APPLICATION OF A LOCALITY PRESERVING DISCRIMINANT ANALYSIS APPROACH TO ASR

Vikrant Singh Tomar, Richard C. Rose

Telecom and Signal Processing Laboratory,
Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada
vikrant.tomar@mail.mcgill.ca, rose@ece.mcgill.ca

ABSTRACT

This paper presents a comparison of three techniques for dimensionally reduction in feature analysis for automatic speech recognition (ASR). All three approaches estimate a linear transformation that is applied to concatenated log spectral features and provide a mechanism for efficient modeling of spectral dynamics in ASR. The goal of the paper is to investigate the effectiveness of a discriminative approach for estimating these feature space transformations which is based on the assumption that speech features lie on a non-linear manifold. This approach is referred to as locality preserving discriminant analysis (LPDA) and is based on the principle of preserving local within-class relationships in this non-linear space while at the same time maximizing separability between classes. This approach was compared to two well known approaches for dimensionality reduction, linear discriminant analysis (LDA) and locality preserving linear projection (LPP), on the Aurora 2 speech in noise task. The LPDA approach was found to provide a significant reduction in WER with respect to the other techniques for most noise types and signal-to-noise ratios (SNRs).

Index Terms: Graph embedding, feature extraction, dimensionality reduction, speech recognition

1. INTRODUCTION

In recent years, the use of feature transformations for dimensionality reduction has emerged as an important area in automatic speech recognition (ASR). Work in this area has been motivated by the importance of characterizing dynamic spectral information in speech to augment static features which model the smoothed spectral envelope over short-time stationary segments. This dynamic information can be captured by concatenating static feature vectors to form high dimensional super-vectors which may represent over 100 milliseconds intervals of speech and have dimensionality of as high as 200. Dimensionality reduction has been applied to these super-vectors to deal with two problems associated with this representation. The first problem is, of course, the high dimensionality of the super-vector, and the second problem is the high degree of correlation between feature vector components.

To deal with these issues, a number of approaches have been applied for estimating linear transformations for feature space dimensionality reduction in ASR. Discriminant transformations, including linear discriminant analysis (LDA) [1] and heteroscadastic discriminant analysis (HDA) [2], are known to reduce feature space dimensionality while maximizing a criterion related to the separability between classes. These techniques are applied to continuous density hidden Markov model (CDHMM) based acoustic models in ASR by associating feature vectors with classes corresponding to HMM states or clusters of states. One problem with discriminant transformations is that they cannot clearly explain the geometric and local distributional structure of the data space. High dimensional data can be considered as a set of geometrically related points resting on or close to the surface of a lower dimensional manifold. The structure of this manifold has been found in some general applications to be important for the classification task [3]. Motivated by this fact, manifold learning approaches, such as locality preserving projections (LPP) [4, 5], deal with this issue by explicitly preserving local relationships among data vectors in the transformed space. However, such unsupervised methods only focus on preserving the manifold data similarities and fail to discover the discriminant structure of the data.

This work investigates a new supervised discriminant approach, referred to as locality preserving discriminant analysis (LPDA), for feature space transformation as applied to a noisy ASR task. The locality preserving discriminant analysis (LPDA) algorithm is motivated by the lack of discriminative power in the manifold based techniques. Building upon the generalized concept of graph embedding (GE) [6], the algorithm attempts to maximize the separability between different classes, while preserving the underlying local sub-manifold based relationships of the feature vectors belonging to the same class. There has been a great deal of work on extending manifold based algorithms with some notion of discriminative power in other application domains. Deng et al. [7] reported significant improvement in a face-recognition task while using a locality preserving discriminant technique. Chen et al. [8], and Yan et. al. [6] also reported gain in face-recognition accuracy with a manifold learning discriminative technique.

