A Family of Discriminative Manifold Learning Algorithms and Their Application to Speech Recognition

Vikrant Singh Tomar, *Student Member, IEEE*, and Richard C. Rose, *Fellow, IEEE*

***Abstract*—This paper presents a family of discriminative manifold learning approaches to feature space dimensionality reduction in noise robust automatic speech recognition (ASR). The specific goal of these techniques is to preserve local manifold structure in feature space while at the same time maximizing the separability between classes of feature vectors. In the manifold space, the relationships among the feature vectors are defined using nonlinear kernels. Two separate distance measures are used to characterize the kernels, namely the conventional Euclidean distance and a cosine-correlation based distance. The performance of the proposed techniques is evaluated on two task domains involving noise corrupted utterances of connected digits and read newspaper text. Performance is compared to existing approaches used for feature space transformations, including linear discriminant analysis (LDA) and locality preserving linear projections (LPP). The proposed approaches are found to provide a significant reduction in word error rate (WER) with respect to the more well-known techniques for a variety of noise conditions. Another contribution of the paper is to quantify the interaction between acoustic noise conditions and the shape and size of local neighborhoods which are used in manifold learning to define local relationships among feature vectors. Based on this analysis, a procedure for reducing the impact of varying acoustic conditions on manifold learning is proposed .**

***Index Terms*—Cosine distances, dimensionality reduction, discriminative manifold learning, feature extraction, graph embedding, speech recognition.**

I. Introduction

HIGH dimensionality of feature vectors is a common issue in pattern recognition, particularly in acoustic feature analysis for automatic speech recognition (ASR). There are many reasons for having high dimensional feature spaces. For instance, static features can be augmented with dynamic spectral information in speech feature extraction. One way of accomplishing this is by combining multiple consecutive Mel-filtered cepstrum coefficients (MFCC) feature vectors to form high dimensional super-vectors that may represent on the order of 100 milliseconds of speech. These super feature vectors can have very high dimensionality, which may lead to significant problems when performing a pattern recognition task [1]. Therefore, it is a good practice to perform some sort of dimensionality reduction before applying a particular pattern recognition algorithm to these features. Intuitively, a good dimensionality reduction algorithm should be able to preserve important information from the original feature space in the low dimensional transformed feature vectors. Thus, the dimensionality reduction problem involves finding a good feature space, where, for example, features belonging to different classes are well separated.

This issue has inspired the use of subspace learning for feature extraction and dimensionality reduction. When estimating projections from an original high dimensional feature space to a low dimensional feature space, subspace learning establishes optimization constraints so that the desired data relations and distributions are preserved. One widely utilized family of such algorithms is supervised discriminative techniques. Linear discriminant analysis (LDA) [2]–[4] and heteroscedastic linear discriminant analysis (HLDA) [5] are two examples of many such algorithms that have been widely used in ASR for reducing feature space dimensionality while maximizing a criterion related to the separability between classes of speech features. However, one common issue with discriminative feature transformations is their inability to capture the geometric and local distributional structure of the data space.

Recent studies have demonstrated that the geometric and local structure of the data space are important for classification [6]. This has motivated the use of manifold learning techniques in ASR. The underlying assumption of these techniques is that the high-dimensional data can be considered as a set of geometrically related points lying on or close to the surface of a smooth low-dimensional manifold embedded in the ambient space [1]. Manifold learning approaches, such as locality preserving projections (LPP) [1], [7] take advantage of the geometric distribution of data points in the high dimensional space, and seek to preserve manifold constrained relationships among data vectors.

Consider, for example, the set of four data points in Fig. 1. Points A, B, C, and D are depicted to be lying on a 1-dimensional manifold represented by a curve. For neighboring points on the manifold, such as C and D, the closeness between the two points can be approximated by the Euclidean distance directly. However, for points that are well separated on the manifold, such as A and D, the direct Euclidean distance measured between the two points will be much different than the distance measured along the manifold curve. A simple dimensionality reduction algorithm may project the points A and D close together in the target space. However, a manifold learning transformation

Manuscript received December 17, 2012; revised October 04, 2013; accepted October 07, 2013. Date of publication October 23, 2013; date of current version November 22, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. James Glass.

The authors are with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC H3H 2G8, Canada (e-mail: vikrant.tomar@mail.mcgill.ca; rose@ece.mcgill.ca).

Digital Object Identifier 10.1109/TASLP.2013.2286906

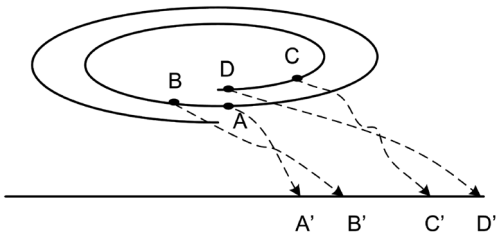

Fig. 1. Illustration of dimensionality reduction for data embedded in a non-linear manifold space with relative position information preserved [1].

would project the points A and D far from each other, thus preserving the manifold based structure as illustrated in the figure. This may be particularly important for speech signals as it has been suggested that the acoustic feature space, constrained by the articulatory dynamics associated with speech production, is confined to lie on one or more low dimensional manifolds [8], [9]. Therefore, a dimensionality reduction technique that explicitly models and preserves the data distribution along the underlying manifold structure should be more effective for ASR feature space transformations. However, manifold learning techniques are inherently unsupervised and non-discriminative. As a result, feature vectors belonging to different classes may not be optimally separated in the transformed space.

Both the discriminative and the manifold learning based dimensionality reduction techniques have been shown to provide transformed features leading to lower word error rates (WER) in ASR [1], [4]. The work in this paper is motivated by the assumption that there could be merit to integrating inter-class discrimination aspect into manifold preservation algorithms in ASR. The proposed framework incorporates a discriminative component to manifold learning techniques by maximizing the separability between different classes while preserving the within-class local manifold constrained relationships of the data points. There has been some work on extending manifold based algorithms with some notion of discriminative power in other application domains. Cai *et al.* [10] reported significant improvement in a face-recognition task while using a locality preserving discriminative technique. Ma *et al.* [11], and Yan *et al.* [12] also reported gain in face-recognition accuracy with a manifold learning discriminative technique.

The discriminative manifold learning framework acts by embedding feature vectors into one or more high-dimensional graphs, and then optimizing the structure of these graphs under a set of constraints. The nodes of the graphs represent the feature vectors. The weight over an edge is a measure of closeness (affinity) between the associated feature vectors [12]. In this work, two different metrics have been used to define the affinity weights, leading to two different approaches. The first, locality preserving discriminant analysis (LPDA), defines affinity between nodes as a Euclidean distance metric [13]. The second, correlation preserving discriminant analysis (CPDA), uses a cosine-correlation distance metric to define the manifold domain affinity between nodes [14]. The use of the cosine-correlation based distance metric is motivated by work where a cosine distance metric has been found to be more robust to noise corruption than Euclidean distances [11], [15], [16]. Thus, CPDA is expected to demonstrate a performance advantage over LPDA for high noise scenarios. Another important contribution of this paper is to analyze the effect of noise on the ASR performance on manifold learning algorithms. The outcome of this analysis is an automated noise-aware manifold learning (NAML) mechanism that can be used to increase the robustness of manifold learning algorithms against noise.

The remainder of this paper presents the theory and application of discriminative manifold learning, and is organized as follows. Section II provides a brief review of existing techniques for feature space dimensionally reduction in ASR. Section III presents the family of discriminative feature space transformations. Experimental studies comparing the ASR performance of LPDA, CPDA and NAML with a number of well-known techniques in terms of word error rate (WER) on a connected digit speech-in-noise task and read newspaper speech-in-noise task are provided in Section IV. Section V presents a discussion and some issues pertaining to the application of this family of algorithms to ASR. Finally, Section VI concludes the paper.

