a speaker- or environment-independent system of the same task. On the other hand, it will be interesting to investigate if the subspace Gaussian tying structure is task-independent. If so, we may derive a "generic" SGTS from one task (e.g. WSJ) and use it to train the SDCHMMs of another task (e.g. ATIS).

6. REFERENCES

1. B. Mak and E. Bocchieri, "Training of Subspace Distribution Clustering Hidden Markov Model," in *Proceedings of ICASSP*, 1998, vol. 2, pp. 673–676.
2. E. Bocchieri and G. Riccardi, "State Tying of Triphone HMM's for the 1994 AT&T ARPA ATIS Recognizer," in *Proceedings of Eurospeech*, 1995, vol. 2, pp. 1499–1502.
3. X. Huang et al., "The SPHINX-II Speech Recognition System: An Overview," *Journal of Computer Speech and Language*, vol. 7, no. 2, pp. 137–148, April 1993.
4. E. Bocchieri and B. Mak, "Subspace Distribution Clustering for Continuous Observation Density Hidden Markov Models," in *Proceedings of Eurospeech*, 1997, vol. 1, pp. 107–110.
5. 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, pp. 339–346.
6. X.D. Huang, Y. Ariki, and M.A. Jack, "Fundamentals of Pattern Recognition," in *Hidden Markov Models for Speech Recognition*, chapter 2, pp. 10–51. Edinburgh University Press, 1990.
7. B. Mak, *Towards A Compact Speech Recognizer: Subspace Distribution Clustering Hidden Markov Model*, Ph.D. thesis, Department of Computer Science, Oregon Graduate Institute of Science and Technology, April 1998.
8. L. Hirschman et al., "Multi-Site Data Collection and Evaluation in Spoken Language Understanding," in *Proceedings of ARPA Human Language Technology Workshop*. 1993, Morgan Kaufmann Publishers.