A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models

Chin-Hui Lee, *Member, IEEE*, Chih-Heng Lin, and Biing-Hwang Juang, *Senior Member, IEEE*

***Abstract*—It is generally agreed that, for a given speech recognition task, a speaker-dependent system usually outperforms a speaker-independent system, as long as a sufficient amount of training data is available. When the amount of speaker-specific training data is limited, however, such a performance gain is not guaranteed. One way to improve the performance is to make use of existing knowledge, contained in a rich speaker-independent (or multispeaker) data base, so that a minimum amount of training data is sufficient to model the new speaker. Such a training procedure is often referred to as speaker adaptation when *a priori* knowledge is derived from a speaker-independent (or multispeaker) data base; and as speaker conversion when the knowledge is derived from a different speaker. We mainly address the speaker adaptation issue here. For a speech recognition system based on continuous density hidden Markov models (CDHMM), speaker adaptation of the parameters of CDHMM is formulated as a Bayesian learning procedure. In this study we present a speaker adaptation procedure which is easily integrated into the segmental *k*-means training procedure for obtaining adaptive estimates of the CDHMM parameters. We report on some results for adapting both the mean and the diagonal covariance matrix of the Gaussian state observation densities of a CDHMM. When testing on a 39-word English alpha-digit vocabulary in isolated word mode, the results indicate that the speaker adaptation procedure achieves the same level of performance of a speaker-independent system, when one training token from each word is used to perform speaker adaptation. It also shows that much better performance is achieved when two or more training tokens are used for speaker adaptation. When compared with the speaker-dependent system, we found that the performance of speaker adaptation is always equal to or better than that of speaker-dependent training using the same amount of training data.**

I. Introduction

ADAPTIVE learning of the values of parameters of a speech model is of great interest for both theoretical and practical purposes. In the area of speech recognition, adaptive learning techniques have been applied to the problem of adapting reference speech patterns or models to handle situations not seen or dealt with during the training phase. This includes effects such as varying channel characteristics, changing environmental noise, and varying transducers. In this study we focus our attention on adaptive learning techniques for dealing with speaker mismatch problems. This type of mismatch arises when only a limited (insufficient) amount of training data is available from a particular speaker for creating speaker-specific speech patterns or models. Even though all the experiment setups discussed in this study are for speaker adaptation, the same formulation can also be used to handle varying channels, noise, and transducer mismatch problems.

Manuscript received March 30, 1989; revised May 7, 1990.

C.-H. Lee and B.-H. Juang are with the Speech Research Department, AT&T Bell Laboratories, Murray Hill, NJ 07974.

C.-H. Lin was with the Speech Research Department, AT&T Bell Laboratories, Murray Hill, NJ 07974 on leave from Telecommunication Laboratories, Chung-Li, Taiwan.

IEEE Log Number 9042264.

It is generally agreed that, for a given speech recognition task, a speaker-dependent (SD) system usually outperforms a speaker-independent (SI) system, as long as a sufficient amount of training data is available to obtain speaker-dependent models. When the amount of speaker-specific training data is limited, however, such a performance gain is not guaranteed. One way to improve the performance, under these conditions, is to make use of existing knowledge, contained in a data set from a rich multispeaker pool, so that a minimum amount of training data is sufficient to model the new speaker. Such a training procedure is often referred to as speaker adaptation (SA) when the prior knowledge is derived from a multispeaker (or speaker-independent) data base.

Speaker adaptation can be formulated in a number of ways, as illustrated in Fig. 1, including 1) *adaptive clustering*, in which an existing set of speaker-independent speech models is updated using a new set of speaker-independent data; 2) *speaker transformation* or *speaker conversion*, in which a well-trained model for one speaker is converted into a model for a new speaker using a small amount of speaker-specific training data; 3) *speaker adaptation*, in which a speaker-independent (or multispeaker) model is adapted to a single speaker using speaker-specific training data from the new talker; and 4) *sequential adaptation*, in which speaker-specific training data are acquired over time, and the speaker-dependent model is adapted sequentially every time that new training data is acquired. These implementations differ only in the ways in which the training data are utilized; the adaptation techniques involved are usually very similar.

A number of speaker adaptation techniques have been described in the literature [1]–[6]. The specific adaptation techniques employed often depend on the way speech patterns are modeled in the speech recognizer. For a template-based recognizer, the adaptation procedure usually involves addition or modification of a new template. For a recognizer based on vector quantization, codebook adaptation techniques are used (e.g., [1]–[3]). For a feature-based recognizer, the adaptation is performed on the feature parameters [4]. For systems based on discrete hidden Markov models (HMM), adaptation often involves modification of the discrete observation distribution (i.e., histogram adaptation) [5]. For model-based recognizers using continuous density HMM (CDHMM), a Bayesian adaptation training procedure has been applied [6] with good success.

Here we focus our attention on model-based recognizers. In particular, the stochastic model we use is the continuous density HMM. Under some reasonable regularity conditions, speaker adaptation of the CDHMM parameters can usually be formulated as a Bayesian learning procedure. In this paper, we present a Bayesian learning algorithm, which is easily integrated into the segmental *k*-means training procedure [7], for obtaining adaptive estimates of the CDHMM parameters.

1053-587X/91/0400-0806$01.00 © 1991 IEEE

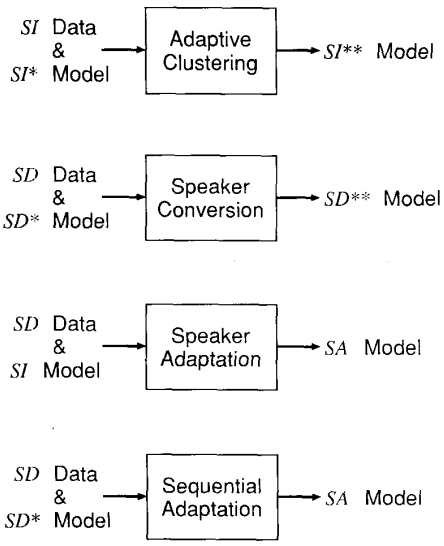

Fig. 1. Block diagrams of four different speaker adaptation setups.