The rest of this paper will present an implementation of LPDA for feature space dimensionality reduction for CDHMM acoustic modelling in ASR. Section 2 will provide a brief review of existing techniques LDA and LPP. Section 3 will present the development of locality preserving discriminant analysis (LPDA) approach with brief introduction to the underlying graph embedding framework, and discuss its application to dimensionality reduction for robust ASR. An experimental study comparing the relative performance of LPDA with the more well known methods LDA and LPP in terms of word error rate (WER) on the Aurora 2 speech corpus in noise task will be presented in Section 4. Finally, Section 5 concludes the paper, and defines the direction of future work.

2. RELATED WORK

The generic problem of feature space dimensionality reduction could be defined as follows. Consider a set of feature vectors, labelled or unlabelled, $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots \boldsymbol{x}_N\}$ in $\mathbb{R}^d$. This set $\boldsymbol{X}$ (with class labels, if available) is referred to as the training data, while the goal is to estimate the parameters of a projection matrix $\boldsymbol{P} \in \mathbb{R}^{d \times m}$, with $m \leq d$, to transform vectors from a d- dimensional feature space onto an m dimensional feature space. Dis-

This work is supported by Google Inc., Natural Sciences and Engineering Research Council of Canada, and McGill University.

978-1-4673-0382-8/12/$31.00 ©2012 IEEE

criminative algorithms like LDA attempt to maximize the discrimination between classes of the projected data. On the other hand, manifold based algorithms like LPP attempt to maximize the local manifold based relationships of the data vectors. Both LDA and LPP are briefly described below.

2.1. LDA

Suppose that all the vectors in $\boldsymbol{X}$ are labeled as belonging to one of a set of classes $c \in \{c_1, c_2, \cdots, c_{N_c}\}$, where N_c is the number of classes. The transformation is performed according to

$$\boldsymbol{y}_i = \boldsymbol{P}^T \boldsymbol{x}_i \qquad \forall i = 1, 2, \cdots, n \tag{1}$$

where $\boldsymbol{x}_i$ is an arbitrary vector in the source space, and $\boldsymbol{y}_i$ is the corresponding transformed vector in the new feature space.

Furthermore, lets assume $\boldsymbol{\mu}$ is the total sample mean of $\boldsymbol{X}$, and that each class c_i contains N_i of the total N vectors and is characterized by its mean vector $\boldsymbol{\mu}_i$, and the covariance matrix $\boldsymbol{\Sigma}_i$. The prior probability of each class is given by $p_i = N_i/N$. We define the following two within and between class scatter matrices [2].

$$\boldsymbol{S}_W = \sum_{i=1}^{N_c} p_i \boldsymbol{\Sigma}_i \tag{2}$$

$$\boldsymbol{S}_B = \frac{1}{N} \sum_{i=1}^{N_c} \left(N_i \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T - \boldsymbol{\mu}\boldsymbol{\mu} \right) \tag{3}$$

LDA aims to attain the highest class-separability by maximizing the following objective function,

$$\boldsymbol{P}_{lda} = \arg\max_{\boldsymbol{P}} \frac{|\boldsymbol{P}^T S_B \boldsymbol{P}|}{|\boldsymbol{P}^T S_W \boldsymbol{P}|} \tag{4}$$

Eq. 4 can be solved as a generalized eigenvector problem. The LDA transformation matrix $\boldsymbol{P}_{lda}$ is formed from the eigenvectors associated with m largest eigenvalues. Further discussion on LDA can be found in [1]. It should be evident that the within-class scatter is a measure of the average variance of the data within each class, while the between-class scatter represents the average distance between the means of the data in each class and the global mean. Thus, LDA aims to preserve the global class relationship between data vectors, although it fails to discover the intrinsic local structure of the data manifold.

