MAXIMUM LIKELIHOOD DISCRIMINANT FEATURE SPACES

George Saon, Mukund Padmanabhan, Ramesh Gopinath and Scott Chen

IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598
E-mail: {saon,mukund,rameshg,schen}@watson.ibm.com, Phone: (914)-945-2985

ABSTRACT

Linear discriminant analysis (LDA) is known to be inappropriate for the case of classes with unequal sample covariances. In recent years, there has been an interest in generalizing LDA to heteroscedastic discriminant analysis (HDA) by removing the equal within-class covariance constraint. This paper presents a new approach to HDA by defining an objective function which maximizes the class discrimination in the projected subspace while ignoring the rejected dimensions. Moreover, we will investigate the link between discrimination and the likelihood of the projected samples and show that HDA can be viewed as a constrained ML projection for a full covariance gaussian model, the constraint being given by the maximization of the projected between-class scatter volume. It will be shown that, under diagonal covariance gaussian modeling constraints, applying a diagonalizing linear transformation (MLLT) to the HDA space results in increased classification accuracy even though HDA alone actually degrades the recognition performance. Experiments performed on the Switchboard and Voicemail databases show a 10%-13% relative improvement in the word error rate over standard cepstral processing.

1. INTRODUCTION

State-of-the-art speech recognition systems use cepstral features augmented with dynamic information from the adjacent speech frames. The standard MFCC+Δ+$\Delta\Delta$ scheme, while performing relatively well in practice, has no real basis of existence from a discriminant analysis point of view. The same argument applies for the computation of the cepstral coefficients from the spectral features: it is not clear that the discrete cosine transform, among all linear transformations, has the best discriminatory properties even if its use is motivated by orthogonality considerations.

Linear discriminant analysis [3, 4] is a standard technique in statistical pattern classification for dimensionality reduction with a minimal loss in discrimination. Its application to speech recognition has shown consistent gains for small vocabulary tasks and mixed results for large vocabulary applications [7, 11, 8]. One reason could be because of the diagonal modeling assumption that is imposed on the acoustic models in most systems: if the dimensions of the projected subspace are highly correlated then a diagonal covariance modeling constraint will result in distributions with large overlap and low sample likelihood. In this case, a maximum likelihood feature space transformation [6, 5] which aims at minimizing the loss in likelihood between full and diagonal covariance models is known to be very effective. Secondly, it is not clear what the best definition for the classes should be: phone, subphone, allophone or even prototype-level classes can be considered [7]. Related to this argument, the class assignment procedure has an impact on the performance of LDA; EM-based approaches which aim at jointly optimizing the feature space transformation and the model parameters have been proposed [11, 8, 5].

Chronologically, the extension of LDA to HDA under the maximum likelihood framework appears to have been proposed first by Schukat-Talamazzini [11] (called maximum likelihood rotation). Kumar [8] studied the case for diagonal covariance modeling and general (not necessarily orthogonal) transformation matrices and made the connection with LDA. Following an argument of Campbell [1], he showed that HDA is a maximum likelihood solution for normal populations with common covariances in the rejected subspace. In [6], a maximum likelihood linear transform (MLLT) was introduced which turns out to be a particular case of Kumar's HDA when the dimensions of the original and the projected space are the same. Interestingly, Gales' global transform for semi-tied covariance matrices [5] is identical to MLLT but applied in the model space (all other cases are feature space transforms). Finally, Demuynck [2] uses a minimum divergence criterion between posterior class distributions in the original and transformed space to estimate an HDA matrix. We will make further references to these approaches and their relation to our work throughout the paper.

The paper is organized as follows: in section 2 we will briefly recall the basics of LDA and introduce the HDA extension. Section 3 will describe the experimental results and section 4 will provide a final discussion.

We would like to acknowledge the support of DARPA under Grant MDA972-97-C-0012 for funding part of this work.

0-7803-6293-4/00/$10.00 ©2000 IEEE.

