Speaker Adaptation in Continuous Speech Recognition via Estimation of Correlated Mean Vectors

William A. Rozzi[1] and Richard M. Stern

Department of Electrical and Computer Engineering
and School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Abstract

This paper describes recent attempts to improve the recognition performance of a semi-continuous version of the CMU SPHINX system [Huang89, KFLee90] (hereafter SPHINX-SC) through the use of speaker adaptation. Our approach to speaker adaptation is to use multivariate parameter estimation procedures to update the mean values of the component densities which comprise the system's codebook, given the speaker-specific observations. While the optimal Bayesian estimation procedure exploits the correlation between features to obtain fast convergence, it incurs a large computational expense. We have developed a least mean square (LMS) algorithm which produces a faster rate of convergence than the Bayesian estimator, at the expense of a finite misadjustment. This estimate is similar in form to an LMS transversal filter, and is computationally more efficient than the Bayesian estimate.

We discuss the general limitations of the Bayesian, LMS, and maximum likelihood estimation procedures in the context of speaker adaptation for the SPHINX-SC system. We compare SPHINX-SC performance using these codebook mean vector adaptation methods with the system using no adaptation. Results show an overall reduction of 2.0 to 3.4% in word error rate due to adaptation for a set of 11 speakers from the DARPA resource management task. Using a distance metric applied to the adapted codebooks, word error rate was reduced on average by 15% for those speakers automatically identified as good candidates for adaptation.

1. Introduction

Estimation of mean vectors is an essential procedure in the training of hidden Markov model-based speech recognition systems. Normal mixture densities are often used to model the states' output probability distributions. In these cases, we are presented with the problem of accurately estimating the parameters of the assumed densities. Also, we can adjust the system to more closely match new speakers by modifying these parameters given speaker-specific observations. The standard maximum likelihood (ML) methods ignore information which is present in knowledge of the correlation between system parameters when estimating or updating these parameters. This paper explores the effects on recognition performance when correlation information is incorporated into the parameter estimation process. Specifically, we investigate speaker adaptation in a semi-continuous version of the CMU SPHINX system [Huang89, KFLee90] through modification of the codebook mean vectors.

This work was motivated in part to develop an estimation method which was more computationally efficient than, yet retained the desirable properties of, the optimal extended maximum *a posteriori* (EMAP) procedure [Lasry84] which was successful in providing speaker adaptation in the FEATURE [Stern87] recognition system. The work was also motivated by a desire to better understand the relationships between adaptive filtering and parameter estimation. We first focus on the derivation of an algorithm for sub-optimal multivariate parameter estimation which retains the fast convergence property of the optimal estimate, and a discussion of some of its properties. We then investigate the effectiveness of these estimators for codebook adaptation in SPHINX-SC. We also suggest a distance metric that can be applied to each speaker's adapted codebook to determine whether he or she a likely candidate for codebook adaptation.

2. Mean Vector Estimation Methods

Consider a pattern-classification problem with C decision classes and D features, or a C-component mixture in D dimensions. Let the set of input data for the j^{th} class be $\{\mathbf{x}_{j,1}, \mathbf{x}_{j,2},..., \mathbf{x}_{j,n_j}\}$. It is assumed that the random vector $\mathbf{x}_j$ is normally distributed about a mean vector μ_j with a covariance matrix Σ_j, [*i.e.* $p(\mathbf{x}_j|\mu_j) \approx N(\mu_j,\Sigma_j)$], and that all data samples are independent. Define the overall mean vector as $\mu=[\mu_1^T, \mu_2^T,..., \mu_C^T]^T$. The CD-dimensional vector to be estimated is assumed to be normally distributed around the known mean vector μ_o with CDxCD covariance matrix Σ_o, so $p(\mu) \approx N(\mu_o,\Sigma_o)$. Denote the set of observations as $\chi=\{\mathbf{x}_{1,1},..., \mathbf{x}_{1,n_1}, \mathbf{x}_{2,1},..., \mathbf{x}_{2,n_2}, \mathbf{x}_{C,1},..., \mathbf{x}_{C,n_C}\}$. The parameter estimation problem considered here is to produce reliable estimates of the parameter μ given the information provided by the samples in χ.

