An Acoustic-Phonetic and a Model-Theoretic Analysis of Subspace Distribution Clustering Hidden Markov Models

Brian Mak (mak@cs.ust.hk)
Department of Computer Science, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

June 14, 2003

Abstract. Recently, we proposed a new derivative to conventional continuous density hidden Markov modeling (CDHMM) that we call "*subspace distribution clustering hidden Markov modeling*" (SDCHMM). SDCHMMs can be created by tying low-dimensional subspace Gaussians in CDHMMs. In tasks we tried, usually only 32–256 subspace Gaussian prototypes were needed in SDCHMM-based system to maintain recognition performance of its original CDHMM-based system — a reduction of Gaussian parameters by one to three orders of magnitude. Consequently, both recognition time and memory were greatly reduced. We also have showed that if the underlying *subspace distribution tying structure* is known, it may be used to train an SDCHMM-based system with as little as eight minutes of speech from *scratch*. All the results suggest that there is substantial redundancy in conventional CDHMM and that SDCHMM is a more compact model. In this paper, we analyze the tying structure from two perspectives: from the acoustic-phonetic perspective showing that the tying structure seems to capture prominent relationship among phones; and, from the model-theoretic perspective showing that SDCHMMs, if properly created from CDHMMs, may be preferred over the latter as they are less complex and have the potential of greater generalization power.

Keywords: distribution clustering, parameter tying, model complexity, Bayesian information criterion

1. Introduction

Despite the long desire to use speech — often the most natural and efficient modality humans use to communicate — for human-machine interaction, the promise of an ubiquitous speech user interface has yet to be fulfilled. One reason is due to the high computational cost of today's state-of-the-art laboratory recognizers, mostly based on hidden Markov modeling (HMM). To arrive at the high recognition accuracy, recognizers are running at one to two orders of magnitude slower than real time, requiring high-end workstations equipped with hundreds of megabytes (MB) of memory. While there are many ways to achieve greater speed *or* less memory usage (Beyerlein and Ullrich, 1995; Bocchieri, 1993; Komori et al., 1995; Padmanabhan et al., 1997; Seide, 1995; Gopalakrishnan and Bahl, 1996), it is less easy to achieve *both* goals *without* sacrificing the accuracy. One exception is the successful technique of parameter tying. In the past, various parameters such as phones (generalized biphones/triphones (Lee et al., 1990)), states (tied-state HMM (Hwang, 1993; Young and Woodland, 1993)), observation distributions (tied-mixture/semi-continuous HMM (Bellegarda and Nahamoo, 1990; Huang and Jack, 1989; Singer and Lippmann, 1992)), and feature parameters (Takahashi and Sagayama, 1995) have all been tied. Recently, we propose to push the technique to an even finer sub-phonetic unit — subspace distributions. Subspace distributions are the projections of the full-space distributions of an HMM in low-dimensional subspaces. The hypothesis is that speech sounds are more alike in acoustic subspaces than in the full acoustic space. We call our novel HMM formulation *"subspace distribution clustering hidden Markov modeling"* (SDCHMM) (Bocchieri and Mak, 2001), and the tying information among subspace distributions of SDCHMMs together with the mappings between them and the full-space distributions the *subspace distribution tying structure* (SDTS).

In (Bocchieri and Mak, 2001), we showed that Gaussian quantization can be very efficient in low dimension, resulting in a reduction of Gaussian parameters by one to three orders of magnitude. As a result, there can be substantial savings in recognition time and memory usage. Since then similar findings were obtained by other researchers of the field (Aiyer et al., 2000; Rigazio et al., 2000; Astrov, 2002). Table I shows the typical computational savings we (on the Resource Management (RM) (Price et al., 1988) and Air Traffic Information System (ATIS) (Hemphill et al., 1990)) and others (an IBM LVASR task (Aiyer et al., 2000)) obtained in some recognition tasks[1]. In the

[1] Since the comparative performance of SDCHMMs and CDHMMs is the main theme here, details of the recognition systems, such as the number of tied states,

table, memory savings (MS) refers only to the memory used by mixture weights (4 bytes), Gaussian means (4 bytes) and variances (4 bytes), and indices used to encode the SDTS (1 byte); and time savings (TS) refers to savings in the total decoding time, not just the computation time of Gaussian likelihoods. Both memory and time savings are computed by comparing the figures in an SDCHMM-based system with respect to those in the corresponding CDHMM-based system. In (Mak and Bocchieri, 2001), we further demonstrated on the ATIS task that if one has *a priori* knowledge of the SDTS of ATIS, one may even train context-independent or context-dependent SDCHMMs for the task from scratch using as little as 8 minutes of ATIS data which perform as well as their CDHMM counterparts. All these results suggest substantial redundancy in our standard CDHMM-based systems, and that the SDTS is a succinct representation of the inter-relationship among phones.

Table I. Typical computational savings by SDCHMMs (Vocab = size of vocabulary in words, PP = grammar perplexity, WER = word error rate (%), G = #Gaussian, K = #subspaces, g = #subspace Gaussian prototypes per subspace, MS = memory savings, TS = time savings)

Task	Monophone /Triphone	Vocab	PP	WER	G	K	g	MS	TS
ATIS	monophone	1,536	20	9.5	2,254	20	128	87.8%	30.0%
ATIS	triphone	1,536	20	5.2	76,154	13	128	94.5%	56.0%
RM	triphone	1,000	60	3.9	5,349	39	64	85.2%	60.8%
IBM	triphone	20,000	160	11	43,444	40	64	86.2%	—

Although tying structures and schemes are not new in speech recognition systems, in the past, parameter tying is generally only treated as a technique to robustly increase model complexity for a given amount of training data. Little analysis is done on the resulting tying structures. In this paper, we examine the SDTS in SDCHMMs from two different perspectives, hoping that it will shed some light on our understanding of these tying structures. Firstly, in Section 3, we will present an acoustic-phonetic analysis of the SDTS. The major outcome is that the SDTS seems to capture prominent relationship among phones.

language models, etc. are not provided. However, from the various word accuracies, one can be assured that these results are generated from reasonably good systems.

Secondly, in Section 4, we will examine SDCHMMs from the model-theoretic perspective, and consider tying as a model order selection problem. The analysis shows that the Bayesian information criterion (BIC) is a fairly good predictor of SDCHMM performance. That is, a less complex SDCHMM-based system — with a smaller BIC value — has a higher recognition accuracy. In addition, SDCHMMs that perform better than CDHMMs (from which they are derived) on testing data have smaller BICs.

In the next section, we will first review the basic theory of SDCHMM.

2. Review of SDCHMM

In this Section, we review the theory of subspace distribution clustering hidden Markov modeling, and briefly outline two ways to train SDCHMMs: one directly from training data if the subspace distribution tying structure is known, and an indirect conversion from a set of CDHMMs.