II. Related Work

Brief summaries of discriminative and manifold learning based feature space transformations are provided here as background for the techniques presented in Section III. Linear discriminant analysis (LDA) [2], [4] and locality preserving projections (LPP) [1], [7] are presented as well known examples of discriminative and manifold based feature space projections, respectively. None of these techniques produce transformed features whose distributions are consistent with the densities of the continuous density hidden Markov models (CDHMM) based ASR. The semi-tied covariance (STC) [17] procedure is presented as a means for reducing the impact of this mismatch.

The general problem of dimensionality reduction in pattern recognition can be defined as follows. Consider a set of labeled or unlabeled feature vectors represented in the form of a matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots \boldsymbol{x}_N]^T$, where each row denotes a vector in the high-dimensional source space $\mathbb{R}^d$. For labeled data, each vector $\boldsymbol{x}_j$ would also be associated with a class/label $C(\boldsymbol{x}_j) \in \{c_1, c_2, \cdots, c_{M_c}\}$, where M_c is the number of classes, and each class c_i contains N_i of the total N vectors. This data, $\boldsymbol{X}$, (with class labels, if available) is referred to as the training set. The goal of a dimensionality reduction task is to estimate the optimal projection matrix $\boldsymbol{P} \in \mathbb{R}^{d \times m}$, with $m \ll d$, to transform vectors from the d-dimensional source space onto an m-dimensional target space. 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 low dimensional vector in the target space.

A. Discriminative Techniques: Linear Discriminant Analysis

Discriminative algorithms, such as LDA and HLDA [5], attempt to maximize the discrimination between classes of feature vectors. While HLDA has in some cases demonstrated performance improvements with respect to LDA, there is some debate as to whether similar effects can be achieved by applying a semitied covariance transform (STC) (discussed

in Section II-C) with LDA [17], [18]. For this reason, LDA, in combination with STC, is selected as a representative of discriminant algorithms in this work. The following discussion provides a brief description of LDA.

Suppose that for the aforementioned training set, each class, c_i, 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 $p_i = N_i/N$. If $\boldsymbol{\mu}$ is the total sample mean of $\boldsymbol{X}$, then the within and between class scatter matrices can be defined as follows [2]:

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

$$\boldsymbol{S}_B = \frac{1}{N} \sum_{i=1}^{N_c} \left(N_i \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T - \boldsymbol{\mu}\boldsymbol{\mu}^T \right). \tag{2b}$$

LDA optimizes a class separability criterion by maximizing the following objective function,

$$\boldsymbol{P}_{lda} = \arg\max_{\boldsymbol{P}} \left\{ tr(|\boldsymbol{P}^T S_W \boldsymbol{P}|^{-1} |\boldsymbol{P}^T S_B \boldsymbol{P}|) \right\}. \tag{3}$$

(3) can be solved as a generalized eigenvector problem as given in (4),

$$\boldsymbol{S}_B \boldsymbol{p}_{lda}^j = \lambda_j \boldsymbol{S}_W \boldsymbol{p}_{lda}^j \tag{4}$$

where $\boldsymbol{p}_{lda}^j$ is the jth column of the LDA transformation matrix $\boldsymbol{P}_{lda}$, which is formed from the eigenvectors associated with the m largest eigenvalues. Further discussion of LDA can be found in [2].

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 relationships; however, it does not capture the intrinsic local structure of the data manifold.

B. Manifold Learning Approaches: Locality Preserving Projections

The underlying idea of manifold learning based feature transformations is to extend the manifold constrained relationships that exist among the input data vectors to the vectors in the projected space. Manifold based relationships can be characterized by a high dimensional graph connecting neighborhoods of feature vectors. This process is referred to as graph embedding (GE) [12]. In this graph, feature vectors, $\boldsymbol{X}$, correspond to nodes of the graph. The graph edge-weights denote the relationships among the nodes and are given by the affinity edge-weight matrix $\boldsymbol{W} = [w_{ij}]_{N \times N}$, where the $\{i, j\}$th element of the affinity matrix, w_{ij}, defines the weight of the edge connecting the nodes $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Such an embedding provides a strong mathematical framework to represent the distribution and geometrical structure of data.

For a generic graph $\mathcal{G}$, the relative scatter measure in the target space can be given by,

$$F_{\mathcal{G}}(\boldsymbol{P}) = \sum_{i \neq j} d\{\boldsymbol{y}_i, \boldsymbol{y}_j\} w_{ij} \tag{5}$$

where $d\{\boldsymbol{y}_i, \boldsymbol{y}_j\}$ is a distance measure between vectors $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ in the transformed space. Depending on whether the goal is to preserve or diminish the concerned graph properties, the optimal projection matrix $\boldsymbol{P}$ can be obtained by minimizing or maximizing the scatter in (5). A detailed study and generalization of various graph embedding based techniques can be found in [12].

This work chooses LPP as an example of manifold based techniques. Following (5), the objective function can be defined as:

$$\arg\min_{\boldsymbol{P}_{lpp}} \left\{ \sum_{i \neq j} \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2 w_{ij} \right\} \tag{6}$$

where $w_{ij} = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2)/\rho$ when $\boldsymbol{x}_j$ is in the near neighborhood of $\boldsymbol{x}_i$, and 0 otherwise. The weight, w_{ij}, is referred to as a Gaussian heat kernel, and ρ is a scale factor controlling the width of the kernel. The vector $\boldsymbol{x}_j$ is said to be in the neighborhood of $\boldsymbol{x}_i$ if it lies within the k-nearest neighbors of $\boldsymbol{x}_i$.

The optimal value of the projection matrix $\boldsymbol{P}_{lpp}$ for minimizing the objective function in 6 can be obtained by solving the following general eigenvalue problem,

$$\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^T \boldsymbol{p}_{lpp}^j = \lambda \boldsymbol{X}\boldsymbol{C}\boldsymbol{X}^T \boldsymbol{p}_{lpp}^j \tag{7}$$

where $\boldsymbol{L} = \boldsymbol{C} - \boldsymbol{W}$ is the Laplacian of the similarity matrix, $\boldsymbol{C}$ is a diagonal matrix whose elements are the corresponding column sums of the matrix $\boldsymbol{W}$. The vector $\boldsymbol{p}_{lpp}^j$ is the jth column of the linear transformation matrix, $\boldsymbol{P}_{lpp}$, which is formed from the eigenvectors associated with the m smallest non-zero eigenvalues. A general discussion of LPP can be found in [7], with ASR specific implementation in [1].

C. Semi-Tied Covariance

A common issue associated with all feature space transformation techniques when applied to ASR is the fact that the transformed features are not guaranteed to be statistically independent. However, because of practical limitations associated with limited training data, most Gaussian mixture based CDHMM ASR systems assume the feature vector dimensions to be "approximately uncorrelated". This is a byproduct of the diagonal covariance assumption imposed on the continuous Gaussian observation densities. Thus, the distribution of feature vectors is mismatched with respect to the densities used for CDHMMs. Therefore, there is a need to incorporate one of a set of procedures that maximizes the likelihood of the data with respect to the diagonal Gaussian models. These procedures, including the maximum likelihood linear transformation (MLLT) [18] and semi-tied covariances (STC) [17], are applied in the transformed feature space. It has been suggested that MLLT is equivalent to a global STC transform [18], [19]. This work adopts the semi-tied covariance (STC) approach for this purpose.

STC approximates full covariance modeling by allowing a number of full covariance matrices to be shared across many Gaussian components, instead of using component specific full covariance matrices. Effectively, each component maintains its own diagonal covariance. Each component consists of two elements, a component specific diagonal covariance matrix, $\boldsymbol{\Sigma}_{diag}^{(m)}$

and a semi-tied regression class dependent non-diagonal matrix, $\boldsymbol{A}^{(r)'}$. The form of the resultant covariance matrix is given by

$$\boldsymbol{\Sigma}^{(m)} = \boldsymbol{A}^{(r)'}\boldsymbol{\Sigma}^{(m)}_{diag}\boldsymbol{A}^{(r)'T} \tag{8}$$

where m specifies the corresponding mixture index, and r refers to the regression class. A detailed discussion of STC can be found in [17].