2.1. Extended Maximum *A Posteriori* Estimation

Bayesian techniques treat unknown parameters as random vectors with a known *a priori* distribution, and maximize the *a posteriori* probability of the parameters given the observations, $p(\mu|\chi)$. Setting the gradient with respect to μ of the natural logarithm of $p(\mu|\chi)$ equal to zero yields the EMAP estimate of the mean μ as in [Lasry84]:

[1]Currently with Ameritech Services Science and Technology.

$$\hat{\mu}_{EMAP}=\Sigma(\Sigma+\mathbf{N}\Sigma_{\mathbf{o}})^{-1}\mu_{\mathbf{o}}+\Sigma_{\mathbf{o}}(\Sigma+\mathbf{N}\Sigma_{\mathbf{o}})^{-1}\mathbf{Na} \quad (1)$$

where **a** is the maximum likelihood estimate of μ (the sample mean), and **N** and Σ are CD-dimensional block diagonal matrices, with $n_j\mathbf{I}$ and Σ_j, respectively, as the DxD diagonal blocks.

The EMAP estimate is an unbiased and consistent estimator, formed as a linear combination of the *a priori* mean vector μ_o and the sample mean of the observed data **a**. When there are few observations or the dogmatism[2] is large, $\hat{\mu}\approx\mu_o$. When **N** becomes large (after many observations), $\hat{\mu}\approx\mathbf{a}$, *i.e.*, the sample mean dominates. This estimate therefore behaves in an intuitively pleasing manner. One disadvantage of the EMAP estimate involves its implementation. As each new observation arrives, $(\Sigma+N\Sigma_o)^{-1}$ must be recomputed; in cases where the product of the number of features and classes (CD) is large, this inversion may be prohibitive. Also, estimation of the covariance matrix Σ_o requires data from more than CD speakers to avoid singularity problems. This requirement created difficulties for the CMU FEATURE system.

2.2. Minimum Mean Square Error Estimation

The MMSE estimate minimizes the expected value of the squared error between the parameter and its estimate. For the vector parameters considered here, the norm of the error vector will be minimized. Let the form of the MMSE estimate of μ be $\hat{\mu}=\mathbf{H}^T\mathbf{a}$, where **a** is the sample mean vector as above and **H** is a CDxCD coefficient matrix to be determined.

The mean square error vector $E\{\varepsilon\}=E\{\|\mu-\hat{\mu}\|^2\}$ is to be minimized with respect to **H**. Rewriting,

$$E\{\varepsilon\}=E\{\mu^T\mu\}-2E\{\mu^T\mathbf{H}^T\mathbf{a}\}+E\{\mathbf{a}^T\mathbf{H}\mathbf{H}^T\mathbf{a}\}$$

Setting the gradient of $E\{\varepsilon\}$ with respect to **H** equal to zero yields

$$-2\nabla_{\mathbf{H}}[E\{\mu^T\mathbf{H}^T\mathbf{a}\}]+\nabla_{\mathbf{H}}[E\{\mathbf{a}^T\mathbf{H}\mathbf{H}^T\mathbf{a}\}] = 0$$

Taking the gradient inside the expectation and using the identities $\nabla_{\mathbf{M}}\mathbf{a}^T\mathbf{M}\mathbf{b}=\mathbf{b}\mathbf{a}^T$ and $\nabla_{\mathbf{M}}\mathbf{a}^T\mathbf{M}^T\mathbf{M}\mathbf{a}=2\mathbf{a}\mathbf{a}^T\mathbf{M}$ results in

$$E\{\mathbf{a}\mu^T\}=E\{\mathbf{a}\mathbf{a}^T\}\mathbf{H}$$

Defining $\Phi_{\mathbf{aa}}=E\{\mathbf{a}\mathbf{a}^T\}$ and $\Phi_{\mathbf{a}\mu}=E\{\mathbf{a}\mu\}$, the optimal **H** is easily shown to be $\mathbf{H}^*=\Phi_{aa}^{-1}\Phi_{\mathbf{a}\mu}$. The MMSE solution is in the form of the traditional Wiener-Hopf equation except that the parameter Φ_{aa} (and therefore $\mathbf{H}^*$) varies with the number of data samples obtained, as will be seen. It is this dependence of the optimal coefficients on sample counts (**N**) that forms the basis of the differences between standard adaptive filtering and our parameter estimation techniques.