2. FROM LDA TO HDA

2.1. Linear discriminant analysis

Consider a set of N independent vectors $\{x_i\}_{1\leq i\leq N}$, $x_i \in \mathbb{R}^n$, each of the vectors belonging to one and only one class $j \in \{1,\ldots,J\}$ through the surjective mapping of indices $l : \{1,\ldots,N\} \to \{1,\ldots,J\}$. Let each class j be characterized by its own mean μ_j, covariance Σ_j, and sample count N_j, where the standard definitions hold:

$$\mu_j = \frac{1}{N_j}\sum_{i\in l^{-1}(j)} x_i, \qquad \Sigma_j = \frac{1}{N_j}\sum_{i\in l^{-1}(j)} x_i x_i^T - \mu_j\mu_j^T$$

and $\sum_{j=1}^{J} N_j = N$. The class information is condensed into 2 scatter matrices called:

- *within-class* scatter: $W = \frac{1}{N}\sum_{j=1}^{J} N_j\Sigma_j$ and
- *between-class* scatter: $B = \frac{1}{N}\sum_{j=1}^{J} N_j\mu_j\mu_j^T - \overline{\mu}\overline{\mu}^T$

The goal of LDA is to find a linear transformation $f : \mathbb{R}^n \to \mathbb{R}^p$, $y = f(x) = \theta x$, with θ a $p \times n$ matrix of rank $p \leq n$, such that the following ratio of determinants is maximized:

$$J(\theta) = \frac{|\theta B\theta^T|}{|\theta W\theta^T|} \tag{1}$$

Even though the objective function in (1) is non-linear, there is a closed form solution given by the transposed eigenvectors corresponding to the p largest eigenvalues of the generalized eigenvalue problem: $Bx = \lambda Wx$ (a proof and an extensive discussion of LDA can be found in [4]).

2.2. Heteroscedastic extension

Let us consider the individual weighted contributions of the classes to the objective function:

$$\prod_{j=1}^{J}\left(\frac{|\theta B\theta^T|}{|\theta\Sigma_j\theta^T|}\right)^{N_j} = \frac{|\theta B\theta^T|^N}{\prod_{j=1}^{J}|\theta\Sigma_j\theta^T|^{N_j}} \tag{2}$$

or, by taking log and rearranging terms, we get:

$$H(\theta) \triangleq \sum_{j=1}^{J} -N_j \log|\theta\Sigma_j\theta^T| + N\log|\theta B\theta^T| \tag{3}$$

H has two very useful properties of invariance. For every nonsingular matrix $\psi \in \mathbb{R}^{p\times p}$, $H(\psi\theta) = H(\theta)$. This means that subsequent feature space transformations of the range of θ will not affect the value of the objective function. Moreover, like LDA, the HDA solution is invariant to linear transformations of the data in the original space.[1] A second remark is that no special provisions have to be made for θ during the optimization of H except for $|\theta\theta^T| \neq 0$; the objective function is invariant to row or column scalings of θ or eigenvalue scalings of $\theta\theta^T$. Using matrix differentiation results from [10], the derivative of H is given by:

$$\frac{\mathrm{d}H(\theta)}{\mathrm{d}\theta} = \sum_{j=1}^{J} -2N_j(\theta\Sigma_j\theta^T)^{-1}\theta\Sigma_j + 2N(\theta B\theta^T)^{-1}\theta B \tag{4}$$

Unfortunately, $H'(\theta) = 0$ has no analytical solution for the stationary points. Instead, we used a quasi-Newton conjugate gradient descent routine from the NAG[2] Fortran library for the optimization of H. Figure 1 illustrates a case where LDA and HDA provide very different answers.