We report on some recognition results for adapting the mean and the diagonal covariance vectors of the Gaussian state observation distribution of a CDHMM. When tested on a 39-word English alpha-digit vocabulary in an isolated word mode, the proposed speaker adaptation procedure, using only one training token from each word for adaptation, achieved a better performance than a speaker-independent system. The performance was further improved when two or more training tokens were incorporated for adaptation. When compared with a fully trained speaker-dependent system, it was found that the performance with speaker adaptive training was always equal to or better than that with speaker-dependent training when the same set of training tokens was used. This shows that the proposed adaptation procedure utilizes training data more effectively than the speaker-dependent training procedure when an equal amount of training data is used for training.

This paper is organized as follows. In Section II, we present an adaptive segmental k-mean algorithm for estimating the parameters of a CDHMM. In Section III, we discuss three specific implementations for adaptive estimation of the parameters of a CDHMM with Gaussian observation distributions. We then apply the algorithm to an English alpha-digit data base. The experiment arrangements and the recognition results are discussed in Section IV. Finally, we summarize our findings in Section V.

II. Adaptive Estimation of CDHMM Parameters

As was mentioned above, our approach to the problem of speaker adaptation is based upon a Bayesian framework. The difference between a maximum likelihood estimation procedure and a Bayesian learning procedure lies in the assumption of an appropriate prior distribution of the parameters to be estimated. Let $\boldsymbol{Y} = \{ y_1, y_2, \cdots, y_T\}$ be a given sequence of observations with a probability distribution function (pdf) $P(\boldsymbol{Y})$, and λ be the parameter set defining the distribution. Given a sequence of training data $\boldsymbol{Y}$, we want to estimate λ. If λ is assumed to be fixed but unknown, the maximum likelihood estimate (MLE) for λ is obtained by solving the likelihood equation, i.e.,

$$\frac{\partial}{\partial \lambda} P(y_1, y_2, \cdots, y_T|\lambda) = 0. \tag{2.1}$$

If λ is assumed random with a priori distribution function $P_0(\lambda)$, then the maximum *a posteriori* (MAP) estimate for λ is obtained by solving

$$\frac{\partial}{\partial \lambda} P(\lambda|y_1, y_2, \cdots, y_T) = 0. \tag{2.2}$$

Using Bayes theorem, we rewrite $P(\lambda|\boldsymbol{Y})$ as

$$P(\lambda|y_1, y_2, \cdots, y_T) = \frac{P(y_1, y_2, \cdots, y_T|\lambda)\, P_0(\lambda)}{P(y_1, y_2, \cdots, y_T)}. \tag{2.3}$$

The optimization criterion in (2.2) thus involves a prior distribution function $P_0(\lambda)$ for the random parameter λ. In most cases of interest, the MAP estimate λ_{MAP} that satisfies (2.2) attains minimum Bayes risk.

A. An Adaptive Segmental MAP Algorithm

Consider an N-state first-order Markov chain with transition probability matrix $A = [a_{ij}]$, $i, j = 1, 2, \cdots, N$. Let $s = (s_0, s_1, \cdots, s_T)$ be a state sequence where $s_t \in \{1, 2, \cdots, N\}$. Clearly, with the initial state probability denoted by $\pi' = [\pi_1, \pi_2, \cdots, \pi_N]$, the probability of observing the state sequence s is simply

$$P(s|\pi, A) = \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t}. \tag{2.4}$$

Each state i of the Markov chain is associated with a stochastic process characterized by a probability distribution function b_i for measurements y_t observed in state i. (Note that since observations are made within a finite period of time, y_t is, in fact, a short section of the stochastic process.) The triplet π, A, and $B = \{b_i\}_{i=1}^{N}$ thus defines a hidden Markov model and is denoted by $\lambda = (\pi, A, B)$. The joint probability for observing the sequence $\boldsymbol{Y}$ along with a state sequence s can be evaluated as

$$P(\boldsymbol{Y}, s|\lambda) = \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} b_{s_t}(y_t). \tag{2.5}$$

The probability for observing the sequence $\boldsymbol{Y}$ is then measured by

$$P(\boldsymbol{Y}|\lambda) = \sum_{\{s\}} P(\boldsymbol{Y}, s|\lambda) \tag{2.6}$$

where the summation is taken over all possible state sequences.

Depending on the choice of the optimization criterion, there are several ways of estimating the model parameter λ. In this paper, we consider maximization of the state-optimized likelihood of the observation sequences in an iterative manner as

$$\hat{\lambda} = \underset{\lambda}{\text{argmax}} \left[\max_{s} P(\boldsymbol{Y}, s|\lambda)\right]. \tag{2.7}$$

To accomplish this objective, the well-studied segmental k-means algorithm [7] can be used. Successful applications of the algorithm to speech recognition problems have been widely reported [8].

In extending the segmental k-means algorithm to the case of adaptive learning under the Bayesian framework, some algorithm revisions are necessary. The development of the MAP

estimate now involves the state sequence s. We obtain the MAP estimate by solving

$$\frac{\partial}{\partial \lambda} P(\lambda, s | Y) = 0. \tag{2.8}$$

Again by Bayes theorem, the above joint probability can be evaluated as

$$P(\lambda, s | Y) = \frac{P(Y, s | \lambda) P_0(\lambda)}{P(Y)} \tag{2.9}$$

where $P_0(\lambda)$ is the prior distribution of the parameter λ. The segmental k-means algorithm with embedded Bayesian adaptation thus consists of the following two steps:

1) Obtain the optimal state segmentation of Y, based on a given model $\hat{\lambda}$, i.e.,

$$\hat{s} = \underset{s}{\text{argmax}}\, P(Y, s | \hat{\lambda}) P_0(\hat{\lambda}). \tag{2.10}$$

2) Based on a state sequence $\hat{s}$, find the MAP estimate

$$\hat{\lambda} = \underset{\lambda}{\text{argmax}}\, P(Y, \hat{s} | \lambda) P_0(\lambda). \tag{2.11}$$

These two steps are iterated until some fixed point solution is reached.

The parameter optimization procedure involved in (2.11) is relatively simple because of (2.10) which allows adaptation of parameters in each individual state without interference from other states of the Markov model. The development of the specific adaptation mechanism for CDHMM with Gaussian state observation distribution is given in Section III.

