A STUDY OF USING LOCALITY PRESERVING PROJECTIONS FOR FEATURE EXTRACTION IN SPEECH RECOGNITION

Yun Tang and Richard Rose

Department of Electrical and Computer Engineering,
McGill University, Montreal, Quebec, Canada
yun.tang3@mail.mcgill.ca, rose@ece.mcgill.ca

ABSTRACT

This paper presents a new approach to feature analysis in automatic speech recognition (ASR) based on locality preserving projections (LPP). LPP is a manifold based dimensionality reduction algorithm which can be trained and applied as a linear projection to ASR features. Conventional manifold based dimensionality reduction algorithms are generally restricted to batch mode implementation and it is difficult in practice to apply them to unseen data. It is argued that LPP can model feature vectors that are assumed to lie on a nonlinear embedding subspace by preserving local relations among input features, so it has a potential advantage over conventional linear dimensionality reduction algorithms like principal components analysis (PCA) and linear discriminant analysis (LDA). Experimental results obtained on the Resource Management (RM) data set showed that when LPP based dimensionality reduction was applied in the context of mel frequency cepstrum coefficient (MFCC) based feature analysis, a significant reduction of word error rate (WER) was obtained with respect to standard MFCC features.

Index Terms— speech recognition, feature extraction, manifold learning, locality preserving projections

1. INTRODUCTION

Interest in feature space dimensionality reduction in ASR is motivated largely by the requirement that ASR feature vectors provide an accurate representation of both static and dynamic information in speech while at the same time contain a minimum of information that would be considered irrelevant or redundant. The most widely used approaches for feature analysis in ASR are based on various forms of cepstrum based feature analysis. Cepstrum feature vectors are obtained for ASR by performing linear predictive or filter bank based spectral analysis over a windowed speech segment, applying some form of non-linear amplitude compression, and then applying a discrete cosine transform. These static feature vectors are estimated over 20 to 30 millisecond windows and updated at a rate of 100 frames per second. In addition to static feature vectors, dynamic information characterizing the temporal change in the vicinity of the analytic window is also included in the feature representation. The most common method for integrating dynamic information is to compute first and second order differences of adjacent static feature vectors and simply augment the static feature vector with the first and second order difference vectors. The disadvantage of this approach is that the components of the augmented feature vector are strongly correlated. Another method for integrating dynamic information is to simply to concatenate as many as ten neighboring static feature vectors into a single feature vector. Of course, this representation also results in strongly correlated vector components and feature vector dimensions of well over 100.

There has been a great deal of work directed at obtaining linear feature space transformations that reduce the dimensionality of these high dimensional representations. These techniques all have the effect of reducing the correlation between vector components while maximizing some definition of class separability [1, 2, 3]. The work described in this paper focuses on the use of manifold based methods for reducing feature space dimensionality ASR. The motivation for this approach are recent results demonstrating that benefits can be derived by coding the acoustic speech signal in a nonlinear manifold space which has "local Euclidean" properties [4, 5, 6].

The potential advantages of this type of representation is illustrated by the simple two dimensional example in Figure 1. Points A, B, C, and D lie in a two-dimensional manifold represented by the curve in Figure 1. For neighboring points on the manifold, such as C and D, the distance between the two points can be approximated by the Euclidean distance directly. However, for points that are widely separated on the manifold, such as A and D, the Euclidean distance measured between the two points will much different than the distance measured within the manifold space. Conventional linear projection methods, such as LDA and PCA, are efficient at describing Euclidean space. However, for data that can be characterized as being embedded in a manifold space, nonlinear methods can be considered for revealing the relationship of data along the manifold. Techniques that have been proposed for this purpose include ISOMAP [7], locally linear embedding (LLE) [8], and Laplacian eigenmaps [9]. All of these methods learn the global structure of nonlinear manifolds by exploiting local mutual relationships among input data vectors and have been applied to dimensionality reduction and data visualization. However, in ISOMAP, LLE and Laplacian eigenmaps, the data projections are modeled by nonlinear algorithms and can only be applied to the data that was used to train the parameters of the projection.