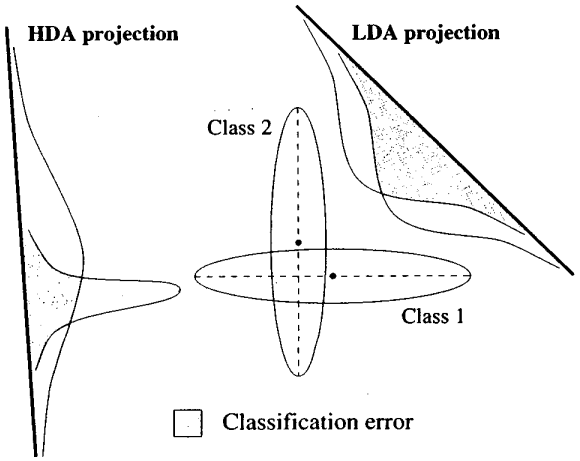

Figure 1: Difference between LDA and HDA.

For $J = 2$ classes with $N_1 = N_2$, there is an interesting connection between H and the Bhattacharyya distance [4] in the range of θ, $\rho(\theta)$, namely $2\rho(\theta) > H(\theta)/N$, where:

$$\rho(\theta) = \frac{1}{2}\operatorname{trace}\left\{(\theta W\theta^T)^{-1}\theta B\theta^T\right\} + \frac{1}{2}\log\frac{|\theta W\theta^T|}{\sqrt{|\theta\Sigma_1\theta^T||\theta\Sigma_2\theta^T|}} \tag{5}$$

Since $e^{-\rho(\theta)}$ is an upper bound on the Bayes classification error rate, $e^{-\frac{H(\theta)}{2N}}$ becomes a (looser) upper bound too. Therefore, maximizing H amounts to minimizing this bound and, hopefully, the error rate.

[1] The invariance is in the following sense: if $x \overset{\phi}{\mapsto} z$, $\phi \in \mathbb{R}^{n\times n}$ nonsingular and $\hat{\theta} = \operatorname{argmax}_\theta H_z(\theta)$ then $\hat{\theta}\phi = \operatorname{argmax}_\theta H_x(\theta)$.

[2] Numerical Algebra Group

2.3. Likelihood interpretation

Consider $\{y_i \in \mathbb{R}^p \mid y_i = \theta x_i,\ 1 \le i \le N\}$, the projected samples onto the HDA space. Assuming a single full covariance gaussian model for each class, the log likelihood of these samples according to the induced ML model $\hat{\mu}_j = \theta\mu_j$ and $\hat{\Sigma}_j = \theta\Sigma_j\theta^T$, $1 \le j \le J$, is:

$$\sum_{j=1}^{J} -\frac{N_j}{2}\log|\hat{\Sigma}_j| - \frac{Np}{2}\log(2\pi) = \sum_{j=1}^{J} -\frac{N_j}{2}\log|\theta\Sigma_j\theta^T| + C \tag{6}$$

It may be seen that the summation in H is related to the log likelihood of the projected samples. Thus, θ can be interpreted as a constrained ML projection, the constraint being given by the maximization of the projected between-class scatter volume.

Next, let us consider the case when diagonal variance modeling constraints are present in the final feature space. The maximum likelihood linear transform [6, 5] aims at minimizing the loss in likelihood between full and diagonal covariance gaussian models. The objective is to find a transformation ψ that maximizes the log likelihood difference of the data, i.e.,

$$\begin{aligned}\hat{\psi} &= \underset{\psi\in\mathbb{R}^{p\times p}}{\operatorname{argmax}} \sum_{j=1}^{J} -\frac{N_j}{2}\left(\log|\operatorname{diag}(\psi\hat{\Sigma}_j\psi^T)| - \log|\psi\hat{\Sigma}_j\psi^T|\right) \\ &= \underset{\psi\in\mathbb{R}^{p\times p}}{\operatorname{argmax}} \sum_{j=1}^{J} -\frac{N_j}{2}\log|\operatorname{diag}(\psi\theta\Sigma_j\theta^T\psi^T)| + N\log|\psi|\end{aligned} \tag{7}$$