The MMSE estimate is the mean of the *a posteriori* density, the conditional mean of $p(\mu|\chi)$. The EMAP estimate is the value of the parameter at which the *a posteriori* density has its maximum. If the *a posteriori* pdf is a unimodal function which is symmetric about the conditional mean, as is the case with the multivariate normal density, then these two estimates are equivalent [VanTrees68]. To demonstrate this equivalence, Φ_{aa} and $\Phi_{a\mu}$ must be expressed in terms of Σ, Σ_o, and μ_o. It is also necessary to add a bias term to the MMSE estimate by appending a constant to the sample mean vector, i.e., $\mathbf{a}'=[1\ \mathbf{a}]^T$, to represent the contribution of the μ_o term in the EMAP expression. It can then be shown that $\Phi_{aa}=\Sigma_o+N^{-1}\Sigma+\mu_o\mu_o^T$, $\Phi_{\mu\mathbf{a}}=\Sigma_o+\mu_o\mu_o^T$, and $E\{\mathbf{a}\} = \mu_o$. Substituting these values into the previous equation and using a block matrix inversion identity yields:

$$\hat{\mu}_{\mathrm{MMSE}} = \Sigma_o N(\Sigma_o N+\Sigma)^{-1}(\mathbf{a}-\mu_o) + \mu_o$$

It can also be shown that $I-\Sigma_oN(\Sigma_oN+\Sigma)^{-1}=\Sigma(\Sigma_oN+\Sigma)^{-1}$, so the coefficients of **a** and μ_o in the MMSE and EMAP estimates are the same.

2.3. Least Mean Square Estimation

Conventional LMS adaptive filtering is derived from MMSE parameter estimation by performing the estimation iteratively using a modified gradient search procedure, and by approximating statistical averages used in the computation by their instantaneous values. The MMSE gradient algorithm is written as

$$\mathbf{H}_{\mathbf{k+1}}=\mathbf{H}_{\mathbf{k}} - \frac{\beta}{2}\nabla_{\mathbf{H}_k}[E\{\varepsilon\}] \quad (2)$$

where β is the step size or adaptation constant. The index **k** is incremented after each new sample $\mathbf{x_k}$ (from any class) is obtained. Borrowing from the above derivation of $E\{\varepsilon\}$,

$$\mathbf{H}_{\mathbf{k+1}} = [I-\beta\Phi_{\mathbf{aa}}]\mathbf{H}_{\mathbf{k}}+\beta\Phi_{\mathbf{a}\mu} \quad (3)$$

Substitution of the stochastic gradient estimate $\nabla_{\mathbf{H}_j}\varepsilon$ for $\nabla_{\mathbf{H}}E\{\varepsilon\}$ in Equation (2) yields

$$\mathbf{H}_{\mathbf{k+1}}=[\mathbf{I}-\beta\mathbf{a_k}\mathbf{a_k}^T]\mathbf{H_k}+\beta\mathbf{a_k}\mathbf{d_k}^T \quad (4)$$

Note that $\mathbf{H_k}$ is updated through the incorporation of sample $\mathbf{x_k}$ into $\mathbf{a_k}$; this sample may be the cepstral data from a new speech frame in an HMM system or an acoustic feature vector in a feature-based recognition system.

As with conventional LMS adaptive filtering, it is necessary to specify several parameters in the LMS adaptive estimate, namely the step size parameter β, desired signal **d**, and initial coefficient matrix $\mathbf{H_0}$. A constant step size was found to be sufficient. Since Φ_{aa} varies with the number of observations, β is chosen to be inversely proportional to the average of the sum of the eigenvalues of the initial and final value of Φ_{aa} (*i.e.* the sum of the traces of these matrices).