III. The Family of Discriminative Manifold Learning Techniques

The proposed family of algorithms incorporate discriminative training into manifold based non-linear locality preservation. In order to formulate an optimality criterion based on manifold preservation and inter-class separability, the feature vectors are assumed to be residing on multiple class-specific manifolds, which are characterized by two undirected weighted graphs, namely the intrinsic graph, $\mathcal{G}_{int} = \{\boldsymbol{X}, \boldsymbol{W}_{int}\}$, and the penalty graph, $\mathcal{G}_{pen} = \{\boldsymbol{X}, \boldsymbol{W}_{pen}\}$. The intrinsic graph refers to the properties of the dataset that are to be preserved by virtue of the transformation, whereas the penalty graph refers to the properties that are to be discarded. In other words, the intrinsic graph defines the relationships between the feature vectors belonging to same class/sub-manifold, whereas the penalty graph defines the relationships between the vectors belonging to different classes/sub-manifolds. The possibility of speech features lying on multiple manifolds have been suggested by many researchers [9].

It is important to note that the characteristics of a graph, including structure, connectivity and compactness, are primarily governed by the weights on the edges of the graph, thus by the affinity matrix $\boldsymbol{W}$. Many different measures can be used to define these graph based relationships. In this work, two different distance metrics are used, the conventional Euclidean distance and a cosine-correlation distance measure. This leads to two different algorithms for feature space transformation and dimensionality reduction, namely locality preserving discriminant analysis (LPDA) and correlation preserving discriminant analysis (CPDA), respectively. These algorithms are described in the following sections.

A. Locality Preserving Discriminant Analysis

In LPDA, the graph based relationships are characterized using the Euclidean distance based Gaussian heat kernel. The elements of the intrinsic and penalty graph weight matrices are defined as,

$$w^{int}_{ij} = \begin{cases} \exp\left(\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\rho}\right) & ; C(\boldsymbol{x}_i) = C(\boldsymbol{x}_j), e(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 \\ 0 & ; \text{Otherwise} \end{cases} \tag{9}$$

and

$$w^{pen}_{ij} = \begin{cases} \exp\left(\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\rho}\right) & ; C(\boldsymbol{x}_i) \neq C(\boldsymbol{x}_j), e(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 \\ 0 & ; \text{Otherwise} \end{cases} \tag{10}$$

where ρ is the kernel scale parameter, $C(\boldsymbol{x}_i)$ refers to the class or label of vector $\boldsymbol{x}_i$. The function $e(\boldsymbol{x}_i, \boldsymbol{x}_j)$ indicates whether $\boldsymbol{x}_i$ lies in the near neighborhood of $\boldsymbol{x}_j$. Closeness to a vector $\boldsymbol{x}_i$ can be measured either by k-nearest neighbors (kNN), or neighbors within certain radius r. In the intrinsic graph, $\mathcal{G}_{int}$, a node $\boldsymbol{x}_i$ is connected to the k_{int} nearest neighbors belonging to the same class $C(\boldsymbol{x}_i)$. Similarly, in the penalty graph, $\mathcal{G}_{pen}$, a node $\boldsymbol{x}_i$ is connected to the k_{pen} largest affinity neighbors *not* belonging to the class $C(\boldsymbol{x}_i)$. In this work, the optimal values of k_{int} and k_{pen} were empirically determined to be $k_{int} = k_{pen} = 200$.

Following (5), a scatter measure for a generic graph $\mathcal{G}$ in LPDA is defined as,

$$F_{\mathcal{G}}(\boldsymbol{P}) = \sum_{i \neq j} \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2 w_{ij} \tag{11a}$$

$$= 2\boldsymbol{P}^T\boldsymbol{X}(\boldsymbol{D} - \boldsymbol{W})\boldsymbol{X}^T\boldsymbol{P} \tag{11b}$$

where $\boldsymbol{D}$ is a diagonal matrix whose elements correspond to the column sum of the affinity matrix $\boldsymbol{W}$, *i.e.,* $\boldsymbol{D}_{ii} = \sum_j w_{ij}$.

The goal in LPDA is to estimate the parameters of a projection matrix $\boldsymbol{P} \in \mathbb{R}^{d \times m}$, with $m \leq d$ (generally $m \ll d$), which maximizes the sub-manifold discrimination in the projected feature space while retaining the within-sub-manifold inherent data relations. Thus, the algorithm should maximize the scatter of the penalty graph $F_{pen}(\boldsymbol{P})$, while at the same time minimize the scatter of the intrinsic graph $F_{int}(\boldsymbol{P})$. To this end, the ratio of $F_{pen}(\boldsymbol{P})$ to $F_{int}(\boldsymbol{P})$ is defined as a measure of class separability and graph-preservation,

$$F(\boldsymbol{P}) = \frac{F_p(\boldsymbol{P})}{F_i(\boldsymbol{P})} = \frac{\boldsymbol{P}^T\boldsymbol{X}(\boldsymbol{D_p} - \boldsymbol{W_p})\boldsymbol{X}^T\boldsymbol{P}}{\boldsymbol{P}^T\boldsymbol{X}(\boldsymbol{D_i} - \boldsymbol{W_i})\boldsymbol{X}^T\boldsymbol{P}}. \tag{12}$$

Thus, an optimal projection matrix is the one to maximize the expression in (12):

$$\boldsymbol{P}_{LPDA} = \arg\max_{\boldsymbol{P}} F(\boldsymbol{P}). \tag{13}$$

Or,

$$\arg\max_{\boldsymbol{P}} \left\{ \mathrm{tr}\left(\left(\boldsymbol{X}(\boldsymbol{D}_i - \boldsymbol{W}_i)\boldsymbol{X}^T\boldsymbol{P}\right)^{-1} \times \left(\boldsymbol{P}^T\boldsymbol{X}(\boldsymbol{D}_p - \boldsymbol{W}_p)\boldsymbol{X}^T\boldsymbol{P}\right) \right) \right\} \tag{14}$$

where the subscripts i and p signify 'intrinsic' and 'penalty' graphs respectively. (14) can be formulated into a generalized eigenvalue problem, given as follows:

$$(\boldsymbol{X}(\boldsymbol{D}_p - \boldsymbol{W}_p)\boldsymbol{X}^T)\boldsymbol{p}^j_{lpda} = \lambda_j(\boldsymbol{X}(\boldsymbol{D}_i - \boldsymbol{W}_i)\boldsymbol{X}^T)\boldsymbol{p}^j_{lpda} \tag{15}$$

where $\boldsymbol{p}^j_{lpda}$ is the jth column of the transformation matrix $\boldsymbol{P}_{lpda} \in \mathbb{R}^{d \times m}$ and is the eigenvector associated with the jth largest eigenvalue. Thus the eigenvectors corresponding to m largest eigenvalues constitute the optimal LPDA projection matrix $\boldsymbol{P}_{lpda}$.

B. Correlation Preserving Discriminant Analysis

CPDA follows the same framework as LPDA, but instead uses a cosine-correlation based distance measure to characterize

the relationships between the graph nodes. The motivation for using a cosine-correlation based distance measure arises from the fact that magnitude of the cepstrum vectors, and hence the Euclidean distances based acoustic models are highly susceptible to ambient noise [20]. It has also been found that the angles between cepstrum vectors are comparatively more robust to noise [21]. This suggests a potential advantage to a feature space transformation estimated using a cosine-correlation based objective function, particularly in the context of noise robust ASR.