2.2. LPP

Unlike LDA, LPP is an unsupervised learning technique. The underlying idea is to extend and preserve the local relationships that exist among the input data vectors to the vectors of projected feature space. In other words, the goal is to minimize the Euclidean distance between two data vectors $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ in the projected feature space given the corresponding vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are closely located in the input feature space. That is,

$$D = \min \sum_{i,j} (\boldsymbol{y}_i - \boldsymbol{y}_j)^T (\boldsymbol{y}_i - \boldsymbol{y}_j) s_{i,j} \tag{5}$$

In Eq. 5, the local relationships among the input data vectors are described by the terms of the similarity matrix, $\boldsymbol{S} = [s_{i.j}]_{N \times N}$, where $s_{i.j} = exp(-||\boldsymbol{x}_i - \boldsymbol{x}_j||^2)/\rho$ when $\boldsymbol{x}_j$ is in the neighbourhood of $\boldsymbol{x}_i$, and 0 otherwise. ρ is the heat-factor of the Gaussian kernel. The data vector $\boldsymbol{x}_j$ can be said to be in the neighbourhood of $\boldsymbol{x}_i$ either if it falls within the k-nearest neighbours of $\boldsymbol{x}_i$ or alternatively if $||\boldsymbol{x}_i - \boldsymbol{x}_j||^2 \leq$ some threshold ϵ.

The objective function in Eq. 5 can be simplified to following general eigenvalue problem,

$$\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^T \boldsymbol{p}_{lpp}^j = \lambda \boldsymbol{X}\boldsymbol{C}\boldsymbol{X}^T \boldsymbol{p}_{lpp}^j \tag{6}$$

where $\boldsymbol{L} = \boldsymbol{C} - \boldsymbol{S}$ is the Laplacian of the similarity matrix, $\boldsymbol{C}$ is a diagonal matrix whose elements are the corresponding column sums of the matrix $\boldsymbol{S}$, and $\boldsymbol{p}_{lpp}^j$ is the j^{th} column of the linear transformation matrix $\boldsymbol{P}_{lpp}$, which is formed from the eigenvectors associated with the m smallest non-zero eigenvalues. A detailed discussion on LPP can be found in [4], with ASR specific implementation in [5].

3. LOCALITY PRESERVING DISCRIMINANT ANALYSIS

This section describes the locality preserving discriminant approach to feature space transformation and dimensionality reduction for a noisy ASR task. Section 3.1 presents the underlying graph embedding framework for characterizing the geometrical relationships between data vectors, and Section 3.2 provides optimization of a discriminative objective function to maximize class separability. Section 3.3 presents a discussion of the issues associated with applying LPDA in CDHMM based acoustic modelling.

3.1. Graph Embedding Framework

The purpose of graph embedding for dimensionality reduction is to represent the data space by an undirected connected graph with the data vectors as its vertices, and then perform dimensionality reduction such that the graph structure and similarity between the connected vertex pairs is preserved. Here, similarity between the nodes of the graph is represented by an affinity matrix that characterizes the statistical and geometrical properties of the data set. Following this framework, the training data-set is embedded into an undirected weighted graph $\mathcal{G} = \{\boldsymbol{X}, \boldsymbol{W}\}$, where the data-set $\boldsymbol{X}$ represents the nodes of the graph, and $\boldsymbol{W} = [w_{ij}]_{N \times N}$ is the affinity edge-weight matrix. The affinity weight, w_{ij}, of an edge between two nodes $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the graph, is given by

$$w_{ij} = \begin{cases} exp(-||\boldsymbol{x}_i - \boldsymbol{x}_j||^2)/\rho & ; \quad e(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 \\ 0 & ; \quad e(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0 \end{cases}, \tag{7}$$

where $e(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is an indicator function signifying whether the concerned node $\boldsymbol{x}_j$ is close to $\boldsymbol{x}_i$ as per the neighbourhood definition given in Section 2.2, and ρ is the Gaussian kernel heat parameter. Thus, a scatter measure for graph $\mathcal{G}$ can be given by,

$$F(\boldsymbol{P}) = \sum_{i \neq j} ||\boldsymbol{y}_i - \boldsymbol{y}_j||^2 w_{ij} \tag{8a}$$

