ELSEVIER

Speech Communication 26 (1998) 283–297

SPEECH COMMUNICATION

Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition

Nagendra Kumar, Andreas G. Andreou *

Electrical and Computer Engineering Department, Center for Language and Speech Processing, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA

Received 13 January 1998; received in revised form 12 August 1998; accepted 17 August 1998

Abstract

We present the theory for heteroscedastic discriminant analysis (HDA), a model-based generalization of linear discriminant analysis (LDA) derived in the maximum-likelihood framework to handle heteroscedastic-unequal variance-classifier models. We show how to estimate the heteroscedastic Gaussian model parameters jointly with the dimensionality reducing transform, using the EM algorithm. In doing so, we alleviate the need for an a priori ad hoc class assignment. We apply the theoretical results to the problem of speech recognition and observe word-error reduction in systems that employed both diagonal and full covariance heteroscedastic Gaussian models tested on the TI-DIGITS database.

Keywords: Heteroscedastic; Discriminant analysis; Speech recognition; Reduced rank HMMs

1. Feature extraction for speech recognition

Feature extraction in speech recognition systems is a two-step process, when considered from a data dimensionality viewpoint. In the first step, the dimensionality of the speech waveform is reduced using cepstral analysis (Davis and Mermelstein, 1980) or some other analysis procedure motivated by human speech perception (Hermansky, 1990). In the second step, the dimensionality of the obtained features vector is increased by the formation of an extended features vector that includes derivative and acceleration information (Furui, 1986).

Despite the widespread use of Cepstrum coefficients as features for speech recognition, the method used to arrive at these features is heuristic. It is not even clear that the linear operator (discrete cosine transform) used in the final step of Cepstrum coefficient extraction, is an optimal choice. The Karhunen–Loeve (KL) transform that is often used as a linear operator for dimension reduction, is an optimum transform for signal representation and speech compression, but not necessarily for speech recognition, which is a classification task. The same is true for any other signal processing scheme, including those that are based on perceptual considerations (Hermansky, 1990) or auditory models (Cohen, 1989) which do not take into account the task objective and thus are strictly bottom up approaches.

* Corresponding author. Tel.: +1 410 516 8361; fax: +1 410 516 8313; e-mail: andreou@jhu.edu.

0167-6393/98/$ – see front matter
PII: S0167-6393(98)00061-2

The second step in the feature extraction process is aimed at incorporating information about the context. Feature derivative and acceleration are widely used for this step, but, their use is nevertheless ad hoc. The obtained features, indeed incorporate context information, but they are not necessarily the most parsimonious representation of information for the intended task, which is speech recognition.

To obtain features suitable for speech sound classification, Hunt (1979) proposed the use of linear discriminant analysis (LDA) (Duda and Hart, 1973; Dillon and Goldstein, 1984; Fukunaga, 1990) and derived features so as to improve separability of syllables.

Brown (1987), almost a decade later, experimented with both principal component analysis (PCA) (Duda and Hart, 1973; Fukunaga, 1990) and LDA to project the features in subspaces of reduced dimensions. Brown's experiments, were done using a discrete hidden Markov model (HMM) classifier, showed that the LDA transform is superior to the PCA transform. Brown incorporated context information by applying LDA on an augmented feature vector formed by concatenating the features from a number of frames around the observation vector. By doing so, context is incorporated selectively based on the best linear combination of observation vectors and thus all components of the feature vector are likely to contribute to better classification. In the years following Brown's work, researchers have applied LDA to discrete (Zahorian et al., 1991) and continuous (Doddington, 1989; Woodland and Cole, 1991; Bocchieri and Wilpon, 1993) HMM speech recognition systems and have reported improved performance on small vocabulary tasks but with mixed results on large vocabulary phoneme-based systems (Yu et al., 1990; Wood et al., 1991).

Applying LDA to the speech recognition problem is more complicated when the classifier is based on continuous HMMs and Gaussian mixtures, where the sample class assignment is not obvious. Various techniques for class assignment have been proposed and used with different degrees of success with continuous HMMs (Haeb-Umbach and Ney, 1992; Haeb-Umbach et al., 1993; Aubert et al., 1993; Roth et al., 1993; Siohan, 1995).

LDA has also been employed successfully to reduce the feature dimensions from large dimensionality auditory representations (Hunt and Lefebvre, 1989; Jankowski, 1992; Kumar et al., 1995) for speech recognition. Adaptive forms of LDA have been also proposed with encouraging results, taking into account individual class distribution and circumvent mismatch between the assumed class distribution and the actual data (Ayer, 1992; Ayer et al., 1993).

Despite its popularity and promise for significant improvements, LDA has not always improved the performance of speech recognition systems. We attribute the lack of robustness to a basic shortcomings in the widely used model-free formulation of LDA. In the original Fisher/Rao model-free formulation (Fisher, 1936, 1938; Rao, 1965), LDA projections are best suited to classifier models where class distributions have equal variance. LDA is not the optimal transform when the class distributions are heteroscedastic.

Our observation is based on recent work by Campbell (1984) who has shown that LDA is related to the maximum-likelihood estimation of parameters for a Gaussian model, with two a priori assumptions on the structure of the model. The first assumption is that all the class-discrimination information resides in a p-dimensional subspace of the n-dimensional feature space. The second assumption is that the within-class variances are equal for all the classes. Hastie and Tibshirani (1994) further generalized this result by assuming that class distributions are a mixture of Gaussians. However, the constraint of common covariance matrices is maintained in both (Campbell, 1984) and (Hastie and Tibshirani, 1994). It should be further noted that heteroscedastic AR models are well studied in the econometric literature, the interested reader may review the state of the art in (Engle, 1995).

We have generalized LDA (Kumar and Andreou, 1996a,b, submitted; Kumar, 1997) to handle heteroscedasticity by dropping the assumption of equal variance in the parametric model. We refer to this generalization as *heteroscedastic discriminant analysis* (HDA). The class assignment problem is the second shortcoming of LDA as used today. We address it by finding the optimum transformation as a joint optimization process when training HMM model parameters using the EM algorithm (Kumar and Andreou,

1996a, submitted; Kumar, 1997). This training procedure does not necessitate an explicit class assignment. Feature dimension reduction using reduced-rank maximum-likelihood estimation for HMMs was also reported recently by Sun (1997).