The true mean vector is the appropriate choice for the desired signal, but it is obviously unavailable. Substitution of the ML estimate for the true mean allows the LMS estimate to asymptotically converge to the sample mean, as does the MAP estimate, but the error is often larger than what is desirable. To obtain a value for **d** which reduces this error without excessive additional computation, **d** is chosen to be a weighted sum of the *a priori* mean and the sample mean with the constant weights derived from the EMAP procedure:

$$\mathbf{d_k}=\Sigma(\Sigma+\mathbf{N_c}\Sigma_o)^{-1}\mu_o+\Sigma_o(\Sigma+\mathbf{N_c}\Sigma_o)^{-1}\mathbf{N_c}\mathbf{a_k}$$

[2]Dogmatism is defined as the ratio of the within-speaker variance to the between-speaker variance, σ/σ_o.

where $\mathbf{N_c} = \eta\mathbf{I}$ and η is a constant.[3] We refer to the LMS estimate using this value for **d** as LMS-C. Using this fixed weight estimate for **d** introduces a misadjustment or bias into the LMS-C estimate, which can be shown to be equal to:

$$M_{LMS-C} = trace[\Sigma(\Sigma+\Sigma_o\mathbf{N_c})^{-1}\Sigma_o(\Sigma+\mathbf{N_c}\Sigma_o)^{-1}\Sigma]$$

Although it has not been proven, it is believed that it is the addition of this misadjustment which allows the LMS-C to obtain a mean-square error which is initially lower than that of the optimal solution. The parameter η controls the tradeoff between this misadjustment and rate of convergence. As η increases, the misadjustment decreases, but so does the contribution of μ_o in **d**. Since giving more weight to the sample mean increases the initial error, it is necessary to strike a balance between these two effects.

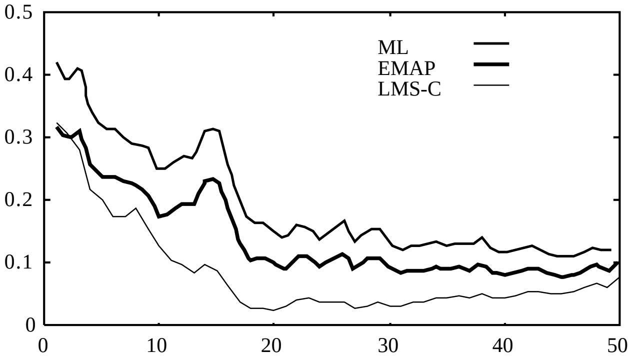

Figure 1. Normalized mean square error of ML, EMAP, and LMS-C estimation algorithms vs. number of samples for a 2 class, 1 feature case.

Appropriate choices for initialization of the coefficient matrix **H** can be determined from an analysis of the expected coefficient error. Define the expected coefficient error matrix as $\mathbf{V_k} = E[\mathbf{H_k} - \mathbf{H_k^*}]$. From Equation (3), we can write $\mathbf{V_{k+1}} = (\mathbf{I} - 2\beta\Phi_{\mathbf{aa}})\mathbf{V_k} - (\mathbf{H^*_{k+1}} - \mathbf{H^*_k})$. We can show that in cases where the dogmatism is small the last term can be ignored and $\Phi_{aa}(k)$ may be assumed to be constant. Therefore, we can write the coefficient error as $\mathbf{V}_k = (I - \beta\Phi_{\mathbf{aa}}(k))^k\mathbf{V}_0$. Obviously we want to choose $\mathbf{H_0}$ to make $\mathbf{V_0}$ as small as possible. We do this by setting $\mathbf{H_0}$ equal to the coefficients which would be specified by the optimal estimate after a single sample from each class, *i.e.* $\mathbf{H_0} = [\Sigma(\Sigma+\Sigma_o)^{-1}\mu_o \mid \Sigma_o(\Sigma+\Sigma_o)^{-1}]$. This choice incorporates more knowledge of the structure of the data into the LMS-C estimate than initializing $\mathbf{H_0}$ to zero or an identity matrix. We have observed in empirical simulations that this choice often reduces the initial mean square error to levels lower than that of the EMAP (see Figure 1). We also have observed in simulations a tendency for the LMS-C estimate to diverge when the data dogmatism is large and the above assumptions are invalid.