2.1. 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 (or block-diagonal covariances[2]). 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

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

[2] For simplicity and clarity, this paper assumes Gaussians with diagonal covariances which are most commonly used in speech recognition. The discussion can easily be generalized to the case with block-diagonal covariances.

The key observation is that a Gaussian with diagonal covariance can be expressed as a product of subspace Gaussians where the subspaces (or streams[3]) 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, $1 \leq k \leq K$, with diagonal covariances. Thus, Eqn. (1) can be rewritten as

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

where $\boldsymbol{O}_k$, $\boldsymbol{\mu}_{imk}$, and $\boldsymbol{\sigma}_{imk}^2$ 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* CDHMMs. Hence, the state observation probability in Eqn. (2) is modified as

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

where $\omega_k(\cdot)$ represents the tying structure: it maps the k-stream of the m-th component of the mixture Gaussian of state i to a subspace Gaussian prototype of the k-th stream.

2.2. SDCHMM Training

SDCHMMs can be trained either directly from speech data or indirectly from a set of already-trained CDHMMs as shown in Fig. 1.

2.2.1. *Indirect Training: Model Conversion from CDHMMs*

The formulation of SDCHMM as of Eqn. (3) suggests that SDCHMMs may be implemented in two steps as shown in the left block in Fig. 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.

[3] In this paper, the two terms, "subspace" and "stream" are used interchangeably to mean a feature space of dimension smaller than that of the full feature space. "Subspace" is clearer mathematically, but "stream" is more common in the speech recognition community.

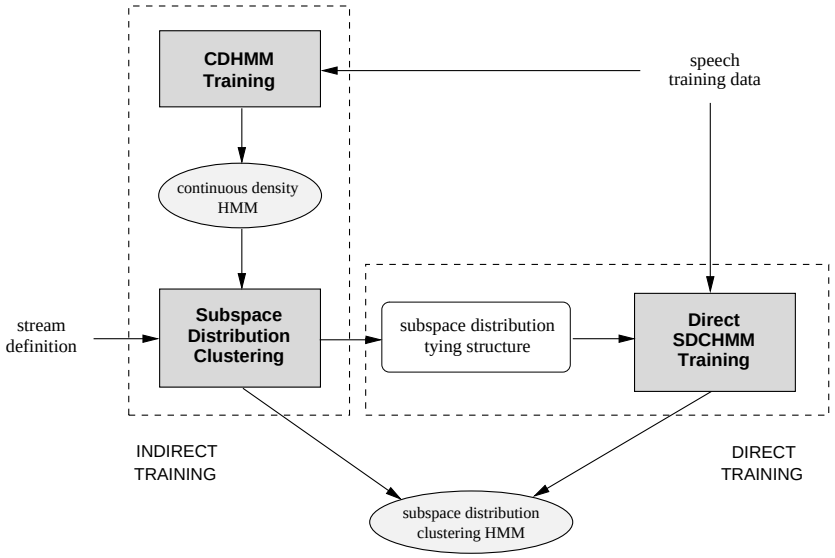


Figure 1. Two methods of training SDCHMMs

(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 (Bocchieri and Mak, 2001).

2.2.2. *Direct Training*

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. It does not make use of the fact that SDCHMMs have significantly fewer Gaussian parameters (mixture weights, means, and variances). Thus, if we have *a priori* knowledge of the *subspace distribution tying structure* (SDTS), one should be able to train SDCHMMs directly from significantly less speech data as shown in the right-block in Fig. 1. Maximum likelihood estimation of SDCHMM parameters may be done in much the same way as CDHMM parameters are estimated using the *Baum-Welch* algorithm (Baum et al., 1970). In fact, the additional constraints imposed by the SDTS only alter the way in which statistics are gathered from the training observations. That is, to estimate the Gaussian parameters of a subspace Gaussian prototype, statistics are collected from all frames of any states of any models that use the prototype. The detailed re-estimation formulas can be found in (Mak and Bocchieri, 2001).

3. Acoustic-Phonetic Analysis of the Subspace Distribution Tying Structure

From Section 2, one way to create SDCHMMs is through converting conventional CDHMMs. The conversion involves a simple Gaussian clustering process which is fully automatic, utilizing only acoustic information from the training data. Yet recognition results on the ATIS task (Table II) show that, for instance, SDCHMMs with 20 streams and 64 subspace Gaussians prototypes per stream are adequate to represent the original context-dependent CDHMMs containing 76,154 full-space Gaussians — a reduction of Gaussian parameters (means and variances) by a factor of more than 1,000. What is even more intriguing is that with only two subspace Gaussian prototypes — or one bit of information — per stream, a 39-stream context-dependent SDCHMM-based system can still achieve a WER of 9.1%. Such efficient tying suggests that the (original) full-space Gaussians are highly redundant. It is therefore interesting to "see" *how* acoustics are similarly realized by *which* speech units from the subspace distribution tying structure.

Table II. Reduction of Gaussian parameters by context-dependent SDCHMMs on ATIS (K = #subspaces, g = #subspace Gaussian prototypes per subspace, WER = word error rate (%), GPR = reduction in Gaussian parameters)

K	g	WER	GPR
1	76,154	5.2	1
4	256	5.8	297
13	128	5.2	595
20	64	5.0	1,189
39	32	5.0	2,380
39†	2	9.1	38,077

† This case used a much larger decoding beam width of 270 while all other cases used a beam width of 170.

With the huge number of combinations of phonetic units, HMM states, and Gaussian components in SDCHMMs, it is hard to visualize the whole subspace distribution tying structure in a single picture. In the following, we present a simple quantitative analysis of the number of subspace Gaussians that are shared by corresponding HMM states of any pair of phones. We hope that the analysis will shed some light on the acoustic-phonetic nature of speech.

3.1. Analysis on ATIS

In order to generate some readable visual plots of the subspace distribution tying structure between phone pairs, we employ a less complex SDCHMM-based system. To do that, we first trained monophone CDHMMs with about 4,000 ATIS training utterances. The CDHMMs have the same HMM configuration as the baseline CDHMMs of Table I except that there are only four Gaussian mixture components per state (instead of 16 in the latter). Twenty-stream monophone SDCHMMs were then derived from the CDHMMs by the model conversion scheme as explained in Section 2.2.1 requiring 64 prototypes per stream. The resulting SDCHMMs have a recognition word error rate (WER) of 12.6%[4]. The SDTS of the 20-stream SDCHMMs is then analyzed.

