Maximum Conditional Mutual Information Projection For Speech Recognition

Mohamed Kamal Omar, Mark Hasegawa-Johnson

Department of Electrical And Computer Engineering,
University of Illinois at Urbana-Champaign,
Urbana, IL 61801

omar,jhasegaw@uiuc.edu

Abstract

Linear discriminant analysis (LDA) in its original model-free formulation is best suited to classification problems with equal-covariance classes. Heteroscedastic discriminant analysis (HDA) removes this equal covariance constraint, and therefore is more suitable for automatic speech recognition (ASR) systems. However, maximizing HDA objective function does not correspond directly to minimizing the recognition error. In its original formulation, HDA solves a maximum likelihood estimation problem in the original feature space to calculate the HDA transformation matrix. Since the dimension of the original feature space in ASR problems is usually high, the estimation of the HDA transformation matrix becomes computationally expensive and requires a large amount of training data. This paper presents a generalization of LDA that solves these two problems. We start with showing that the calculation of the LDA projection matrix is a maximum mutual information estimation problem in the lower-dimensional space with some constraints on the model of the joint conditional and unconditional probability density functions (PDF) of the features, and then, by relaxing these constraints, we develop a dimensionality reduction approach that maximizes the conditional mutual information between the class identity and the feature vector in the lower-dimensional space given the recognizer model.

Using this approach, we achieved 1% improvement in phoneme recognition accuracy compared to the baseline system. Improvement in recognition accuracy compared to both LDA and HDA approaches is also achieved .

1. Introduction

One of the main objectives of speech signal analysis in ASR systems is to produce a parameterization of the speech signal that reduces the amount of data that is presented to the speech recognizer, and captures salient characteristics suited for discriminating among different speech units. Most ASR systems use cepstral features augmented with dynamic information from the adjacent speech frames. The algorithms for cepstral features estimation use concepts based on human speech perception like Mel-frequency scaling and critical band filters to simulate the front-end of the human auditory system. Even with additional techniques for speaker normalization and combating environmental noise, incorporating properties of human speech production and auditory perception is not necessarily the optimal approach to feature extraction for speech recognition, as they are not optimized to discriminate among speech units.

Most dimensionality reduction techniques applied to speech recognition are variants or extensions of linear discriminant analysis (LDA) [1]. The results reported on the application of LDA to speech recognition show consistent gain for small vocabulary tasks and mixed results for large vocabulary applications [2]. This can be attributed mainly to making assumptions about the problem that are unrealistic like equal class-conditional covariance matrices, and using an optimality criterion that is not necessarily consistent with the objective of minimizing the recognition error. It was shown that linear discriminant analysis is related to the maximum likelihood estimation of parameters for a Gaussian model, with *a priori* assumptions on the structure of the model [3]. This result is further generalized by assuming that class distributions are a mixture of Gaussians [4]. In [2], LDA is generalized to the case of classes of different covariance matrices and this generalization is referred to as heteroscedastic discriminant analysis (HDA). An alternative interpretation of HDA as a constrained maximum likelihood projection for a Gaussian model is introduced in [5].

The objective function in all these methods is not directly related to minimizing the recognition error, and therefore does not necessarily minimize the discrimination loss due to dimensionality reduction. LDA transformation, for example, tends to preserve distances of already well-separated classes [6]. Maximizing the mutual information between the features and the class is more intuitively related to minimizing the recognition error, and therefore we argue that it is a better objective for discriminant analysis than maximizing the likelihood under some model assumptions or constraints.

In this paper, we show that calculating the LDA transformation matrix is a maximum conditional mutual information estimation (MCMIE) problem with constraints on both the class-conditional and the unconditional PDFs. By relaxing these constraints, we present a generalization of LDA to MCMI projection (MCMIP), and describe an algorithm that calculates the MCMIP transform given the recognizer model. This generalization has three advantages: it maximizes the *a posteriori* probability of the model corresponding to the training data given the data which is closely related to minimizing the training data recognition error, it is calculated in the lower-dimensional space, and it takes into consideration the assumptions of the recognizer model. In the next section, discriminant analysis approaches are discussed and the LDA approach is formulated as an MCMIE problem. The MCMIP method is described in section 3 and an iterative algorithm is introduced to estimate the MCMIP transform and the parameters of the recognizer. Then, recognition experiments are described in section 4. Finally, section 5 provides discussion of the results and a summary of this work. In this paper, a superscript is used as an index of a realization of the random vector. Capital letters are used to denote the random variables and the corresponding small letters to denote their realizations.