Recall that, based on our earlier argument, HDA is invariant to subsequent feature space transformations, hence the objective function (3) is the same for the composite transform $\psi\theta$ as for θ. We will refer to this composite transform as the *maximum likelihood discriminant* (or MLD) projection. An important observation is that ψ in (7) does not necessarily have to be square. By means of (3) and (7), one could combine an HDA and an MLLT-type projection through the following function (to be maximized):

$$G(\theta) = \sum_{j=1}^{J} -N_j\log|\operatorname{diag}(\theta\Sigma_j\theta^T)| + N\log|\theta B\theta^T| \tag{8}$$

We will refer to this scheme as the diagonal HDA (or DHDA) projection. Related to this, Kumar [8] defined the following feature space transformation: $g : \mathbb{R}^n \to \mathbb{R}^n$, $y = g(x) = \bar{\theta}x$, where $\bar{\theta} = \left[\theta_{(p)}^T \theta_{(n-p)}'^T\right]^T$ is partitioned into two matrices corresponding respectively, to the projected and the rejected subspace. The objective is to maximize $K(\bar{\theta})$, the likelihood of the transformed samples subject to diagonal modeling constraints and common covariance in the rejected subspace:

$$K(\bar{\theta}) = \sum_{j=1}^{J} -N_j\log|\operatorname{diag}(\theta\Sigma_j\theta^T)| - N\log|\operatorname{diag}(\theta' T\theta'^T)| + 2N\log|\bar{\theta}| \tag{9}$$

where $T = W + B$, represents the total scatter of the data. Even though (8) bears a strong resemblance to (9), they provide different answers in practice because maximizing the DHDA objective function is directly related to maximizing the between class separation in the projected space, $|\theta B\theta^T|$, whereas for Kumar's HDA the assumption is that this is done implicitly by minimizing $|\operatorname{diag}(\theta' T\theta'^T)|$. Indeed, the two quantities are related through the following chain of inequalities[3]:

$$|\operatorname{diag}(\theta' T\theta'^T)| \ge |\theta' T\theta'^T| \ge |\theta' B\theta'^T| \ge \frac{|\bar{\theta}|^2|B|}{|\theta B\theta^T|} \tag{10}$$

It follows that $G(\theta) \ge K(\bar{\theta}) + N|B|$ for all $\bar{\theta} = [\theta^T\theta'^T]^T$.

3. EXPERIMENTS AND RESULTS

The first set of experiments was conducted on a voicemail transcription task [9]. The baseline system has 2313 context dependent states and 134K diagonal gaussian mixture components. The test set consists of 86 messages (approximately 7000 words). The baseline system uses 39-dimensional frames (13 cepstral coefficients plus deltas and double deltas computed from 9 consecutive frames). For the LDA and HDA versions, every 9 consecutive 24-dimensional cepstral vectors were spliced together forming 216-dimensional feature vectors which were then clustered to make possibly multiple full covariance models for each state (totaling around 3K gaussians). Subsequently, a 39×216 transformation, θ, was computed using the objective functions for LDA (1), HDA (3), and DHDA (8), which projected the models and feature space down to 39 dimensions. As mentioned in [7], it is not clear what the most appropriate class definition for LDA and HDA should be. The best results were obtained by considering each individual gaussian as a separate class, with the priors of the gaussians summing up to one for every state. After the computation of the LDA or HDA projection, the vectors were reclustered in the projected space to form a new set of 39-dimensional full covariance models (13.5K). An MLLT transformation, ψ, was then computed to maximize the objective function (7), leading to a

[3] Here we have used Hadamard's inequality $|\operatorname{diag}(A)| \ge |A|$ valid for any symmetric positive definite (SPD) matrix A. The second and third inequalities follow from observing that $|A + B| \ge |A|$ for SPD matrices A and B and by representing $|\bar{\theta}B\bar{\theta}^T| = |\theta B\theta^T||\theta' B\theta'^T - \theta' B\theta^T(\theta B\theta^T)^{-1}\theta B\theta'^T|$ (according to [10]).