3.2. Methodology

For the corresponding states of any two phonetic SDCHMMs with the same number of HMM states, which are modeled as mixture Gaussian densities, the constituent subspace Gaussians of their full-space Gaussians are compared. Specifically, for each stream, the number of common subspace Gaussians at the corresponding states of the two SDCHMMs are counted *irrespective* to which mixture components the subspace Gaussians come from. The procedure may be expressed in pseudo-code as follows:

```
for each pair of phones (P, Q) with the same number of states
    for each state
    {
        num_common_subgaussian = 0
        for each stream
        {
            P.list = subspace Gaussians from all mixture components of phone P
                     in this state projected onto this stream
            Q.list = subspace Gaussians from all mixture components of phone Q
                     in this state projected onto this stream
            num_common_subgaussian += Common_Gaussian(P.list, Q.list)
        }
        print(num_common_subgaussian)
    }

Common_Gaussian(list_1, list_2)
{
    find the number of common subspace Gaussians between list_1 and list_2
}
```

[4] The result is worse than the one in Table I. It is mainly due to the reduced model complexity.

Since each subspace Gaussian may be represented by its prototype index, a full-space Gaussian in a 20-stream SDCHMM can be represented by a tuple of 20 prototype indices, one for each stream. For example, the 4-mixture densities of the third state of the phones "s" and "z" are represented as:

$$\begin{aligned}\{ &< 2,\ 4,\ 3,\ 2,\ 9, 46,\ 2, 52,\ 2,\ 2, 33, 13, 46, 37, 13, 21, 46, 60, 42,\ 2 >,\\ &< 2, 24, 12, 24,\ 2, 46, 24, 16, 13, 21, 47, 12, 46, 46, 46, 46,\ 2, 48, 28,\ 2 >,\\ &< 0, 24, 31, 34, 28,\ 2, 28, 35, 46, 37, 46, 46, 33, 46, 37, 46, 46, 48, 24, 21 >,\\ &< 4, 37, 12, 25, 34, 46,\ 4, 52, 31, 21, 16, 25, 12, 51, 44, 24,\ 5, 25, 12,\ 4 > \}\end{aligned}$$

and

$$\begin{aligned}\{ &< 46,\ 4, 47,\ 2, 47, 46, 13, 52,\ 2,\ 2, 33, 46, 46, 41, 13, 21, 46, 13, 24,\ 2 >,\\ &<\ 0, 24, 31, 34, 28, 12, 28, 35, 27, 37, 46, 12, 33, 46, 37, 21, 46, 48, 13, 21 >,\\ &< 46, 24, 46, 24,\ 2, 46, 24, 16, 13, 21, 47, 52, 33, 46, 46, 46,\ 2, 57, 28,\ 2 >,\\ &<\ 0,\ 4, 46, 44, 28, 13, 47, 37, 25,\ 1,\ 5,\ 4, 25, 51, 35, 21,\ 5, 25, 25, 25 > \}\end{aligned}$$

respectively. Thus, to determine the number of common subspace Gaussians in the fourth stream of the third state of "s" and "z", the two lists {2, 24, 34, 25} and {2, 34, 24 44} are compared and the result is three. Note that the order of indices is ignored. The computation is repeated for every stream and the counts are accumulated for each state.

3.3. Results

Forty-five phones (excluding three noise models) are used in our ATIS system, each having three HMM states. The number of common subspace Gaussians between any pairs of the 45 phones can be computed for each of their three states. The phones are further divided into two major categories: 18 vowels and 27 consonants[5]. Histograms of counts of the number of common subspace Gaussians between any two phones within each category and across the two categories are shown in Fig. 2 together with some of their statistics.

In addition, Fig. 3–5 provides a visualization of the SDTS between three pairs of phones belonging to various phonetic categories:

- vowel-vowel pair: "ae" and "eh"
- consonant-consonant pair: "s" and "z"
- consonant-vowel pair: "t" and "iy".

[5] The vowels are: aa, ae, ah, ao, aw, ax, axr, ay, eh, er, ey, ih, ix, iy, ow, oy, uh, and uw; and the consonants are: b, ch, d, dh, dx, el, en, f, g, hh, jh, k, l, m, n, ng, nx, p, r, s, sh, t, th, v, w, y, and z.

In each of the three figures, the abscissas are stream indices ranging from 1 to 20, while the ordinates are the subspace Gaussian prototype indices for each stream. For each stream of the 4-mixture Gaussians of a state, the subspace Gaussian prototype indices of the first phone in the pair are represented by the four letters "a", "b", "c", and "d". Subspace Gaussians symbolized by the same letter belong to the same full-space Gaussian component. Thus, if one connects all the letter "a"'s together across the 20 streams, one obtains the "trajectory" of a full-space Gaussian encoded by its subspace Gaussian prototypes. On the other hand, the subspace Gaussian prototype indices of the second phone in the pair are represented indiscriminately by square boxes. A match of subspace Gaussians between the two phones occurs when any of the four letters is "captured" by a box. (Due to the low resolution on the ordinate, only when a letter is right in the middle of a square box is there actually a match.) Specifically, the number of matches in the three figures, from the first state to the third state are:

- between "ae" and "eh": 21, 26, 27
- between "s" and "z": 25, 28, 48
- between "t" and "iy": 0, 0, 5

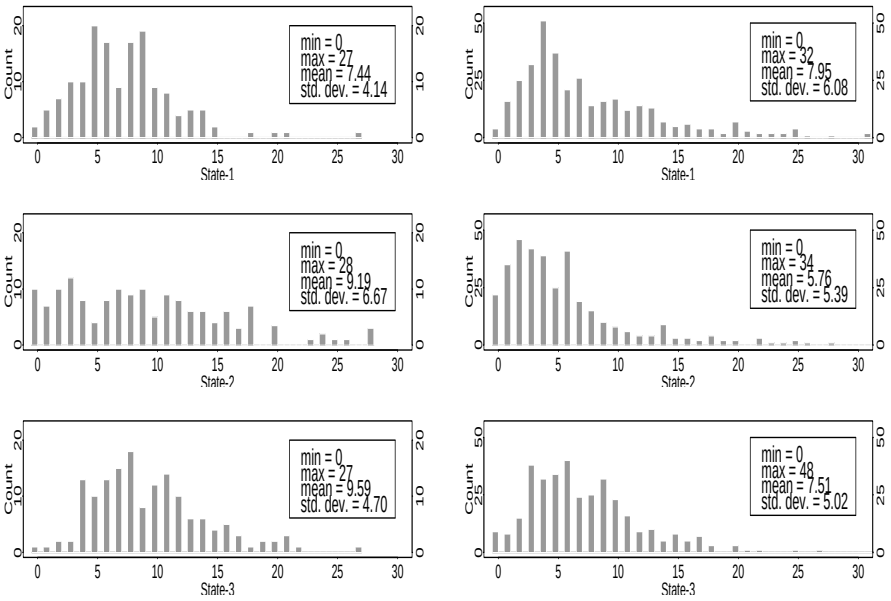
while the maximum possible number of matches is $20 \times 4 = 80$.

3.4. Discussion

The computed figures should be compared with the expected number of common subspace Gaussians between two 4-mixture SDCHMM states should the matches occur by pure chance, which is found to be 0.24 per stream. Thus, the expected number of common subspace Gaussians between two 20-stream SDCHMM states is $20 \times 0.24 = 4.8$ if the matches occur by chance.