2. Discriminant Analysis

An interpretation of LDA that relates LDA to the conditional mutual information between the lower-dimensional feature vector and the class identity is introduced in this section.

2.1. Linear Discriminant Analysis

The linear discriminant analysis (LDA) technique tries to improve the linear separability of the classes by finding the linear transform that maximizes the ratio of the determinant of between-class covariance and the determinant of the average within-class covariance [1]. Given a set of N independent n-dimensional observation vectors $\{x^i\}_{1 \leq i \leq N}, x^i \in \Re^N$, each of them belongs to only one class $j \in 1, \cdots, J$. Let each class j be characterized by its mean μ_j, covariance matrix Σ_j, and observation count N_j, where

$$\mu_j = \frac{1}{N_j} \sum_{i:c^i=c_j} x^i,$$

$$\Sigma_j = \frac{1}{N_j} \sum_{i:c^i=c_j} x^i {x^i}^T - \mu_j \mu_j^T,$$

c^i is the class corresponding to the ith frame. The within-class scatter is given by

$$W \quad = \quad \frac{1}{N} \sum_{j=1}^{J} N_j \Sigma_j, \tag{1}$$

and the between-class scatter is given by

$$B \quad = \quad \frac{1}{N} \sum_{j=1}^{J} N_j \mu_j \mu_j^T - \mu \mu^T, \tag{2}$$

where μ is the global mean of the observations. The goal of LDA is to find a linear transformation characterized by the $p \times n$ matrix θ, for $p < n$, such that

$$J(\theta) \quad = \quad \frac{|\theta B \theta^T|}{|\theta W \theta^T|}, \tag{3}$$

is maximized.

The maximization can be formulated as principal component analysis of the Fisher covariance matrix or as a maximum likelihood estimation problem [2]. This is achieved by using the work by Campbell [3] who has shown that linear discriminant analysis is related to the maximum likelihood estimation of parameters for a Gaussian model, with *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 where the LDA mapping is represented by $p \times n$ matrix. The second assumption is that the within-class variances are equal for all classes.

2.2. Maximum Mutual Information Interpretation of LDA

There are several possible class separability measures. One of the most general measures of the ability of the features to discriminate among classes is its mutual information with the classes. Mutual information is an invariant measure under any one-to-one transformation. Therefore, for a full-rank linear transform of the $n \times 1$ feature vector x,

$$\begin{bmatrix} y \\ z \end{bmatrix} \quad = \quad \begin{bmatrix} \theta \\ \psi \end{bmatrix} x, \tag{4}$$

where y is $p \times 1$ vector, z is $(n-p) \times 1$ vector, θ is $p \times n$ matrix, and ψ is $(n-p) \times n$ matrix,

$$I(Y, C) \quad \leq \quad I(X, C), \tag{5}$$

with equality if and only if $I(Z, C) = 0$. This happens if and only if the feature vector Z is statistically independent of the class identity C [7]. Therefore, we should expect that getting rid of these features will have negligible effect on the recognizer performance or even improve it, if it has a negligible mutual information with the class identities. The mutual information between the feature vector Y and the set of classes C is

$$I(Y, C) \quad = \quad E_{P(Y,C)} \left[\log \frac{P(y|c)}{P(y)} \right], \tag{6}$$

where $\{P(y|c_j)\}_{j=1}^{J}$, and $P(y)$ are the class-conditional and the unconditional PDFs respectively. Since we do not have the true PDFs, we calculate an estimate of the mutual information, which is the conditional mutual information given a maximum likelihood estimate of the parameters Λ of both $\{P(y|c_j)\}_{j=1}^{J}$, and $P(y)$

$$\hat{I}(Y, C|\Lambda) \quad = \quad \sum_{i=1}^{N} \log \frac{P(y^i|c^i, \Lambda)}{P(y^i|\Lambda)}, \tag{7}$$

where N is the number of training frames, and c^i is the class corresponding to the ith frame.