Diagonal covariance (134K prototypes)		
System	Impr. obj. fn.	WER
Baseline (MFCC)	–	39.61%
MFCC+MLLT	5.35%	37.33%
LDA	–	39.60%
LDA+MLLT	2.01%	36.63%
HDA	1.14%	40.22%
HDA+MLLT (MLD)	3.94%	35.62%
DHDA	7.34%	37.11%
Full covariance (16.5K prototypes)		
System	WER	
Baseline (MFCC)	37.72%	
LDA	39.68%	
HDA	36.22%	

Table 1: Word error rates and objective function improvements for voicemail.

System	Dim.	Impr. obj. fn.	WER
Baseline (MFCC)	39	–	45.80%
LDA	60	–	43.16%
LDA+MLLT	60	2.10%	40.46%
HDA	60	6.23%	54.89%
HDA+MLLT (MLD)	60	15.26%	39.67%
DHDA	60	8.67%	40.66%

Table 2: Word error rates and objective function improvements for Switchboard.

composite LDA+MLLT and HDA+MLLT (MLD) feature space. The HDA and the DHDA optimizations were initialized with the LDA matrix. The parameters of the baseline system (with 134K gaussians) were then re-estimated in the tranformed spaces. Table 1 summarizes the improvements in the objective functions and the word error rates for the different systems. In order to assess the effectiveness of the HDA transform alone, we have also trained full covariance systems in the different spaces whose results are indicated in the bottom half of Table 1.

The second set of experiments was performed on the Switchboard database. The baseline system has 2801 context dependent states, 175K diagonal covariance prototypes and was trained on 70 hours of data (the '95 training set). Among the notable differences with the voicemail system is the dimensionality of the feature vectors in the transformed space (60 versus 39) and the use of right context across word boundaries during the search. The test set contains 10 randomly selected conversation sides from the CLSP WS'97 dev test set which has 25 conversations in all. Table 2 provides a comparison between the different techniques.

4. DISCUSSION

Based on the previous results, the following conclusions may be drawn: (i) Considering the individual covariances of the classes in the objective function leads to better discrimination. (ii) However, the clusters are skewed in the HDA space and it is necessary to "rectify" them by computing a subsequent diagonalizing transformation. (iii) Applying a maximum likelihood transform after the HDA projection is more efficient than incorporating the diagonal modeling assumption in the DHDA objective function.

5. REFERENCES

[1] N. A. Campbell. Canonical variate analysis - a general model formulation. *Australian Journal of Statistics*, 26(1):86–96, 1984.

[2] K. Demuynck, J. Duchateau and D. V. Compernolle. Optimal feature sub-space selection based on discriminant analysis. *Proceedings of EUROSPEECH'99*, Budapest, Hungary, 1999.

[3] R. O. Duda and P. B. Hart. Pattern classification and scene analysis. *Wiley*, New York, 1973.

[4] K. Fukunaga. Introduction to statistical pattern recognition. *Academic Press*, New York, 1990.

[5] M. J. F. Gales. Semi-tied covariance matrices for hidden Markov models. *IEEE Transactions on Speech and Audio Processing*, 7:272–281, 1999.

[6] R. A. Gopinath. Maximum likelihood modeling with gaussian distributions for classification. *Proceedings of ICASSP'98*, Seattle, 1998.

[7] R. Haeb-Umbach and H. Ney. Linear Discriminant Analysis for improved large vocabulary continuous speech recognition. *Proceedings of ICASSP'92*, volume 1, pages 13–16, 1992.

[8] N. Kumar and A. G. Andreou. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Communcation*, 26:283–297, 1998.

[9] M. Padmanabhan, G. Saon, S. Basu, J. Huang and G. Zweig. Recent improvements in voicemail transcription. *Proceedings of EUROSPEECH'99*, Budapest, Hungary, 1999.

[10] S. R. Searle. Matrix algebra useful for statistics. *Wiley Series in Probability and Mathematical Statistics*, New York, 1982.

[11] E. G. Schukat-Talamazzini, J. Hornegger and H. Niemann. Optimal linear feature space transformations for semi-continuous hidden Markov models. *Proceedings of ICASSP'95*, 369–372, 1995.