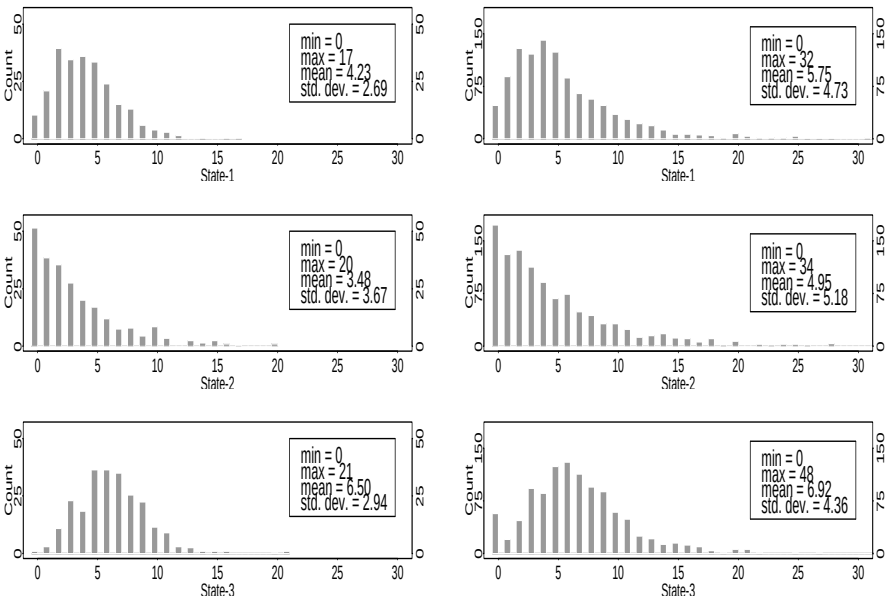
By comparing the expected number of matches of 4.8 and the figures shown in Fig. 2–5, we have the following observations:

- The extent of sharing of subspace Gaussians splits along broad phonetic categories (vowels and consonants; and within consonants, along sub-categories of fricatives, plosives, nasals and approximants (Ladefoged, 1993)). That is, there is more sharing of subspace Gaussians between two vowels or two consonants than between a vowel and a consonant; and, within consonants (from results not shown here due to space limitation), there is more sharing between two fricatives, two plosives, etc. For example, the mean number of shared subspace Gaussians between two vowels



(a) Between any two vowels

(b) Between any two consonants



(c) Between a vowel and a consonant

(d) Between any two phones

Figure 2. Counts of the number of common subspace Gaussians between phones of different broad categories in an ATIS SDCHMM-based system

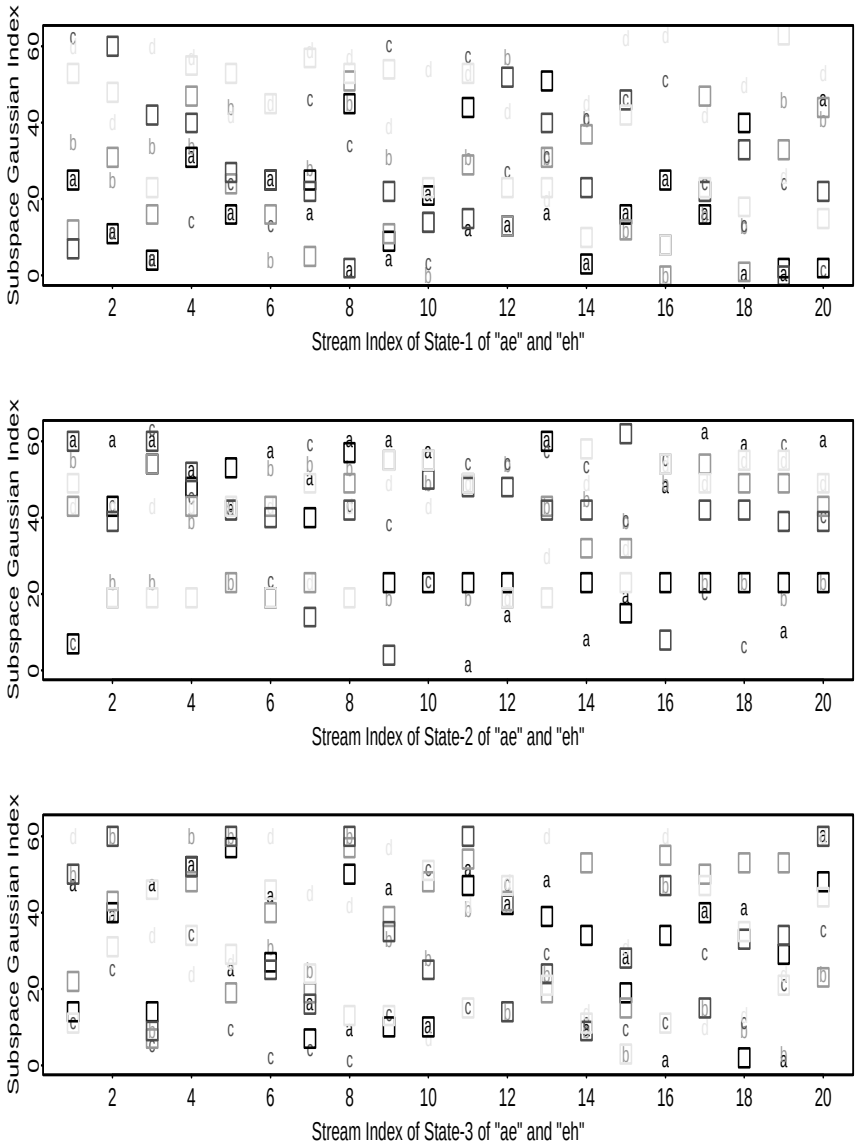


Figure 3. Subspace distribution tying structure between “ae” and “eh” (number of matches from the 1st to the 3rd state are 21, 26, 27)