In this paper, we present the theoretical framework for HDA and apply it to the problem of feature selection and dimensionality reduction for speech recognition. In Section 2 we begin with a brief discussion of LDA in the original Fisher/Rao model-free formulation followed in Section 3 with a description of HDA. The extension to HMMs is derived in Section 4. The speech experiments are discussed in Sections 5 and 6. Section 7 concludes the paper.

2. Linear discriminant analysis

The problem of dimensionality reduction through linear projections is formulated as follows.

Let x be an n-dimensional feature vector. We seek a linear transformation $\mathbb{R}^n \to \mathbb{R}^p$ $(p < n)$ of the form $y_p = \theta_p^{\mathrm{T}} x$, where θ_p is an $n \times p$ matrix. Let θ be a nonsingular $n \times n$ matrix used to define the linear transformation $y = \theta^{\mathrm{T}} x$. Let us partition as

$$\theta = [\theta_p \theta_{n-p}] = [\vec{\theta}_1 \ldots \vec{\theta}_n], \tag{1}$$

where θ_p consists of the first p columns of θ, θ_{n-p} consists of the remaining $n-p$ columns, and $\vec{\theta}_i$ is the ith column of θ. Then, feature dimension reduction can be viewed as a two-step procedure. First we apply a nonsingular linear transformation to x to obtain $y = \theta^{\mathrm{T}} x$. Second, we retain only the first p rows of y to give y_p. This notation may seem superfluous at the moment. However, it will be useful when we consider more complex models in subsequent sections. The objective of LDA is to choose the linear transformation θ_p such as to retain the maximum amount of class-discrimination information in the reduced feature space.

Let there be a total of J classes, and let $g(i) \to \{1 \ldots J\}$ indicate the class that is associated with x_i. Let $\{x_i\}$ be the set of training examples available. Then, $\sum_{g(i)=j} 1 = N_j$ (here, $g(i) = j$ is a notation used to define the set of all i for which $g(i) = j$) is the total number of training examples associated with class j, and $\sum_{j=1}^{J} N_j = N$ is the total number of training examples. Let $\bar{X}$ be the sample mean,

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{2}$$

The total normalized sum of squares and products (SSQP) $\bar{T}$ is defined as

$$\bar{T} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{X})(x_i - \bar{X})^{\mathrm{T}}. \tag{3}$$

The class means $\bar{X}_j$, the normalized class SSQP $\bar{W}_j$, and the overall pooled and normalized SSQP $\bar{W}$ are defined as

$$\bar{X}_j = \frac{1}{N_j} \sum_{g(i)=j} x_i, \quad j = 1 \ldots J, \tag{4}$$

$$\bar{W}_j = \frac{1}{N_j} \sum_{g(i)=j} (x_i - \bar{X}_j)(x_i - \bar{X}_j)^{\mathrm{T}}, \quad j = 1 \ldots J, \tag{5}$$

$$\bar{W} = \frac{1}{N} \sum_{j=1}^{J} N_j \bar{W}_j. \tag{6}$$

In the model-free formulation, the two-class LDA method chooses a single projection that maximizes the ratio of the overall SSQP to the within-class SSQP (Fisher, 1936, 1938; Rao, 1965),

$$\hat{\vec{\theta}}_1 = \arg\max_{\vec{\theta}_1} \frac{\vec{\theta}_1^{\mathrm{T}} \bar{T} \vec{\theta}_1}{\vec{\theta}_1^{\mathrm{T}} \bar{W} \vec{\theta}_1}. \tag{7}$$

The solution to Eq. (7) corresponds to the right eigenvector of $\bar{W}^{-1}\bar{T}$ that has the largest eigenvalue (Dillon and Goldstein, 1984). If the number of classes is more than 2, we can find a p-dimensional linear transformation ($p < n$) for classification. To get a p-dimensional transformation, we maximize the ratio

$$\hat{\theta}_p = \arg\max_{\theta_p} \frac{|\theta_p^{\mathrm{T}} \bar{T} \theta_p|}{|\theta_p^{\mathrm{T}} \bar{W} \theta_p|}. \tag{8}$$

To obtain $\hat{\theta}_p$, we choose those eigenvectors of $\bar{W}^{-1}\bar{T}$ that correspond to the largest p eigenvalues, and let $\hat{\theta}_p$ be an $n \times p$ matrix of these eigenvectors. The p-dimensional features thus obtained $y = \hat{\theta}_p^{\mathrm{T}} x$ are uncorrelated.

3. Generalizations of LDA in the maximum-likelihood framework

In this section, we present a brief derivation of the necessary equations for HDA, the generalization of LDA. We begin by noting that the objective of LDA is to reduce the feature dimension by choosing a linear p-dimensional subspace of the feature space and by rejecting an $(n-p)$-dimensional subspace. Since the final objective is classification, the implicit assumption is that the rejected subspace does not carry any classification information. For Gaussian models, the assumption of lack of classification information is equivalent to the assumption that the means and the variances of the class distributions are the same for all classes, in the rejected $(n-p)$-dimensional subspace. We explicitly write a model that makes these assumptions, and use maximum likelihood to find the optimal transformation under these assumptions.

Let θ be a nonsingular $n \times n$ matrix that we use to define a linear transformation that maps the data variables x into new variables y. Let us assume that only the first p components of y carry any class-discrimination information. In other words, let the full rank linear transformation θ be such that the first p columns of θ span the p-dimensional subspace in which the class means, and probably the class variances, are different. Since the data variables x are Gaussian, their linear transformation y are also Gaussian. Let the parameters that describe the class means and the variances of y be μ_j and Σ_j, respectively. For notational convenience, we partition the parameter space of the means μ_j and variances Σ_j as follows:

$$\mu_j = \begin{bmatrix} \mu_{j,1} \\ \vdots \\ \mu_{j,p} \\ \mu_{0,p+1} \\ \vdots \\ \mu_{0,n} \end{bmatrix} = \begin{bmatrix} \mu_j^p \\ \mu_0 \end{bmatrix}, \tag{9}$$

$$\Sigma_j = \begin{bmatrix} \Sigma^p_{j(p\times p)} & 0 \\ 0 & \Sigma^{(n-p)}_{(n-p\times n-p)} \end{bmatrix}. \tag{10}$$

Here, μ_0 is common to all the class means, and the μ^p_j are different for each class. The Σ_j have also been partitioned in the corresponding manner, such that $\Sigma^{(n-p)}$ is common for all the classes, whereas Σ^p_j are different for different classes.