$$= 2\boldsymbol{P}^T \boldsymbol{X}(\boldsymbol{D} - \boldsymbol{W})\boldsymbol{X}^T \boldsymbol{P} \tag{8b}$$

where $\boldsymbol{y}_i = \boldsymbol{P}^T \boldsymbol{x}_i$ is the target space vector corresponding to the vector $\boldsymbol{x}_i$ in the source space, $\boldsymbol{D}$ is a diagonal matrix whose elements are corresponding column sum of the edge-affinity matrix $\boldsymbol{W}$, *i.e.,* $\boldsymbol{D}_{ii} = \sum_j w_{ij}$. Depending on whether the goal is to preserve or eliminate the concerned graph structure, the optimal projection matrix $\boldsymbol{P}$ can be obtained by minimizing or maximizing the scatter in Eq. 8a, respectively. A detailed study of the graph embedding framework can be followed in [6].

3.2. LPDA - Algorithm Formulation

In order to formulate an optimality criterion based on class separability, two undirected weighted graphs, *viz.*, the intrinsic graph $\mathcal{G}_i = \{\boldsymbol{X}, \boldsymbol{W}_i\}$, and the penalty graph $\mathcal{G}_p = \{\boldsymbol{X}, \boldsymbol{W}_p\}$ are defined. In the intrinsic graph, a node $\boldsymbol{x}_i$ is connected to the k_i nearest neighbours, in affinity sense, belonging to the same class c_i. Similarly, in the penalty graph, a node $\boldsymbol{x}_i$ is connected to the k_p largest affinity neighbours *not* belonging to the class c_i. The

elements of the intrinsic and penalty graph weight matrices are defined as,

$$w_{ij}^{intrinsic} = \begin{cases} w_{ij} & ; \quad e_c(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 \\ 0 & ; \quad e_c(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0 \end{cases} \tag{9}$$

and,

$$w_{ij}^{penalty} = \begin{cases} w_{ij} & ; \quad e_c(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0 \\ 0 & ; \quad e_c(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 \end{cases} \tag{10}$$

where, w_{ij} is defined in Eq. 7, and $e_c(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is an indicator function signifying whether $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ lie within the same class.

As the names suggest, the goal is to penalize the properties inherent to the penalty graph (maximize its scatter $F_p(\boldsymbol{P})$), while at the same time preserve the properties inherent to the intrinsic graph (minimize its scatter $F_i(\boldsymbol{P})$). It is thus justified to define the ratio of the two as a measure of class separability and graph-preservation.

$$\begin{aligned} F(\boldsymbol{P}) &= \frac{F_p(\boldsymbol{P})}{F_i(\boldsymbol{P})} \\ &= \frac{\boldsymbol{P}^T \boldsymbol{X}(\boldsymbol{D^p} - \boldsymbol{W^p})\boldsymbol{X}^T \boldsymbol{P}}{\boldsymbol{P}^T \boldsymbol{X}(\boldsymbol{D^i} - \boldsymbol{W^i})\boldsymbol{X}^T \boldsymbol{P}} \end{aligned} \tag{11}$$

Thus, an optimal projection matrix is the one to maximize the expression in Eq. 11, namely,

$$\arg\max_{\boldsymbol{P}} \operatorname{tr}\left((\boldsymbol{X}(\boldsymbol{D}_i - \boldsymbol{W}_i)\boldsymbol{X}^T\boldsymbol{P})^{-1}(\boldsymbol{P}^T\boldsymbol{X}(\boldsymbol{D}_p - \boldsymbol{W}_p)\boldsymbol{X}^T\boldsymbol{P})\right) \tag{12}$$

where the subscripts i and p signify 'intrinsic' and 'penalty' graphs respectively. Eq. 12 can be translated into a generalized eigenvalue problem as in the following.

$$(\boldsymbol{X}(\boldsymbol{D}_p - \boldsymbol{W}_p)\boldsymbol{X}^T)\boldsymbol{p}_{lpda}^j = \lambda_j(\boldsymbol{X}(\boldsymbol{D}_i - \boldsymbol{W}_i)\boldsymbol{X}^T)\boldsymbol{p}_{lpda}^j, \tag{13}$$

where $\boldsymbol{p}_{lpda}^j$ is the j^{th} column of the linear transformation matrix $\boldsymbol{P}_{lpda} \in \mathbb{R}^{d\times m}$, and is the eigenvector associated with the j^{th} largest eigenvalue.