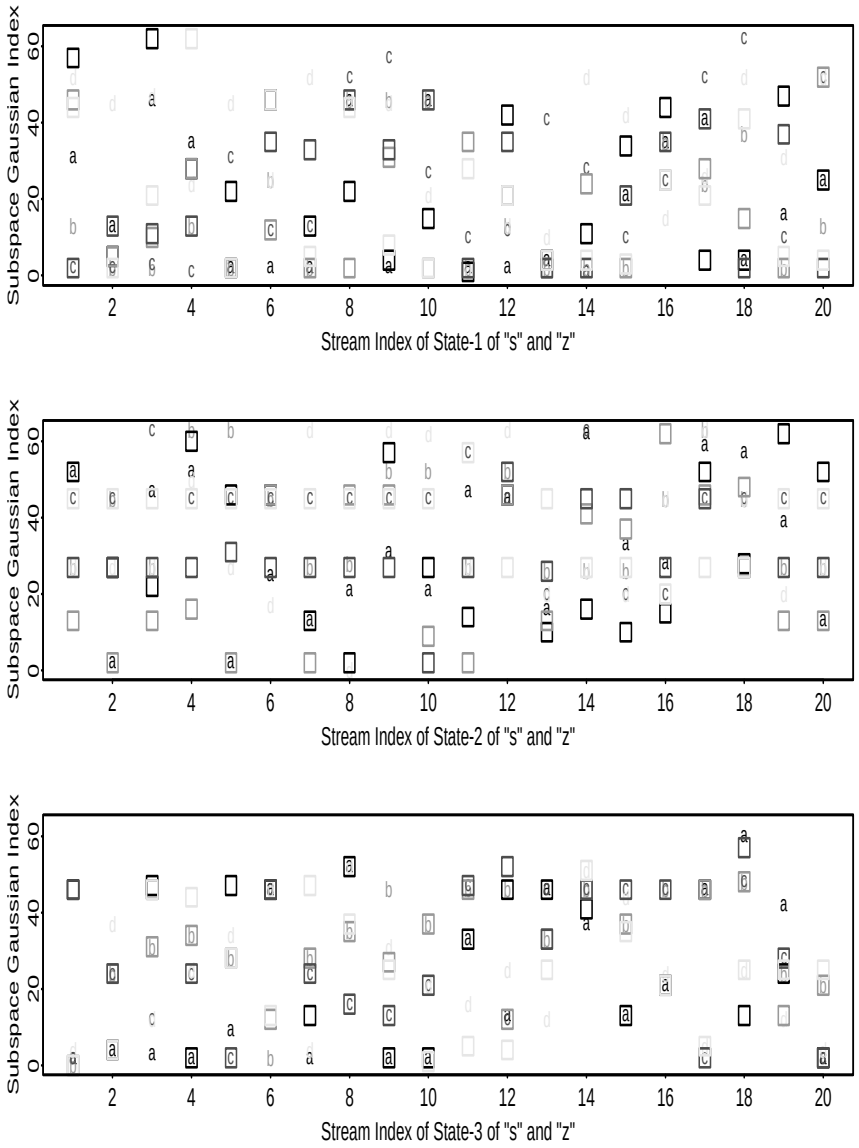


Figure 4. Subspace distribution tying structure between "s" and "z" (number of matches from the 1st to the 3rd state are 25, 28, 48)

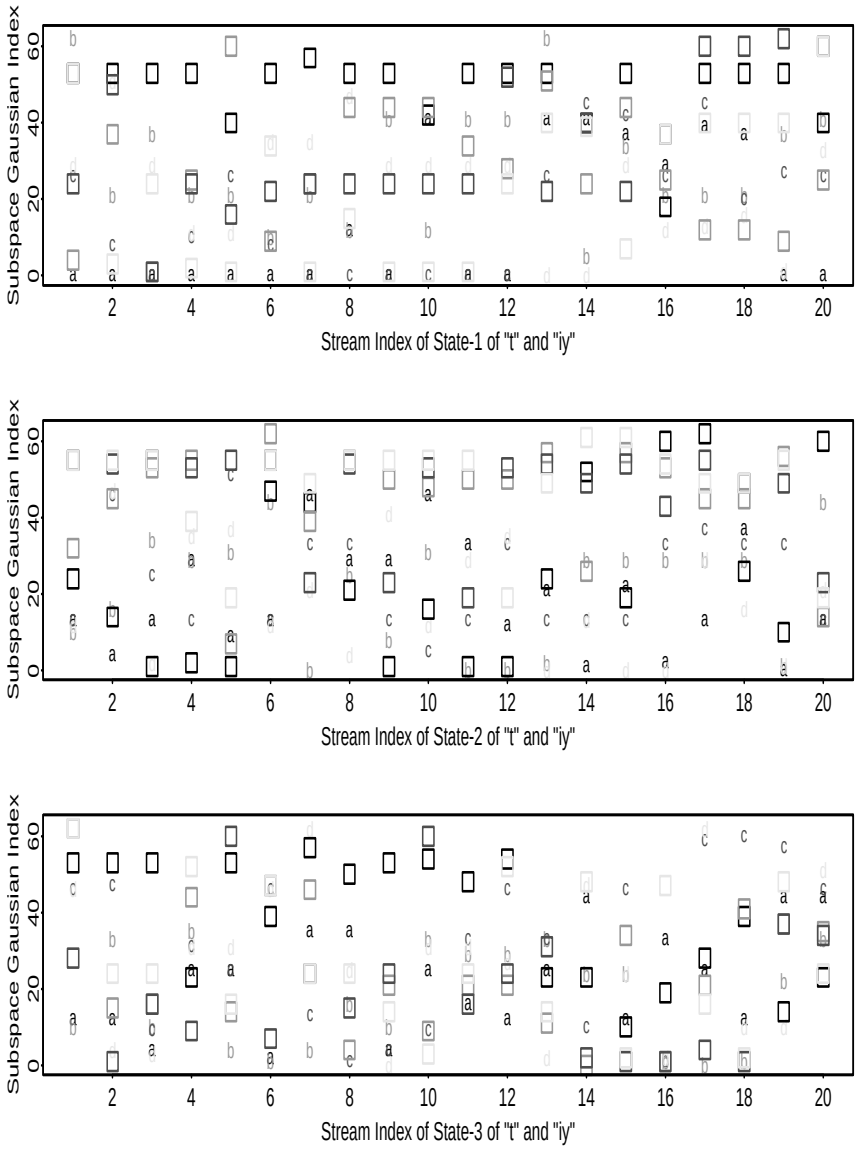


Figure 5. Subspace distribution tying structure between "t" and "iy" (number of matches from the 1st to the 3rd state are 0, 0, 5)

or consonants are well above 4.8, the expected number of matches by chance; however, that between vowels and consonants are below 4.8 for the beginning and middle states. The effect is conspicuously illustrated by Fig. 3–5 in which vowel pair "ae"–"eh" and consonant pair "s"–"z" have 25–60% of their subspace Gaussians shared in all three states; whereas there is basically no sharing between the consonant-vowel pair "t"–"iy".

– In the mid-states, where the coarticulatory effect is weaker and the identity of a phone is better preserved, there is much less sharing of subspace Gaussians between vowel-consonant pairs, while vowel-vowel pairs exhibit more sharing. In fact, the average number (3.48) of common subspace Gaussians between vowel-consonant pairs is well below the expected value of 4.8. The histogram for the case of vowel-vowel pairs is also more uniform than that of consonant-consonant pairs. This may be attributed by the more gradual differences in the articulations of the vowels. In contrast, the articulations of different categories of consonants are very different (c.f. nasals vs. plosives).

– On average, there is more sharing between two vowels than between two consonants. This again confirms the greater resemblance between vowels.