Our goal here is to show that LDA is equivalent to the problem of finding the linear transformation matrix θ that maximize the conditional mutual information between the lower-dimensional feature vector Y and the class identity C with *a priori* assumptions on the structure of the model. Let each class-conditional PDF in the lower-dimensional space be modeled by a Gaussian PDF with all of them sharing the same covariance matrix

$$P(y|c_j) \quad = \quad \frac{1}{(2\pi)^{\frac{n}{2}} |W_y|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (y - \mu_j)^T W_y^{-1} (y - \mu_j) \right), \\ \text{for } j = 1, \cdots J, \tag{8}$$

where W_y is the maximum-likelihood estimate (MLE) of the class-conditional covariance matrix, μ_j is the MLE of the mean. Let also the unconditional PDF in the lower-dimensional space be modeled by a Gaussian PDF

$$P(y) \quad = \quad \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_y|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (y - \mu)^T \Sigma_y^{-1} (y - \mu) \right), \tag{9}$$

where Σ_y is the maximum-likelihood estimate of the covariance matrix, μ is the MLE of the global mean.

Then maximizing the conditional mutual information given the maximum-likelihood estimate of these models with respect to θ is equivalent to maximizing

$$V \quad = \quad \log |\Sigma_y| - \log |W_y| \,. \tag{10}$$

Using the following relations

$$\Sigma_y \quad = \quad \theta^T (W + B) \theta, \tag{11}$$

$$W_y \quad = \quad \theta^T W \theta, \tag{12}$$

and that the logarithm is a monotonic function, the objective function to be maximized becomes

$$O \quad = \quad \frac{|\theta^T (W + B) \theta|}{|\theta^T W \theta|}. \tag{13}$$

The $p \times n$ transformation θ^* that maximize the objective function in Equation 13 is the matrix consisting of the p eigenvectors of the Fisher covariance matrix $W^{-1}B$ corresponding to the largest p eigenvalues, and therefore is the solution of the LDA maximization also.

It should be noted that the assumption that $P(y)$ is Gaussian is inconsistent with the assumption that $\{P(y|c_j)\}_{j=1}^{J}$ are Gaussian, as in general if $\{P(y|c_j)\}_{j=1}^{J}$ are Gaussian PDFs, then $P(y)$ is a Gaussian mixture PDF. This explicit modeling of $P(y)$ that is inconsistent with the models for $\{P(y|c_j)\}_{j=1}^{J}$ is a serious limitation of LDA. It is the main reason that the LDA solution in many cases does not correspond to minimizing the recognition error.

2.3. Heteroscedastic Discriminant Analysis

Heteroscedastic discriminant analysis (HDA) is an extension to LDA that removes the equal covariance constraint [2]. HDA was first formulated as a maximum likelihood estimation problem for normal populations with common covariance matrix in the rejected subspace. An alternative interpretation of HDA as a constrained maximum likelihood projection for a full-covariance Gaussian model is introduced in [5]. It maximizes the objective function

$$J(\theta) \quad = \quad \frac{\left|\theta B \theta^T\right|^N}{\prod_{j=1}^{J} \left|\theta \Sigma_j \theta^T\right|^{N_j}}. \qquad (14)$$

This approach can be related to the maximization of the conditional mutual information in the lower dimensional space by removing the equal class-conditional covariance from the previous derivation for LDA. The assumption that $P(y)$ is Gaussian is still inconsistent with the assumption that $\{P(y|c_j)\}_{j=1}^{J}$ are Gaussian. Using the convexity of the relative entropy [7], it can be shown that this assumption underestimates the conditional mutual information as opposed to calculating $P(y)$ from the class-conditional PDFs $\{P(y|c_j)\}_{j=1}^{J}$, i.e.

$$\hat{I}_{DA}(Y, C|\Lambda) \quad \leq \quad \hat{I}(Y, C|\Lambda), \qquad (15)$$

where $\hat{I}_{DA}(Y, C|\Lambda)$ is the conditional mutual information estimated with an explicit Gaussian model of $P(y)$, and $\hat{I}(Y, C|\Lambda)$ is the conditional mutual information estimated by calculating $P(y)$ from the class-conditional PDFs $\{P(y|c_j)\}_{j=1}^{J}$.

3. Maximum Conditional Mutual Information Projection

In the following, we will relax the constraints of discriminant analysis to develop the maximum conditional mutual information projection (MCMIP) approach.