The probability density of x_i under the preceding model is given as

$$P(x_i) = \frac{|\theta|}{\sqrt{(2\pi)^n |\Sigma_{g(i)}|}} \exp\left(-\frac{(\theta^{\mathrm{T}} x_i - \mu_{g(i)})^{\mathrm{T}} \Sigma^{-1}_{g(i)} (\theta^{\mathrm{T}} x_i - \mu_{g(i)})}{2} \right), \tag{11}$$

where x_i belongs to the group $g(i)$. Note that although the Gaussian distribution is defined on the transformed variable y_i, we are interested in maximizing the likelihood of the original data x_i. The term $|\theta|$ in Eq. (11) comes from the Jacobian of the linear transformation $y = \theta^{\mathrm{T}} x$,

$$\left| \frac{\partial y}{\partial x} \right| = |\theta^{\mathrm{T}}| = |\theta|. \tag{12}$$

Eq. (11) assumes that $|\theta| > 0$. If $|\theta|$ is negative, we can easily find an alternate θ that satisfies the condition by multiplying any column of θ by -1.

3.1. Full rank Σ^p_j

The log-likelihood of the data $L_{\mathrm{F}} = \sum_{i=1}^{N} \log P(x_i)$ under the linear transformation θ and under the constrained Gaussian model assumption for each class is

$$\log L_{\mathrm{F}}(\mu_j, \Sigma_j, \theta; \{x_i\}) = -\frac{1}{2} \sum_{i=1}^{N} \left\{ (\theta^{\mathrm{T}} x_i - \mu_{g(i)})^{\mathrm{T}} \Sigma^{-1}_{g(i)} (\theta^{\mathrm{T}} x_i - \mu_{g(i)}) + \log((2\pi)^n |\Sigma_{g(i)}|) \right\} + N \log|\theta|. \tag{13}$$

We use the subscript F to remind us that the Σ^p_j (which is a part of $\Sigma_{g(i)}$) are different and full covariance matrices. The likelihood function in Eq. (13) can now be maximized with respect to its parameters.

Doing a straightforward maximization with respect to various parameters is computationally intensive. We simplify it considerably by first calculating the values of the mean and variance parameters that maximize the likelihood in terms of a fixed linear transformation θ. Differentiating the likelihood equation with respect to the parameters μ_j and Σ_j, and finding the point where the partial derivatives are zero, gives us the mean and variance estimates:

$$\hat{\mu}^p_j = \theta^{\mathrm{T}}_p \bar{X}_j, \quad j = 1 \ldots J, \tag{14}$$

$$\hat{\mu}_0 = \theta^{\mathrm{T}}_{n-p} \bar{X}, \tag{15}$$

$$\hat{\Sigma}^p_j = \theta^{\mathrm{T}}_p \bar{W}_j \theta_p, \quad j = 1 \ldots J, \tag{16}$$

$$\hat{\Sigma}^{n-p} = \theta^{\mathrm{T}}_{n-p} \bar{T} \theta_{n-p}. \tag{17}$$

Note that the μ_j, $j = 1 \ldots J$, can be calculated if θ is known, and Σ_j, $j = 1 \ldots J$, can be calculated if μ_j, $j = 1 \ldots J$, and θ are known. Therefore, we would first like to solve for θ. Substituting the values of the optimized μ_j, $j = 1 \ldots J$, and Σ_j, $j = 1 \ldots J$, in Eq. (13), we obtain the log-likelihood of the data ($L_{\mathrm{F}}(\theta; \{x_i\})$) in terms of θ.

$$\log L_{\mathrm{F}}(\theta;\{x_i\}) = \frac{-Nn}{2}\log 2\pi - \frac{N}{2}\log|(\theta_{n-p}^{\mathrm{T}}\bar{T}\theta_{n-p})| - \sum_{j=1}^{J}\frac{N_j}{2}\log|(\theta_p^{\mathrm{T}}\bar{W}_j\theta_p)|$$
$$-\frac{1}{2}\sum_{j=1}^{J}\sum_{g(i)=j}((x_i-\bar{X}_j)^{\mathrm{T}}\theta_p(\theta_p^{\mathrm{T}}\bar{W}_j\theta_p)^{-1}\theta_p^{\mathrm{T}}(x_i-\bar{X}_j))$$
$$-\frac{1}{2}\sum_{j=1}^{J}\sum_{g(i)=j}((x_i-\bar{X})^{\mathrm{T}}\theta_{n-p}(\theta_{n-p}^{\mathrm{T}}\bar{T}\theta_{n-p})^{-1}\theta_{n-p}^{\mathrm{T}}(x_i-\bar{X})) + N\log|\theta|. \tag{18}$$

We can simplify $L_{\mathrm{F}}(\theta;\{x_i\})$ and then maximize it with respect to θ to give

$$\hat{\theta}_{\mathrm{F}} = \arg\max_{\theta}\left\{-\frac{N}{2}\log|(\theta_{n-p}^{\mathrm{T}}\bar{T}\theta_{n-p})| - \sum_{j=1}^{J}\frac{N_j}{2}\log|(\theta_p^{\mathrm{T}}\bar{W}_j\theta_p)| + N\log|\theta|\right\}, \tag{19}$$

where $\hat{\theta}_{\mathrm{F}}$ is the estimate of the parameter θ.

Since there is no closed-form solution for maximizing the likelihood with respect to θ, the maximization in Eq. (19) has to be performed numerically. In our experiments, we performed the required optimization using quadratic programming algorithms such as those available in MATLAB™ optimization tool-box. We used the LDA solution as an initial guess for $\hat{\theta}$. We explicitly provided the analytic expressions (which can be found in (Kumar, 1997; Kumar and Andreou, submitted)) for the derivatives of the likelihood to the optimization routines. Even though we use quadratic-optimization techniques, the likelihood surface is not strictly quadratic, and the optimization algorithms occasionally fail. We achieve robustness by using steepest descent. The amount of time it takes to compute $\hat{\theta}$ depends on the particular problem, and it has been our experience that the quadratic optimization routines, when they work, are about 2 orders of magnitude faster than steepest-descent. After optimization, we use only the first p columns of $\hat{\theta}_{\mathrm{F}}$ to obtain the dimension-reduction transformation.