All the observations are well in accord with our knowledge about the phones. The analysis provides some understanding of the efficiency of subspace distribution clustering hidden Markov modeling in encoding the phonetic information.

4. Model-Theoretic Analysis of SDCHMM

Before statistical pattern classification can be performed, mathematical models are first built from observations; and in the case of speech recognition, they are acoustic models. An immediate question is: What is the "best" model that can be estimated from a given set of data? This is a question of great controversy and different schools of model complexity have their own meaning of what is "best" (if they agree on whether there is a "best" at all). In terms of Kolmogorov complexity, it is even an unsolvable problem (Li and Vitanyi, 1997). Nevertheless, if we fix the family of models and for a finite set of models, there are still an arsenal of complexity measures for model selection that have been found useful in practice. That is, given a set of data X, and a set of models $\lambda \in \Lambda$ with model parameters θ_λ that explain X, determine

the "best"model $\hat{\lambda}$ among the models $\lambda \in \Lambda$. In this Section, we will borrow an analytical tool from the model complexity community to compare various SDCHMMs and investigate its predictive power of the models' recognition performance.

4.1. Measures of Model Complexity

Model complexity measures can be classified into two categories:

1. Measures directly derived from the posterior probability of a model. That is, the best model is

$$\begin{aligned} \hat{\lambda} &= \underset{\lambda \in \Lambda}{\operatorname{argmax}}\, P(\lambda|X) \\ &= \underset{\lambda}{\operatorname{argmax}}\, P(X|\lambda)P(\lambda) \\ &= \underset{\lambda}{\operatorname{argmin}}\, [-\log P(X|\lambda) - \log P(\lambda)] \ . \end{aligned} \tag{4}$$

   Common measures of this type include (minimum) Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Schwarz, 1978). Their main differences lie on how they approximate the priors.

2. Measures derived from coding theory to find the shortest string that encodes the posterior probability of Eqn.(4), most notably the Minimum Description Length (MDL) (Rissanen, 1978) and Minimum Message Length (MML) (Wallace and Boulton, 1968). They differ mainly on the coding schemes and their emphasis on the priors.

A detailed treatment of the topic is beyond the scope of this paper and interested readers are referred to a survey by A. D. Lanterman (Lanterman, 2001).

While these model complexity measures differ in details, they all may be expressed as penalized log-likelihoods:

$$\begin{aligned} &\text{model} \quad \text{complexity} \\ &= \quad -\text{log likelihood of data} + \text{penalty due to complexity} \\ &= \quad -\log P(X|\theta_\lambda, \lambda) + C(X, \theta_\lambda|\lambda) \ . \end{aligned} \tag{5}$$

In this paper, we choose the Bayesian Information Criterion (BIC) as our metric to compare the complexity of various SDCHMMs. BIC has been used in other fields with some success (Liang et al., 1992; Wax

and Kailath, 1985) and is also not unfamiliar to the speech community (Chan et al., 2000; Chen and Gopalakrishnan, 1998).

4.2. Complexity Analysis of SDCHMMs Using BIC

Since we are interested only in the *difference* in model complexity between conventional CDHMMs and our SDCHMMs, there is no need to compute the BIC for any common information between the two models. Specifically, our SDCHMMs are converted from CDHMMs by keeping the same HMM topologies (number of states, transitions, and number of Gaussian mixtures for each state), transition probabilities, as well as Gaussian mixture weights, which are thus common to both models and they can be factored out from our BIC calculation. The only differences between SDCHMMs and CDHMMs are their Gaussian parameters and the addition of the subspace distribution tying structure in SDCHMMs. Therefore, we simplify the BIC of CDHMMs and SDCHMMs as follows:

$$BIC(\lambda_{CDHMM}) = -\log P(X|\tilde{\lambda}_{CDHMM}) + \frac{D_{CDHMM}}{2}\log N \ . \quad (6)$$

$$BIC(\lambda_{SDCHMM}) = -\log P(X|\tilde{\lambda}_{SDCHMM}) + \frac{D_{SDCHMM}}{2}\log N \ . (7)$$

where, $\tilde{\lambda}_{CDHMM}$ is the maximum likelihood estimate of the set of CDHMMs and $\tilde{\lambda}_{SDCHMM}$ is the set of SDCHMMs converted from $\tilde{\lambda}_{CDHMM}$; D_{CDHMM} is the sum of all Gaussian parameters in CDHMMs; D_{SDCHMM} is the sum of all Gaussian parameters *plus* the encoding of SDTS in SDCHMMs; and N is the number of training speech frames. Furthermore, if we let

K = number of streams
g = number of subspace Gaussian prototypes per stream
G = number of full-space Gaussians
D = dimension of each feature vector

then Eqn.(6) and Eqn.(7) become

$$BIC(\lambda_{CDHMM}) = -\log P(X|\tilde{\lambda}_{CDHMM}) + DG\log N. \quad (8)$$

$$BIC(\lambda_{SDCHMM}) = -\log P(X|\tilde{\lambda}_{SDCHMM}) + \left(Dg + \frac{GK}{2}\right)\log N. (9)$$

4.3. Analysis of SDCHMMs on Resource Management

SDCHMMs of various complexities were generated from the Resource Management task (RM) (Price et al., 1988) for this analysis. 39-dimensional

feature vectors, consisting of 12 MFCCs and normalized energy plus their first- and second-order derivatives, were extracted every 10ms. Training data come from 3990 speaker-independent training utterances in the feb91 RM corpus, and the testing data is comprised of 300 utterances from the same corpus. There are totally 1,369,977 training speech frames.

4.3.1. *Procedure*
From the RM training corpus, CDHMMs and SDCHMMs of various complexities were derived as follows:

Step 1. Speaker-independent and context-dependent CDHMMs were trained using the HTK Toolkit (Young et al., 1999). There are 2,279 triphones and 5,349 Gaussians.

Step 2. SDCHMMs with various numbers of streams (1, 13, 20, 39) and subspace Gaussian prototypes (64, 128, 256, 512) were obtained by the conversion method described in Section 2.2.1.

Step 3. For each set of HMMs, forced alignment was performed over all training data and the sum of acoustic likelihoods was recorded.

Step 4. Each set of HMMs was used to decode all training and test data separately to get their word recognition accuracies with a very large beam-width to mitigate the effect of pruning the search space.