The adaptive segmental k-means algorithm uses the same prior distribution $P_0(\lambda)$ in (2.10) and (2.11) on the same data Y in all iterations. So far, we have only discussed the cases where the training data are processed as a batch. We can also arrange the data in separate smaller batches and process them one at a time. This adaptation procedure is often called sequential adaptation as previously mentioned. We will not address the sequential adaptation procedure in this study.

It should be noted that the formulation we presented here about the use of prior distributions offers an added flexibility in the estimation of HMM parameters especially in the framework of segmental k-means training algorithm.

B. The Choice of Prior Distributions

The prior distribution characterizes the statistics of the parameters of interest before any measurement was made. It can be used to impose constraints on the values of the parameters. If the parameter is fixed but unknown and is to be estimated from the data, then there is no preference to what the value of the parameter should be. In such a case, the prior distribution $P_0(\lambda)$ is often called a noninformative prior which is a constant for the entire parameter space. The MAP estimate obtained by solving (2.2) is therefore equal to the MLE obtained by solving (2.1).

If we have knowledge about the parameters to be estimated, we can incorporate such prior knowledge into the prior distribution. Such a prior is often called an informative prior. In general, the choice of prior distribution depends on the acoustic models used to characterize the data. The choice is made based on 1) previous experience, 2) physical significance of the data, or 3) mathematical attractiveness. In this study we focus our attention on the use of conjugate priors [9]. A conjugate prior for a random vector is defined as the prior distribution for the parameters of the pdf of the random vector, such that the state-specific posterior distribution $\hat{P}(\lambda | Y)$ derived from (2.9) and the state-specific prior distribution $P_0(\lambda)$ belong to the same distribution family for any sample size n and any values of the observation samples. Since the use of conjugate priors has been studied extensively in the statistical literature (e.g., [9]), analytical forms of a number of conjugate priors for parameters of some of the most useful distributions are readily available. For example, the conjugate prior for the mean of Gaussian density is well known to be a Gaussian density. Due to this mathematical attractiveness, we select our prior distributions based on the concept of conjugate priors. Our conjecture is that the adaptation results will not be much different so long as a reasonably good informative prior is used.

III. Bayesian Adaptation of the Parameters of Gaussian Distributions

In the following, we formulate the specific Bayesian adaptation mechanism for adaptive estimation of the parameters of a CDHMM with Gaussian state observation distributions. To simplify our discussion, we assume the HMM used throughout the rest of the paper is a left-to-right HMM [10], such that the state sequence always starts from the first state and progresses from left to right in an increasing state order. An example of a 5-state left-to-right HMM is shown in Fig. 2. To reduce the number of parameters being adapted, we also assume the transition matrix for each HMM is fixed and known. Therefore, the estimation problem reduces to the estimation of the mean and the covariance matrix of the Gaussian distribution. In this study we consider only a Gaussian distribution with a diagonal covariance matrix. Without loss of generality, all the formulations derived in the following deal with only one component of the random vector within the same state of the Markov chain.

Let μ and σ^2 be the mean and the variance parameters of one component of a state observation distribution. Bayesian adaptation can then be formulated for either the mean μ or the variance σ^2. Adaptive learning can also be formulated for both the mean and the precision $\theta = 1/\sigma^2$ if the joint prior distribution of the parameters is specified. We now discuss three specific adaptation implementations.

A. Bayesian Adaptation of the Gaussian Mean

Assume the mean μ is random with a prior distribution $P_0(\mu)$, and the variance σ^2 is known and fixed. It can be shown that the conjugate prior for μ is also Gaussian with mean ν and variance τ^2 [11]. If we use the conjugate prior for the mean to perform Bayesian adaptation, then the MAP estimate for the parameter μ is solved by [11]

$$\hat{\mu}_{\text{MAP}} = \frac{n\tau^2}{\sigma^2 + n\tau^2} \bar{y} + \frac{\sigma^2}{\sigma^2 + n\tau^2} \nu \tag{3.1}$$

where n is the total number of training samples observed in the corresponding HMM state, and $\bar{y}$ is the sample mean. It is noted that the MAP estimate of the mean μ is a weighted average of the prior mean ν and the sample mean $\bar{y}$. The weights are functions of the parameters. When n is equal to zero, i.e., no additional training data are used, the estimate is simply the prior mean. When a large number of training samples are used (i.e., $n \to \infty$), the MAP estimate in (3.1) converges to the MLE (i.e., $\bar{y}$) asymptotically. It is also noted that if the value of the

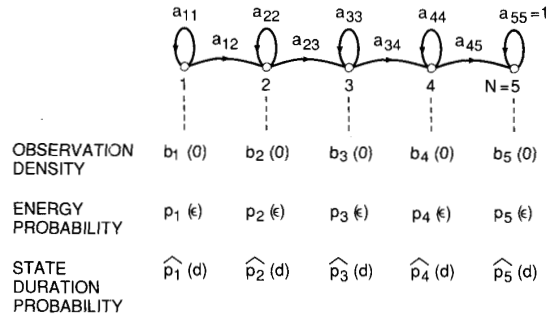

Fig. 2. A 5-state left-to-right hidden Markov model.

prior variance τ^2 is chosen to be relatively large, e.g., τ^2 is much larger than σ^2/n, then the MAP estimate obtained is (3.1) is approximately equal to the MLE, $\bar{y}$, which corresponds to the case of using noninformative priors.

We have assumed that the values of the prior parameters, ν, τ^2, and σ^2 are known in (3.1). In practice, these prior parameters have to be estimated from a collection of speaker-dependent (or multispeaker) models, or from a single speaker-independent model with a number of mixtures per state. For example, the mean and the variance of the prior distribution, can be estimated by the weighted mean and weighted variance of the following form:

$$\nu = \sum_{m=1}^{M} w_m \nu_m \tag{3.2}$$

and

$$\tau^2 = \sum_{m=1}^{M} w_m (\nu_m - \nu)^2 \tag{3.3}$$

where w_m is the weight assigned to the mth model (or mixture component), ν_m is the mean of the mth model (or mixture component), respectively. In the case of using a given speaker-independent Gaussian mixture model to estimate the parameters of the prior distribution, the weight w_m used in (3.2) and (3.3) is basically the mixture gain for the mth mixture component, and the estimates obtained in (3.2) and (3.3) can be interpreted as the MLE's of the mean and variance parameters of the random variable μ before any speaker-specific training data were observed. As for the fixed state observation variance σ^2 it can be estimated by a state-independent minimum variance $\sigma^2_{\min}$ [12], or by a state-independent grand variance [13]. We can also estimate the variance σ^2 by a fixed state-independent variance σ^2_{fix}, using a weighted variance of the following form

$$\sigma^2 = \sum_{m=1}^{M} w_m \sigma_m^2 \tag{3.4}$$

where σ_m^2 is the variance of the mth model (or mixture component).