The first step of training the coefficients of a CPDA transformation is to project the feature vectors onto the surface of a unit hypersphere. This has the effect of discarding magnitude information while retaining the correlation based relationships between data vectors. CPDA provides a projection of the features from the source d-dimensional hypersphere to the target m-dimensional hypersphere. The projected features are given by, $\boldsymbol{y}_i = \boldsymbol{P}^T\boldsymbol{x}_i/\|\boldsymbol{P}^T\boldsymbol{x}_i\|$. Note that such a projection is non-linear. The rest of the algorithm formulation for CPDA is very similar to that for LPDA. The feature vectors are embedded into two undirected graphs, namely, $\mathcal{G}_{int} = \{\boldsymbol{X}, \boldsymbol{W}_{int}\}$ and $\mathcal{G}_{pen} = \{\boldsymbol{X}, \boldsymbol{W}_{pen}\}$. Similarity between the nodes is represented by their cosine-correlation. The elements of the intrinsic and penalty graph weight matrices are defined as,

$$w_{ij}^{int} = \begin{cases} \exp\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle - 1}{\rho}\right) & ; C(\boldsymbol{x}_i) = C(\boldsymbol{x}_j), e_c(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 \\ 0 & ; \text{Otherwise} \end{cases} \tag{16}$$

and

$$w_{ij}^{pen} = \begin{cases} \exp\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle - 1}{\rho}\right) & ; C(\boldsymbol{x}_i) \neq C(\boldsymbol{x}_j), e_c(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 \\ 0 & ; \text{Otherwise} \end{cases} \tag{17}$$

where variables in (16) and (17) follow the same nomenclature as those in (9) and (10), albeit the nearest neighbors are defined in cosine-correlation sense.

Following (5), a generalized scatter measure for a graph $\mathcal{G}$ in the transformed space can be given by

$$F_{\mathcal{G}}(\boldsymbol{P}) = \sum_{i\neq j} \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2 w_{ij} \tag{18a}$$

$$= \sum_{i\neq j} \left\| \frac{\boldsymbol{P}^T\boldsymbol{x}_i}{\|\boldsymbol{P}^T\boldsymbol{x}_i\|} - \frac{\boldsymbol{P}^T\boldsymbol{x}_j}{\|\boldsymbol{P}^T\boldsymbol{x}_j\|} \right\|^2 w_{ij} \tag{18b}$$

$$= 2\sum_{i\neq j} \left(1 - \frac{f_{ij}}{f_i f_j}\right) w_{ij} \tag{18c}$$

where, for two arbitrary vectors $\boldsymbol{x}_u$ and $\boldsymbol{x}_v$, $f_u = \sqrt{\boldsymbol{x}_u^T\boldsymbol{P}\boldsymbol{P}^T\boldsymbol{x}_u}$, and $f_{uv} = \boldsymbol{x}_u^T\boldsymbol{P}\boldsymbol{P}^T\boldsymbol{x}_v$. Similar to LPDA formulation, the goal is to maximize the scatter of the penalty graph $F_{pen}(\boldsymbol{P})$, while at the same time minimize the scatter of the intrinsic graph $F_{int}(\boldsymbol{P})$. It is important to note here that because of the non-linear nature of the projection, it is not possible to formulate the CPDA objective function as the ratio of the scatter measures of the two graphs. As a result, the optimal projection matrix cannot be obtained by solving a generalized eigenvalue problem. For this reason, the difference of the scatter measures is defined as a measure of manifold separability and graph-preservation,

$$F(\boldsymbol{P}) = F_{pen}(\boldsymbol{P}) - F_{int}(\boldsymbol{P}) = 2\sum_{i\neq j}\left(1 - \frac{f_{ij}}{f_i f_j}\right)\cdot w_{ij}^{p-i} \tag{19}$$

where $w_{ij}^{p-i} = w_{ij}^{pen} - w_{ij}^{int}$. The optimal projection matrix is the one to maximize the above function, *i.e.*,

$$\boldsymbol{P}_{CPDA} = \arg\max_{\boldsymbol{P}} F(\boldsymbol{P}). \tag{20}$$

To find the optimal CPDA projection matrix, the gradient ascent rule can be utilized as follows:

$$P := P + \alpha\nabla_P F, \quad \text{with}$$
$$\nabla_P F = 2\sum_{i\neq j}\left[\frac{f_{ij}\boldsymbol{x}_i\boldsymbol{x}_i^T}{f_i^3 f_j} + \frac{f_{ij}\boldsymbol{x}_j\boldsymbol{x}_j^T}{f_i f_j^3} - \frac{\boldsymbol{x}_i\boldsymbol{x}_j^T + \boldsymbol{x}_j\boldsymbol{x}_i^T}{f_i f_j}\right]\boldsymbol{P}\cdot w_{ij}^{p-i} \tag{21}$$

where α is the gradient scaling factor, and $\nabla_P F$ represents the gradient of (19) with respect to $\boldsymbol{P}$.

Unfortunately, since $F(\boldsymbol{P})$ is a non-linear and non-convex function, the gradient ascent is susceptible to converge to a local optima. Therefore, a good initialization is critical for achieving global optima. To this end, an initial projection is obtained by neglecting the normalization of the transformed features to approximate to a linear solution, *i.e.*, by setting $\boldsymbol{y}_i = \boldsymbol{P}^T\boldsymbol{x}_i$. A closed-form solution for the initialization is then achieved by solving the generalized eigenvalue problem [12], [13],

$$(\boldsymbol{X}(\boldsymbol{D}_p - \boldsymbol{W}_p)\boldsymbol{X}^T)\boldsymbol{p}_j = \lambda_j(\boldsymbol{X}(\boldsymbol{D}_i - \boldsymbol{W}_i)\boldsymbol{X}^T)\boldsymbol{p}_j, \tag{22}$$

where $\boldsymbol{D}$ is a diagonal matrix whose elements correspond to the row sum of the matrix $\boldsymbol{W}$. Note that the matrix $\boldsymbol{X}$ here represents the normalized feature vectors. The subscript i and p signifies 'intrinsic' and 'penalty' matrices respectively. The vector $\boldsymbol{p}_j$ indicates the jth column of the *initial* transformation matrix $\boldsymbol{P}$.

C. Noise Aware Manifold Learning

All of the manifold based approaches for estimating linear projections rely on edge-weights for characterizing local neighborhoods in the feature space. These local neighborhoods are defined by the shape and size of the exponential kernels as given in (9)-(10) and (16)-(17). The ASR performance of the manifold learning systems is reasonably robust with respect to the choice of kernel size across varying speaker populations and task domains. However, the interaction between the kernel size and background noise has a significant impact on word accuracy. Experimental results supporting these arguments are shown in Section IV-B–IV-D. This section elaborates on this dependence of kernel size on background noise level and suggests an automated mechanism, referred to as noise-aware manifold learning (NAML), for increasing the noise robustness of manifold learning methods.

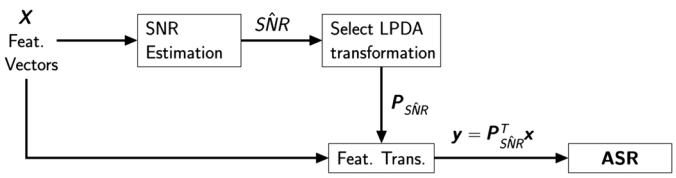

Fig. 2. Illustration of application of noise aware manifold learning to ASR by the example of LPDA.

The kernel governs the compactness of the neighborhood graphs and the smoothness of the manifolds. The shape and size of the kernel is determined by its scale parameter ρ. The selection of this parameter has a crucial effect on the definition of local neighborhoods, and consequently on the characteristics of the linear feature space transformation [1], [22]–[24]. Using a value of the scale parameter which is too large would tend to flatten the kernel leading to a graph where all data pairs are considered equally important. On the other hand, using a value which is too small would result in a graph which lacks sufficient smoothing of the manifold, thus resulting in a kernel which is overly sensitive to noise. Thus, the optimal choice of the kernel size and the governing scale parameter, ρ, is dependent on the SNR level of the speech signal. These claims are supported by experimental results presented in Section IV-D.