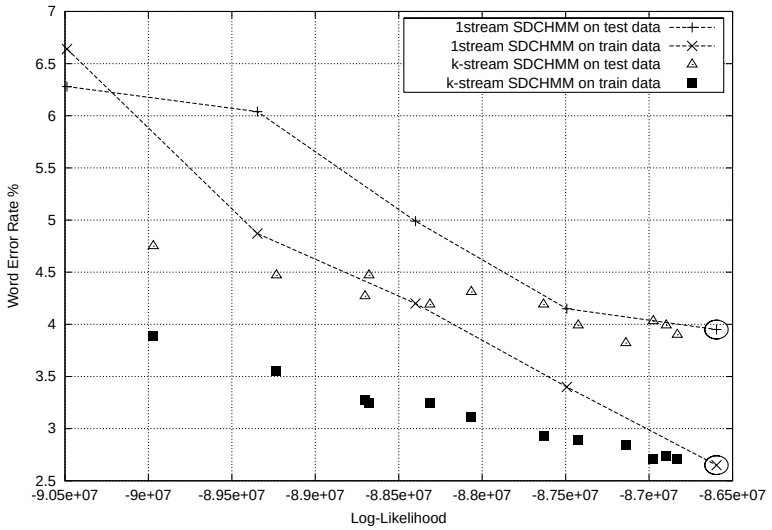
Step 5. The Bayesian information criterion of each set of CDHMMs and SDCHMMs was computed by Eqn.(8) and Eqn.(9) respectively.

The relation between the log-likelihoods on training data, BIC and recognition performance of each set of HMMs is depicted in Table III and Fig. 6.

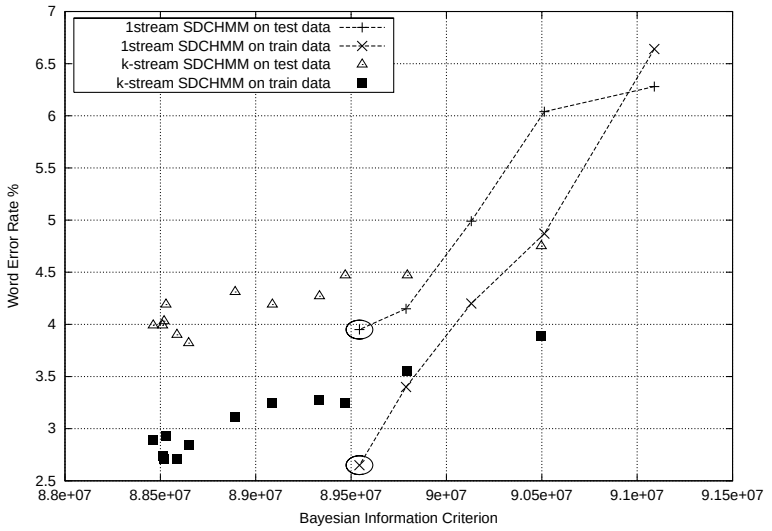
4.4. Discussion

From Table III, we have the following observations on the model likelihoods:

- the baseline CDHMMs have the highest log-likelihood. This is expected since all other SDCHMMs are approximations to the baseline CDHMMs.
- for a given number of streams, the log-likelihood increases monotonically with the number of subspace Gaussian prototypes in the SDCHMMs. This again agrees with the fact that since SDCHMMs



(a) log-likelihood



(b) BIC

Figure 6. BIC analysis on various HMMs (Performance of the baseline CDHMMs are circled.)

Table III. Model complexity analysis of HMMs using BIC (K = #stream; g = #subspace Gaussian prototypes per stream; LL = log-likelihood over all training data; P = penalty before multiplied by $\log N/2$; WER1/WER2 = word error rate on training/testing data)

Model	K	g	$LL(\times 10^7)$	$P(\times \log N/2)$	BIC $(\times 10^7)$	WER1	WER2
SDCHMM	1	1,024	-9.04880	85,221	9.10901	6.64	6.28
SDCHMM	1	2,048	-8.93470	165,093	9.05134	4.87	6.04
SDCHMM	1	3,072	-8.83998	244,965	9.01305	4.20	4.99
SDCHMM	1	4,096	-8.74931	324,837	8.97881	3.40	4.15
CDHMM	1	5,349	-8.65962	417,222	8.95439	2.65	3.95
SDCHMM	13	64	-8.99707	74,529	9.04973	3.89	4.75
SDCHMM	13	128	-8.92331	79,521	8.97949	3.55	4.47
SDCHMM	13	256	-8.87021	89,505	8.93345	3.27	4.27
SDCHMM	13	512	-8.83138	109,473	8.90872	3.25	4.19
SDCHMM	20	64	-8.86780	111,972	8.94691	3.25	4.47
SDCHMM	20	128	-8.80656	116,964	8.88920	3.11	4.31
SDCHMM	20	256	-8.76324	126,948	8.85293	2.93	4.19
SDCHMM	20	512	-8.74248	146,916	8.84628	2.89	3.99
SDCHMM	39	64	-8.71393	213,603	8.86484	2.84	3.82
SDCHMM	39	128	-8.69752	218,595	8.85196	2.71	4.03
SDCHMM	39	256	-8.68971	228,579	8.85120	2.74	3.99
SDCHMM	39	512	-8.68312	248,547	8.85872	2.71	3.90

are converted from the baseline CDHMMs and better approximation is obtained with more prototypes, resulting in smaller quantization errors and thus higher likelihoods.

– for a given number of prototypes, the model likelihood also increases with the number of streams proving that Gaussian clustering is more effective in lower dimensional spaces.

Likelihoods of the models are highly correlated with their recognition performance. In general, except for the 1-stream SDCHMMs, the recognition performance of the various models on both the training and testing data agrees well with their likelihoods: higher the likelihood is, smaller the recognition error will be. However, the 1-stream SDCHMMs and the rest of SDCHMMs seem to take on two different courses so that, for example, although a 1-stream SDCHMM and 20-stream SDCHMM may have the same likelihood, the latter always has a lower WER.

Now, when we look at the BICs of the models, most of the observations with model likelihoods also holds for the model BICs. This is not surprising as the log-likelihood accounts for about 98.6% of the value of a BIC on average. However, if all models are taken into consideration, the recognition performance of the models correlate better with their BICs than their likelihoods. From Fig. 6, except for the baseline CDHMMs, the WERs of all the other SDCHMMs more or less fall with their BIC values. In this aspect, the BIC is a better predictor of a model's performance. According to Table III, 10 SDCHMMs (starting from the one with 13 streams and 256 subspace Gaussian prototypes and downwards) have smaller BICs than the baseline CDHMMs. Among them, two models actually have a smaller WER on the test data than the baseline CDHMMs while half of them are within (relatively) 2.0% of the baseline WER (and some of these differences are actually not statistically significant). Hence we see that even though the log-likelihoods dominate overly in the BIC value of the models, they may not be a good metric for the models' predictive power. Instead, model complexity measure such as BIC, though imperfect as shown in our case, can be a better indicator.

