Direct Training of Subspace Distribution Clustering Hidden Markov Model

Brian Mak, *Member, IEEE*, and Enrico Bocchieri, *Member, IEEE*

Abstract

It generally takes a long time and requires a large amount of speech data to train hidden Markov models for a speech recognition task of a reasonably large vocabulary. Recently we proposed a compact acoustic model called *"subspace distribution clustering hidden Markov model"* (SDCHMM) with an aim to save some of the training effort. SDCHMMs are derived from tying continuous density hidden Markov models (CDHMMs) at a finer sub-phonetic level, namely the subspace distributions. Experiments on the ATIS (Airline Travel Information System) task show that SDCHMMs with significantly fewer model parameters — by one to two orders of magnitude — can be converted from CDHMMs with no loss in word accuracy [1], [2]. With such compact acoustic models, one should be able to train SDCHMMs directly from significantly less speech data (without intermediate CDHMMs). In this paper, we devise a direct SDCHMM training algorithm, assuming an *a priori* knowledge of the subspace distribution tying structure. On the ATIS task, it is found that *both* a context-independent and a context-dependent speaker-independent 20-stream SDCHMM system trained with 8 minutes of speech perform as well as their corresponding CDHMM system trained with 105 minutes and 36 hours of speech respectively.

Keywords

Subspace distribution clustering hidden Markov modeling, direct training, subspace distribution tying structure.

I. Introduction

One of the major components of an automatic speech recognition (ASR) system is its acoustic models. Acoustic modeling for a reasonably large vocabulary is a time-consuming exercise, requiring a large amount of speech training data in order to accommodate the great variabilities in speech. With the availability of many large speech corpora nowadays, more accurate acoustic

Brian Mak is with the Department of Computer Science, the Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong. This work began when he was a PhD candidate of Oregon Graduate Institute of Science and Technology, Portland, Oregon, USA, and a Research Consultant at the AT&T Labs – Research. E-mail: mak@cs.ust.hk. Phone: +852 2358-7012. Fax: +852 2358-1477.

Enrico Bocchieri is with the AT&T Labs – Research, Florham Park, NJ 07932, USA. E-mail: enrico@research.att.com.

models can be built with more data covering different variabilities in speech. Thus, it is not uncommon that it takes days to build acoustic models for a recognition task of medium to large vocabulary. This has greatly hampered the development of ASR systems in practice.

The lengthy and data-intensive training of acoustic models can be attributed to:

(1) the large number of acoustic model parameters. For example, many state-of-the-art laboratory recognizers contain millions of model parameters [3], [4]. In general, more model parameters will require more training data to generate robust estimates of the parameters.

(2) the brute-force data-driven training scheme. Acoustic modeling is reduced to pure parameter estimation of some density functions (as in ASR based on hidden Markov models) or non-parametric classifiers (such as artificial neural networks) without utilizing other knowledge such as the acoustic-phonetic relationship.

Recently we proposed a new derivative of the continuous density hidden Markov modeling (CDHMM) methodology which we call "subspace distribution clustering hidden Markov modeling" (SDCHMM) in order to build more compact acoustic models. It has been shown that the subspace distribution clustering hidden Markov models (SDCHMMs) can capture the acoustic-phonetic information efficiently with significantly fewer parameters — by one to two orders of magnitude — than similar continuous density hidden Markov models (CDHMMs) [1], [2]. Consequently, SDCHMM systems run faster with a smaller memory footprint than similar CDHMM systems, and yet they are as accurate as the latter. In the past, K-stream SDCHMMs are derived from a set of CDHMMs with mixture Gaussian densities and diagonal covariances by a simple model conversion procedure in three steps:

Step 1. Decompose the feature space into K orthogonal (disjoint) subspaces or streams.

Step 2. Project all Gaussians of the CDHMMs onto those orthogonal subspaces.

Step 3. Tie the subspace Gaussians from *all* states and *all* phone models (CDHMMs) in each subspace. This is done by clustering the subspace Gaussians into a small number of Gaussian prototypes in each subspace (stream).

One may consider SDCHMMs as CDHMMs tied at the sub-phonetic level of subspace Gaussians. We refer to the tying information among the subspace Gaussians of SDCHMMs together with the mappings between them and the full-space Gaussians of CDHMMs as the *subspace Gaussian tying structure* (SGTS), or generally *subspace distribution tying structure* (SDTS) when the type of distribution is immaterial for the discussion. By exploiting the combinatorial effect of subspace distribution encoding, all the original full-space distributions can be closely approximated by some combinations of a small number of subspace distribution prototypes. Now with

the significantly fewer model parameters in SDCHMMs, one should be able to train SDCHMMs *directly* from significantly less speech data. Since acoustically similar subspace Gaussians are tied, the SGTS efficiently represents the acoustic inter-relationship among the phones as supported by an acoustic-phonetic analysis of the SDCHMMs in [5]. The presumption of an SGTS should therefore be considered as a utilization of acoustic-phonetic knowledge in designing our acoustic models, resulting in fewer model parameters and theoretically requiring less training data.

In this paper, we propose a novel direct SDCHMM training algorithm and demonstrate that by making use of

- the small number of parameters in the compact subspace distribution clustering hidden Markov models; and,
- the *a priori* knowledge of the acoustic-phonetic relationship encapsulated in a subspace Gaussian tying structure,

SDCHMMs can be trained directly from significantly less speech data — one to two orders of magnitude — than those required for equally accurate CDHMMs. Specifically, on the ATIS (Air Travel Information System) [6] task, by progressively reducing the amount of training data, we study the training data requirement for SDCHMMs and compare that with the data requirement for training CDHMMs of various complexities. Both context-independent and context-dependent SDCHMMs are trained.

The organization of this paper is as follows. In Section II, we first review the theory of SDCHMM and an indirect method to generate SDCHMMs from a set of CDHMMs. In Section III, the direct SDCHMM training algorithm is presented together with the reestimation formulas of the various SDCHMM parameters. Section IV describes the experimental set-up and methodology used to evaluate the direct SDCHMM training algorithm on the ATIS task. Direct training of context-independent SDCHMMs is evaluated in Section V, while that of context-dependent SDCHMMs is evaluated in Section VI with progressively more adaptation information. Finally, we draw our conclusions in Section VII and points to some future directions in this work.

II. Review of SDCHMM

In this Section, we review the theory of subspace distribution clustering hidden Markov modeling, and briefly outline a conversion procedure that derives SDCHMMs from a set of CDHMMs.