The state-specific posterior distribution $P(\mu | \boldsymbol{Y})$ is also Gaussian, with mean $\hat{\nu} = \hat{\mu}_{\text{MAP}}$ as shown in (3.1) and variance of the form

$$\hat{\tau}^2 = \frac{\sigma^2}{\sigma^2 + n\tau^2}\,\tau^2 \tag{3.5}$$

which is always no greater than the prior variance τ^2. As the sample size n increases, the posterior distribution is more concentrated around the sample mean (i.e., $\hat{\tau}^2 \to \sigma^2/n \approx 0$). If sequential adaptation is desired, we can replace the prior distribution by the posterior distribution, i.e., substitute ν by $\hat{\nu}$ as in (3.1) and τ^2 by $\hat{\tau}^2$ as in (3.5), and proceed the same way as in the case of batch adaptation when additional training data are ready to be incorporated.

The extension of the adaptation procedure to vector case is trivial when the covariance matrix is known. In the current study, we restrict ourselves to the case of uncorrelated vector input, and therefore each component of the parameter vector can be adapted independently.

B. Bayesian Adaptation of the Gaussian Variance

Variance adaptation can be accomplished by assuming that the mean parameter μ is fixed but unknown, and the a priori distribution for the variance parameter σ^2 is an informative prior $P_0(\sigma^2)$. In this study, we use the following prior distribution

$$P_0(\sigma^2) = \begin{cases} \text{constant} & \text{if } \sigma^2 \geq \sigma^2_{\min} \\ 0 & \text{otherwise} \end{cases} \tag{3.6}$$

where $\sigma^2_{\min}$ is estimated from a large collection of speech data [11]. The mean parameter μ is estimated by the sample mean $\bar{y}$, since no prior knowledge about the mean parameter is assumed (noninformative prior). For the variance parameter σ^2, the MAP estimate is solved by

$$\max_{\sigma^2} \left\{ -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \bar{y})^2 \right\}. \tag{3.7}$$

Let S_y^2 be the sample variance computed from the sample data and define the following function:

$$q(\sigma^2) = \log \sigma^2 + \frac{S_y^2}{\sigma^2}. \tag{3.8}$$

Solving (3.7) is then equivalent to solving

$$\max_{\sigma^2} \left\{ -\frac{n}{2} q(\sigma^2) \right\}. \tag{3.9}$$

If $S_y^2 \geq \sigma^2_{min}$, then $\hat{\sigma}^2_{\text{MAP}} = S_y^2$. If $S_y^2 \leq \sigma^2_{\min}$, then for any estimate $\hat{\sigma}^2 \geq \sigma^2_{\min}$, consider the difference

$$\begin{aligned} \Delta &= -\frac{n}{2} \left\{ q(\sigma^2_{\min}) - q(\hat{\sigma}^2) \right\} \\ &= -\frac{n}{2} \left\{ \frac{S_y^2}{\sigma^2_{\min}} (1 - \sigma^2_{\min}/\hat{\sigma}^2) + \log (\sigma^2_{\min}/\hat{\sigma}^2) \right\} \\ &\geq -\frac{n}{2} \left\{ (1 - u) + \log u \right\} \end{aligned} \tag{3.10}$$

where u is the ratio $\sigma^2_{\min}/\hat{\sigma}^2$ which is always greater than zero and less than or equal to 1. The difference Δ is greater than or equal to zero, and the minimum difference (which is equal to zero) in (3.10) is attained only when $\hat{\sigma}^2$ is equal to $\sigma^2_{\min}$. Therefore, the MAP estimate for σ^2 is simply

$$\hat{\sigma}^2_{\text{MAP}} = \begin{cases} S_y^2 & \text{if } S_y^2 \geq \sigma^2_{\min} \\ \sigma^2_{\min} & \text{otherwise.} \end{cases} \tag{3.11}$$

Although the above result looks trivial, we found it most effective in cases where not enough samples were available for estimating the variance parameter. This adaptation procedure has

been applied successfully in estimating CDHMM parameters for both speaker-dependent and speaker-independent applications. The reader is referred to [12] for a detailed description of this variance adaptation and the effectiveness of this procedure.

Prior distributions other than the one in (3.6) can also be used. For example, the conjugate prior for the precision parameter $\theta = 1/\sigma^2$, which has a Gamma distribution [9], can be incorporated to obtain adaptive estimate of the variance parameter. The conjugate prior formulation is similar to the one for adaptation of both the mean and the precision parameters which we will be discussing in the following.

C. Bayesian Adaptation of Both the Gaussian Mean and Precision

Consider the case in which both the mean and the precision parameters are assumed to be random. It can be shown [9] that the joint conjugate prior $P_0(\mu, \theta)$ is a normal-gamma distribution, defined as follows: the conditional distribution of μ given θ is a normal distribution with mean ν and variance $\tau^2 = 1/\omega\theta$, and the marginal distribution of θ is a gamma distribution with parameters $\alpha > 0$ and $\beta > 0$, i.e.,

$$P_0(\mu, \theta) = \frac{\sqrt{\omega\theta}}{\sqrt{2\pi}} \exp\left[-\frac{\omega\theta}{2}(\mu - \nu)^2\right] \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta). \tag{3.12}$$