3.2. Diagonal Σ_j^p

In speech recognition, we often assume that the within-class variances are diagonal. Therefore, we also consider the optimal projections for this case. Let us assume that Σ_j^p and $\Sigma^{(n-p)}$ are diagonal matrices such that the diagonal elements of Σ_j are $\{\sigma_j^1 \ldots \sigma_j^p \sigma^{p+1} \ldots \sigma^n\}$. Then, in terms of the matrix partitions in Eqs. (9) and (10), the log-likelihood of the data can be written as

$$\log L_{\mathrm{D}}(\mu_j,\Sigma_j,\theta;\{x_i\}) = \frac{-Nn}{2}\log 2\pi + N\log|\theta| - \frac{N}{2}\sum_{k=p+1}^{n}\log|\sigma^k|$$
$$-\sum_{j=1}^{J}\frac{N_j}{2}\sum_{k=1}^{p}\log|\sigma_j^k| - \frac{1}{2}\sum_{j=1}^{J}\sum_{g(i)=j}\sum_{k=1}^{p}\frac{(\vec{\theta}_k^{\mathrm{T}}x_i-\mu_{j,k})^2}{\sigma_j^k}$$
$$+\frac{1}{2}\sum_{j=1}^{J}\sum_{g(i)=j}\sum_{k=p+1}^{n}\frac{(\vec{\theta}_k^{\mathrm{T}}x_i-\mu_{0,k})^2}{\sigma^k}. \tag{20}$$

Using the same method as before, and maximizing the likelihood with respect to μ_j and Σ_j, $j = 1 \ldots J$, we get

$$\hat{\mu}_j^p = \theta_p^{\mathrm{T}}\bar{X}_j, \quad j = 1 \ldots J, \tag{21}$$

$$\hat{\mu}_0 = \theta_{n-p}^{\mathrm{T}}\bar{X}, \tag{22}$$

$$\hat{\Sigma}_j^p = \text{Diag}\,(\theta_p^{\mathrm{T}} \bar{W}_j \theta_p), \quad j = 1 \ldots J, \tag{23}$$

$$\hat{\Sigma}^{n-p} = \text{Diag}\,(\theta_{n-p}^{\mathrm{T}} \bar{T} \theta_{n-p}). \tag{24}$$

Substituting values of the maximized mean and variance parameters in Eq. (20) gives the maximized likelihood of the data in terms of θ,

$$\begin{aligned}\log L_{\mathrm{D}}(\theta; \{x_i\}) = {} & \frac{-Nn}{2} \log 2\pi - \frac{N}{2} \log |\text{Diag}\,(\theta_{n-p}^{\mathrm{T}} \bar{T} \theta_{n-p})| - \sum_{j=1}^{J} \frac{N_j}{2} \log |\text{Diag}\,(\theta_p^{\mathrm{T}} \bar{W}_j \theta_p)| \\ & - \frac{1}{2} \sum_{j=1}^{J} \sum_{g(i)=j} ((x_i - \bar{X}_j)^{\mathrm{T}} \theta_p \,\text{Diag}\,(\theta_p^{\mathrm{T}} \bar{W}_j \theta_p)^{-1} \theta_p^{\mathrm{T}} (x_i - \bar{X}_j)) \\ & - \frac{1}{2} \sum_{j=1}^{J} \sum_{g(i)=j} ((x_i - \bar{X})^{\mathrm{T}} \theta_{n-p} \,\text{Diag}\,(\theta_{n-p}^{\mathrm{T}} \bar{T} \theta_{n-p})^{-1} \theta_{n-p}^{\mathrm{T}} (x_i - \bar{X})) + N \log |\theta|. \end{aligned} \tag{25}$$

We can now find the maximum-likelihood estimate for θ by maximizing this likelihood numerically. We can simplify this maximization to the following:

$$\hat{\theta}_{\mathrm{D}} = \arg \max_{\theta} \left\{ -\frac{N}{2} \log |\text{Diag}\,(\theta_{n-p}^{\mathrm{T}} \bar{T} \theta_{n-p})| - \sum_{j=1}^{J} \frac{N_j}{2} \log |\text{Diag}\,(\theta_p^{\mathrm{T}} \bar{W}_j \theta_p)| + N \log |\theta| \right\}, \tag{26}$$

where $\hat{\theta}_{\mathrm{D}}$ is the estimator for θ_{D}; that is, it is the optimal transformation if we wish to use diagonal-variance Gaussian models for each class.

3.3. Σ with equal parameters

We finally consider the case where $\forall j,\ \Sigma_j = \Sigma$. Then, the maximum-likelihood parameter estimates can be written as follows:

$$\hat{\mu}_j^p = \theta_p^{\mathrm{T}} \bar{X}_j, \quad j = 1 \ldots J, \tag{27}$$

$$\hat{\mu}_0 = \theta_{n-p}^{\mathrm{T}} \bar{X}, \tag{28}$$

$$\hat{\Sigma}^p = \text{Diag}\,(\theta_p^{\mathrm{T}} \bar{W} \theta_p), \tag{29}$$

$$\hat{\Sigma}^{n-p} = \text{Diag}\,(\theta_{n-p}^{\mathrm{T}} \bar{T} \theta_{n-p}), \tag{30}$$

$$\hat{\theta}_{\mathrm{E}} = \arg \max_{\theta} \left\{ -\frac{N}{2} \log |\text{Diag}(\theta_{n-p}^{\mathrm{T}} \bar{T} \theta_{n-p})| - \frac{N}{2} \log |\text{Diag}(\theta_p^{\mathrm{T}} \bar{W} \theta_p)| + N \log |\theta| \right\}. \tag{31}$$

We obtain Eq. (31) by inserting the values of μ_j and Σ that maximize the likelihood, and then dropping the constant terms in the log-likelihood. We have shown (Kumar, 1997; Kumar and Andreou, submitted) that the solution that we obtain by taking the eigenvectors corresponding to largest p eigenvalues of $\bar{W}^{-1}\bar{T}$ also maximizes the expression in Eq. (31), thus asserting the claim that LDA is the maximum-likelihood parameter estimate of a constrained model.