A. Theory of SDCHMM

The theory of SDCHMM is derived from that of continuous density hidden Markov model (CDHMM) in which state-observation distributions are estimated as mixture Gaussian densities with M components and diagonal covariances. Using the following notations (where bold-faced quantities represent vectors):

$\boldsymbol{O}$: an observation vector of dimension D
$P_i(\boldsymbol{O})$: output probability of state i given $\boldsymbol{O}$
c_{im} : weight of the m-th mixture component of state i
$\boldsymbol{\mu}_{im}$: mean vector of the m-th component of state i
$\boldsymbol{\sigma}^2_{im}$: variance vector of the m-th component of state i
$\mathcal{N}(\cdot)$: Gaussian pdf

the observation probability of the i-th state of a CDHMM is given by

$$P_i^{CDHMM}(\boldsymbol{O}) = \sum_{m=1}^{M} c_{im}\mathcal{N}(\boldsymbol{O};\boldsymbol{\mu}_{im},\boldsymbol{\sigma}^2_{im}),\ \sum_{m=1}^{M} c_{im} = 1. \tag{1}$$

The key observation is that a Gaussian with diagonal covariance can be expressed as a product of subspace Gaussians where the subspaces (or streams) are orthogonal and together span the original full feature vector space. To derive K-stream SDCHMMs from a set of CDHMMs, we first partition the feature set with D features into K disjoint feature subsets with d_k features, $\sum_{k=1}^{K} d_k = D$. Each of the original full-space Gaussians is projected onto each feature subspace to obtain K subspace Gaussians of dimension d_k, $1 \le k \le K$, with diagonal covariances. Thus, Equation (1) can be rewritten as

$$P_i^{CDHMM}(\boldsymbol{O}) = \sum_{m=1}^{M} c_{im}\left(\prod_{k=1}^{K}\mathcal{N}(\boldsymbol{O}_k;\boldsymbol{\mu}_{imk},\boldsymbol{\sigma}^2_{imk})\right) \tag{2}$$

where $\boldsymbol{O}_k$, $\boldsymbol{\mu}_{imk}$, and $\boldsymbol{\sigma}^2_{imk}$ are the projection of the observation $\boldsymbol{O}$, and mean and variance vectors of the m-th mixture component of the i-th state onto the k-th subspace respectively.

For each stream, we tie the subspace Gaussians across *all* states of *all* CDHMM acoustic models. Hence, the state observation probability in Equation (2) is modified as

$$P_i^{SDCHMM}(\boldsymbol{O}) = \sum_{m=1}^{M} c_{im}\left(\prod_{k=1}^{K}\mathcal{N}^{tied}(\boldsymbol{O}_k;\boldsymbol{\mu}_{imk},\boldsymbol{\sigma}^2_{imk})\right). \tag{3}$$

B. Indirect SDCHMM Training Algorithm: Model Conversion from CDHMMs

The formulation of SDCHMM as of Equation (3) suggests that SDCHMMs may be implemented in two steps as shown in Figure 1:

(1) Train CDHMMs for all the phonetic units (possibly with tied states), wherein state observation distributions are estimated as mixture Gaussian densities with diagonal covariances.
(2) Convert the CDHMMs to SDCHMMs by tying the subspace Gaussians in each stream. Details of the stream definitions and the clustering algorithm can be found in [2].

By exploiting the combinatorial effects of subspace Gaussian encoding, the original large number of full-space Gaussians in the CDHMMs can be represented by a few subspace Gaussians in each stream of the SDCHMMs. For instance, on the ATIS task, 32 to 128 subspace Gaussians per stream are found adequate. Subsequent ATIS recognition with a set of 20-stream context-dependent SDCHMMs runs twice as fast as that with CDHMMs, and consumes 13 times less memory and 80 times fewer model parameters [2].

III. Direct SDCHMM Training Algorithm

Although the indirect training scheme of SDCHMMs through model conversion of CDHMMs is simple and runs fast, it requires an amount of training data as large as CDHMM training since the scheme requires intermediate CDHMMs. Evaluation of the indirect SDCHMM training scheme on the ATIS task shows that if the subspace Gaussian tying structure (SGTS)[1] is ignored, SDCHMMs have significantly fewer model parameters (mixture weights, Gaussian means, and variances) — by one to two orders of magnitude — than their parent CDHMMs [1], [2]. Thus, if we have *a priori* knowledge of the SGTS, one should be able to train SDCHMMs directly from significantly less speech data as shown in Figure 2.

One should notice that a subspace Gaussian tying structure encapsulates a lot of information about the acoustic-phonetic relationship, and if such information is applicable to the task on hand, it will take fewer data to train the new SDCHMMs. The *a priori* SGTS may be obtained from the conversion of a generic set of CDHMMs (trained with a large amount of speech data) to SDCHMMs, or from speaker-independent SDCHMMs when speaker-specific SDCHMMs are to be trained, and so forth, depending on the application.

In the following, we will present the reestimation formulas of SDCHMM parameters.

[1]Subspace Gaussian tying structure is defined in Section I

A. Maximum Likelihood Estimation of SDCHMM Parameters

SDCHMM parameters may be estimated in much the same way as CDHMM parameters are estimated using the *Baum-Welch* (BW) algorithm [7]. In fact, the additional constraints imposed by the subspace distribution tying structure (SDTS) only alter the way in which statistics are gathered from the observations in the estimation of the distribution parameters.

Let us denote the whole set of SDCHMMs of all speech units by Λ. Each N-state SDCHMM $\lambda \in \Lambda$ is defined by three sets of parameters:

- initial-state probabilities $\boldsymbol{\pi}^\lambda = [\pi_1^\lambda, \pi_2^\lambda, \ldots, \pi_N^\lambda]$
- state-transition probability matrix $\boldsymbol{a}^\lambda = \{a_{ij}^\lambda\}, 1 \le i, j \le N$
- state observation pdf's $\boldsymbol{b}^\lambda = [b_1^\lambda, b_2^\lambda, \ldots, b_N^\lambda]$.

Also assume that for each SDCHMM, there is a sequence of training observation $\boldsymbol{O}^\lambda = \boldsymbol{o}_1^\lambda \boldsymbol{o}_2^\lambda \cdots \boldsymbol{o}_T^\lambda$ (where $\boldsymbol{o}_t^\lambda$ is the observation vector at time t) of T frames.

A.1 Reestimation of $\boldsymbol{\pi}$ and $\boldsymbol{a}$ in SDCHMM

It is clear that from the theory of SDCHMM (Equation (3)) that only the state observation pdf $b_i^\lambda(\cdot)$ of the CDHMM is modified, 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 still be estimated for each SDCHMM in the same way as those of conventional CDHMM.

A.2 Reestimation of $\boldsymbol{b}$ in SDCHMM

According to the theory of SDCHMM, the state observation pdf $b_i^\lambda(\cdot)$ of state i of a K-stream SDCHMM λ is assumed to be a mixture density with M 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 pdf's $b_{imk}^\lambda(\cdot)$, $1 \le k \le K$, of the same functional form. That is,

$$b_i^\lambda(\boldsymbol{o}_t^\lambda) = \sum_{m=1}^{M} c_{im} b_{im}^\lambda(\boldsymbol{o}_t^\lambda), \quad \sum_{m=1}^{M} c_{im} = 1 \tag{4}$$

$$= \sum_{m=1}^{M} \left(c_{im} \prod_{k=1}^{K} b_{imk}^\lambda(\boldsymbol{o}_{tk}^\lambda) \right) \tag{5}$$

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 of $\boldsymbol{b}$ depends on the functional form of the state observation pdf. Here, we will consider only the two cases when the state output distribution is either a single

Gaussian distribution or a mixture Gaussian density.

Case I: Single Gaussian Output Distribution

Let us first look at the special case when there is only one Gaussian in the mixture density. Equation (5) may then be simplified to

$$b_i^\lambda(\boldsymbol{o}_t^\lambda) = \prod_{k=1}^{K} b_{ik}^\lambda(\boldsymbol{o}_{tk}^\lambda) \tag{6}$$

by dropping the mixture weight of unity and the mixture component subscript m.