One could compensate for this dependence of the optimal choice of kernel scale factor on SNR level by using multiple scale parameters, each specific to a noise condition. As a demonstration of how the relationship between ρ and environmental noise can be exploited, an approach involving the use of multiple scale factor dependent linear transformations is proposed. This is referred to as noise-aware LPDA (N-LPDA).

During the training phase, multiple SNR dependent LPDA projections and CDHMM models are obtained. This procedure is carried out in three steps. First, an ensemble of LPDA projection matrices are trained from the intrinsic and penalty matrices estimated from an ensemble of kernel scale factors, ρ. Second, the optimal value of ρ and hence a specific LPDA transformation, that maximizes ASR performance for a given SNR level, is identified heuristically. Finally, separate CDHMM models are trained from the features obtained by using this ensemble of projection matrices.

Note that an intermediate step of estimating SNR for each speech utterance is involved here. Recent research has produced a number of highly accurate SNR estimation algorithms [25], [26]. This work utilizes a hybrid SNR estimation algorithm based on [25] and [26] to automatically estimate the SNRs for the noise corrupted utterances in the Aurora-2 corpus. The hybrid approach correctly estimates the SNR levels for 85% of the utterances in Aurora-2 test set.

Given an ensemble of multiple SNR dependent LPDA transformations, Fig. 2 describes the steps involved in applying N-LPDA to ASR. SNR is estimated for each test utterance, and the corresponding LPDA transformation matrix is identified. Then feature-space transformation is performed using the chosen LPDA matrix. Finally, the corresponding CDHMM model is used for recognition.

IV. Experimental Study

This section describes the experimental study performed to evaluate the ASR performance of features obtained using the discriminative manifold learning approaches. Two speech-in-noise corpora, a connected digit task and a read newspaper task, are presented along with a summary of the CDHMM ASR systems used in the experiments. The study compares the word error rate (WER) obtained using LPDA and CPDA to that obtained for the more well known techniques, LDA and LPP, over a range of noise types and signal-to-noise ratios (SNRs). The impact of the cosine-correlation based distance measure as compared to Euclidean distance is also considered by comparing the WER obtained using CPDA to that obtained using LPDA. Finally, ASR performance of NAML extension of LPDA (N-LPDA) is compared with that of LPDA.

A. Task Domain and System Configuration

The ASR experiments in this work are conducted on two different datasets. The first is the Aurora-2 connected digit speech in noise corpus. The standard Aurora-2 mixed-condition set is used for training [27]. The training set contains a total of 8440 noise corrupted utterances collected from 55 male and 55 female speakers. For this corpus, the ASR system was configured using whole word CDHMM models with 16 states per word-model, 3 states for the silence model and 1 state for the short pause model. There were 11 word-based CDHMM models. Each state was modeled by a mixture of 3 Gaussians. This corresponds to the standard baseline configuration for the Aurora-2 ASR system specified in [27]. The test dataset consists of a total of 3003 utterances artificially corrupted by three different noise types (subway, car, and exhibition hall) at signal-to-noise ratios (SNR) ranging from 5 to 20 dB, and clean speech. The WERs obtained for the baseline system configuration agrees with that obtained in [27].

The second is the Aurora-4 read newspaper text speech-in-noise corpus, which is obtained by adding noise to the Wall Street Journal (WSJ0) read text corpus [28]. This corpus corresponds to a large vocabulary continuous speech recognition task. The size of vocabulary is 5000 words. The standard Aurora-4 bigram language model is used with a perplexity of 147. The standard Aurora-4 mixed-noise training set is used for training. It contains about 14 hours of speech consisting of a total of 7138 utterances from 83 speakers. The ASR system is configured using cross-word triphone CDHMM models. The system has three state silence models and a single state model for inter-word silence. Each CDHMM state was modeled by a mixture of 16 Gaussians. The test dataset consists of seven subsets, where each subset contains 330 utterances from 8 different speakers. One subset corresponds to uncorrupted high SNR speech, and the remaining six are artificially corrupted by different noise types (car, babble, restaurant, street, airport and train station) at signal-to-noise ratios (SNR) ranging from 5 to 20 dB. The baseline system configuration and WER performance agrees with that given in [28].

Both of these corpora were created by adding noise from multiple environments to speech utterances spoken in a quiet environment. Hence, they represent a simulation of actual speech in

noise domains, and one must be careful about generalizing these results to other speech-in-noise tasks.

For both sets of experiments, the baseline ASR systems are configured using 12-dimensional static Mel-frequency cepstrum coefficient (MFCC) features augmented by normalized log energy, difference cepstrum, and second difference cepstrum resulting in 39-dimensional vectors. The ASR performance is reported in terms of word error rate (WER).

The feature space transformations are estimated using 117 dimensional super-vectors obtained by concatenating 9 frames of MFCCs augmented with log energy. For discriminative training of feature space transformations, class labels are defined as the states of the continuous density hidden Markov models (CDHMM). There are a total of 180 speech classes for the Aurora-2 corpus. For the Aurora-4 corpus, multiple triphone states are tied together based on the central phone to give a total of 120 speech classes. The projection matrix $\boldsymbol{P}$ is used to transform the 117-dimensional training and test vectors into a 39 dimensional space. For both datasets, Aurora-2 and Aurora-4, a neighborhood size of $k = k_{int} = k_{pen} = 200$ is chosen for estimating the intrinsic and penalty graph weights from Section III. Values of the Gaussian kernel heat factor, ρ, have been empirically chosen separately for each of the three manifold learning approaches. The values for ρ were empirically determined using a development speech corpus. The values used for ρ are: 900 for LPP, 1000 for intrinsic and 3000 for penalty graph for LPDA, and 10^{-2} for the intrinsic and penalty graphs of CPDA. Note that the same choice of the kernel scale factor is used for the both datasets. Semitied Covariance transformations (STC) are applied prior to recognition to account for the correlation introduced to the transformed features by the LDA, LPP, LPDA and CPDA projections, as described in Section II-C and [17].

B. Results for Aurora-2 Connected Digit Corpus

Table I compares the WER of CPDA and LPDA transformed features with that of LDA and LPP transformed features for three noise types and four SNRs ranging from 5 dB to 20 dB. Four separate tables are displayed, one for each noise type (subway, exhibition hall and car), and one for average over all noise types. 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 LDA projection matrix to the concatenated MFCC feature vectors as described in Section IV-A. The third row, labeled "LPP" corresponds to the features obtained as a result of the LPP approach. The fourth row "LPDA" corresponds to features obtained by applying the LPDA transformation to the concatenated super-vectors. The final row, labeled "CPDA", corresponds to the ASR WER when CPDA is used as the feature space transformation technique.

For all but baseline features, STC transformations are performed to minimize the impact of the data correlation resulting from the application of feature space transformations. It is important to note that, without applying STC, ASR performance degrades for all of the above transformations. For example, without STC, the WER for LDA features increased by 40% relative for 20 dB subway noise case [13]. Therefore, all experiments in this work have used STC after feature space transformations. This is consistent with the discussion in Section II-C, and results that other researchers have obtained when comparing the performance of feature space transformations with and without STC [13], [18], [19], [29].