In this paper, locality preserving projections (LPP) [10, 11] are applied to the ASR dimensionality reduction problem described above. High dimensional vectors obtained by concatenating consecutive static feature vectors are projected to a low dimensional subspace. LPP is an extension of Laplacian eigenmaps. The LPP based projection from a high dimensional space to a low dimensional space is described by a transformation matrix instead of using a nonlinear mapping method defined on the training set. Hence, it is easy to apply the transformation to unseen data while, at the same time, LPP is still able to preserve the local relationships between input data after

This work was performed in collaboration with the DIVINES FP6 project and supported under NSERC Program Number 307188-2004

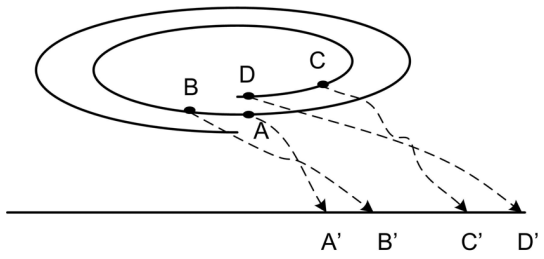

Fig. 1. *Illustration of dimensionality reduction for two-dimensional data embedded in a nonlinear manifold space with relative position information reserved.*

being projected onto a low dimensional subspace.

The rest of the paper is organized as follows. A brief introduction to LDA-based feature extraction is given in Section 2. In Section 3, the LPP algorithm and LPP based feature extraction for ASR is presented. The results of an experimental study comparing LDA and LPP based feature extraction on the Resource Management (RM) task is described in Section 4. Summary and conclusions are provided in the last section.

2. LDA BASED FEATURE EXTRACTION

LDA is a widely used dimensionality reduction approach in speech recognition which acts to preserve the discriminating characteristics of input data in the lower-dimensional transformed space [1, 12, 13]. Consider a set of d_h dimensional input data vectors $X = \{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_N\}$, $\vec{x}_i \in \mathcal{R}^{d_h}$, where each vector belongs to one of P classes. The d_l dimensional representation $Y = \{\vec{y}_1, \vec{y}_2, \cdots, \vec{y}_N\}$ of X is obtained by the linear transformation,

$$\vec{y}_i = W'\vec{x}_i \tag{1}$$

where $\vec{y}_i \in \mathcal{R}^{d_l}$, $d_l < d_h$, and W is a $d_h \times d_l$ matrix.

The matrix W is obtained by LDA so that the linear class separability is maximized by solving the following objective function,

$$\max_{\vec{w}} \frac{\vec{w}' S_B \vec{w}}{\vec{w}' S_W \vec{w}}, \tag{2}$$

where the scatter matrices S_B and S_W are obtained from the class labeled data vectors,

$$S_B = \sum_{p=1}^{P} N_p (\vec{\mu}^{(p)} - \vec{\mu})(\vec{\mu}^{(p)} - \vec{\mu})', \tag{3}$$

$$S_W = \sum_{p=1}^{P} \Big(\sum_{i=1}^{N_p} (\vec{x}_i^{(p)} - \vec{\mu}^{(p)})(\vec{x}_i^{(p)} - \vec{\mu}^{(p)})' \Big). \tag{4}$$

In Equations 3 and 4, the matrix S_W is the within-class scatter matrix, S_B is the between-class scatter matrix, $\vec{x}_i^{(p)}$ is the i-th vector in the class p, N_p is the total number of input data in class p, $\vec{\mu}^{(p)}$ is the mean of class p and $\vec{\mu}$ is the mean of all input vectors. Equation 2 can be solved as a generalized eigenvector problem. The transformation matrix W is formed from the eigenvectors associated with d_l largest eigenvalues.