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

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

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

$$h_{kl}(\boldsymbol{o}_{tk}) = N(\boldsymbol{o}_{tk}; \boldsymbol{\mu}_{kl}, \boldsymbol{\Sigma}_{kl})$$

then the new model is

$$\hat{\boldsymbol{\mu}}_{kl} = \frac{\sum_{\lambda\in\Lambda} \sum_{i\ :\ b_{ik}^\lambda \equiv h_{kl}} \sum_{t=1}^{T} \gamma_t^\lambda(i) \cdot \boldsymbol{o}_{tk}^\lambda}{\sum_{\lambda\in\Lambda} \sum_{i\ :\ b_{ik}^\lambda \equiv h_{kl}} \sum_{t=1}^{T} \gamma_t^\lambda(i)} \tag{7}$$

$$\hat{\boldsymbol{\Sigma}}_{kl} = \frac{\sum_{\lambda\in\Lambda} \sum_{i: b_{ik}^\lambda \equiv h_{kl}} \sum_{t=1}^{T} \gamma_t^\lambda(i)(\boldsymbol{o}_{tk}^\lambda - \hat{\boldsymbol{\mu}}_{kl})(\boldsymbol{o}_{tk}^\lambda - \hat{\boldsymbol{\mu}}_{kl})'}{\sum_{\lambda\in\Lambda} \sum_{i: b_{ik}^\lambda \equiv h_{kl}} \sum_{t=1}^{T} \gamma_t^\lambda(i)}. \tag{8}$$

where

$$\gamma_t^\lambda(i) \stackrel{\text{def}}{=} P(q_t = i \mid \boldsymbol{O}^\lambda, \lambda) \tag{9}$$

is the probability of being in state i at time t, which can be efficiently computed by the *forward-backward* algorithm [8].

Case II: Mixture Gaussian Output Distribution

Since an HMM state with a mixture density is equivalent to a multi-state HMM with single-mixture densities [9], the reestimates of $\boldsymbol{b}$ are similar to those of Case I except that the quantity

$\gamma_t^\lambda(i)$ is modified as $\gamma_t^\lambda(i,m)$ which is the probability of being in state i and the m-th mixture component at time t, given the model λ and the observation sequence $\boldsymbol{O}^\lambda$. Hence,

$$\hat{c}_{im}=\frac{\sum_{t=1}^{T}\gamma_t^\lambda(i,m)}{\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_t^\lambda(i,m)} \tag{10}$$

$$\hat{\boldsymbol{\mu}}_{kl}=\frac{\sum_{\lambda\in\Lambda}\sum_{i,m\,:\,b_{imk}^\lambda\equiv h_{kl}}\sum_{t=1}^{T}\gamma_t^\lambda(i,m)\cdot\boldsymbol{o}_{tk}^\lambda}{\sum_{\lambda\in\Lambda}\sum_{i,m\,:\,b_{imk}^\lambda\equiv h_{kl}}\sum_{t=1}^{T}\gamma_t^\lambda(i,m)} \tag{11}$$

$$\hat{\boldsymbol{\Sigma}}_{kl}=\frac{\sum_{\lambda\in\Lambda}\sum_{i,m:b_{imk}^\lambda\equiv h_{kl}}\sum_{t=1}^{T}\gamma_t^\lambda(i,m)(\boldsymbol{o}_{tk}^\lambda-\hat{\boldsymbol{\mu}}_{kl})(\boldsymbol{o}_{tk}^\lambda-\hat{\boldsymbol{\mu}}_{kl})'}{\sum_{\lambda\in\Lambda}\sum_{i,m:b_{imk}^\lambda\equiv h_{kl}}\sum_{t=1}^{T}\gamma_t^\lambda(i,m)}. \tag{12}$$

IV. Evaluation: Set-up

The Air Travel Information System (ATIS) task [6] is chosen for the evaluation of the direct SDCHMM training algorithm. The evaluation may be rephrased as follows:

"If the subspace Gaussian tying structure for the acoustic models of the ATIS task is known, how much training data is required to directly train SDCHMMs for the task?"

Both context-independent (CI) and context-dependent (CD) SDCHMMs will be trained and evaluated. Nonetheless, more emphasis is put on the CI models simply because the simpler and fewer CI models allow us to train and test many CDHMMs and SDCHMMs of various complexities in a manageable amount of time. Moreover, CI modeling tends to be more stable as there is usually ample coverage of training data for the CI phones. In contrast, CD modeling requires delicate fine-tuning effort to obtain a good balance between training data and model accuracy, which may complicate our main research goal here.

A. Experimental Set-up

Speech features are extracted at a frame rate of 10ms. Twelve MFCCs (after mean subtraction) and power, together with their first and second order time derivatives are computed from a frame of 20ms speech producing a 39-dimensional feature vector. Each phone model is a 3-state left-to-right HMM with the exception of one noise model which has only one state. The testing conditions (test dataset, vocabulary, pronunciation models, language models, decoding algorithm, and beam-width) are shown in Table I.

Lastly, the number of streams is fixed to 20 for all SDCHMMs trained below. This follows from the conclusion in [1], [2] which suggests that 20 streams give a good balance between accuracy, computation time, and model memory on the ATIS task.

B. Methodology

To evaluate the effectiveness of direct SDCHMM training, its training data requirement is compared with that for CDHMM training. The evaluation procedure consists of the following basic steps:

Step 1. Generate N data subsets $\mathcal{D}_i$, $1 \leq i \leq N$, from all the given training data by progressively cutting the data in half. That is, the amount of data in $\mathcal{D}_{i+1}$ is half of that in $\mathcal{D}_i$.

Step 2. Train CDHMM acoustic models with *all* available training data in $\mathcal{D}_1$.

Step 3. Convert the CDHMMs to SDCHMMs as described in Section II-B.

Step 4. Deduce the subspace Gaussian tying structure (SGTS) from the converted SDCHMMs.

Step 5. For each data subset ($\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$, ..., $\mathcal{D}_N$), repeat Steps 6 and 7.

Step 6. Train CDHMMs of different model complexities by varying the number of components in each state mixture density.

Step 7. Train SDCHMM acoustic models using the direct SDCHMM training algorithm as shown in Figure 2 with the SGTS obtained in Step 4.

Step 8. Compare the recognition performance of all CDHMMs and SDCHMMs obtained in the above steps.

C. Preparation of Training Datasets

A collection of 16,896 utterances from the ATIS-2 [6] and ATIS-3 [10] corpora are employed in this study. They are divided into 16 datasets of roughly 1,000 utterances each, denoted as S1, S2, S3, ..., to S16, so that data from the five sites are spread out into each dataset as evenly as possible. The 100 longest utterances from S16 are selected for bootstrapping HMMs and this set is denoted as dataset A. Other smaller datasets are derived as follows:

- dataset S0 contains 500 utterances from dataset S1
- dataset B contains 50 utterances from dataset A
- dataset C contains 25 utterances from dataset B
- dataset D contains 12 utterances from dataset C
- E-sets comprise 10 datasets denoted as E1, E2, ..., E10, and each contains 15 different utterances from dataset S15, three from each of the five collecting sites.
- F-sets comprise 10 datasets denoted as F1, F2, ..., F10, which are sub-sampled from the corresponding E-sets such that each contains five utterances, one from each of the five collecting sites.