4. Extension to hidden Markov models

We have proposed the theoretical framework to obtain optimal features for heteroscedastic Gaussian models in the previous sections, but we have not tackled yet the problem of assigning the feature vectors to

classes. In this section we propose a method to do so by extending the maximum-likelihood framework of feature-dimension reduction to HMMs. Such formulation is necessary in speech recognition, where the data form a time series of vectors.

To distinguish the time series from the independent identically distributed (i.i.d.) observations x, let us denote the time-series by o_j, $o_j = o_1, \ldots, o_T$, where T is the index of final time step. The initial state is assumed to be 1. The final state is assumed to be N. At a discrete time step, the model can make a transition from state S_i to state S_j based on the transition probability a_{ij}. When a state S_j is entered, an n-dimensional feature vector is emitted, based on an underlying distribution $P_j(O)$, which is assumed to be Gaussian. We use the expectation-maximization algorithm (Baum et al., 1970; Dempster et al., 1977) to find the parameters of an HMM. This algorithm is an iterative procedure that can be used if the observations are viewed as incomplete data. The underlying complete data should come from an exponential family. In the case of HMM, consider the state transitions as the missing data. Gaussian distributions are members of the exponential family of distributions. If the state transitions are known, we can compute the parameters of these distributions. Therefore, the EM algorithm can be applied to HMM.

If we wish to reduce the feature dimension, we have to make the assumption that *there is an* $(n-p)$*-dimensional linear subspace in which the* $P_j(O)$ *are exactly the same for all the states* S_j. For Gaussian distributions, this assumption is the same as that described in Section 3. The underlying distributions are still exponential. Therefore, the EM algorithm can still be used.

We will step through the algorithm. Let us assume, w.l.o.g., that the initial state and the final state are fixed and known. Then, such an HMM is specified completely by the state-transition probability matrix $A = [a_{ij}]$, and the state probability distributions $\{P_j(O)\}$ are specified by the set $\{\theta, \mu_0, \mu_j^p, \Sigma_j^p, \Sigma^{(n-p)}\}$. We find a maximum-likelihood estimate of the HMM parameters by embedding the optimization for θ in the maximization step of the EM algorithm (Baum et al., 1970; Dempster et al., 1977). In particular, for the models discussed in Section 3, the parameters to be estimated are $\{\theta, \mu_j, \Sigma_j\}$. Now there are additional parameters to be estimated: namely, $a_{i,j}$ the transition probabilities from state i to state j. If the state-transition sequence were known for the observed data, then it would be easy to estimate the transition probabilities. Since the state-transition sequence is not known, it is termed *missing data*. The expanded set of parameters is estimated in an iterative manner. We make an initial guess of the model parameters. We then estimate the probabilities of being in a particular state at each time step (missing data) on the basis of the training data, and using our initial guess of the values of $\{a_{i,j}, \theta, \mu_j, \Sigma_j\}$. This step is the *expectation* step. Then, we reestimate $\{a_{i,j}.\theta, \mu_j, \Sigma_j\}$ so as to maximize the likelihood of the training data plus the transition probabilities. This step is the *maximization* step.

We calculate the expected value of the transitions at any time using the forward–backward algorithm (Baum et al., 1970). Let the forward probability $\alpha_j(t)$ for some model M be defined as

$$\alpha_j(t) = P(o_1, \ldots, o_t, s(t) = j|M); \tag{32}$$

that is, it is the joint probability of observing the first t speech vectors and being in the state j at time t. This forward probability is calculated efficiently by the recursion

$$\alpha_j(t) = \left[\sum_{i=\text{all states}} \alpha_i(t-1)a_{ij}\right] b_j(o_t). \tag{33}$$

The backward probability is defined as

$$\beta_i(t) = P(o_{t+1}, \ldots, o_T|s(t) = j, M). \tag{34}$$

This backward probability can be computed efficiently by the backward recursion,

$$\beta_i(t) = \sum_{j=\text{all states}} a_{ij}b_j(o_{t+1})\beta_j(t+1). \tag{35}$$

Then, based on the forward and backward probabilities,

$$\alpha_j(t)\beta_j(t) = P(O, s(t) = j|M). \tag{36}$$

The likelihood of being in state j at time t is then

$$L_j(t) = P(s(t) = j|O, M) = \frac{P(O, s(t) = j|M)}{P(O|M)} = \frac{\alpha_j(t)\beta_j(t)}{P(O|M)}. \tag{37}$$

The normalizing constant $P(O|M)$ satisfies the condition $\sum_{\text{all states}} L_j(t) = 1$, and it can be shown that $P(O|M) = \alpha_N(T)$.

Once the likelihood of being in any state is known, the weighted means and sum of squares and products are calculated as

$$N_j = \sum_{t=1}^{T} L_j(t), \tag{38}$$

$$N = \sum_{\text{all states}} N_j, \tag{39}$$

$$\bar{N}_j = \frac{N_j}{N}, \tag{40}$$

$$\bar{X} = \frac{\sum_{t=1}^{T} o_t}{N}, \tag{41}$$

$$\bar{X}_j = \frac{\sum_{t=1}^{T} L_j(t) o_t}{N_j}, \tag{42}$$

$$\bar{W}_j = \sum_{t=1}^{T} \frac{L_j(t)}{N_j} (o_t - \bar{X}_j)(o_t - \bar{X}_j)^{\mathrm{T}}, \tag{43}$$

$$\bar{W} = \sum_{\text{all states}} \bar{N}_j \bar{W}_j, \tag{44}$$

$$\bar{T} = \sum_{t=1}^{T} \frac{1}{N} (o_t - \bar{X})(o_t - \bar{X})^{\mathrm{T}}. \tag{45}$$

We can then use the statistics in the above to perform the constrained optimization of Section 3 to find θ, μ_0, $\Sigma^{(n-p)}$ and $\{\mu_j^p, \Sigma_j^p\}$ $\forall S_j$. This procedure is repeated until convergence.