TABLE I
WER FOR MIXED NOISE TRAINING AND NOISY TESTING ON AURORA-2 SPEECH CORPUS FOR BASELINE, LDA, LPP, LPDA AND CPDA. THE BEST PERFORMANCE HAS BEEN HIGHLIGHTED FOR EACH NOISE TYPE PER SNR LEVEL

Noise	Technique	SNR (dB)			
		20	15	10	5
Subway	Baseline	2.99	4.00	6.21	11.89
	LDA	2.25	2.93	5.29	12.32
	LPP	2.33	3.50	5.71	13.26
	LPDA	**2.18**	3.29	5.28	11.73
	CPDA	2.30	**2.91**	**4.54**	**11.24**
Exhibition	Baseline	3.34	3.83	6.64	12.72
	LDA	2.63	3.37	6.67	14.29
	LPP	2.56	4.23	8.55	16.91
	LPDA	**2.22**	3.64	6.66	13.85
	CPDA	2.30	**2.95**	**5.37**	**12.59**
Car	Baseline	2.77	3.36	5.45	**12.31**
	LDA	2.83	3.45	5.69	15.92
	LPP	2.71	3.61	6.08	14.97
	LPDA	**2.30**	**2.77**	**5.19**	12.73
	CPDA	2.51	3.52	5.70	14.23
Average	Baseline	3.03	3.73	6.10	**12.31**
	LDA	2.57	3.25	5.88	14.18
	LPP	2.53	3.78	6.78	15.05
	LPDA	**2.23**	3.23	5.71	12.77
	CPDA	2.37	**3.13**	**5.20**	12.68

The baseline WERs displayed for all conditions in Table I agree with those reported in [27], [30]. Table I shows that LDA transformation provides a consistent reduction in WER across all noise types at high SNRs. This is consistent with results reported for this task in [31].

The results in Table I demonstrate advantages of the proposed discriminative manifold learning approaches, LPDA and CPDA, over conventional techniques, LDA and LPP. A number of observations can be made from this table. First, all techniques provide reduced WER at high SNRs for most noise conditions. Second, all techniques show smaller relative reduction in WER at low SNR than at high SNR for most noise conditions. Third, LPDA and CPDA perform better than the conventional techniques, LDA and LPP, in most noise conditions with a relative WER improvement ranging from 6 to 30%. This result appears to support the assertion that the combination of manifold constraints and discriminative learning results in a transformed feature space that is well behaved and robust. Fourth, by comparing the fourth and the fifth rows of Table I, it is clear that CPDA provides a larger reduction in WER than the Euclidean counterpart LPDA at all but the highest SNR for most noise types. This result demonstrates the noise robustness of the cosine-correlation based distance measure as compared to Euclidean distance in

TABLE II
WER FOR CLEAN TRAINING AND CLEAN TESTING ON AURORA-2 SPEECH CORPUS FOR LDA, LPP, LPDA AND CPDA. THE BEST PERFORMANCES HAVE BEEN HIGHLIGHTED IN BOLD

Technique	Avg. WER	(Rel. Improvement)
Baseline	1.07	–
LDA	0.93	(13.08)
LPP	0.90	(15.89)
LPDA	**0.83**	**(22.42)**
CPDA	**0.82**	**(23.36)**

discriminative manifold learning. The noise type "Car" is a notable exception where all the feature space transformation techniques were found to be less effective.

The statistical significance of the differences in WERs for selected system pairs in Table I are reported using the Gillick and Cox matched-pairs significance test [32]. The WER obtained using LPDA features and LDA features for 20 dB subway noise is found to be statistically significant at a confidence level of 99.5%. Furthermore, the performance gains in WERs reported for LPDA with respect to LDA systems are found to be statistically significant for all conditions except for the subway and exhibition hall noise types at 10 dB SNR. For LPDA and CPDA performance comparisons, the difference in WERs are found to be statistically significant with a confidence level of 99.99% for all conditions except for the subway and exhibition hall 20 dB SNR.

Another important observation that can be made from Table I is that relatively high error rates are obtained for LPP as compared to all other techniques for most conditions. This may appear to contradict earlier results reported by Tang and Rose in [1]. However, it should be noted that the work in [1] reports ASR performance using LPP for a task involving relatively clean training and test conditions, whereas the results in Table I correspond to mixed noisy condition training and noisy testing scenarios. Table II presents the results for a clean training and clean test scenario as an average over clean subsets from Aurora-2. Along with WERs, the table also presents relative WER improvements with respect to the baseline. Note that, in this case, LPP performs better than LDA, however, LPDA and CPDA report even higher performance gains. These results are in agreement with those reported in [1]. While highlighting the importance of LPDA over LPP, these experiments suggest that, though the local geometry of the data plays an important role for clean testing, it is the discriminative training that becomes important in the presence on noise.

C. Results for Aurora-4 Read News Corpus

Table III compares the recognition performance of LPDA transformed features with that of LDA transformed features and the baseline system configuration for the Aurora-4 large vocabulary task. The six noisy test scenarios consists of utterances with SNRs ranging from 5 dB to 20 dB. Note that the performance trends in Table III are similar to those in Table I.

The first column in Table III displays the labels of different test cases. The second column gives ASR WER performance for the baseline system, when no feature transformation is performed. The third column, labeled "LDA", corresponds to application of the LDA projection matrix to the concatenated MFCC feature vectors as described in Section IV-A. The fourth column, labeled "LPDA" corresponds to the application of the LPDA transformation to the concatenated super-vectors. The last column also displays the relative WER reductions obtained using LPDA features with respect to LDA features. Similar to the Aurora-2 experiments, for both LDA and LPDA, STC transformations are performed to minimize the impact of the data correlation resulting from the application of feature space transformations.

TABLE III
WER FOR MIXED NOISE TRAINING AND NOISY TESTING ON AURORA-4 SPEECH CORPUS FOR BASELINE, LDA, AND LPDA AND (WER IMPROVEMENT RELATIVE TO LDA). THE BEST PERFORMANCE HAS BEEN HIGHLIGHTED FOR EACH NOISE TYPE

Noise Type	Technique		
	Baseline	LDA	LPDA (rel. LDA)
Clean	15.34	15.09	**13.97** (7.44)
Car	15.90	16.34	**14.53** (11.08)
Babble	26.62	25.37	**21.56** (15.02)
Restaurant	28.28	28.77	**24.51** (14.81)
Street	31.59	29.87	**27.46** (8.07)
Airport	23.65	23.65	**18.96** (19.83)
Train Stn.	32.08	29.96	**28.60** (4.54)
Average	24.78	24.15	**21.37** (11.51)

The results in Table III demonstrate that the effectiveness of discriminative manifold learning LPDA over conventional LDA for a large vocabulary continuous speech recognition task. Similar to the results presented for the Aurora-2 corpus, LPDA obtains improved WER performance over conventional LDA across all noise types. The relative WER improvement of LPDA with respect to LDA ranges from 4.54 to 19.83%. The results presented in Table III are tested for the Gillick and Cox matched-pairs significance test [32]. The WER improvements using LPDA transformed features with respect to LDA features are found to be statistically significant at a confidence level of 99.98% for all conditions. The performance of CPDA transformations was not evaluated on this task for practical reasons. However, based on the similar performance trends observed for the Aurora-2 and Aurora-4 corpora, one might expect reductions in WER on this corpus that are similar to those observed in Table I for Aurora-2.

The results in Table III demonstrate that the relative performance gains obtained for the discriminative manifold learning approaches generalize across task domains and speaker populations. Furthermore, as mentioned in Section IV-A, it is empirically determined that the optimal settings of neighborhood sizes and kernel scale factors used also generalize reasonably well across task domains. This is important when considering the application of these techniques to new corpora. However, while these parameters were found to be robust with respect to task domains, the next section demonstrates that they are not robust with respect to noise level.