The joint posterior distribution is also a normal-gamma distribution, such that the conditional distribution of μ, given θ and the sample data, is a normal distribution with mean $\hat{\nu}$ and variance $\hat{\tau}^2$ such that

$$\hat{\nu} = \frac{\omega\nu + n\bar{y}}{\omega + n} \tag{3.13}$$

and

$$\hat{\tau}^2 = \frac{1}{(\omega + n)\theta} \tag{3.14}$$

and the marginal distribution of θ given the sample data is also a gamma distribution with parameter $\hat{\alpha}$ and $\hat{\beta}$, such that

$$\hat{\alpha} = \alpha + n/2 \tag{3.15}$$

and

$$\hat{\beta} = \beta + \frac{n}{2} S_y^2 + \frac{n\omega(\bar{y} - \nu)^2}{2(\omega + n)}. \tag{3.16}$$

It is noted that there is no joint distribution in the normal-gamma family such that μ has a normal distribution, θ has a gamma distribution, and μ and θ are independent. Even if the prior distribution is specified so that μ and θ are independent, their posterior distribution would specify that they are dependent after a single value has been observed. The marginal prior and posterior distributions of μ can be shown to have a t distribution [9]. For our purposes, we are more interested in obtaining the joint MAP estimate of μ and σ^2, which can be derived as [9]

$$\hat{\mu}_{\text{MAP}} = \hat{\nu} = \frac{\omega\nu + n\bar{y}}{\omega + n} \tag{3.17}$$

and

$$\hat{\sigma}^2_{\text{MAP}} = \hat{\beta}/\hat{\alpha}. \tag{3.18}$$

The prior parameters ν, ω, α, and β, needed for evaluating (3.12)–(3.18), can either be assigned arbitrarily or be estimated from a richer data base or from a set of existing models. In our study, we simply use the set of mixture components in our speaker independent model to estimate the two sets of prior parameters as follows:

$$\nu = \sum_{m=1}^{M} w_m \nu_m \tag{3.19}$$

$$\alpha = 1/\sigma^2 = \frac{1}{\sum_{m=1}^{M} w_m \sigma_m^2} \tag{3.20}$$

$$\omega = \sigma^2/\tau^2 = \frac{1}{\alpha \sum_{m=1}^{M} w_m (\nu_m - \nu)^2} \tag{3.21}$$

$$\beta = 1 \tag{3.22}$$

where ν is estimated as the weighted average of means over a set of mixture components (same as (3.2)), α is estimated as the precision parameter such that the corresponding variance estimate is obtained from the weighted average of the variance parameters σ_m^2 over a set of mixture components, β is fixed to be one, and ω is estimated as the ratio of the variance estimate $(1/\alpha)$ and the variance of the mean parameter obtained as the weighted average of squared deviations from the estimated mean parameter ν over a set of mixture components.

For readers interested in applying the same Bayesian adaptation procedure to a multivariate Gaussian density with a full covariance matrix, a vector version of the conjugate prior distribution for the mean vector and the covariance matrix of a multivariate Gaussian density is also available in [9, ch. 9]. It can be shown that both the joint conjugate prior and posterior densities are from a normal-Wishart distribution family. The reader is referred to [9] for a thorough discussion of the formulations and applications.

IV. Experimental Setup and Recognition Results

To study the effect of speaker adaptation on recognition performance, we choose a vocabulary of 39 words consisting of the 26 letters of the English alphabet (A–Z), 3 command words (stop, error, and repeat) and the ten English digits (0–9) for all recognition experiments. Two data sets are needed to set up speaker adaptation experiments. One is a rich multispeaker data base to train the speaker-independent models for estimating the parameters of the prior distributions needed for Bayesian adaptation. The other is a speaker-independent set consisting of one session of speaker-specific training data for adaptation and the other session for testing. The multispeaker training set we used consists of one occurrence of each of the 39 words, uttered by 100 talkers (50 females and 50 males). The speaker-dependent set consists of utterances from four talkers (2 females and 2 males). For speaker adaptive training, we used 5 training utterances per word from each of the male talkers and 7 utterances per word from each of the female talkers, respectively. For testing, we used 10 utterances per word per speaker which gives a total of 390 testing utterances for each speaker. All of the data were recorded over local dialed-up telephone lines, band-pass filtered, and digitized at a sampling rate of 6.67 kHz. The two data sets were collected at a difference time and the recording environments as well as the channel conditions were quite dif-

ferent. A number of experiments have been conducted on these two data sets. The reader is referred to [14] for a summary of some recognition performance benchmarks.

Throughout this study, each of the 39 words in the vocabulary is modeled by a single left-to-right 5-state HMM. The feature extraction performed in both training and testing is an eighth-order linear prediction analysis with a 45-ms Hamming window and a 15-ms frame shift. The feature used is a vector of 24 elements consisting of 12 bandpass-liftered cepstral coefficients [15] and 12 corresponding cepstral time derivatives [16].

The speaker-independent word model was obtained using the segmental k-means training procedure [6]. The observation distribution for each state of the HMM was modeled by a multivariate Gaussian mixture distribution [17], in which the maximum number of mixture components in each state is limited to 9, and each of the mixture components has a diagonal covariance matrix.

The speaker-dependent word models used in all recognition experiments were also 5-state HMM (the same topology as the SI model). However, all the state transition probabilities were fixed to be 0.5 for all experiments (no performance difference was observed when the estimated transition matrix was used in a preliminary experiment). The state observation distribution was assumed to be a multivariate Gaussian distribution with a diagonal covariance matrix.

To set up a baseline speaker-independent performance on the speaker-dependent data set, we first conducted two sets of experiments using two different speaker-independent models (before any speaker-specific training data were incorporated). The first set used the Gaussian mixture SI model described above, and the second set used a single Gaussian distribution for every state of the HMM, by combining the mixture components in a state. In this study, the Gaussian distribution corresponding to each state is specified by choosing the mean to be the mean of the corresponding Gaussian mixture distribution, i.e.,

$$\mu = \sum_{m=1}^{M} w_m \nu_m \tag{4.1}$$

and the variance to be the variance of the corresponding Gaussian mixture distribution, i.e.,

$$\sigma^2 = \sum_{m=1}^{M} w_m \sigma_m^2. \tag{4.2}$$

In the above, w_m is the mixture gain, ν_m is the mean, and σ_m^2 is the variance of the mth mixture component, and M is the actual number of mixture components used in each state.

