A study on speaker adaptation of continous density HMM parameters

Conference Paper *in* Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on · May 1990

DOI: 10.1109/ICASSP.1990.115559 · Source: IEEE Xplore

3 authors, including:

Chin-Hui Lee
Georgia Institute of Technology
471 PUBLICATIONS **11,708** CITATIONS

SEE PROFILE

S3.4

A STUDY ON SPEAKER ADAPTATION OF CONTINUOUS DENSITY HMM PARAMETERS

Chin-Hui Lee, Chih-Heng Lin† and Biing-Hwang Juang

Speech Research Department
AT&T Bell Laboratories
Murray Hill, New Jersey 07974

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 data is available for training. 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 multi-speaker) database, so that a small amount of training data is sufficient to model the new speaker using speaker adaptive training. For a speech recognition system based on a continuous density hidden Markov model (CDHMM), we show that speaker adaptation of the parameters of CDHMM can be formulated as a Bayesian learning procedure and it can be integrated into the segmental *k*-means training algorithm. 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 tested on a 39-word English alpha-digit vocabulary in isolated word mode, the results indicate that the speaker adaptation procedure achieves better performance than that of a speaker-independent system, when only one training token from each word is used to perform speaker adaptation. It is also shown that much better performance can be achieved when two or more training tokens are used for speaker adaptation.

1. 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.

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 multi-speaker 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 multi-speaker (or speaker-independent) database.

A number of speaker adaptation techniques have been described in the literature [1-4]. The specific adaptation techniques employed often depend on the way speech patterns are modeled in the speech recognizer. For a VQ-based recognizer, codebook adaptation techniques are used (e.g., [1]). For a feature-based recognizer, the adaptation is performed on the feature parameters [2]. For systems based on discrete hidden Markov models (HMM), adaptation often involves modification of the discrete observation distribution (i.e. histogram adaptation) [3]. For model-based recognizers using continuous density HMM (CDHMM), a Bayesian adaptation training procedure has been used [4] with good success. In this paper, we present a Bayesian learning algorithm, which is easily integrated into the segmental *k*-means training procedure [5], for obtaining adaptive estimates of the CDHMM parameters.

† Chih-Heng Lin is now with Telecommunication Laboratories, Chung-Li, Taiwan

2. ADAPTIVE ESTIMATION OF CDHMM PARAMETERS

As 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 $\boldsymbol{Y}=\{y_1,y_2,\ldots,y_T\}$ 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. 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

$$\frac{\partial}{\partial \lambda} P(\lambda \mid \boldsymbol{Y}) = \frac{\partial}{\partial \lambda} \frac{P(\boldsymbol{Y} \mid \lambda) P_0(\lambda)}{P(\boldsymbol{Y})} = 0. \tag{2.1}$$

The optimization criterion in (2.1) 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 solves (2.1) attains minimum Bayes risk.

2.1 An Adaptive Segmental *K*-Means Algorithm

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 using the segmental *k*-means training algorithm [5]. 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 $\boldsymbol{s}$. We obtain the MAP estimate by solving

$$\frac{\partial}{\partial \lambda} P(\lambda, \boldsymbol{s} \mid \boldsymbol{Y}) = 0. \tag{2.2}$$

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

1. For a given model $\hat{\lambda}$, find the optimal state segmentation $\hat{\boldsymbol{s}}$

$$\hat{\boldsymbol{s}} = \underset{\boldsymbol{s}}{\operatorname{argmax}}\ P(\boldsymbol{Y}, \boldsymbol{s} \mid \hat{\lambda})\, P_0(\hat{\lambda}). \tag{2.3}$$

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

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}}\, P(\boldsymbol{Y}, \hat{\boldsymbol{s}} \mid \lambda)\, P_0(\lambda). \tag{2.4}$$

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

CH2847-2/90/0000-0145 $1.00 © 1990 IEEE

The parameter optimization procedure involved in (2.4) is relatively simple because of (2.3) which provides and 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 3.

2.2 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 *non-informative prior* which is a constant for the entire parameter space. The MAP estimate obtained by solving (2.1) is therefore equal to the MLE solution.

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* [6]. 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 $P(\lambda,s \mid Y)$ 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. Analytical forms of a number of conjugate priors for parameters of some of the most useful distributions are readily available in the statistical literature (e.g. [6]). Due to this mathematical attractiveness, we select our prior distributions based on the concept of conjugate priors.

3. BAYESIAN ADAPTATION OF GAUSSIAN PARAMETERS

In the following, we formulate the specific Bayesian adaptation mechanisms for adaptive estimation of the parameters of a CDHMM with Gaussian state observation distributions. To simplify our discussion, we use a left-to-right HMM throughout the rest of the paper. 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 diagonal covariance matrix of the Gaussian distribution. 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.