Note that mixture models are a special case of HMMs: they contain non-emitting initial and final states, and as many additional states as there are mixtures in the model (K). The transition probability from the initial state to a mixture state is a_{1j}, with the constraint that $\sum_{j=1}^{K} a_{1j} = 1$. Each of the mixture states makes a non-emitting transition into the final state. Due to this relationship, the generalization holds for mixture models as well.

5. Application of discriminant analysis to speech recognition

In this section, we discuss practical issues that arise when applying discriminant analysis to the speech recognition problem.

5.1. Feature vector to class assignment

If we approximate the largest value of $L_j(t)$ for any given t to 1, and the rest of the values to 0, then Eqs. (39)–(44) approximate to assigning the feature vector at time t, to the most likely state at time t. In case of mixture distributions, the assignment would be to the most likely mixture component. Due to some computational constraints in finding $\hat{\theta}_D$ and $\hat{\theta}_F$ for every iteration, we have used this approximation to the method developed in Section 4. Also, the computation of $\hat{\theta}_D$ and $\hat{\theta}_F$ is not performed every iteration.

Techniques closely related to the one suggested by the above approximation have been used previously by other researchers (Brown, 1987; Haeb-Umbach and Ney, 1992; Aubert et al., 1993; Siohan, 1995). Now, we describe here the method that we have used for assigning each feature vector to a class.

1. Estimate the HMM model parameters by using the n-dimensional training data and applying the Baum/Welch algorithm (Baum et al., 1970).
2. Use the model parameters and the segment transcriptions to determine the most likely sequence of HMM states for the training data.
3. Treat each state as a different class. Assign each feature to a class corresponding to the state from which it (most probably) came.
4. Now that the class labels are available, perform HDA to reduce the feature dimension. Apply the linear transformation θ_p, thus obtained, to the feature vectors. Also, use θ_p to transform the features at the time of recognition.
5. Estimate the HMM parameters for the new transformed and reduced features.
   We use the transformed features and the corresponding HMM parameters to evaluate performance on the test data.

5.2. Incorporating context information

Information about the dynamics of features, such as the feature derivative and accelerations, is known to improve the performance of speech recognition systems (Furui, 1986). These features are computed using standard numerical methods of polynomial regression. In the HDA framework, derivative and acceleration computation are treated as one of the possible linear operations. Let us define the *context size* C as the number of feature vectors before and after the current feature vector (each feature vector is n-dimensional) that are used to incorporate the dynamic information. We create a new feature vector of size $(2C + 1)n$ by concatenating all the feature vectors within the context size of the current feature. We then perform discriminant analysis on this new feature. Thus HDA makes the choice regarding what aspects of the dynamics are important, without the constraints of assuming the velocity and the acceleration to be the only relevant terms.

5.3. Selection of the feature dimension p

There is no formal method for choosing the feature dimension that would optimize the speech recognition performance. One good way to choose the appropriate feature dimension is through cross validation. This approach requires that we set aside a portion of the training data, and evaluate the model parameters on the basis of remaining data. Then we evaluate the recognizer performance on the set-aside data, and repeat this process of training and testing for different values of p. We select the dimension that gives the best recognition performance, and evaluate the model parameters for that value of p using the entire training data. Another possibility is to use a model selection criterion such as AIC or BIC (Akaike, 1974; Schwarz, 1978; Rissanen, 1989). These methods are good, but require large amount of computation to find the optimum HMM parameters, for each model that is considered.

Here, we attempt to use a method that does not require repeated HMM parameter estimation. Let us assume that all the features x_i are independent. Then, Bartlett has shown that the statistic

$$V = \left((N-1) - \frac{1}{2}(n+J)\right) \sum_{i=p+1}^{n} \log \hat{\lambda}_i \tag{46}$$

is approximately distributed as a χ^2 random variable with $(n-p)(J-p-1)$ degrees of freedom, where N is the total number of independent features, n the feature dimension, J the total number of classes and $\hat{\lambda}_i$ are the eigenvalues of $W^{-1}T$. He used the null hypothesis that the subspace spanned by the right eigenvectors of $W^{-1}T$ corresponding to the smallest eigenvalues λ_i (in the set $i = \{p+1 \ldots n\}$ does not contain any discrimination information and has a multivariate Gaussian distribution (Bartlett, 1947; Dillon and Goldstein, 1984).

We use this result to construct a hypothesis test for selecting the feature dimension. The null hypothesis H_0 states that the subspace that we wish to reject follows a multivariate Gaussian distribution, and does not contain any class-discrimination information (class means and variances are the same for all classes). Starting from the subspace corresponding to the smallest eigenvalue ($p = n - 1$), we perform a series of hypothesis tests, rejecting subspaces of increasingly larger dimension until the null hypothesis H_0 is rejected.

This method is suitable if all the class distributions were Gaussian with equal variance. However, in our case, this method has two problems. First, the class distributions may not be Gaussian. Second and more seriously, the consecutive feature vectors are not independent for speech signals. However, the distinct advantage of using this approach is that repeated parameter estimation and cross validation are not required. It is our conjecture that if the value of N in Eq. (46) is chosen to be a fraction of the total number of feature vectors, reasonable estimates of the best feature dimension may be obtained.

In our experiments we adopt the following approach for selecting the feature dimension p. Initially, without using any context information, we reduce the feature dimension to about 13 by applying HDA. We then use the class assignment obtained from this model to perform HDA using a context size $C = 1$. We choose N in Eq. (46) as *one-fourth* of the total number of feature vectors. This decision was made a priori, and was not derived from the test results presented here. We perform hypothesis tests, select a feature dimension, and evaluate HMM parameters for the selected feature dimension. We then use the class assignment obtained from this model to perform HDA for context size $C = 4$, and again, select a feature dimension.

This method does not require repeated estimation of HMM parameters – that task would be necessary if cross validation were to be used. Therefore, we can test quickly to evaluate the merit of any particular set of features. However, there is no reason to believe that the value of p obtained by this method gives the best recognition performance and therefore we also evaluate the recognition performance for several other values of feature dimension p, and we shall present the results in the form of a figure in Section 6.