The word recognition rates for each of the four test speakers (F1, F2, M1, and M2), for each experiment, along with the average rates over the four speakers are listed in Table I. For reference purposes, we also listed, in the row labeled "SD," the results obtained using a fully trained speaker-dependent model, which used two Gaussian mixture components per state per word [17]. This fully trained model gave the best average performance and it provided a performance upper bound for this speaker-dependent data base. It is noted, in the SI testing, that the average performance with 9 Gaussian mixture components is significantly better than that using only a Gaussian distribution. However, for the second male talker M2, the single Gaussian distribution case did slightly better in performance. The average performance for using the Gaussian mixture distribution is also much worse than the average performance of 90% reported in [17] when tested on a different SI data base (of the same vocabulary) using a model of 10 states per word and 9 mixture components per state. This is because of a serious mismatch in channel and recording conditions between the SI training set and the speaker-dependent testing set. We will show, in the following, that speaker adaptation can be used to reduce this mismatch and to improve the performance even if we use only one single Gaussian density per state to characterize the state observation density.

TABLE I
RECOGNITION RESULTS USING SPEAKER-INDEPENDENT MODELS AND A FULLY TRAINED SPEAKER-DEPENDENT MODEL

Setup	State Density	F1	F2	M1	M2	AVG
SI	9-Mixture	81.0	79.7	78.5	84.9	81.5
SI	Gaussian	74.9	72.6	66.9	86.2	75.1
SD	2-Mixture	96.9	94.1	98.2	98.2	96.9

TABLE II
SUMMARY OF ADAPTATION RESULTS FOR SPEAKER M1

tokens	EXP1	EXP2	EXP3	EXP4	EXP5
1	62.05	86.41	90.00	86.41	98.46
2	89.23	95.64	95.64	94.36	96.41
3	91.28	95.64	96.92	96.15	97.69
4	94.10	96.41	97.18	97.69	97.18
5	95.90	97.18	97.44	96.67	98.46

TABLE III
SUMMARY OF ADAPTATION RESULTS FOR SPEAKER M2

tokens	EXP1	EXP2	EXP3	EXP4	EXP5
1	56.67	93.85	95.13	95.13	95.13
2	87.44	96.15	96.41	95.90	96.15
3	93.59	95.38	96.15	96.67	96.15
4	95.13	97.69	97.18	97.18	96.92
5	95.64	96.92	97.44	97.69	96.92

To examine the effect of the various mean adaptation schemes, five sets of recognition experiments were conducted. The word recognition rates, for each individual talker, are listed in Tables II-V. The five experimental setups were:

- EXP1: using an SD mean and an SD variance (regular MLE),
- EXP2: using an SD mean and a fixed variance estimate (3.4),
- EXP3: using an SA mean (3.1) with prior parameters (3.2) and (3.3) and a fixed variance (3.4),
- EXP4: using an SD mean and an SA variance (3.11),
- EXP5: using SA mean and variance (3.17) and (3.18), with prior parameters (3.19)–(3.22).

The rows in Tables II–V correspond to the number of training tokens used for either speaker-dependent or speaker-adaptive cases. In all other training setups, up to 5 or 7 tokens were incorporated.

The results in Tables II–V clearly show that the regular MLE training procedure (EXP1) is inadequate when the amount of available training data is limited. The performance for EXP1 improves as the number of training tokens increases. The problem with EXP1 is that the variance cannot be properly estimated with only a small number of samples. This mismatch between

TABLE IV
SUMMARY OF ADAPTATION RESULTS FOR SPEAKER F1

tokens	EXP1	EXP2	EXP3	EXP4	EXP5
1	42.56	85.90	86.92	81.54	86.41
2	78.72	89.49	90.51	89.49	90.00
3	83.85	92.82	94.87	92.56	93.08
4	89.49	94.36	93.33	92.82	93.59
5	89.74	94.36	94.87	92.82	93.85
6	90.26	93.33	94.36	92.82	93.85
7	93.08	95.13	94.87	94.10	94.61

TABLE V
SUMMARY OF ADAPTATION RESULTS FOR SPEAKER F2

tokens	EXP1	EXP2	EXP3	EXP4	EXP5
1	40.77	75.38	81.79	77.44	82.05
2	70.00	85.38	87.95	84.10	88.21
3	79.49	86.67	87.69	88.47	87.95
4	84.10	88.46	90.00	89.74	89.74
5	89.49	91.03	93.08	91.28	93.59
6	90.26	92.05	93.33	91.79	93.59
7	90.26	92.56	93.59	91.53	94.61

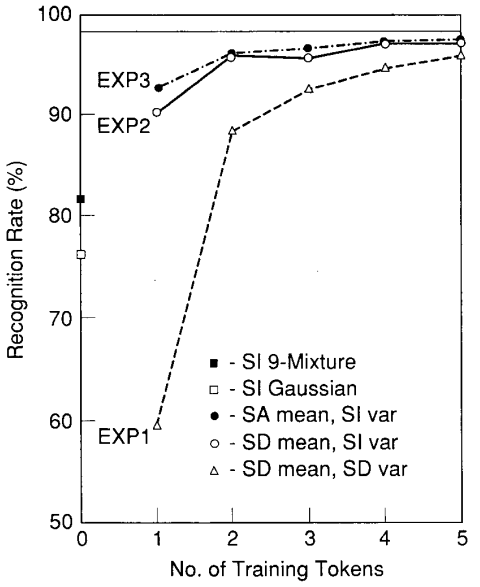

Fig. 3. Mean adaptation performance for male speakers.