3.3. Discussion

The advantages of LPDA arises from the fact that it combines graph embedding based within class sub-manifold learning with inter-class discrimination. There are several components to LPDA that contribute to its performance gain. First is the use of soft-weights when building the affinity edge-weight matrices. One clear advantage of using soft-weights over hard-weights[1] is that, even within the k-nearest neighbours, the nodes which are closely located in the source feature space are naturally given more importance when evaluating the scatter measure. This further enforces the notion of locality preservation.

A second important aspect of the algorithm can be observed by analyzing the optimization expression in Eq. 8a separately for the intrinsic and penalty graphs, as given in the Eq. 14a and 14b.

$$\max_{\boldsymbol{P}} \left\{ F_p(\boldsymbol{P}) = \sum_{i\neq j} ||\boldsymbol{y}_i - \boldsymbol{y}_j||^2 w_{ij}^{penalty} \right\} \tag{14a}$$

$$\min_{\boldsymbol{P}} \left\{ F_i(\boldsymbol{P}) = \sum_{i\neq j} ||\boldsymbol{y}_i - \boldsymbol{y}_j||^2 w_{ij}^{intrinsic} \right\} \tag{14b}$$

One should notice that (a) the intrinsic objective function is penalized if the vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are located far apart in the projected space despite actually being in the same class, similarly (b) the penalty graph objective function is penalized if the neighbouring vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are mapped close even though they belong to separate classes, and (c) because of the soft weights, the closer the two vectors the higher the penalty upon a misclassification. Thus maximizing the ratio of penalty graph scatter to intrinsic graph scatter ensures that if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are close and in the same class, then $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ would be close as well; and even if two nodes from different classes are close, their projections would be far apart.

There are several more general statements that can be made concerning the potential advantages of the LPDA transform relative to the LDA based approaches. First, LPDA not only maximizes class separability as is done in all discriminant transformations, but it also preserves local relationships that exist among input data vectors. These local relationships are described by the intrinsic graph $\mathcal{G}_i$. Second, LPDA does not make assumptions about the underlying distribution of data as is done, for example, in LDA where it is assumed that the class conditional distribution of the data is Gaussian. Third, it can preserve non-linear structure in the data as exemplified by popular low-dimensional manifold learning applications [3]. All of these properties are shared with other graph embedding approaches [4, 6, 7].

In comparison to LPP, LPDA is has two distinct advantages. First, clearly LPDA takes advantage of the discriminant structure of the data, which LPP entirely neglects. This is done by maximizing the scatter of the penalty graph $\mathcal{G}_p$. Second, the inherent assumption in LPP is that data vectors belonging to different classes are already well separated on the surface of the manifold. Hence, it attempts to preserve the manifold based structure of the entire dataset without considering distribution of the data among various classes. This assumption may not be valid in practical scenarios. LPDA makes no such assumption, and only preserves within-class manifold based geometrical structures as illustrated by the affinity weight matrices $\boldsymbol{W}_i$ and $\boldsymbol{W}_p$ in Eq. 9 and 10.

It is well known that all sub-manifold learning approaches to feature space transformation have extremely high computational complexity when compared to other discriminant feature transformations [9, 10]. The task domain described in Section 4 involves 180 states and a training corpus of 1.4 million 117-dimensional feature vectors. This is a far larger task than addressed in the application domains described in [9, 10, 7]. This represents a definite disadvantage of LPDA transformations when applied to speech processing tasks.