The success of the EMAP and LMS-C procedures is highly dependent on a number of properties of the data, the most notable being the degree of correlation and the dogmatism. For example, the mean-square error of the MMSE and small-dogmatism LMS-C estimates can be expressed as $trace[\Sigma(\Sigma+\Sigma_o\mathbf{N})^{-1}\Sigma_o]$. This expression is plotted in Figure 2 for different values of dogmatism. It can readily be seen that when the dogmatism is much greater than 1 these estimation algorithms are ineffective. Similar comparisons [Lasry84] have shown that the estimates converge more rapidly when there is greater correlation among features and/or decision classes.

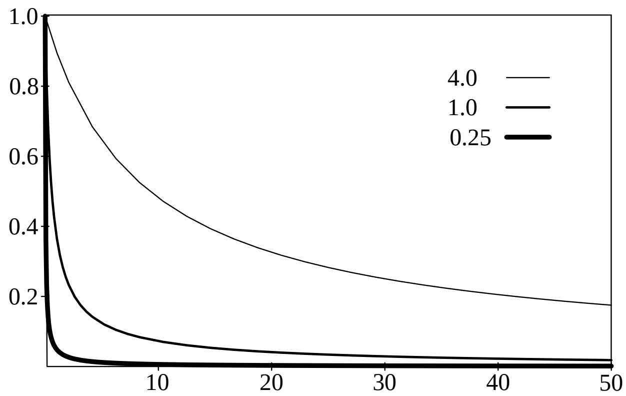

Figure 2. Normalized expected mean square error of LMS-C and MMSE estimates for dogmatism values of 0.25, 1.0, and 4.0 vs. number of samples for a 2 class, 1 feature case.

3. Application to Stochastic Model-Based Recognition Systems

The original CMU SPHINX system [KFLee90] demonstrated a speaker-independent word accuracy of better than 94% on a 1000-word task with a perplexity of 60. This system used a hidden Markov model with discrete state output probability densities and a vector quantization (VQ) codebook. Other researchers have investigated the use of continuous gaussian mixtures as the output probability distributions, eliminating the need for quantization of the speech frame data [Huang89, CHLee90]. Huang has investigated an approach in which VQ codebook entries become the mean vectors of mixture density components, and each state's discrete output probabilities become the mixture coefficients. This semi-continuous HMM (SCHMM) reduces the number of parameters to be estimated with respect to the continuous-density HMM systems. We attempted to determine how effective the above estimation techniques were in reducing SCHMM word error rates through codebook adaptation, and to demonstrate that LMS-C estimation performs as well as Bayesian techniques in an HMM framework.

An early version of a semi-continuous form of SPHINX[4] was modified to allow ML, EMAP, or LMS-C adaptation of codebook mean vectors after observation of a set of adaptation sentences from a given speaker. SPHINX-SC uses a set of three codebooks:

[3] The weights for μ_o and $\mathbf{a_k}$ in $\mathbf{d_k}$ are those which would be specified in the EMAP procedure if η samples had been obtained from each class.

[4] There are a number of differences between the SPHINX-SC system used in the present evaluation and other semi-continuous versions of SPHINX described elsewhere in these proceedings.

cepstral, differenced cepstral, and power cepstral data, with 256 codewords each. Each codeword has an associated mean vector, covariance matrix, and determinant for use in output probability calculations. The cepstral and differenced cepstral vectors are 12-dimensional, and power cepstral vectors have 2 elements. Since the dimension of the *a priori* adaptation statistics is equal to the product of the number of classes and dimensions, at least 3072 training set speakers would have been required to estimate these parameters. Because of the limited availability of training data it was necessary to reduce this estimation problem to a set of parallel problems by considering only the correlation between a given codeword's features.

Adaptation Type	Word Error Rate	Percent Reduction
Unadapted	6.46%	----
ML	6.33%	2.0%
EMAP	6.24%	3.4%
LMS-C	6.33%	2.0%

Table 1: Recognition results for SPHINX-SC with codeword adaptation.