3. LOCALITY PRESERVING PROJECTIONS BASED FEATURE EXTRACTION

This section describes the theory of locality preserving projections and their application to feature analysis for ASR. First, LPP is presented as a technique that is able to identify lower dimensional spaces that preserve local relationships among data vectors in the transformed space even when the data is assumed to be embedded in a nonlinear manifold space. Second, issues relating to the estimation of LPP based transforms on ASR tasks with large amounts of training data are discussed.

3.1. Locality preserving projections

The optimality criterion used for LPP is based on extending the local mutual relationships that exist among the input data vectors to the vectors of the projected subspace. That is,

$$D_N = \min \sum_{i,j} (\vec{y}_i - \vec{y}_j)'(\vec{y}_i - \vec{y}_j) s_{i,j}, \tag{5}$$

In Equation 5, the local relationships among the input data vectors are described by the terms of the similarity matrix, $S = \{s_{i,j}\}_{N \times N}$, where the similarity relationship is defined as follows,

$$s_{i,j} = \begin{cases} \exp(-\|\vec{x}_i - \vec{x}_j\|^2/\rho), & e(\vec{x}_i, \vec{x}_j) = 1 \\ 0, & e(\vec{x}_i, \vec{x}_j) = 0 \end{cases} . \tag{6}$$

In Equation 7, $e(\vec{x}_i, \vec{x}_j)$ is an indicator function designating whether $\vec{x}_i$ and $\vec{x}_j$ are neighbors and ρ is the heat kernel factor . The neighborhood of a given input vector, $\vec{x}_i$, can be defined as the K-nearest vectors to $\vec{x}_i$ or, alternatively, as the set of vectors that fall within a maximum distance defined by threshold ϵ from $\vec{x}_i$. In this paper, the neighborhood of $\vec{x}_i$ is defined by its K-nearest neighbors where $K = 100$.

Suppose $y_i = \vec{w}'\vec{x}_i$ is one-dimensional representation of original feature $\vec{x}_i$, and substituting $\vec{y}_i$ in Equation 5 by y_i, the optimization criterion in Equation 5 can be rewritten as

$$\begin{aligned} \frac{1}{2} D_N &= \frac{1}{2} \sum_{i,j} (y_i - y_j)^2 s_{i,j} \\ &= \frac{1}{2} \sum_{i,j} (\vec{w}'\vec{x}_i - \vec{w}'\vec{x}_j)^2 s_{i,j} \\ &= \sum_i \vec{w}'\vec{x}_i (\sum_j s_{i,j}) \vec{x}_i'\vec{w} - \sum_{i,j} \vec{w}'\vec{x}_i (s_{i,j}) \vec{x}_j'\vec{w} \\ &= \vec{w}' X L X' \vec{w} \end{aligned} \tag{7}$$

where $L = C - S$ is the Laplacian matrix. The matrix C is a diagonal matrix whose entries are the column sums of S, $c_{i,i} = \sum_j s_{i,j}$. In order to get an unique solution, a constraint is imposed on the magnitude of the transformed vectors,

$$\vec{w}' X C X' \vec{w} = 1. \tag{8}$$

Hence, minimizing the objective function given in Equation 8 under the constraints specified in Equation 9 corresponds to solving a generalized eigenvalue problem

$$X L X' \vec{w} = \lambda X C X' \vec{w}. \tag{9}$$

The linear project matrix W is formed from the eigenvectors associated with d_l smallest non-zero eigenvalues. Further discussion of the LPP algorithm can be found in [10].

3.2. Class based LPP for ASR tasks

The main issue associated with estimating the transformation matrix, W, using the LPP method described in Section 3.1 for ASR applications is estimating the $N \times N$ similarity matrix, S, defined in Equation 5. For typical large vocabulary ASR tasks, it is not unusual for training utterances to consist of tens of millions of feature frames. Furthermore, the input vectors, $\vec{x}_i, i = 1, \ldots, N$, are generally formed by concatenating multiple adjacent feature frames, resulting in input vector dimension, d_h, of well over 100. This makes it impractical to compute and store the N^2 similarity terms, $s_{i,j}$.