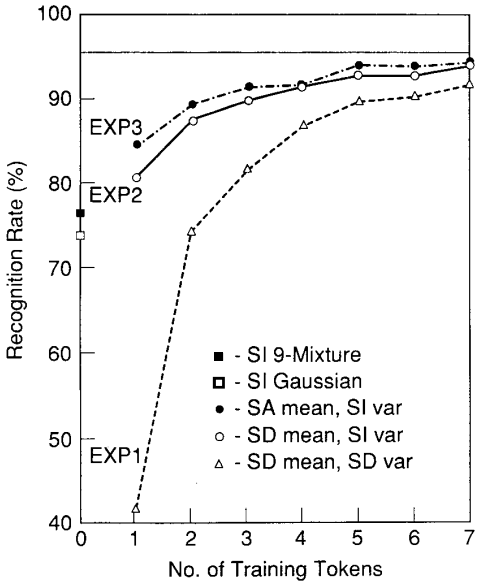

Fig. 4. Mean adaptation performance for female speakers.

the training and testing data caused a great deal of performance degradation. The performance can be improved by simply replacing the speaker-dependent variance with an appropriate SI variance estimate (EXP2). The best results, among the first four set of experiments, were achieved when a speaker adaptive mean was used (EXP3). The performance improvement from EXP1 to EXP3 is more significant when the number of training tokens used is very small. For variance adaptation using (3.11), the results are listed in the columns labeled EXP4. We found, in our particular setup, that EXP2 achieves better results than that obtained in EXP4. The best results were obtained using adaptive training for both the mean and the variance parameters (EXP5). An average word recognition rate of 96.1% over all four talkers was achieved using the MAP estimates of (3.17) and (3.18).

To summarize the performance of mean adaptation for both male and female talkers, we plot, in Figs. 3 and 4, the average word recognition rates versus the number of training tokens for EXP1, 2, and 3. The average performance shown in Table I, using a fully trained speaker-dependent model, is plotted as a horizontal straight line indicating a performance upper bound. The average word recognition rates for the two SI experiments shown in Table I are also plotted in the figures for comparison. From the performance curves, it is clear that the speaker-dependent performance is not as good as the speaker-independent performance when the number of training tokens is limited. However, when one additional training token is available for speaker adaptation, the speaker-adaptive models always outperform the speaker-independent models. Much better performance can also be achieved when more training tokens are incorporated in adaptation. It is noted that, when using the same amount of training data, speaker-adaptive training outperforms speaker-dependent training in all cases tested. This implies that speaker-adaptive training utilizes training data more effectively than speaker-dependent training, especially in cases of insufficient training data. As expected, the speaker-adaptive performance quickly becomes equivalent to the speaker-dependent performance when the number of training tokens increases. In our experimental setup we do not have enough training data and the asymptotic performance is still not achieved with speaker-dependent training. All the results obtained using the three adaptive training schemes, namely EXP3, EXP4, and EXP5, are summarized in Figs. 5 and 6. In Fig. 5, we plot the average word recognition rates versus the number of training (or adaptation) tokens for male talkers. In Fig. 6 we plot the average word recognition rates for female talkers. For all the three adaptation schemes evaluated, the performance differences were not very significant when more training tokens were used in adaptation. However, for the case with very small amount of

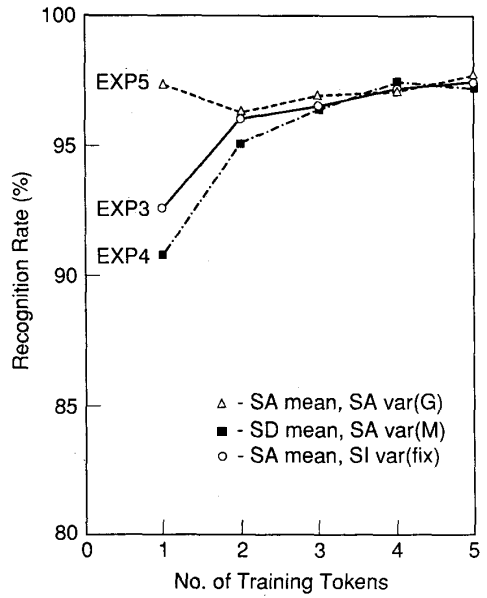

Fig. 5. The results of three adaptation schemes for male speakers.

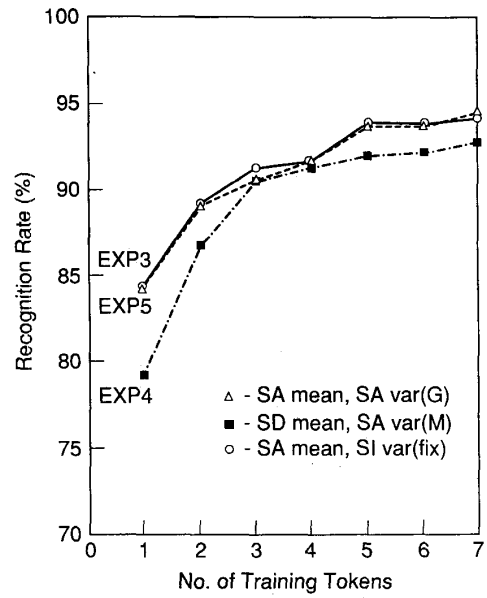

Fig. 6. The results of three adaptation schemes for female speakers.

data, EXP3 and EXP5 are more effective than EXP4. The effects of adaptation on recognition performance are clearly demonstrated when the adaptation results are compared with the results obtained from speaker-dependent training (i.e., EXP1).

As shown in Figs. 5 and 6, the average performance for male speakers is much better than that for female speakers. The rates of convergence in speaker adaptation are also shown in the performance curves in Figs. 5 and 6. It is noted that for male speakers the adaptation results using only two tokens per word is almost as good as the results obtained using all 5 training tokens. For female speakers, the adaptation rate seems to be slower than that for male speakers in that it takes about 5 tokens from each word for the average performance to saturate.

V. Summary

In this study we present a Bayesian framework for adaptive estimation of the parameters of the continuous density hidden Markov models. We show that the algorithm can be embedded in the segmental k-means training algorithm by replacing the MLE with the MAP estimate. We also present formulations for obtaining adaptive estimates of the parameters of the Gaussian observation distribution. The Bayesian adaptation procedure takes advantage of the conjugate prior formulation readily available in the statistical literature. Because of the use of sufficient statistics and informative prior information in the formulation, the Bayesian procedure uses training data more effectively compared to the conventional codebook mapping approaches. It can also be shown, both theoretically and in practice, that the Bayesian adaptive estimates converge to speaker-dependent estimates of the CDHMM parameters when the number of training tokens increase. This is a desirable property which is not easily attainable in the conventional codebook mapping approaches.