3.1 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. If we use the conjugate prior for the mean to perform Bayesian adaptation, then the MAP estimate for the parameter μ is simply

$$\hat{\mu}_{MAP} = \frac{n\tau^2}{\sigma^2+n\tau^2}\,\bar{y} + \frac{\sigma^2}{\sigma^2+n\tau^2}\,\nu \qquad (3.1)$$

where n is the total number of training samples observed in the corresponding HMM state, and $\bar{y}$ is the sample mean. When a large number of training samples are used, 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 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 non-informative 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 multi-speaker) 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

$$\nu = \sum_{m=1}^{M} w_m \nu_m, \quad \tau^2 = \sum_{m=1}^{M} w_m(\nu_m-\nu)^2 \qquad (3.2)$$

where w_m is the weight assigned to the m^{th} model (or mixture component), ν_m is the mean of the m^{th} 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) is basically the mixture gain for the m^{th} mixture component, and the estimates obtained in (3.2) 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, we use a weighted variance of the following form

$$\sigma^2 = \sum_{m=1}^{M} w_m {\sigma_m}^2 \qquad (3.3)$$

where ${\sigma_m}^2$ is the variance of the m^{th} model (or mixture component).

3.2 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

$$P_0(\sigma^2) = \begin{cases} \text{constant} & \text{if } \sigma^2 \geq {\sigma_{min}}^2 \\ 0 & \text{otherwise} \end{cases} \qquad (3.4)$$

where ${\sigma_{min}}^2$ is estimated from a large collection of speech data [7]. The mean parameter μ is estimated by the sample mean $\bar{y}$, since no prior knowledge about the mean parameter is assumed (non-informative prior). For the variance parameter σ^2, the MAP estimate is solved by

$$\hat{\sigma}_{MAP}^{\,2} = \begin{cases} {S_y}^2 & \text{if } {S_y}^2 \geq {\sigma_{min}}^2 \\ {\sigma_{min}}^2 & \text{otherwise} \end{cases} \qquad (3.5)$$

where ${S_y}^2$ is the sample variance. Although the above result looks trivial, we found it most effective in cases where not enough samples are 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 [7] for a detailed description of this variance adaptation and the effectiveness of this procedure.

Prior distributions other than the one in (3.5) can also be used. For example, the conjugate prior for the *precision parameter* $\theta=1/\sigma^2$, which is a Gamma distribution [6], 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.

3.3 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 [6] 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.,

$$P_0(\mu,\theta) = \frac{\sqrt{\omega\theta}}{\sqrt{2\pi}}\exp[-\frac{\omega\theta}{2}(\mu-\nu)^2]\,\frac{\beta^\alpha}{\Gamma(\alpha)}\,\theta^{\alpha-1}\exp(-\beta\theta) \qquad (3.6)$$

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 [6]. For our purposes, we are more interested in obtaining the joint MAP estimate of μ and σ^2, which can be derived as [6]

$$\hat{\mu}_{MAP} = \frac{\omega\nu + n\bar{y}}{\omega + n}, \quad \hat{\sigma}_{MAP}^2 = \frac{\beta + \frac{1}{2}S_y^2 + n\omega(\bar{y}-\nu)^2}{(\alpha + n/2)}. \tag{3.7}$$

The prior parameters ν, ω, α and β, needed for evaluating (3.7), can either be assigned arbitrarily or be estimated from a richer database 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:

$$\nu = \sum_{m=1}^{M} w_m \nu_m, \quad \alpha = 1/\sigma^2 = 1/[\sum_{m=1}^{M} w_m \sigma_m^2] \tag{3.8}$$

$$\omega = 1/[\alpha \sum_{m=1}^{M} w_m(\nu_m - \nu)^2], \quad \beta = 1. \tag{3.9}$$

4. EXPERIMENTAL SETUP AND RECOGNITION RESULTS

To study the effect of speaker adaptation on recognition performance, we chose 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 multi-speaker database to train the speaker-independent models for estimating the parameters of the prior distributions needed for Bayesian adaptation. The other is a speaker-dependent set consisting of one session of speaker-specific training data for adaptation and the other session for testing. The multi-speaker 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 different time and the recording environments as well as the channel conditions were quite different. A number of experiments have been conducted on these two data sets. The reader is referred to [8] for a summary of some recognition performance benchmarks.

The signal analysis used in this study is an 8^{th} order linear prediction analysis with a 45 msec Hamming window and a 15 msec frame shift. The feature used is a vector of 24 elements consisting of 12 bandpass-liftered cepstral coefficients and 12 corresponding cepstral time derivatives.

The speaker-independent word model was obtained using the segmental k-means training procedure [5]. The observation distribution for each state of the HMM was modeled by a multivariate Gaussian mixture distribution, 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) with Gaussian state observation distributions.

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 and the variance to be the mean the variance of the corresponding Gaussian mixture distribution respectively (as shown in (3.2) and (3.4)).

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 given below in Table 1. 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. This fully-trained model gave the best average performance and it provided a performance upper bound for this speaker-dependent database. 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 [8] when tested on a different SI database (of the same vocabulary) using a model with 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.