Another important issue with many data classification approaches is choosing the neighbourhood size k, and the Gaussian kernel heat factor ρ. Optimal choices of these two parameters have been shown to critically affect the performance of most classification algorithms, and is an independent research problem in itself [11, 12]. In this work, both sub-manifold based approaches LPP and LPDA are affected by the choice of these two parameters.

A final issue associated with all feature space transformations is the fact that the dimensions of the projected feature space can be highly correlated. However, the diagonal Gaussian covariance parametrization used in CDHMM based ASR systems poses the assumption that feature vector dimensions are "approximately uncorrelated." Therefore, there is a need to incorporate one of a set of procedures that maximizes the likelihood of the data under the constraint that feature vectors be uncorrelated. These procedures, including the maximum likelihood linear transformation (MLLT) [2] and semi-tied covariances (STC) [13], are applied in the transformed feature space.

4. EXPERIMENTAL STUDY

This section describes the experimental study performed to evaluate the ASR performance obtained using the LPDA approach

[1] Hard weights:

$$w_{ij} = \begin{cases} 1 & ; \quad e(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 \\ 0 & ; \quad e(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0 \end{cases}$$

to feature space transformations with respect to the performance obtained using the more well known LDA and LPP based approaches. Performance is reported as the word error rate obtained for the Aurora 2 speech in noise task domain. Section 4.1 describes the Aurora 2 task domain and the baseline ASR system configuration. Section 4.2 provides the experimental results across a range of noise types and signal-to-noise ratios.

4.1. Task domain and system configuration

All systems described in Section 4.2 are evaluated on the European Telecommunications Standards Institute (ETSI) Aurora-2 speech corpus. The corpus was created by adding noise from multiple environments to connected digit utterances spoken in a quiet environment. All HMM models and feature space transformations were trained using the 8440 utterances that Aurora 2 training set had collected from 55 male and 55 female speakers. The multi-condition training scenario, involving noise corrupted training utterances, was used for training all models. The results reported in Tables 1 and 2 were evaluated over the standard Aurora 2 test set consisting of 4004 utterances artificially corrupted by four different noise types at SNRs ranging from 5 to 20 dB.

The baseline ASR system was configured using whole word CDHMM models. MFCC features were used for the baseline system which included 12 static coefficients, normalized log energy, Δ-cepstrum, and $\Delta\Delta$-acceleration appended to create a 39 dimensional feature vector. The input features for both the LDA and LPDA based feature transformations are supervectors consisting of 9 concatenated vectors of MFCC coefficients augmented with log energy. This corresponds to an input feature vector dimensionality of 117. All projection matrices were trained using the multi-condition training scenario. For the supervised algorithms concerned, LDA and LPDA, classes are defined as the states of the CDHMM models. There are 180 states overall for the 11 word-based CDHMM models. The resultant matrix $\boldsymbol{P}$ was then used to perform discriminant feature space transformation on the test data and project the features to a 39 dimensional space. STC matrices were estimated as described in Section 3.3 to account for the correlation introduced to the transformed features by all LDA, LPP and LPDA approaches [13]. Furthermore, as mentioned in Section 3.3, the choice of neighbourhood size and Gaussian kernel heat-factor affect the performance of both LPP and LPDA. The optimal values of these parameters depend on the level as well as the nature of noise [11, 12]. For the current set of experiments a neighbourhood size of $k = k_i = k_p = 200$ is taken for all the simulations. The Gaussian kernel heat factor is taken to be equal to 900 for LPP, and 1000 and 3000 for intrinsic and penalty graphs of LPDA algorithm, respectively.