We applied the adaptive segmental k-mean algorithm to speaker adaptation problems using a 39-word English alphadigit vocabulary. When compared with recognition results obtained using speaker-independent models, the adaptive training procedure achieved better performance. It was also found that much better performance can be achieved when two or more training tokens were incorporated in adaptation. When all the training tokens were incorporated to obtain adaptive estimates for both the mean and the variance parameters, we achieved an average word recognition of 96.1%, which is the best performance reported using a single Gaussian distribution for every HMM state in the vocabulary. We also compared the adaptive training procedure with the speaker-dependent training procedure. The results indicate that adaptive training procedure is highly effective when a very limited amount of speaker-specific training data is available. It is interesting to note that our speaker-adaptive results are always equal to or better than that obtained using speaker-dependent models.

The Bayesian speaker adaptation techniques discussed in this paper can easily be applied to other adaptation problems such as noise, channel, and transducer adaptation. The results obtained in this study are very encouraging. A continuing studying on extending the Bayesian adaptation technique to speaker adaptation and context adaptation problems for large vocabulary speech recognition is under way.

References

[1] Y. Shiraki and M. Honda, "Speaker adaptation algorithms for segment vocoder," *IEICE*, vol. SP87-67, pp. 49–56, Oct. 1987 (in Japanese).

[2] K. Shikano, K.-F. Lee, and R. Reddy, "Speaker adaptation through vector quantization," in *Proc. ICASSP86* (Tokyo, Japan), Apr. 1986, pp. 2643–2646.

[3] S. Furui, "Unsupervised speaker adaptation method based on hierarchical spectral clustering," in *Proc. ICASSP89* (Glasgow, Scotland), May 1989, pp. 286–289.

[4] R. Schwartz, Y. L. Chow, and F. Kubala, "Rapid speaker adaptation using a probabilistic spectral mapping," in *Proc. ICASSP87* (Dallas, TX), Apr. 1987, pp. 633–636.

[5] R. M. Stern and M. J. Lasry, "Dynamic speaker adaptation for feature-based isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, no. 6, June 1987.

[6] P. F. Brown, C.-H. Lee, and J. C. Spohrer, "Bayesian adaptation in speech recognition," in *Proc. ICASSP83* (Boston, MA), Apr. 1983, pp. 761–764.
[7] L. R. Rabiner, J. G. Wilpon, and B.-H. Juang, "A segmental *k*-means training procedure for connected word recognition," *AT&T Tech. J.*, vol. 65, no. 3, pp. 21–32, May–June 1986.
[8] L. R. Rabiner, J. G. Wilpon, and F. K. Soong, "High performance connected digit recognition using hidden Markov models," in *Proc. ICASSP88* (New York, NY), Apr. 1988, pp. 119–122.
[9] M. H. DeGroot, *Optimal Statistical Decisions*. New York: McGraw-Hill, 1970.
[10] L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," *IEEE ASSP Mag.*, vol. 3, no. 1, pp. 4–16, Jan. 1986.
[11] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
[12] L. R. Rabiner, C.-H. Lee, B.-H. Juang, D. B. Roe, and J. G. Wilpon, "Improved training procedure for hidden Markov models," *J. Acoust. Soc. Amer.*, suppl. 1, vol. 84, S61, Fall 1988.
[13] R. P. Lippman, E. A. Martin, and D. B. Paul, "Multistyle training for robust isolated-word speech recognition," in *Proc. ICASSP87* (Dallas, TX), Apr. 1987, pp. 705–709.
[14] L. R. Rabiner and J. G. Wilpon, "Some performance benchmarks for isolated word recognition systems," *Comput., Speech, Language*, vol. 2, nos. 3/4, pp. 343–357, Dec. 1987.
[15] B.-H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, no. 7, pp. 947–954, July 1987.
[16] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, no. 6, pp. 871–879, June 1988.
[17] B.-H. Juang, "Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains," *AT&T Tech. J.*, vol. 64, no. 4, pp. 1235–1249, 1985.

Chin-Hui Lee (S'81–M'82) received the B.S. degree from National Taiwan University, Taipei, Taiwan, in 1973, the M.S. degree from Yale University, New Haven, CT, in 1977, and the Ph.D. degree from the University of Washington, Seattle, in 1981, all in electrical engineering.

In 1981, he joined Verbex Corporation, Bedford, MA, and was involved in research work on connected word recognition. In 1984, he became affiliated with Digital Sound Corporation, Santa Barbara, CA, where he engaged in research work in speech coding, speech recognition, and signal processing for the development of the DSC-2000 Voice Server. Since 1986, he has been with AT&T Bell Laboratories, Murray Hill, NJ. His current research interests include speech modeling, speech recognition, and signal processing.

Chih-Heng Lin was born in February 1955. He received the B.S. degree in control engineering from National Chiao-Tung University, Hsintsu, in 1979, and the M.S. degree in electrical engineering from Ohio University, Athens, in 1983. Currently he is pursuing the Ph.D. degree in electrical engineering at National Taiwan University, Taipei, Taiwan.

In 1983 he joined Telecommunication Laboratories, Taiwan, and was involved in research work on optical character recognition and speech recognition. In 1988 he visited AT&T Bell Laboratories, Murray Hill, NJ, for one year.

Biing-Hwang Juang (S'79–M'80–SM'87) received the B.Sc. degree in electrical engineering from the National Taiwan University, Taipei, in 1973 and the M.Sc. and Ph.D. degrees in electrical and computer engineering from the University of California, Santa Barbara, in 1979 and 1981, respectively.

In 1978, he joined the Speech Communications Research Laboratory, Santa Barbara, and was involved in research work on vocal tract modeling. In 1979, he became affiliated with Signal Technology, Inc., Santa Barbara, where his research work was in the areas of speech coding and speech interference suppression. Since 1982, he has been with AT&T Bell Laboratories, Murray Hill, NJ. His current research interests include speech recognition, coding, enhancement, and stochastic processes.

Dr. Juang is a member of the IEEE Signal Processing Technical Committees on DSP, Speech, and Audio and Acoustics. He was an Associate Editor of the IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING during 1986–1988 and served as the Technical Program Chairman of the 1986 ASSP Workshop on Applications of DSP to Audio and Acoustics.