6. Experiments on TI-DIGITS

We performed speech recognition experiments on the TI-DIGITS database using 12th order mel-cepstral features (with signal energy as an additional feature). We used only the isolated digits portion of the database. We removed the silence portions of the beginning and the end of each digit, by using the automatic silence detection algorithm proposed by Rabiner and Sambur (1975). Five state left-to-right word HMM models with single mixture component were used to model each digit. We used full covariance matrices to model each Gaussian distribution. We performed all the listed experiments, using the normal split of the database.

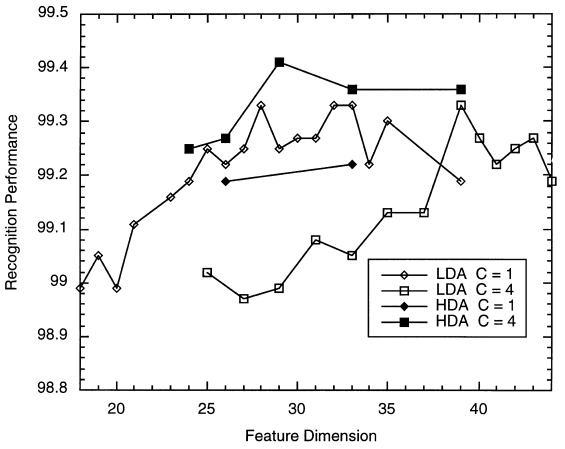

Fig. 1. Recognition performance on LDA and HDA transformed Mel-cepstrum features and full covariance matrix, unequal variance models. The context size is indicated by the number in the legend.

The standard technique, which uses the features and the derivative and acceleration information, gave a recognition performance of 99.33%. The feature vectors for this experiment were 39-dimensional (13 + 13 + 13).

The experimental results that we achieved by applying LDA and HDA are shown in Fig. 1. Reducing the features dimension by applying LDA did not improve the recognition performance. However, the same recognition performance was achieved with a lower feature dimension. The best error rates for the curves in Fig. 1 are summarized in Table 1. When HDA was applied using $C=4$, the recognition performance improved to 99.41% with only 29 features. This improvement corresponds to about a 12% reduction in error rate (21 errors instead of 24 in 3586 examples), and to about 25% fewer features.

When the method described in Section 5.3 is applied to the data to obtain the optimal feature dimension, the estimated feature dimension p is 26. Given that the context size is 4, the value of p could be any number, up to 117. Therefore, we must conclude that this method of dimension selection, does help in reducing our search for the optimal feature dimension.

We have also performed an experiment using diagonal variance matrices (Table 2). We test diagonal variance models because such models are employed in large vocabulary continuous speech recognition systems. When the standard technique of using derivatives and accelerations is applied, the recognition performance is only 96.54%, corresponding to an error rate of 3.36%. By applying the standard LDA method, we reduced the feature dimension to 29, and reduce the error rate to 2.29%. When we applied HDA, we could further reduce the error rate to 1.65%, which is less than half of the original error rate, and corresponds to about 30% fewer errors as compared to using LDA. In these experiments, the dimension choice of 29 came from using the method described in Section 5.3.

Table 1
Best recognition performances on LDA and HDA transformed Mel-cepstrum features and full covariance, unequal variance models

Method used	Feature dimension	Recognition error (%)
Δ and acceleration	39	0.67
LDA ($C=1$)	28	0.67
LDA ($C=4$)	39	0.67
HDA ($C=1$)	29	0.78
HDA ($C=4$)	29	0.59

Table 2
Best recognition performance on LDA and HDA transformed Mel-cepstrum features and diagonal covariance matrix, unequal variance models

Method used	Feature dimension	Recognition error (%)
Δ and acceleration	39	3.36
LDA ($C=4$)	29	2.29
HDA ($C=4$)	29	1.65

7. Discussion

The performance of a classifier depends on two important factors. The first factor is how well the classifier model is capable of representing the data. The second factor is the amount of data available to reliably estimate the parameters of the model.

Among the models used in Section 6, the models using full covariance matrices are the more general compared to the model that uses diagonal covariance matrices. Therefore, the full covariance models are capable of providing better performance, provided we make good estimates of the parameters. With the full covariance models, any linear transformation $y = \theta^{\mathrm{T}}x$ will only change the value of the means and the variances of the transformed variables. However, it cannot increase the likelihood of the model. We notice that when the feature dimension is reduced to a very small number (25 for example), the performance of the recognizer is worse than what it is in the original 39-dimensional feature space. We can justify this phenomenon by noticing that the corresponding models for reduced feature dimension are more restrictive, since they assume equal means and variances for all states in the rejected subspace. However, the performance improves slightly if HDA is used to reduce the feature dimension to 29. A possible explanation for this phenomenon may be that there is not enough classification information in the rejected subspace (means and variances of different classes are close to each other), and, therefore, by using a more constrained model, we are able to get better parameter estimates, and hence better performance. In this situation, HDA does a better job of finding the right subspace for rejection, as compared to LDA.

The situation is different when the models use diagonal covariance matrices. In this case, if the features are correlated within a class, the correlation may be reduced by introducing a linear transformation defined by θ, thus increasing the likelihood of the model. Therefore, when HDA is applied, much greater improvements in performance are observed. A model selection criterion that weighs the increase in likelihood against the number of free parameters can also be used to choose between models, and is discussed in (Kumar, 1997).

8. Conclusions

We have presented the formal technique of HDA and investigated its application to continuous speech recognition using HMM models. Our experiments on the TI-DIGITS are encouraging and they warrant further investigation on larger and more difficult tasks, for the potential of improvement in performance, and reduction of the computational load through reduced feature dimension.

Acknowledgements

Chalapathy Neti has shown us the way to discriminant analysis for speech recognition, Carey Priebe has helped on the way and, fortunately, Fred Jelinek has stayed out of our way. Hynek Hermansky has pointed

us to the not so well-known paper by Melvyn Hunt (Hunt, 1979). G.L. Meyer has critically reviwed the theoretical sections of the paper and the comments of the three anonymous reviewers helped improve the clarity of the paper.

This research was supported by Lockheed Martin; the personal interest and support of Dr. John Damoulakis is gratefully acknowledged. Additional support was provided by the Center for Language and Speech Processing at Johns Hopkins University, and by the National Science Foundation Neuromorphic Engineering Research Center at Caltech.

References

Akaike, H., 1974. A new look at the statistical identification model. IEEE Transactions on Automatic Control 19, 716–723.