D. Results for Noise Aware Manifold Learning

The impact of the kernel scale factor on neighborhood shape and the influence of the environmental noise level on the choice

TABLE IV
COMPARISON OF LPDA ASR PERFORMANCE IN TERMS OF %-WER FOR THREE DIFFERENT VALUES OF ρ, THAT IS, 800, 900, AND 1000 ON AURORA-2. THE BEST CASES HAVE BEEN HIGHLIGHTED IN BOLD

Noise	ρ	Clean	20 dB	15 dB	10 dB	5 dB
Sub.	800	**1.69**	**2.27**	3.65	6.02	13.11
	900	1.70	2.32	3.59	5.54	12.69
	1000	1.83	2.43	**3.29**	**5.25**	**11.82**
Exh.	800	**1.08**	**2.56**	**3.61**	6.79	16.17
	900	1.23	2.87	3.62	6.10	14.78
	1000	1.38	2.56	3.72	**6.08**	**14.04**
Car	800	**1.73**	2.74	3.40	6.83	15.99
	900	1.82	2.48	3.07	5.25	15.52
	1000	2.19	**2.27**	**3.02**	**5.04**	**15.33**

TABLE V
ASR %-WER FOR MIXED NOISE TRAINING AND NOISY TESTING ON AURORA-2 SPEECH CORPUS FOR LPDA USING $\rho = 1000|3000$ AND N-LPDA

Noise	Approach	SNR (dB)				
		20	15	10	5	Avg.
Sub.	LPDA	2.18	3.29	5.28	11.73	5.62
	N-LPDA	2.18	3.25	5.25	11.44	5.53
Exh.	LPDA	2.22	3.64	6.66	13.85	6.59
	N-LPDA	2.28	3.36	6.08	13.85	6.39
Car	LPDA	2.30	2.77	5.19	12.73	5.75
	N-LPDA	2.36	2.92	5.04	12.60	5.74
Avg.	LPDA	2.23	3.23	5.71	12.77	5.99
	N-LPDA	2.25	3.18	5.46	12.63	5.88

of this factor was discussed in Section III-C. This section begins by presenting experimental evidence that demonstrates this influence. Following this, an implementation of noise aware LPDA (N-LPDA) is presented and evaluated on the Aurora-2 task.

The results in Table IV show how the ASR WER over a range of SNRs is influenced by the choice of ρ. ASR WER's for the Aurora-2 corpus using the multi-noise mixed condition training scenario with the LPDA transformed features are given for three different noise types (Sub.=subway, Exh.=exhibition hall, and car). For each noise type, LPDA transformations were trained using a range of kernel scale parameter, from $\rho = 800$ to $\rho = 1000$. Each noise type has five SNR levels (clean, 20 dB, 15 dB, 10 dB, and 5 dB).

It can be observed from the results in Table IV that a smaller ρ value gives better performance in the case of clean speech and high SNR compared to the case when a larger ρ value is used. However, using a kernel with larger ρ value results in better performance for low SNR conditions. This general trend is apparent for all noise types.

N-LPDA was presented in III-C as a mechanism for compensating for this dependence on the choice of ρ by using multiple scale factors, each specific to a given noise level. To demonstrate this mechanism, an ensemble of LPDA transformations are trained using affinity and penalty matrices relying on five different sets of kernel scale parameters: $\{800, 800|900, 900, 1000, 1000|3000\}$. The values in the format '$a|b$' refer to the two different scaling parameters used for the intrinsic and penalty graph kernels, respectively. These values were empirically chosen based on ASR performance obtained on a development set across a range of SNRs. The results of this approach are given in Table V for the various noise conditions described earlier. Results for clean testing have been omitted. For each noise type, ASR %-WERs are compared for LPDA and N-LPDA. Note that the WERs given in Table V for LPDA are identical to those shown in Table I since the settings $\rho = 1000|3000$ are the same. The last column in the table lists ASR WER averaged over all listed SNR levels for each noise condition. The last two rows in Table V, labeled "Avg.", display the ASR WERs for LPDA and N-LPDA averaged across the different noise types. It is apparent from the results in Table V that N-LPDA produces slightly better average ASR performance for most conditions as compared to any single ρ choice. These results suggest that choosing an optimal value of ρ by selecting from an ensemble of alternative transforms may be a plausible approach for reducing the impact of this dependence.

V. DISCUSSION AND ISSUES

There are several aspects of the discriminative manifold learning algorithms that lead to their apparent advantages over the more well known approaches for feature space transformation. The primary factor contributing toward these advantages is the fact that these techniques combine within class sub-manifold learning with inter-class discrimination, as characterized by the intrinsic and penalty graphs, $\mathcal{G}_{int}$ and $\mathcal{G}_{pen}$. LPDA and CPDA essentially use non-linear mapping functions for feature extraction and, therefore, have a greater capability for exploiting geometrical knowledge of the feature space than is possible for linear techniques for feature extraction. There are a number of other factors contributing to the effectiveness of the proposed discriminative manifold learning approaches. This section highlights some of the important factors and issues affecting these techniques.

A. Cosine-Correlation Distance Measure

The improvement in noise robustness reported in Section IV for CPDA relative to LPDA is supported by many previous studies. The fact that adding noise to clean speech results in distortion in the magnitude of cepstrum feature vectors but has a relatively small effect on the angular affinity between cepstrum vectors is well known [20]. Cosine-correlation based distance measures have also been implemented in CDHMM based ASR decoders and were found to achieve lower WERs on speech in noise tasks than the standard Euclidean based measure [21]. It has also been shown in other application domains that cosine-correlation based distance metrics outperform Euclidean or L1-distance metrics for classification tasks [11], [16]. Thus, the CPDA results obtained in Section IV are consistent with previous work.

B. Graph Embedding

There are two distinct advantages that graph embedding provides to the discriminative manifold learning algorithms. The

first is that it enables a mathematical representation of the distribution and geometrical structure of data. The structure of these graphs can be exploited to obtain well behaved feature-space transformations. The second is that by formulating an ASR feature analysis problem in terms of graph structures and scatters one can avoid making any assumption about the distribution of data. This is important as common dimensionality reduction approaches, namely principal components analysis (PCA) and linear discriminant analysis (LDA), work under the assumption of class conditional Gaussian distribution of the data [2]. The high degree of variability in speech production results in a much more complex distribution. Graph embedding avoids such assumptions.

C. Computational Complexity

It is well known that all manifold learning approaches to feature space transformation have extremely high computational complexity when compared to other discriminant feature transformations [11], [16]. The complexity primarily arises from computing the affinity matrices $\boldsymbol{W}_{int}$ and $\boldsymbol{W}_{pen}$. The Aurora-2 task described in Section IV-A involves 180 states and a training corpus of 1.4 million 117-dimensional feature vectors. The Aurora-4 task involves 6 million feature vectors each having dimensionality of 117. These are far larger tasks than those addressed in the application domains described in [10], [11], [16]. This represents a definite disadvantage of the discriminative manifold learning algorithms when applied to the generally very large corpora associated with most speech processing tasks. One mechanism currently being investigated for reducing computational complexity is locality sensitive hashing (LSH) [33], [34]. LSH enables fast nearest neighbor search in high-dimensional spaces, thus allowing for fast computation of affinity matrices $\boldsymbol{W}_{int}$ and $\boldsymbol{W}_{pen}$ [35], [36].

VI. Conclusion

This paper has presented a family of discriminative manifold learning techniques for locality preserving feature space transformation and demonstrated their performance on two ASR tasks. The proposed approaches attempt to preserve the within class manifold based local relationships while at the same time maximizing the separability between classes. This is achieved by embedding feature vectors into undirected graphs by using nonlinear kernels and preserving or penalizing the local structure of the graphs. Two approaches were presented which rely on two different kernels that are based on Euclidean and cosine-correlation distance measures. The performance of the proposed techniques was evaluated on two speech in noise tasks. When compared to well-known approaches such as LDA and LPP, the discriminative manifold learning algorithms demonstrated up to 30% reduction in WER. It was also shown that the use of the cosine-correlation based distance measures was more robust than those based on Euclidean distances when speech is corrupted by noise. Furthermore, these performance gains generalized across task domains and speaker populations.

The effect of acoustic noise conditions on manifold learning approaches has also been investigated. This investigation led to a multi-model approach for improving the robustness of manifold learning based feature space transformations, referred to as noise aware manifold learning (NAML). The approach was shown to provide reduced WER across a range of acoustic conditions with respect to the LPDA transformation implemented without incorporating knowledge of background conditions.

The effectiveness of the discriminative manifold learning based techniques should encourage widespread adoption of these techniques over a range of ASR task domains. Existing work is directed towards the issue of computational complexity of manifold based techniques using locality sensitive hashing (LSH) techniques [35], [36]. It is expected that this should facilitate the application of these techniques to large speech databases.