All the various datasets are summarized in Table II. Datasets A, B, C, D, the E-sets, and the F-sets are all phonetically labeled using AT&T's 1994 ATIS recognizers [11]. Both context-independent and context-dependent phone labeling are performed.

D. Hybrid Viterbi/Baum-Welch Training Procedure

We adopt a combination of Viterbi training (VT) and Baum-Welch (BW) reestimation to train all acoustic models, with an additional step of segmental k-means (SKM) training for CDHMM training. The hybrid VT/BW training procedure takes advantage of the simplicity of Viterbi training and the accuracy of Baum-Welch reestimation. The procedures for training CDHMMs and SDCHMMs are schematically depicted in Figure 3 and Figure 4 respectively.

V. Evaluation: Direct Training of Context-Independent SDCHMM

Following the methodology described in Section IV-B, the effectiveness, the data requirement, and the variability of the direct SDCHMM training algorithm are evaluated on training context-independent SDCHMMs.

A. Experiment I: Effectiveness of Direct SDCHMM Training

We first check, for the same amount of training data, whether SDCHMMs trained by the direct SDCHMM training algorithm achieve the same recognition performance as that of the SDCHMMs converted from CDHMMs. Only CI models are trained in this experiment, and the SGTS from the converted SDCHMMs is used for direct SDCHMM training.

A.1 Procedure

Following the CDHMM training procedure of Figure 3, 16- and 32-mixture CDHMMs are trained with dataset S1–4 (meaning a combination of S1, S2, S3, and S4). The CDHMMs are then converted to 20-stream SDCHMMs with 16, 32, 64, and 128 subspace Gaussian prototypes per stream. Recognition on the ATIS test data determines the best SDCHMMs in each case of model complexity: 128 prototypes for the 16-mixture SDCHMMs and 64 prototypes for the 32-mixture SDCHMMs. SGTS's are derived from the best 16-mixture and 32-mixture SDCHMMs and are denoted as CI-SGTS-M16-n128 and CI-SGTS-M32-n64 respectively. Finally, a new set of SDCHMMs are trained directly from the two SGTS's with dataset S1–4 according to the SDCHMM training procedure of Figure 4.

A.2 Result and Discussion

The recognition results of the following three sets of acoustic models on the ATIS test data are shown in Table III:

- CI CDHMMs trained from the dataset S1–4
- CI SDCHMMs converted from the CDHMMs (converted SDCHMMs)
- CI SDCHMMs directly trained from the dataset S1–4 using the SGTS of the converted SDCHMMs (trained SDCHMMs)

Although the recognition accuracies of the converted SDCHMMs and the directly-trained SDCHMMs are slightly lower than that of their parent CDHMMs, they have very similar performance. The result demonstrates the effectiveness of our novel direct SDCHMM training algorithm: if one is only given the SGTS and the training data of a set of converted SDCHMMs, the SDCHMMs can be "recovered" by our direct SDCHMM training algorithm.

B. Experiment II: Data Requirement for Training Context-Independent SDCHMM

Once the effectiveness of direct SDCHMM training is established, we go a step further to investigate the data requirement for training CI SDCHMMs as compared to that for training CI CDHMMs using the methodology described in Section IV-B.

B.1 Procedure

CDHMMs of various model complexities are trained using five different datasets: A only, S0 only, S1 only, S1–2, and S1–4. Dataset A is used to bootstrap all models. The maximum number of mixtures[2] in each state density varies from one to 32 in powers of two.

Similarly, SDCHMMs with the two SGTS's, CI-SGTS-M16-n128 and CI-SGTS-M32-n64, are trained directly from the five datasets. In addition, we also train SDCHMMs with the smaller datasets: B only, C only, and D only. These latter SDCHMMs are bootstrapped with the training data under study in each case (and *not* with dataset A).

B.2 Result and Discussion

The recognition accuracies of all CDHMMs and SDCHMMs trained above are shown in Figure 5.

As the model complexity (measured in terms of the number of Gaussians) decreases, the accuracy or resolution power of HMMs is compromised. The recognition performance of all

[2]Note that the final number of mixtures in a density produced by the segmental k-means algorithm [12] can be fewer than what the user specifies, when there are too few training data in the state.

CDHMMs with different number of mixtures falls off when they are presented with fewer than 197 minutes of training speech (dataset S1–2). In contrast, the recognition performance of the 20-stream SDCHMMs trained with CI-SGTS-M16-n128 or CI-SGTS-M32-n64 using the direct SDCHMM training algorithm does not start to fall significantly until there are less than 8.3 minutes of training speech (dataset B). Moreover, the performance of these two sets of SDCHMMs, trained with only 8.3 minutes of speech, is unmatched by any CDHMMs (with the same or simpler model complexity) trained with less than 197 minutes of speech in this study. This is a roughly 20-fold reduction in the amount of training data for SDCHMMs. The result should be attributed to the fewer model parameters (mixture weights, Gaussian means, and variances) of SDCHMMs — the ratios of the number of model parameters in the two SDCHMMs to that in their parent CDHMMs are 1:14 (for CI-SGTS-M16-n128) and 1:36 (for CI-SGTS-M32-n64).

Furthermore, as the amount of training data is reduced, the performance of SDCHMMs degrades gracefully whereas the performance of CDHMMs drops sharply. For example, when the amount of training data is pared down from 374 minutes (dataset S1–4) to 17 minutes (dataset A) the word error rates (WERs) of the 16-mixture and 32-mixture CDHMMs increases by almost 100%. On the other hand, the WERs of the corresponding SDCHMMs trained using CI-SGTS-M16-n128 and CI-SGTS-M32-n64 drop by only ~20% when the amount of training data is slashed from 374 minutes (dataset S1–4) to 2.1 minutes (dataset D). At first sight, this does not seem to be possible: For instance, when the 32-mixture SDCHMMs are trained with CI-SGTS-M32-n64 and the dataset D, there are only 12421 frames of speech to train the 4,086 Gaussians of the 48 monophones. That is, on average, there are about only 259 training frames per phone or three training frames per Gaussian! Even worse is the fact that some phones rarely occur, or do not even appear in the small training dataset D as shown in the frame distribution over the phones in Figure 6. For example, phones "hh" and "oy" do not occur in dataset D, and consonants like "el", "g", "jh", "nx", "th", and "uh" are rare. However, if one looks at the frame distribution over the 64 subspace Gaussians of each stream of the SDCHMMs in Figure 7, one should be convinced that there are ample estimation data for most of the subspace Gaussians (194 frames on average), and there is full coverage for all of them. Thus, the efficient sharing of Gaussian parameters in the SDCHMMs plays an equally important role in reducing the training data requirements.

C. Experiment III: Performance Variability with Little Training Data

C.1 Procedure

When the amount of training data is small, the effect of random sampling of training data may become important. To check the performance variability of SDCHMM training with little training data, we repeat the SDCHMM training procedure of Experiment II with 20 even smaller datasets: E1 only, E2 only, ..., E10 only, F1 only, F2 only, ..., and F10 only. Each of the E-sets contains 15 utterances, and each of the F-sets contains five utterances, with durations ranging from 13.35 seconds to 97.82 seconds of speech. Both CI-SGTS-M16-n128 and CI-SGTS-M32-n64 are tried.