3.1. MCMIP Formulation

Given a set of class-conditional probabilistic models used by the classifier or the recognizer, the goal of MMICP is to find a p-dimensional subspace of an n-dimensional feature space that retains the discrimination information contained in the original high-dimensional space by maximizing an estimate of the conditional mutual information between the features and the class identity. In other words, MMICP searches for the $p \times n$ linear transformation or projection θ^* of the features that maximize the conditional mutual information $\hat{I}(Y, C|\Lambda)$, i.e.

$$\theta^* \quad = \quad \arg\max_{\theta} \hat{I}(Y, C|\Lambda), \qquad (16)$$

where $y = \theta x$. From the previous discussion of discriminant analysis, the feature vector Y achieved by MMICP has a higher conditional mutual information with the class identities given the classifier's set of class-conditional probabilistic models than the one obtained by discriminant analysis approaches. From Equation 7, it can be easily shown that maximizing $\hat{I}(Y, C|\Lambda)$ is equivalent to maximizing the *a posteriori* probability of the model corresponding to the training data given the data which is closely related to minimizing the training data recognition error.

3.2. Implementation of MCMIP For Speech Recognition

Applying the MCMIP approach for dimensionality reduction to an HMM-based speech recognizer requires the estimation of the conditional mutual information given the HMM parameters. The parameters of the HMM recognizer can be calculated using maximum likelihood estimation or discriminant approaches like maximum mutual information. We choose to use the expectation maximization (EM) algorithm to get maximum likelihood estimates of the HMM parameters [8]. Using these estimates of the parameters, the empirical estimate of the mutual information to be maximized is

$$\begin{aligned} \hat{I}(Y, C|\Lambda) \quad &= \quad \sum_{i=1}^{N} \log P_\Lambda(y^i|c^i) \\ &\quad - \sum_{i=1}^{N} \log\left(\sum_{j=1}^{J} P_\Lambda(y^i|c_j) P_\Lambda(c_j)\right), \qquad (17) \end{aligned}$$

where c^i is the maximum likelihood state assignment for the ith frame from the training data, N is the number of frames in the training data, and

$$\begin{aligned} P_\Lambda(y^i|c_j) \quad &= \quad \sum_{k=1}^{K} H_{jk} \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_{jk}|^{\frac{1}{2}}} \\ &\quad \exp\left(-\frac{1}{2}(y - \mu_{jk})^T \sigma_{jk}^{-1} (y - \mu_{jk})\right), \qquad (18) \end{aligned}$$

$$\sum_{k=1}^{K} H_{jk} \quad = \quad 1$$

for all $j = 1, 2, \cdots, J$, where H_{jk} is the weight of the kth Gaussian PDF in the Gaussian mixture of state j, K is the number of Gaussian PDFs in the Gaussian mixture, μ_{jk} is the mean of the kth Gaussian PDF in the mixture, and Σ_{jk} is the covariance matrix of the kth Gaussian PDF in the mixture.

To use a gradient-based algorithm to maximize our empirical estimate of the conditional mutual information, $\hat{I}(Y, C|\Lambda)$, with respect to the linear transform θ, we calculate the derivative of the objective function with respect to θ

$$\begin{aligned} \frac{d\hat{I}(Y, C|\Lambda)}{d\theta} \quad &= \quad \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{H_{c^i k}}{P_\Lambda(y^i|c^i)} \Sigma_{c^i k}^{-1} (\mu_{c^i k} - y^i) x^{i^T} \\ &\quad - \sum_{i=1}^{N} \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{P_\Lambda(y^i|c_{jk}) P_\Lambda(c_{jk})}{P_\Lambda(y^i)} \Sigma_{jk}^{-1} \\ &\quad (\mu_{jk} - y^i) x^{i^T}. \qquad (19) \end{aligned}$$

The steps of the iterative algorithm to update the transformation matrix θ and the HMM parameters are

1. Initialize the transformation matrix θ.

2. Calculate the feature vectors y using the relation $y = \theta x$, where x is the input acoustic feature vector.

3. Using the EM algorithm, estimate the HMM parameters and segment the training data.

4. Using the current HMM parameters and training data segmentation, estimate θ that maximizes the conditional mutual information, $\hat{I}(Y, C|\Lambda)$, using the conjugate-gradient algorithm.

5. Iterate (starting from 2) until convergence.

4. EXPERIMENTS AND RESULTS

The MCMIP algorithm described in section 3 is used to study the optimal feature subspace for diagonal-covariance Gaussian mixture HMM modeling of the TIMIT database.