4.2. Results

Table 1 compares the recognition performance of LDA with the baseline for five different noise types and four SNRs ranging from 5dB to 20dB. Five separate tables are displayed, one for each noise type (subway, car, exhibition hall, airport, and subway(mirs)) where each of these tables contains ASR WER for five different systems. For each of these tables, the first row displays the baseline ASR WER obtained using mixed condition HMM training when no feature transformation is performed. The second row, labeled "LDA", corresponds to application of the 39 by 117 LDA discriminant projection matrix to the concatenated MFCC feature vectors described above. In the third row of these tables, labeled "LDA+STC", ASR WER is reported for the case where a STC transform is estimated to minimize the impact of the data distributions resulting from the LDA projection. The fourth row of these tables, labeled "LPP+STC" corresponds to the ASR WER performance when using LPP as the feature space transformation technique. The fifth row, labeled "LPDA+STC", refers to the ASR WER when using LPDA as the feature space transformation technique. For all these results, an STC transformation is performed to minimize the impact of the data correlation resulting from the LDA or LPDA projection, respectively.

Table 1. WER for mixed noise training and noisy testing on Aurora-2 speech corpus for LDA, LPP and LPDA. The best performance has been highlighted for each noise type per SNR level.

Noise Type	Technique	SNR (dB)			
		20	15	10	5
Subway	Baseline	2.99	4.0	6.21	11.89
	LDA	3.19	4.14	7.68	14.00
	LDA+STC	2.25	**2.93**	5.29	12.32
	LPP+STC	2.33	3.50	5.71	13.26
	LPDA+STC	**2.18**	3.29	**5.28**	**11.73**
Car	Baseline	2.77	3.36	5.45	12.31
	LDA	3.82	4.26	6.74	17.15
	LDA + STC	2.83	3.45	5.69	15.92
	LPP+STC	2.71	3.61	6.08	14.97
	LPDA+STC	**2.30**	**2.77**	**5.19**	**12.73**
Exhibition	Baseline	3.34	3.83	6.64	12.72
	LDA	3.39	4.63	7.47	15.15
	LDA+STC	2.63	**3.37**	6.67	14.29
	LPP+STC	2.56	4.23	8.55	16.91
	LPDA+STC	**2.22**	3.64	**6.66**	**13.85**
Airport	Baseline	3.42	4.88	8.49	16.58
	LDA	5.67	7.07	10.26	19.83
	LDA+STC	3.18	4.11	7.72	15.65
	LPP+STC	4.35	6.95	10.38	21.15
	LPDA+STC	**3.10**	**4.09**	**7.49**	**15.09**
Subway(mirs)	Baseline	3.26	4.74	7.02	18.37
	LDA	3.53	4.54	8.07	18.08
	LDA+STC	2.68	3.26	6.43	**15.91**
	LPP+STC	2.61	3.62	7.46	19.99
	LPDA+STC	**2.30**	**2.95**	**6.05**	16.49

The main observations that can be made from Table 1 are as follows. The most important observation is that LPDA+STC outperforms LDA+STC, and LPP + STC in most cases by a margin of 6-27% relative improvement in WER for different noise conditions. These results reasserts the importance of exploiting discriminant as well as manifold based local structure of the dataset when performing dimensionality reduction. Second, though the recognition performance for LDA alone is less than the baseline performance for almost all conditions, a training pass by STC improves the performance significantly. This decrease in performance, when the discriminant LDA transformation is applied without any de-correlating transform, is consistent with our discussion in Section 3.3, and results presented elsewhere [2]. The trend is also observed for LPP and LPDA based dimensionality reduction, if STC is not used. These results have not been reported as the LDA results suffice the point.

Yet another important observation that can be made from Table 1 is that LPP + STC almost always reports low performance as compared to LDA + STC. At first, this may appear to be contradicting with earlier results as reported by Tang and Rose in [5]. However, it should be noted that the work in [5] reports ASR performance for a clean testing case, whereas the results in Table 1 reflects the performance when testing in a noisy environment. When the recognition performance was measured for a clean testing scenario, LPP+STC performs consistently better than LDA+STC, however, LPDA+STC still reports the best performance. The results of this experiment are presented in Table 2

for five different clean test sets. These five sets are subsets from various Aurora-2 test cases as labelled by noise types in Table 1, and explained in [14] [2]. Clearly, these results are in agreement with those reported in [5]. While highlighting the importance of LPDA over LPP, these experiments perhaps suggest that though the local geometry of the data plays an important role for clean testing, it is the discriminative structure of the data vectors that becomes important in the presence on noise.