C.2 Result and Discussion

Figure 8 shows the scatter plots of the recognition accuracies of SDCHMMs trained with each of the two SGTS's over each of the 20 datasets. Superimposed on each scatter plot is a cubic B-spline fit generated by the statistical software S-PLUS [13]. The performance of the CI-SGTS-M32-n64 SDCHMMs degrades more slowly than that of the CI-SGTS-M16-n128 SDCHMMs when the amount of training data decreases. This is clearly due to the fact that there are even fewer model parameters and more sharing among the subspace Gaussians of the CI-SGTS-M32-n64 SDCHMMs. Nonetheless, it is observed that the 20 individual recognition results for each set of SDCHMMs fit well into the curve-fitting spline with only small fluctuations. Combining these results with those of Experiment II, we see a consistent trend that SDCHMMs can be trained with significantly fewer data over different samples of training sets.

VI. Evaluation: Direct Training of Context-Dependent SDCHMM

Since Experiment II already shows that context-independent SDCHMMs require much less training data than CDHMMs, we next only investigate if context-dependent (CD) SDCHMMs also require little training data. Thus, only CD SDCHMMs are trained.

A. Experiment IV: Data Requirement for Training Context-Dependent SDCHMM

As mentioned before, CD modeling requires more fine tuning to control the phonetic coverage (e.g. through using other parameter tying techniques such as state tying). In order not to let other factors possibly complicate our main research goal here, we start from the CDHMMs of AT&T's 1994 context-dependent ATIS recognizer [11] (hereafter referred to as the baseline CD recognizer/system). There are 9,769 triphone CDHMMs with a total of 76,154 full-space Gaussians, each having a maximum of 20 mixtures.

A.1 Procedure

The baseline CD CDHMMs is converted to a set of 20-stream SDCHMMs with 64 subspace Gaussian prototypes per stream, from which the subspace Gaussian tying structure, denoted as CD-SGTS-M20-n64, is extracted. The baseline CD CDHMM system has a WER of 5.2% on the official test set, while the converted CD SDCHMM system has a WER of 5.0%.

The subspace Gaussian tying structure, CD-SGTS-M20-n64, is used for all subsequent CD SDCHMM training experiments. We also have all training data phonetically labeled by the baseline CD CDHMM recognizer. To save training computation, subsequent SDCHMM training will *not* re-segment any training data.

Following the direct SDCHMM training procedure of Figure 4, SDCHMMs are trained from datasets: D, C, B, A, S0, S1, S1–2, S1–4, S1–8, and S1–16 respectively. For training with datasets larger than A, CD SDCHMMs are initialized with CD-SGTS-M20-n64 using the phonetically transcribed dataset A. Whereas for other smaller datasets, bootstrapping is done with the dataset under study.

A.2 Result and Discussion

The first two curves from the top of Figure 9 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 (2.1 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, obtained with more than 735 minutes of training speech (dataset S1–8) does not meet the performance of neither the baseline CDHMMs nor the converted SDCHMMs (WERs of 5.5% vs. 5.2% or 5.0%). One possible explanation is the insufficient triphone coverage: Even with all the data from S1–16, about 8% of the triphones are unrepresented. In the baseline system, all triphones appearing even once in all of the ATIS corpora are modeled. To do that, it is not only trained with more ATIS data, but also with 8000 additional utterances from the Wall Street Journal Corpus [14] to increase the coverage for the rare triphones.

B. Experiment V: With Prior Knowledge of SGTS and Mixture Weights

Analysis of Experiment IV shows that:

- Although there is inadequate triphone coverage when the amount of training data is limited, there is still high coverage of the subspace Gaussians of the CD SDCHMMs (full coverage in

all our experiments).

• When a speech unit is not observed in the training data, the main effect on SDCHMM training is that the mixture weights of its SDCHMM are not learned — they stay at their initial values of $1/M$ (where M is the number of Gaussian mixtures in the state density) and are not reestimated in subsequent VT/BW training cycles.

Hence, to confirm our conjecture that the poor performance of CD SDCHMM training is due to poor triphone coverage in the given training data, we repeat Experiment IV by borrowing the mixture weights from the baseline CD CDHMMs, and by fixing them during direct SDCHMM training. The result is presented in the third curve from the top in Figure 9. By incorporating additional *a priori* knowledge of the mixture weight (on top of the SGTS, CD-SGTS-M20-n64), the CD SDCHMMs (which have a model complexity of 76,154 Gaussians), 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 CD CDHMMs, even when only 14% of the triphones are observed in the training data.

C. Experiment VI: With Prior Knowledge of SGTS, Mixture Weights, and Gaussian Variances

The incorporation of known mixture weights does not totally eliminate the gap between the asymptotic performance of the directly trained SDCHMMs and that of the converted CD SDCHMMs, but only greatly reduces it. Since 8% of the triphones are unrepresented in the training data S1–16, some acoustics of the unseen triphones are probably still missing. To further account for the missing acoustics of the unseen triphones, Experiment V is repeated by borrowing the Gaussian variances from the converted CD SDCHMMs as well. The result is shown in the bottom curve in Figure 9. Now even with only 2.1 minute of speech, the performance is almost the same as that of the baseline CD CDHMMs (WERs of 5.3% vs. 5.2%) and it reaches that of the converted CD SDCHMMs with 59 minutes of training data.

VII. Conclusion

In this paper, we successfully train SDCHMMs directly from much less data without training intermediate CDHMMs. Such great reduction in the amount of training data is attributed to the significantly fewer model parameters in SDCHMMs as well as to the effective tying of subspace Gaussians among the models. While the fewer model parameters, in theory, require less estimation data, should the tying of subspace Gaussians not be effective, SDCHMM training would have required even sampling of the phones in the training data. However our experiments show that even when many phones are under-represented in the training data (Figure 6 or

Figure 9), there is still a good coverage of the subspace Gaussians (Figure 7); hence, good estimation of SDCHMMs is still possible.

When the amount of training data is small (say, less than 8 minutes of speech on the ATIS task), the performance of the ensuing SDCHMMs degrades gracefully. However, over-training readily occurs in this case. In this study, we exhaustively search for the best BW iteration to stop using the test data. In practice, cross-validation using unseen data may be employed to determine the best Baum-Welch iteration. Additionally, one could also investigate the use of model selection criteria [15], [16], [17] to address the problem of model complexity.

Direct SDCHMM training requires *a priori* knowledge of, at least, the subspace Gaussian tying structure. Although in our experiments, the tying structure is derived from an existing recognizer on the same task, our results are still significant. One possible application is speaker enrollment — using a speaker-independent SGTS to train speaker-specific SDCHMMs with little enrollment data.

Results of Experiment IV, V, and VI also suggest that if more *a priori* information is available, even less training data may be sufficient. For instance, we may also incorporate the mixture weights and/or Gaussian variances in addition to the SGTS from the converted SDCHMMs (from which the SGTS is derived), and fix them during SDCHMM training. This may be found useful in speaker (environment) adaptation.

Of course, we still need one set of CDHMMs from which to derive the SGTS for SDCHMM training. It will be interesting to investigate if the SGTS is task independent so that one may deduce a "generic" SGTS from a set of "generic" CDHMMs and apply it to SDCHMM training in other tasks.