Training data for the generation of models and adaptation statistics were 40 sentences from approximately 100 speakers in the TIRM database. The adaptation data were 40 sentences from 11 additional speakers in the TIRM database. Evaluation test data were another 25 sentences from these 11 speakers. Evaluations were based on comparisons of word recognition rates between systems using speaker independent codebooks and adapted codebooks, as shown in Table 1. Although significant reductions (on the order of 25%) in error rate were observed for some speakers, the aggregate results are much lower. In each experiment, about half of the speakers showed improvement due to adaptation while the remainder showed no change or an increased error rate. Results may be improved if an automatic method of identifying speakers which are good candidates for adaptation could be found.

We attempted to automatically identify those speakers that could most benefit from adaptation by calculating the change in the sum of the pairwise Euclidean distances between all codewords for the adapted and unadapted codebooks. Reasoning that if the codebook expands the decision classes should become less confusable, we chose to use the speaker-adapted codebook only when this distance metric was positive. Experimental results when this decision criterion is applied are given in Table 2 which lists the number of adaptation candidates (out of 11) identified by this method (C), the number of those identified which actually showed improvement (I), the number of improved speakers which were missed (M), and the adapted (A) and unadapted (U) error rates when only the adapted speakers are included.

Standard statistical analysis [Gillick89] of the experiments reported in Table 2 showed that while the adapted speaker recognition scores were significantly different from the unadapted with a confidence of 90% to 95%, the adapted systems were not

Adaptation Type	C / I / M	Word Error Rate A / U / % Reduction
ML	3 / 3 / 3	4.99%/5.96%/16.2%
EMAP	4 / 3 / 2	4.26%/5.01%/15.0%
LMS-C	3 / 3 / 2	5.15%/5.96%/13.6%

Table 2: Recognition results for automatically selected speakers.

significantly different from each other. The fact that the EMAP and LMS-C error rates are not significantly lower than the ML error rate is not suprising when the following facts are considered. First, the average dogmatism was estimated to be around 2.0 for the cepstral codebook and 3.0 for the differenced cepstral codebook. Second, only the within-codeword correlation and not the between-codeword correlation information was available for use in the EMAP and LMS-C procedures. Under these conditions, much more adaptation data is necessary to provide better adaptation performance.

Acknowledgements

This research was sponsored by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 5167, under contract number N00039-85-C-0163, and by Ameritech Services. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the US Government. The authors wish to thank Xuedong Huang for his guidance and help with the semi-continuous SPHINX system, and Robert Wohlford and Ameritech Services for their support of portions of this work.

References

[Lasry84] M. J. Lasry and R. M. Stern (1984). A Posteriori Estimation of Correlated Jointly Gaussian Mean Vectors, IEEE Trans. on Pattern Analysis and Machine Intelligence PAMI-6, 530-535.

[Stern87] R. M. Stern and M. J. Lasry (1987). Dynamic Speaker Adaptation for Feature-Based Isolated Letter Recognition, IEEE Trans. on Acoustics, Speech, and Signal Processing 35, 751-763.

[VanTrees68] H. L. Van Trees (1968). Detection, Estimation, and Modulation Theory, Part I, Wiley, New York.

[KFLee90] K. F. Lee, H. W. Hon, and R. Reddy (1990). An Overview of the SPHINX Speech Recognition System, IEEE Trans. on Acoustics, Speech, and Signal Processing 38, No. 1, January 1990.

[Huang89] X. D. Huang, H. W. Hon, and K. F. Lee (1989). On Semi-Continuous Hidden Markov Modeling, Proc. of ICASSP89, 689-692.

[CHLee90] C. H. Lee, C. H. Lin, and B. H. Juang (1990). A Study on Speaker Adaptation of Continuous Density HMM Parameters, Proc. of ICASSP90, 145-148.

[Gillick89] L. Gillick and S. J. Cox (1989). Some Statistical Issues in the Comparison of Speech Recognition Algorithms, Proc. of ICASSP89, 532-535.