5. CONCLUSION

This paper presents a discriminant analysis algorithm for feature space dimensionality reduction which utilizes graph embedding to preserve the within class local relationships, while at the same time maximizes the separability between classes. Use of the generalized framework of graph embedding facilitates elimination of dependency on the data distribution. When compared to the traditional algorithms such as LDA and LPP, the presented algorithm promises better performance for speech recognition with experimental results showing an improvement of 6-27% in WER as compared to LDA+STC based ASR.

Performance of LPDA is critically dependent on neighbourhood size, and the value of Gaussian kernel heat factor. The task of identifying the optimal set of these parameter is left for future work.

6. REFERENCES

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley Interscience, 2000.

[2] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, May 2000.

[3] L. Saul and S. Roweis, "Think globally, fit locally: Unsupervised learning of low dimensional manifolds," *J. of Machine Learning Research*, vol. 4, pp. 119 – 155, 2003.

[4] X. He and P. Niyogi, "Locality preserving projections," in *ICASSP: IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[5] Y. Tang and R. Rose, "A study of using locality preserving projections for feature extraction in speech recognition," in *ICASSP: IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008.

[2] Note that the noise types here are only labels for a specific test set from the Aurora-2 speech corpus, and do not imply any noise corruption of the test utterances. The reader should refer to Aurora-2 documentation for detailed description of the corpus structure [14].

Table 2. WER for mixed noise training and clean testing on Aurora-2 speech corpus for LDA, LPP and LPDA. The best performance has been highlighted for each test set. Note that the noise types here are only labels for a specific test set from the Aurora-2 speech corpus [14], and do not imply any noise corruption of the test utterances.

Technique	Noise Type				
	Sub.	Car	Exh.	Airport	Sub.(mirs)
Baseline	1.76	1.99	1.89	1.99	1.67
LDA	2.43	3.22	2.38	3.22	2.36
LDA+STC	1.82	2.29	1.83	2.29	1.73
LPP+STC	1.66	1.88	1.76	1.88	**1.54**
LPDA+STC	**1.57**	**1.52**	**1.23**	**1.52**	**1.54**

[6] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A generalized framework for dimensionality reduction," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40 – 51, Jan 2007.

[7] D. Cai, X. He, and et al., "Locality sensitive discriminant analysis," in *Intl. Joint Conference on Artifical Intelligence*, 2007.

[8] H.-T. Chen, H.-W. Chang, and T.-L. Liu, "Local discriminant embedding and its variants," in *IEEE Conference on Computer Vision and Patter Recognition*, vol. 1, no. 5, 2005, pp. 846–853.

[9] Y. Fu, S. Yan, and T. S. Huang, "Correlation metric for generalized feature extraction," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2229 – 2235, Dec 2008.

[10] Y. Ma, S. Lao, E. Takikawa, and M. Kawade, "Discriminant analysis in correlation similarity measure space," in *24th International Conference on Machine Learning*, vol. 227, Corvallis, OR, 2007, pp. 577 – 584.

[11] H. Xiong, M. N. S. Swamy, and M. O. Ahmad, "Optimizing the kernel in the empirical feature space." *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, vol. 16, no. 2, pp. 460–74, Mar. 2005.

[12] J. Wang, H. Lu, K. Plataniotis, and J. Lu, "Gaussian kernel optimization for pattern classification," *Pattern Recognition*, vol. 42, no. 7, pp. 1237–1247, Jul. 2009.

[13] M. J. F. Gales, "Semi-tied covariance matrices for hidden markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272 – 281, May 1999.

[14] H. Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *Speech Recognition: Challenges for the*, 2000.