References

[1] E. Bocchieri and B. Mak, "Subspace Distribution Clustering for Continuous Observation Density Hidden Markov Models," in *Proceedings of the European Conference on Speech Communication and Technology*, 1997, vol. 1, pp. 107–110.

[2] 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.

[3] E. Bocchieri and G. Riccardi, "State Tying of Triphone HMM's for the 1994 AT&T ARPA ATIS Recognizer," in *Proceedings of the European Conference on Speech Communication and Technology*, 1995, vol. 2, pp. 1499–1502.

[4] 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.

[5] 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.

[6] 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.

[7] L.E. Baum, T. Petrie, G. Soules, and N. Weiss, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *Annals of Mathematical Statistics*, vol. 41, pp. 164–171, 1970.

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

[9] B.H. Juang, S.E. Levinson, and M.M. Sondhi, "Maximum Likelihood Estimation for Multivariate Mixture Observations of Markov Chains," *IEEE Transactions on Information Theory*, vol. 32, no. 2, pp. 307–309, March 1986.

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

[11] E. Bocchieri, G. Riccardi, and J. Anantharaman, "The 1994 AT&T ATIS CHRONUS Recognizer," in *Proceedings of ARPA Spoken Language Systems Technology Workshop*. 1995, pp. 265–268, Morgan Kaufmann Publishers.

[12] 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*, vol. 38, no. 9, pp. 1639–1641, September 1990.

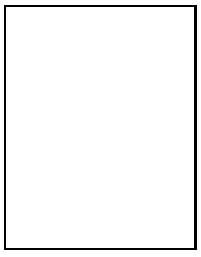
[13] StatSci, a Division of MathSoft, Inc., *S-PLUS Guide to Statistics and Mathematical Analysis*, pp. 6–52, StatSci, Seattle, Washington, 3.2 edition, 1993.

[14] D.B. Paul and J.M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," in *Proceedings of the International Conference on Spoken Language Processing*, 1992, vol. 2, pp. 899–902.

[15] L.R. Bahl and M. Padmanabhan, "A Discriminant Measure for Model Complexity Adaptation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1998, pp. 453–456.

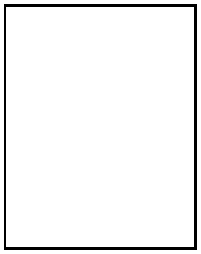
[16] S. S. Chen and P. S. Gopalakrishnan, "Clustering via the Bayesian Information Criterion with Applications in Speech Recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1998, pp. 645–648.

[17] S. S. Chen et al., "Recent Improvements to IBM's Speech Recognition System for Automatic Speech Recognition," in *Proceedings of the DARPA Speech Recognition Workshop*, 1999.



Brian Kan-Wing Mak received the B.Sc. degree in Electrical Engineering from the University of Hong Kong in 1983, the M.S. degree in Computer Science from the University of California, Santa Barbara, USA in 1989, and the Ph.D. degree in Computer Science from Oregon Graduate Institute of Science and Technology, Portland, Oregon, USA in 1998. From 1990 until 1992, he was a research programmer at the Speech Technology Laboratory of Panasonic Technologies Inc. in Santa Barbara and worked on endpoint detection research in noisy environment. From 1997 until his Ph.D. graduation in 1998, he was also a research consultant at the AT&T Labs – Research, Florham Park, New Jersey, USA. Since April 1998, he has been an assistant professor

at the Department of Computer Science, the Hong Kong University of Science and Technology. His interests include speech recognition, spoken language understanding, dialog modeling, and machine learning.



Enrico Bocchieri received the Laurea with Honors in Electrical Engineering from the University of Pavia, Italy in 1979, and the M.S. and Ph.D. degrees in Electrical Engineering from the University of Florida in 1981 and 1983 respectively. He was a Member of the Technical Staff with the Central Research Laboratories of Texas Instruments from 1984 to 1987, and with Bell Communication Research from 1987 to 1990. He then joined the Information Principle Research Laboratories of AT&T Bell Laboratories, and since the AT&T-Lucent spinoff he is with AT&T Research in the Speech and Image Processing Services Research Lab, as a Principal Technical Staff Member. His research interests include speech recognition and understanding, signal processing and software engineering.

TABLE I

ATIS: Testing conditions

TESTING CONDITION	CI SYSTEM	CD SYSTEM
#Test Sentences	981 (1994 ARPA-ATIS evaluation set)	
Vocabulary	1,532 words	
Language Model	word-sequence bigram (perplexity ≈ 20)	
Search	one-pass Viterbi beam search	
Lexical Structure	lexical tree	linear lexicon
Beam-Width	100	170
#HMMs	48	9,769
#States	142	3,916 (tied)
Max. #Mixtures per State	32	20

TABLE II

ATIS: TRAINING DATASETS (∗ DATASETS ARE PHONETICALLY LABELED BY AT&T'S ATIS RECOGNIZERS. † FIGURES ARE AVERAGES.)

DATASET	#FRAMES	DURATION (min.)	DESCRIPTION
X	13,000,205	2,167	training data for AT&T's CD ATIS recognizer
Y	6,444,959	1,074	training data for AT&T's CI ATIS recognizer
Test	545,642	91	981 (1994 ARPA's official) test utterances
S1–16	8,883,240	1,480	16,896 utterances
S1–4	2,140,470	357	4,226 utterances
S1–2	1,080,650	180	2,114 utterances
S1	527,599	88	1,055 utterances
S0	249,565	42	500 utterances from subset S1
A∗	101,309	17	100 utterances from subset S16
B∗	49,616	8.3	50 utterances from subset A
C∗	27,811	4.6	25 utterances from subset B
D∗	12,421	2.1	12 utterances from subset C
E1–E10∗	7,758†	1.29†	15 utterances from subset S15
F1–F10∗	2,702†	0.45†	5 utterances from subset S15

TABLE III

ATIS: COMPARISON OF RECOGNITION ACCURACIES AMONG CI CDHMMS, CI SDCHMMS CONVERTED FROM THE CDHMMS, AND CI SDCHMMS ESTIMATED BY DIRECT SDCHMM TRAINING USING THE SGTS OF THE CONVERTED SDCHMMS

#MIXTURES PER STATE	TOTAL #GAUSSIAN COMPONENTS	#PROTOTYPES PER STREAM	WORD ERROR RATE (%)		
			CDHMM	CONVERTED SDCHMM	TRAINED SDCHMM
16	2143	128	9.0	9.5	9.3
32	4086	64	8.5	8.7	8.7