Aubert, X., Haeb-Umbach, R., Ney, H., 1993. Continuous mixture densities and linear discriminant analysis for improved context-dependent acoustic models. In: Proc. of ICASSP, Vol. 2, pp. 648–651.

Ayer, C.M., 1992. A discriminatively derived transform capable for improved speech recognition accuracy. Ph.D. Thesis, University of London.

Ayer, C.M., Hunt, M.J., Brookes, D.M., 1993. A discriminately derived linear transform for improved speech recognition. In: Proc. Eurospeech 93, Vol. 1, pp. 583–586.

Bartlett, M.S., 1947. Multivariate analysis. J. Roy. Statist. Soc. B 9, 176–197.

Baum, L.E., Petrie, T., Soules, G., Weiss, N., 1970. A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Stat. 41, 164–171.

Bocchieri, E.L., Wilpon, J.G., 1993. Discriminative feature selection for speech recognition. Computer Speech and Language 7, 229–246.

Brown, P.F., 1987. The acoustic-modelling problem in automatic speech recognition. Ph.D. Thesis, Carnegie Mellon University.

Campbell, N., 1984. Canonical variate analysis – a general formulation. Australian Journal of Statistics 26, 86–96.

Cohen, J.R., 1989. Application of an auditory model to speech recognition. J. Acoust. Soc. Amer. 85, 2623–2629.

Davis, S.B., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentances. IEEE Transactions on Acoustics, Speech, and Signal Processing 28 (4), 357–366.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via EM algorithm. J. Roy. Statist. Soc. 1–38.

Dillon, W.R., Goldstein, M., 1984. Multivariate Analysis. Wiley, New York.

Doddington, G., 1989. Phonetically sensitive discriminants for improved speech recognition. In: Proceedings 1989 ICASSP, no. S10 b.11, pp. 556–559.

Duda, R.O., Hart, P.B., 1973. Pattern Classification and Scene Analysis. Wiley, New York.

Engle, R.F., 1995. ARCH: Selected Readings. Oxford Univ. Press, Oxford.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 7, 179.

Fisher, R.A., 1938. The statistical utilization of multiple measurements. Ann. Eugen. 8, 376.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition. Academic Press, New York.

Furui, S., 1986. Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Transactions on Acoustics, Speech, and Signal Processing 34, 52–59.

Haeb-Umbach R., Ney H., 1992. Linear discriminant analysis for improved large vocabulary continuous speech recognition. In: Proc. ICASSP, Vol. 1, pp. 13-16.

Haeb-Umbach, R., Geller, D., Ney, H., 1993. Improvement in connected digit recognition using linear discriminant analysis and mixture densities. In: Proceedings of ICASSP, pp. 239–242.

Hastie, T., Tibshirani, R., 1994. Discriminant analysis by gaussian mixtures. Tech. Rep., AT&T Bell Laboratories.

Hermansky, H., 1990. Perceptual linear predictive (plp) analysis of speech. J. Acoust. Soc. Amer. 87, 1738–1752.

Hunt, M., 1979. A statistical approach to metrics for word and syllable recognition. In: 98th Meeting of the Acoustical Society of America, November.

Hunt M.J., Lefebvre C., 1989. A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. In: Proc. ICASSP, Vol. 1, pp. 262-265.

Jankowski Jr., C.R., 1992. A comparison of auditory models for automatic speech recognition. Master's Thesis, MIT.

Kumar, N., 1997. Investigation of silicon auditory models and generalization of linear discriminant analysis for improved speech recognition. Ph.D. Thesis, Johns Hopkins University, http://olympus.ece.jhu.edu/archives/phd/nkumar97/index.html.

Kumar, N., Andreou, A., 1996a. On generalizations of linear discriminant analysis. Tech. Rep., Electrical and Computer Engineering Technical Report-96-07, April.

Kumar, N., Andreou, A., 1996b. Generalization of linear discriminant analysis in maximum likelihood framework. In: Proceedings of Joint Meeting of American Statistical Association, Chicago, IL, August.

Kumar, N., Andreou, A., submitted. Heteroscedastic discriminant analysis: maximum likelihood feature extraction for heteroscedastic models. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Kumar, N., Neti, C., Andreou, A., 1995. Application of discriminant analysis to speech recognition with auditory features. In: Proceedings of the 15th Annual Speech Research Symposium, Johns Hopkins University, Baltimore, MD, pp. 153–160, June.

Rabiner, L., Sambur, M., 1975. An algorithm for determining the endpoints of isolated utterances. Bell Syst. Tech. J. 54, 297–315.

Rao, C.R., 1965. Linear Statistical Inference and its Applications. Wiley, New York.

Rissanen, J., 1989. Stochastic Complexity in Statistical Inquiry. Series in Computer Science, Vol. 15. World Scientific, Singapore.

Roth R., Baker J.K., Baker J.M., Gillick L., Hunt M.J., Ito Y., Loewe S., Orloff J., Peskin B., Scattone F., 1993. Large vocabulary continuous speech recognition of wall street journal data. In: Proc. ICASSP, Vol. 2, pp. 640–643.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6 (2), 461–464.

Siohan O., 1995. On the robustness of linear discriminant analysis as a preprocessing step for noisy speech recognition. In: Proc. ICASSP, Vol. 1, pp. 125–128.

Sun, D., 1997. "Feature dimensionality reduction using reduced-rank maximum likelihood estimation for hidden Markov models." In: International Conference on Language and Speech, pp. 244–247.

Wood L., Pearce D., Novello F., 1991. Improved vocabulary-independent sub-word HMM modelling. In: Proc. ICASSP, Vol. 1, pp. 181–184.

Woodland P.C., Cole D.R., 1991. Optimising hidden markov models using discriminative output distribution. In: Proc. ICASSP, Vol. 1, pp. 545–548.

Yu, G., Russell, W., Schwartz, R., Makhoul, J., 1990. Discriminant analysis and supervised vector quantization for continuous speech recognition. In: Proceedings of ICASSP, pp. 685–688, April.

Zahorian S.A., Qian D., Jagharghi A.J., 1991. Acoustic-phonetic transformations for improved speaker-independent isolated word recognition. In: Proc. ICASSP, Vol. 1, pp. 561–564.