The baseline 26-feature vector consists of 12 MFCC coefficients, energy and their deltas. The input to the MCMIP algorithm consists of 5 of these feature vectors centered at the target frame. This 130-feature vector is then transformed using the MCMIP algorithm to a 26-feature vector. In each iteration, the new feature vector is calculated using the current MCMIP transformation parameters, then the maximum likelihood estimates of the HMM model parameters are calculated. Then, the MCMIP transformation matrix is calculated using the conjugate-gradient algorithm. After the iterative algorithm converges to a set of locally optimal HMM and MCMIP parameters, the training data are transformed by the MCMIP matrix yielding the final MCMIP feature vector.

In our experiments, the 61 phonemes defined in the TIMIT database are mapped to 48 phoneme labels for each frame of speech as described in [9]. These 48 phonemes are collapsed to 39 phonemes for testing purposes as in [9]. A three-state left-to-right model for each triphone is trained. The number of mixtures per state was fixed to four. The parameters of the recognizer and the MCMIP transform are trained using the training portion of the TIMIT database. The parameters of the triphone models are then tied together using the same approach as in [10].

To compare the performance of the proposed algorithm with other approaches, we generated acoustic features using LDA, and HDA. We used the same 130-feature vector input to MCMIP with both LDA and HDA and kept the dimensions of the output of LDA and HDA the same as the MCMIP output.

Testing this recognizer, using the test data in the TIMIT database, we get the phoneme recognition results in table 1. These results are obtained by using a bigram phoneme language model and by keeping the insertion error around 10% as in [9]. The table compares MCMIP recognition results to the ones obtained by the baseline MFCC, LDA, and HDA.

Table 1: Phoneme Recognition Accuracy

Acoustic Features	**Recognition Accuracy**
MFCC	73.7%
LDA	73.8%
HDA	74.1%
MCMIP	74.7%

5. DISCUSSION

In this work, we described a framework for discriminant analysis for speech recognition. This framework is an extension of current approaches by relaxing the constraints imposed on the model in LDA and HDA approaches. Our approach maximizes the conditional mutual information between the feature vector and the HMM states which is closely related to recognition error, as opposed to maximizing the likelihood in LDA and HDA approaches that is not directly related to recognition error. We introduced also an iterative algorithm to calculate the MCMIP matrix for an HMM-based recognizer. Phoneme recognition experiments using features generated by this algorithm show significant improvement compared to previous dimensionality reduction transforms like LDA, and HDA.

6. ACKNOWLEDGMENT

This work was supported by NSF award number 0132900. Statements in this paper reflect the opinions and conclusions of the authors, and are not endorsed by the NSF.

7. References

[1] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification,* Wiley, New York, NY, 2000.

[2] N. Kumar, *Investigation of silicon-auditory models and generalization of linear discriminant analysis for improved speech recognition,* Ph.D. dissertation, John Hopkins Univ., Baltimore, MD, 1997.

[3] N. Campbell, “Canonical variate analysis - a general formulation” *Australian Journal of Statistics*, Vol. 26, pp. 86-96, 1984.

[4] T. Hastie, and R. Tibshirani, *Discriminant Analysis by Gaussian Mixtures,* Technical Report, AT&T Bell Laboratories, 1994.

[5] George Saon, Mukund Padmanabhan, Ramesh Gopinath, and Scott Chen, “Maximum Likelihood Discriminant Feature Spaces,” *IEEE Proceedings of ICASSP*, Istanbul, Turkey, 2000.

[6] Marco Loog, and Reinhold Haeb-Umbach, “Multi-Class Linear Dimension Reduction By Generalized Fisher Criteria,” *Proc. of Int. Conf. of Spoken Language Processing*, Beijing, China, 2000.

[7] Thomas M. Cover, and Joy A. Thomas, *Elements of Information Theory,* Wiley, New York, NY, 1997.

[8] Todd K. Moon, “The expectation maximization algorithm,” *IEEE Signal Processing magazine*, November 1996.

[9] Kai-Fu Lee, and Hsiao-Wuen Hon, “Speaker Independent Phone Recognition Using Hidden Markov Models,” *IEEE Trans. on ASSP*, vol. 37, pp. 1641-1648, November 1989.

[10] S. Young, and P. Woodland, “State Clustering in hidden Markov model continuous speech recognition,” *Computer, Speech, and Language*, vol. 8, pp. 369-383, October 1994.