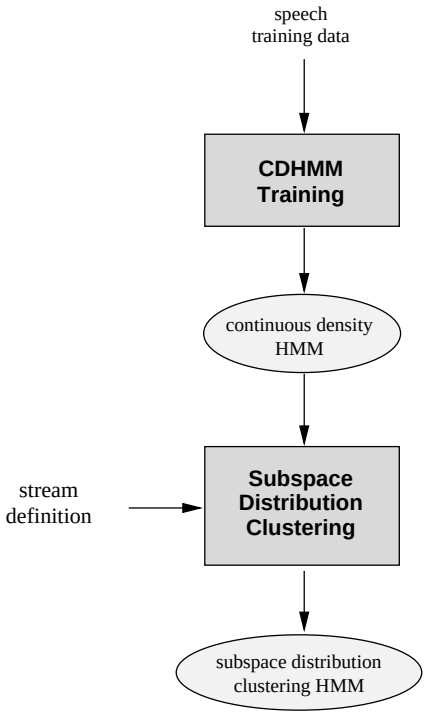


Fig. 1. Indirect SDCHMM training scheme

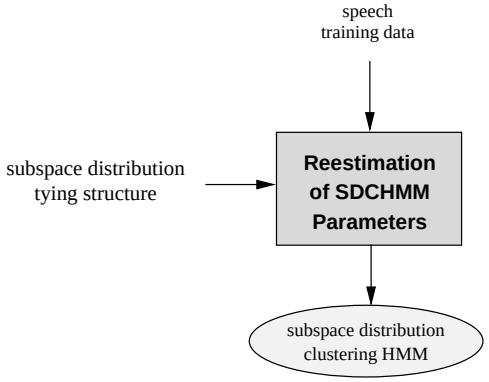


Fig. 2. Direct SDCHMM training scheme

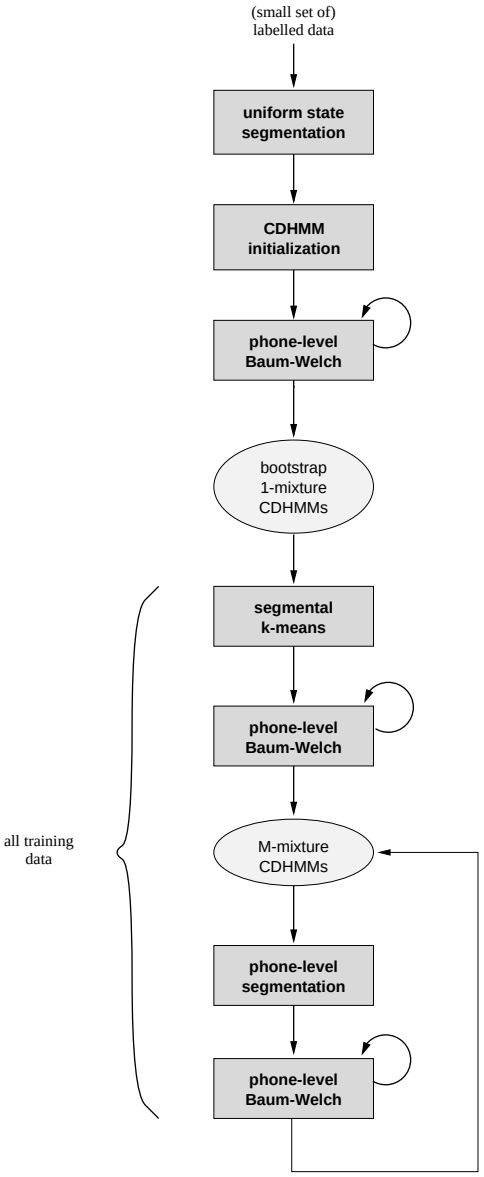


Fig. 3. Hybrid Viterbi/Baum-Welch CDHMM training procedure

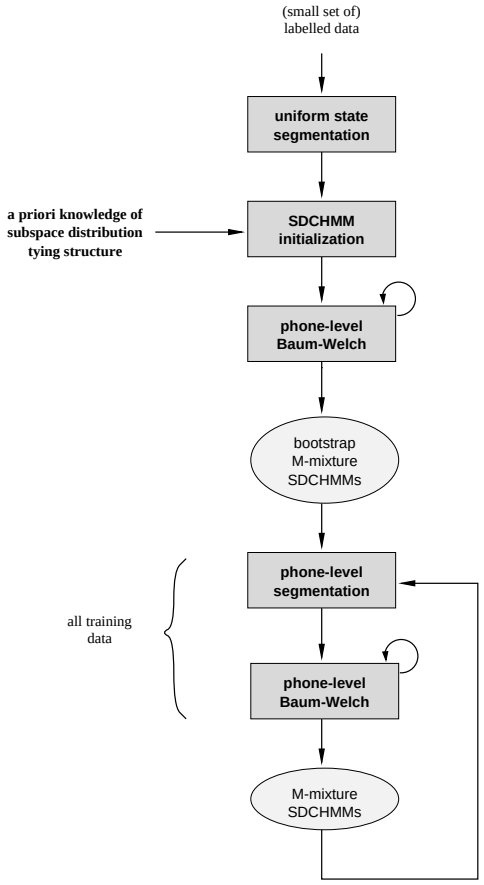


Fig. 4. Hybrid Viterbi/Baum-Welch SDCHMM training procedure

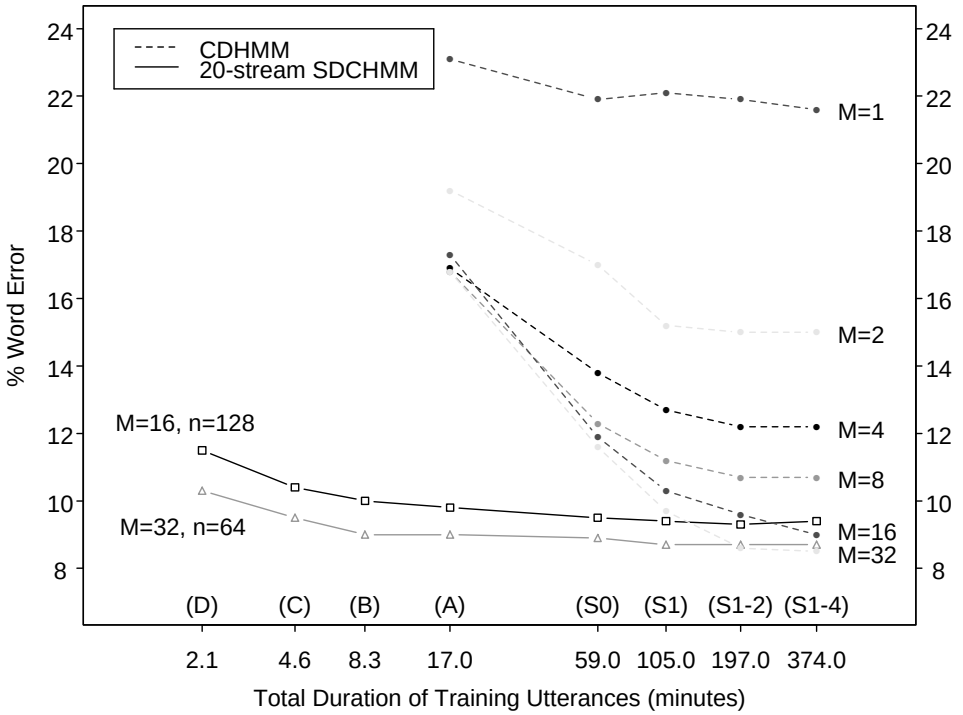


Fig. 5. ATIS: Comparison between the amount of training data required for CDHMM training and direct SDCHMM training (M = #mixtures and n = #subspace Gaussian prototypes per stream)