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 1. Recognition results using SI models and a fully-trained SD model

To examine the effect of the various mean adaptation schemes, five sets of recognition experiments were conducted. The average word recognition rates, over the two male and two female talkers, are given in Tables 2-3 respectively. The five experimental setups were:

EXP1 : SD mean and an SD variance (regular MLE)
EXP2 : SD mean and a fixed variance estimate (3.3)
EXP3 : SA mean (3.1) with prior parameters (3.2)-(3.3)
EXP4 : SD mean and an SA variance (3.5)
EXP5 : SA estimates (3.7), with prior parameters (3.8)-(3.9)

The rows in Tables 2-3 correspond to the number of training tokens used for either speaker-dependent or speaker-adaptive training. In all training setups, up to 5 or 7 tokens were used. The results in Tables 2-3 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 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 (3.5), 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 (3.7).

To summarize the performance of mean adaptation for the female talkers, we show, in Figure 1, the average word recognition rates versus the number of training tokens for EXP1, 2 and 3. The average performance shown in Table 1, 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 figure 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 don't have enough training data and the asymptotic performance is still not achieved with speaker-dependent training.

The average word recognition rates over the two male talkers, obtained using the three adaptive training schemes (as discussed in Sections 3.1-3.3, and tested in EXP3, EXP4 and EXP5), are shown in Figure 2.

tokens	EXP1	EXP2	EXP3	EXP4	EXP5
1	59.4	90.1	92.6	90.8	96.8
2	88.3	95.9	96.0	95.1	96.3
3	92.4	96.5	96.5	96.4	96.9
4	94.6	97.1	97.2	97.4	97.1
5	95.8	97.1	97.4	97.2	97.7

Table 2. Summary of adaptation results for male talkers

tokens	EXP1	EXP2	EXP3	EXP4	EXP5
1	41.7	80.6	84.3	79.5	84.2
2	74.4	87.4	89.2	86.8	89.1
3	81.7	89.7	91.3	90.5	90.5
4	86.9	92.4	91.7	91.3	91.7
5	89.6	92.7	94.0	92.1	93.7
6	90.3	93.7	93.8	92.3	93.7
7	91.7	93.8	94.2	92.8	94.6

Table 3. Summary of adaptation results for female talkers

5. 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.

We applied the adaptive segmental *k*-mean algorithm to speaker adaptation problems using a 39-word English alpha-digit 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 incorporateed 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 on this database 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 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 underway.

REFERENCES

[1] K. Shikano, K.-F. Lee, and R. Reddy, "Speaker Adaptation through Vector Quantization," *Proc. ICASSP86*, pp. 2643-2646, Tokyo, April 1986.

[2] R. Schwartz, Y. L. Chow, and F. Kubala, "Rapid Speaker Adaptation using a Probabilistic Spectral Mapping," *Proc. ICASSP87*, pp. 633-636, Dallas, April 1987.

[3] R. M. Stern and M. J. Lasry, "Dynamic Speaker Adaptation for Feature-Based Isolated Word Recognition," *IEEE Trans. on ASSP*, Vol. ASSP-35, No. 6. June 1987.

[4] P. F. Brown, C.-H. Lee, and J. C. Spohrer, "Bayesian Adaptation in Speech Recognition," *Proc. ICASSP83*, pp. 761-764, Boston, April 1983.

[5] L. R. Rabiner, J. G. Wilpon, and B.-H. Juang, "A Segmental *k*-Means Training Procedure for Connected Word Recognition," *AT&T Tech. Journal.*, Vol. 65, No. 3, pp. 21-32, May-June 1986.

[6] M. H. DeGroot, *Optimal Statistical Decisions*, McGraw-Hill, New York, 1970.

[7] L. R. Rabiner, C.-H. Lee, B.-H. Juang, D. B. Roe and J. G. Wilpon, "Improved Training Procedures for Hidden Markov Models," J. Acous. Soc. Am. Suppl. 1, Vol. 84, S61, Fall 1988.

[8] L. R. Rabiner and J. G. Wilpon, "Some Performance Benchmarks for Isolated Word Recognition Systems," *Computer, Speech & Language*, Vol. 2, Nos. 3/4, pp. 343-357, December 1987.

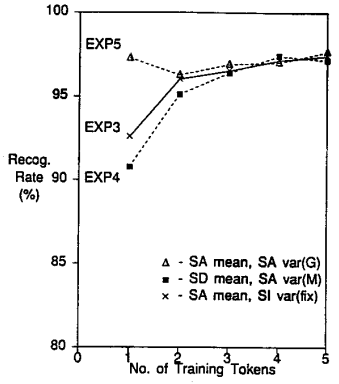

Figure 1. Mean adaptation performance for female speakers

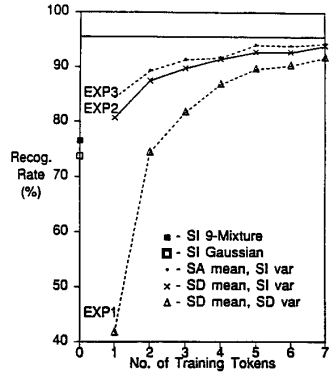

Figure 2. Adaptation performance summary for male speakers