5. Conclusion

Recently, we developed the subspace distribution clustering hidden Markov models (SDCHMM) to improve system performance in terms of speed and memory usage. Other researchers were able to repeat our experience in their own laboratories. Yet there lacks a formal analysis on SDCHMMs. In this paper, we attempt to investigate the effectiveness of subspace distribution tying in SDCHMMs from two different perspectives. From an acoustic-phonetic analysis, we conclude that the subspace distribution tying structure can capture prominent acoustic relationship among the phones. For instance, phones belonging to the same broad phonetic category are tied to a much greater extent than those belonging to different categories. From a model-theoretic analysis using the Bayesian information criterion (BIC) as a measure of model complexity, we find that less complex HMMs generally result in higher recognition accuracy, and the BIC is a better predictor of their recognition performance (on testing data) than their likelihoods (computed from the training data). Furthermore, SDCHMM-based systems with a lower recognition error rate than the reference CDHMM-based system all have smaller BIC values.

In summary, SDCHMMs are more compact models and can give better system performance than their parent CDHMMs from which they are derived.

References

Aiyer, A., M. Gales, and M. Picheny: 2000, 'Rapid Likelihood Calculation of Subspace Clustered Gaussian Components'. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. pp. 1519–1522.

Akaike, H.: 1974, 'A New Look at Statistical Model Identification'. *IEEE Transactions on Automatic Control* **19**(6), 716–723.

Astrov, S.: 2002, 'Memory Space Reduction for Hidden Markov Models in Low-resource Speech Recognition Systems'. In: *Proceedings of the International Conference on Spoken Language Processing*. pp. 1585–1588.

Baum, L., T. Petrie, G. Soules, and N. Weiss: 1970, 'A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains'. *Annals of Mathematical Statistics* **41**, 164–171.

Bellegarda, J. and D. Nahamoo: 1990, 'Tied Mixture Continuous Parameter Modeling for Speech Recognition'. *IEEE Transactions on Acoustics, Speech and Signal Processing* **38**(12), 2033–2045.

Beyerlein, P. and M. Ullrich: 1995, 'Hamming Distance Approximation for a Fast Log-Likelihood Computation for Mixture Densities'. In: *Proceedings of the European Conference on Speech Communication and Technology*, Vol. 2. pp. 1083–1086.

Bocchieri, E.: 1993, 'Vector Quantization for the Efficient Computation of Continuous Density Likelihoods'. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2. pp. 692–695.

Bocchieri, E. and B. Mak: 2001, 'Subspace Distribution Clustering Hidden Markov Model'. *IEEE Transactions on Speech and Audio Processing* **9**(3), 264–275.

Chan, Y. C., M. Siu, and B. Mak: 2000, 'Pruning of State-Tying Tree using Bayesian Information Criterion with Multiple Mixtures'. In: *Proceedings of the International Conference on Spoken Language Processing*, Vol. IV. Beijing, China, pp. 294–297.

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

Gopalakrishnan, P. and L. Bahl: 1996, 'Fast Match Techniques'. In: C. Lee, F. Soong, and K. Paliwal (eds.): *Automatic Speech and Speaker Recognition (Advanced Topics)*. Kluwer Academic Publishers, Chapt. 17, pp. 413–428.

Hemphill, C., J. Godfrey, and G. Doddington: 1990, 'The ATIS Spoken Language Systems Pilot Corpus'. In: *Proceedings of the DARPA Speech and Natural Language Workshop*. Morgan Kaufmann Publishers.

Huang, X. and M. Jack: 1989, 'Semi-continuous Hidden Markov Models for Speech Signals'. *Journal of Computer Speech and Language* **3**(3), 239–251.

Hwang, M.: 1993, 'Shared Distribution Hidden Markov Models for Speech Recognition'. *IEEE Transactions on Speech and Audio Processing* **1**(4), 414–420.

Komori, Y., M. Yamada, H. Yamamoto, and Y. Ohora: 1995, 'An Efficient Output Probability Computation for Continuous HMM Using Rough and Detail Mod-

els'. In: *Proceedings of the European Conference on Speech Communication and Technology*, Vol. 2. pp. 1087–1090.

Ladefoged, P.: 1993, *A Course in Phonetics*. Harcourt Brace Jovanovich College Publishers, 3rd edition.

Lanterman, A. D.: 2001, 'Schwarz, Wallace, and Rissanen: Intertwining Themes in Theories of Model Selection'. *International Statistical Review*.

Lee, K., S. Hayamizu, H. Hon, C. Huang, J. Swartz, and R. Weide: 1990, 'Allophone Clustering for Continuous Speech Recognition'. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2. pp. 749–752.

Li, M. and P. Vitanyi: 1997, *An Introduction to Kolmogorov Complexity and Its Applications*. New York: Springer-Verlag, 2nd edition.

Liang, Z., R. Jaszczak, and R. Coleman: 1992, 'Parameter Estimation of Finite Mixtures Using the EM Algorithm and Information Criteria with Application to Medical Image Processing'. *IEEE Transactions on Nuclear Science* **39**, 1126 –1133.

Mak, B. and E. Bocchieri: 2001, 'Direct Training of Subspace Distribution Clustering Hidden Markov Model'. *IEEE Transactions on Speech and Audio Processing* **9**(4), 378–387.

Padmanabhan, M., D. N. L.R. Bahl, and P. de Souza: 1997, 'Decision-Tree Based Quantization of the Feature Space of a Speech Recognizer'. In: *Proceedings of the European Conference on Speech Communication and Technology*. pp. 147–150.

Price, P., W. Fisher, J. Bernstein, and D. Pallett: 1988, 'The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition'. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. pp. 651–654.

Rigazio, L., B. Tsakam, and J. Junqua: 2000, 'An Optimal Bhattacharyya Centroid Algorithm for Gaussian Clustering with Applications in Automatic speech Recognition'. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3. pp. 1599–1602.

Rissanen, J.: 1978, 'Modeling by Shortest Data Description'. *Automatica* **14**, 465–471.

Schwarz, G.: 1978, 'Estimating the Dimension of a Model'. *Annals of Statistics* **6**(2), 461–464.

Seide, F.: 1995, 'Fast Likelihood Computation for Continuous-Mixture Densities Using a Tree-Based Nearest Neighbor Search'. In: *Proceedings of the European Conference on Speech Communication and Technology*, Vol. 2. pp. 1079–1082.

Singer, E. and R. Lippmann: 1992, 'A Speech Recognizer Using Radial Basis Function Neural Networks in an HMM Framework'. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. pp. 629–632.

Takahashi, S. and S. Sagayama: 1995, 'Four-Level Tied-Structure for Efficient Representation of Acoustic Modeling'. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. pp. 520–523.

Wallace, C. and D. Boulton: 1968, 'An Information Measure for Classification'. *The Computer Journal* **11**(2), 195–209.

Wax, M. and T. Kailath: 1985, 'Detection of Signals by Information Theoretic Criteria'. *IEEE Transactions on ASSP* **33**, 387–392.

Young, S. et al.: 1999, *The HTK Book (for HTK Version 2.2)*. Entropic Ltd.

Young, S. and P. Woodland: 1993, 'The Use of State Tying in Continuous Speech Recognition'. In: *Proceedings of the European Conference on Speech Communication and Technology*, Vol. 3. pp. 2203–2206.