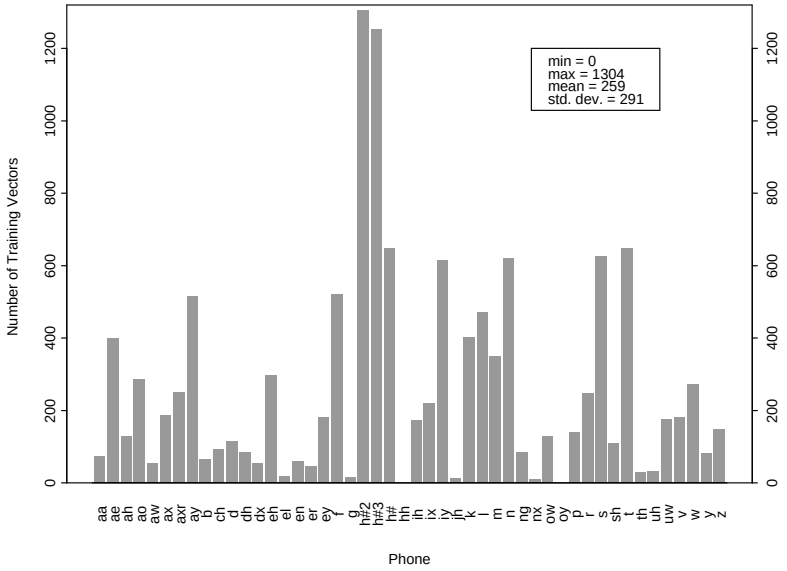


Fig. 6. Frame distribution of the phones in the training dataset D (2.1 minutes, 12421 frames of speech)

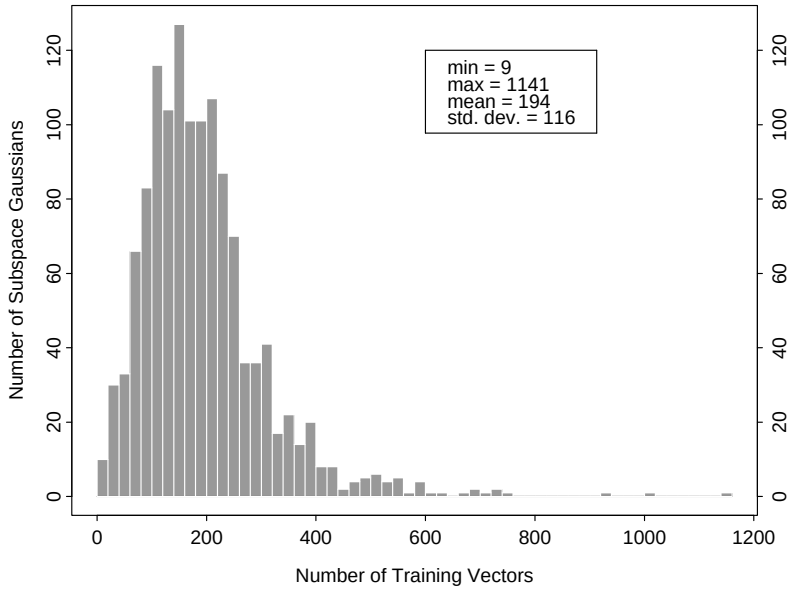


Fig. 7. Frame distribution of the subspace Gaussians in the training dataset D (2.1 minutes, 12421 frames of speech)

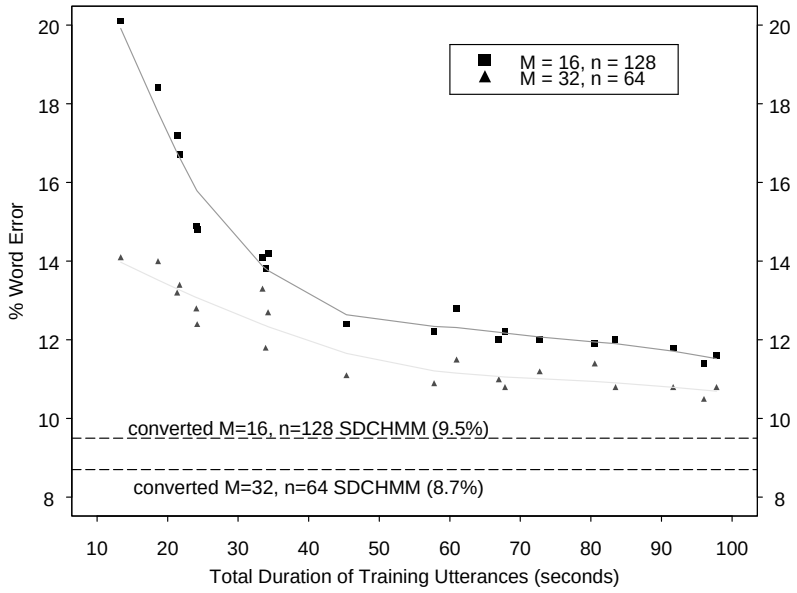


Fig. 8. ATIS: Variability with few training data (M = #mixtures and n = #subspace Gaussian prototypes per stream)

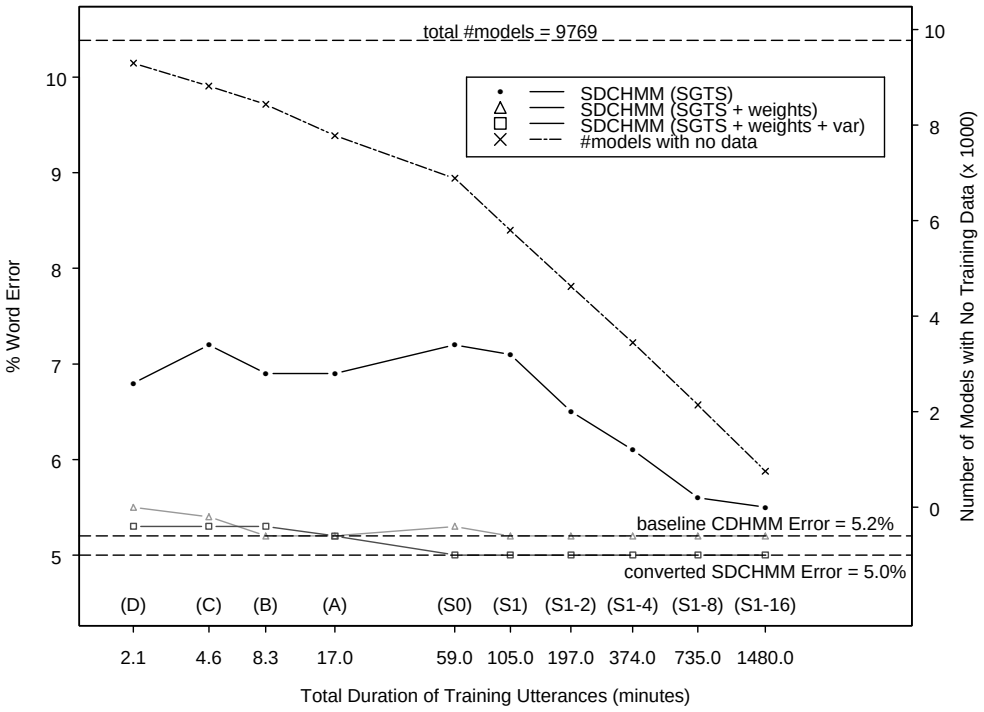


Fig. 9. ATIS: Data requirement for training CD SDCHMMs