References

[1] Y. Tang and R. Rose, "A study of using locality preserving projections for feature extraction in speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, NV, USA, Mar. 2008, 2013, pp. 1569–1572.

[2] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. ed. New York, NY, USA: Wiley Interscience, 2000.

[3] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proc. ICASSP*, 1992, vol. 1, pp. 13–16.

[4] K. Beulen, L. Welling, and H. Ney, "Experiments with linear feature extraction in speech recognition," in *Proc. Eur. Conf. Speech Commun. Technol*, 1995.

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

[6] L. Saul and S. Roweis, "Think globally, fit locally: Unsupervised learning of low dimensional manifolds," *J. Mach. Learn. Res.*, vol. 4, pp. 119–155, 2003.

[7] X. He and P. Niyogi, "Locality preserving projections," *Neural Inf. Process. Syst. (NIPS)*, 2002.

[8] K. N. Stevens, *Acoustic Phoenetics*. Cambridge, MA, USA: MIT Press, 1998.

[9] A. Jansen and P. Niyogi, "Intrinsic Fourier analysis on the manifold of speech sounds," in *Proc. ICASSP*, 2006, pp. 241–244.

[10] D. Cai, X. He, K. Zhou, and J. Han, "Locality sensitive discriminant analysis," in *Int. Joint Conf. Artif. Intell.*, 2007, no. 60633070, pp. 708–713.

[11] Y. Ma, S. Lao, E. Takikawa, and M. Kawade, "Discriminant analysis in correlation similarity measure space," in *Proc. 24th Int. Conf. Mach. Learn. (ICML '07)*, 2007, no. 1, pp. 577–584.

[12] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[13] V. S. Tomar and R. C. Rose, "Application of a locality preserving discriminant analysis approach to ASR," in *Proc. 11th Int. Conf. Inf. Sci., Signal Process. Their Applicat. (ISSPA)*, Montreal, QC, Canada, Jul. 2012, pp. 103–107.

[14] V. S. Tomar and R. C. Rose, "A correlational discriminant approach to feature extraction for robust speech recognition," in *Proc. Interspeech*, Portland, OR, USA, 2012.

[15] H. Tang, S. M. Chu, and T. S. Huang, "Spherical discriminant analysis in semi-supervised speaker clustering," in *Proc. Human Lang. Technol.: Annu. Conf. North Amer. Chapter Assoc. Comput. Linguist., Companion Vol.: Short Papers (NAACL '09)*, Jun. 2009, p. 57.

[16] Y. Fu and T. Huang, "Correlation embedding analysis," in *Proc. 15th IEEE Int. Conf. IEEE Image Process. (ICIP '08)*, 2008, pp. 1696–1699.

[17] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 3, pp. 272–281, May 1999.

[18] G. Saon and M. Padmanabhan, "Maximum likelihood discriminant feature spaces," in *Proc. ICASSP*, 2000, pp. 1129–1132.

[19] C. Breslin, "Generation and combination of complementary systems for automatic speech recognition," Ph.D. dissertation, Cambridge Univ., Cambridge, U.K., 2008.

[20] D. Mansour and B. Juang, "A family of distortion measures based upon projection operation for robust speech recognition," *IEEE Speech Signal Process.*, vol. 37, no. 11, pp. 4–7, Nov. 1989.

[21] B. A. Carlson and M. A. Clements, "A projection-based likelihood measure for speech recognition in noise," *Audio*, vol. 2, no. 1, 1994.

[22] H. Xiong, M. N. S. Swamy, and M. O. Ahmad, "Optimizing the kernel in the empirical feature space," *IEEE Trans. Neural Netw.*, vol. 16, no. 2, pp. 460–74, Mar. 2005.
[23] V. S. Tomar and R. C. Rose, "Noise aware manifold learning for robust speech recognition," in *Proc. ICASSP*, 2013, pp. 7087–7090.
[24] J. Wang, H. Lu, K. Plataniotis, and J. Lu, "Gaussian kernel optimization for pattern classification," *Pattern Recogn.*, vol. 42, no. 7, pp. 1237–1247, Jul. 2009.
[25] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," *In Practice*, pp. 2598–2601, 2008.
[26] M. Vondrasek and P. Pollak, "Methods for speech SNR estimation: Evaluation tool and analysis of VAD dependency," *Radioengineering*, vol. 14, no. 1, p. 7, 2005.
[27] H. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *Challenges for the Speech Recogn.*, 2000.
[28] N. Parihar and J. Picone, "Aurora working group: DSR front end LVCSR evaluation," Eur. Telecomm. Standards Inst., Tech. Rep., 2002.
[29] Y. Shekofteh, "Comparison of linear based feature transformations to improve speech recognition performance," in *Proc. 19th Iranian Conf. Elect. Eng. (ICEE)*, 2011.
[30] Q. Zhu and A. Alwan, "Non-linear feature extraction for robust speech recognition in stationary and non-stationary noise," *Comput. Speech Lang.*, vol. 17, pp. 381–402, 2003.
[31] G. Saon, J. Huerta, and E. Jan, "Robust digit recognition in noisy environments: The IBM Aurora 2 system," in *Proc. INTERSPEECH*, no. 1, pp. 0–3.
[32] S. Cox, "The Gillick Test: A method for comparing two speech recognisers tested on the same data," NASA STI/Recon Tech. Rep. N, Tech. Rep., 1988.
[33] S. Har-Peled, P. Indyk, and R. Motwani, "Approximate nearest neighbors: Towards removing the curse of dimensionality," *Theory of Comput.*, vol. 8, no. 1, pp. 321–350, 2012.
[34] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proc. 20th Annu. Symp. Comput. Geometry (SCG '04)*, 2004, p. 253.
[35] V. S. Tomar and R. C. Rose, "Efficient manifold learning for speech recognition using locality sensitive hashing," in *Proc. ICASSP*, 2013, pp. 6995–6999.
[36] V. S. Tomar and R. C. Rose, "Locality sensitive hashing for fast computation of correlational manifold learning based feature space transformations," in *Proc. Interspeech*, 2013.

Vikrant Singh Tomar (S'07) is currently a Ph.D. candidate in the Department of Electrical Engineering at McGill University, QC, Canada. He has been working with Dr. Richard Rose. His research has been in the general area of feature extraction and acoustic modeling for automatic speech recognition. In particular, he has been investigating nonlinear feature extraction techniques, such as manifold learning and deep neural networks based approaches, for noise robust automatic speech recognition. Prior to joining McGill in 2010 Vikrant worked as a research fellow at the Center of Excellence in Telecommunication Engineering at IIT Bombay, Mumbai, India. He obtained his undergraduate degree in Information and Communication Technology from Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, GJ, India in 2008.

Richard C. Rose (F'13) is an Associate Professor and Graduate Program Director of Electrical and Computer Engineering at McGill University in Montreal, QC, Canada. His major area of research is in speech and language processing. His recent research contributions have been in acoustic modeling for speech recognition, robust speech recognition, computer aided human language translation, and computer aided speech therapy. Over his career, he has published over 120 articles in refereed international journals and conference proceedings. He has served as Adjunct Research Scientist at the Human Language Technology Center of Excellence in Baltimore and as Adjunct Professor of ECE at Johns Hopkins University.

Prof. Rose is an IEEE Fellow. Before coming to McGill in 2004 Prof. Rose was a senior member of technical staff at AT&T Labs Research where he contributed to AT&T's speech enabled services and was inventor or co-inventor on twelve patents. His professional service has included General Chair of the IEEE Automatic Speech Recognition and Understanding Workshop, membership in the IEEE Speech Technical Committee, elected membership on the IEEE Signal Processing Society Board of Governors, associate editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, associate editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, and founding editor of the IEEE Speech Technical Committee Newsletter. Prof. Rose is a member of Tau Beta Pi, Eta Kappa Nu, and Phi Kappa Phi.