There are many ways to address this issue. In this work, the similarity terms in Equation 5 were computed only for input feature vectors that were labeled as belonging to the same class, p, of the P classes defined in Section 2. The class label assigned to a given input vector can be defined in many ways. These might include the HMM state that the associated feature frame was assigned to, the subword model associated with this state, or any number of other alternatives. Suppose the complete set of input feature vectors, X, are segmented into class specific subsets, $X_p = \{\vec{x}_1^{(p)}, \vec{x}_2^{(p)}, \ldots, \vec{x}_{N_p}^{(p)}\}$, so that $X = \{X_1, X_2, \ldots, X_P\}$. Restricting the computation of the similarity terms to be between input vectors of the same class results in a block diagonal similarity matrix,

$$S = \begin{pmatrix} S_1 & 0 & \ldots & 0 \\ 0 & S_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & S_P \end{pmatrix}. \quad (10)$$

Each S_p is an $N_p \times N_p$ dimension matrix whose elements correspond to the similarity terms for input vectors labeled as belonging to class p. Similarly, the matrices C and L in Equation 10 are also block diagonal with class specific matrices, C_p and L_p, corresponding to class p. Therefore, XLX' and XCX' in Equation 10 can be computed from the class specific submatrices as

$$XCX' = \sum_{p=1}^{P} X_p C_p {X_p}' \quad (11)$$

$$XLX' = \sum_{p=1}^{P} X_p L_p {X_p}'. \quad (12)$$

Both the storage and computational requirements for the block diagonal similarity matrix in Equation 11 are far less than for the general similarity matrix defined in Section 3.1. Furthermore, it is worth noting that the optimization criterion used in LPP is strictly based on preserving local relationships that exist among input data vectors and does not incorporate any notion of class separability. It is reasonable to argue that restricting the similarity measure given in Equation 7 to be computed among data vectors that are nearest neighbors within the same class may potentially improve the statistical robustness of the procedure. However, since evaluating the more general assumption of class independent similarity was considered to be intractable in this work, this issue has been left as an open question.

4. EXPERIMENTS

This section describes an experimental study that was performed to evaluate the ability of the feature space transformations obtained using locality preserving projections to reduce ASR word error rate (WER) on the Resource Management (RM) task. LPP is applied to the scenario described in Section 3, where multiple adjacent cepstrum feature vectors are concatenated and a linear feature space transformation is applied to reducing the dimensionality of the concatenated freatures. Performance of LPP based dimensionality reduction is compared with the well known LDA approach that was descirbed in Section 2.

4.1. RM Task and Baseline ASR System

This section briefly describes the RM speech corpus and the baseline ASR system that was configured for this task. The standard RM SI-109 training set with 3990 utterances and 3.3 hours of speech spoken by 109 speakers was used to train context dependent subword HMM acoustic models. WER was evaluated on the standard DARPA test sets which include Feb'89, Oct'89, Feb'91, and Sep'92. Each test set consists of 300 utterances spoken by 10 speakers.

Feature analysis in the baseline MFCC based system includes 12 mel frequency cepstrum coefficients, normalized log energy, and their first and second difference coefficients all concatenated to obtain a 39-dimension feature vector. All HMM systems evaluated in this section are based on subword models formed from left-to-right 3-state HMMs with 6 diagonal Gaussians per state. The standard RM word-pair grammar is used as the language model for ASR.

4.2. Feature Space Transformations

The input vectors for all feature space transformations, implemented using LPP or LDA approaches, consist of 5 to 11 concatenated MFCC +Energy static feature vectors. This corresponds to an input dimension ranging from 65 to 143. The transformed vectors for all the experiments described in this section are 39 dimensional.

There are two issues that affect the behavior of the data similarity terms in Equation 7. The first is the definition of the K-nearest neighbor region surrounding an input vector which is associated with the indicator function, $e()$. A value of $K = 100$ is used for all of the LPP experiments. The second issue is the choice of ρ in Equation 7. Empirical estimation of ρ will be discussed below.

Finally, a class, p, for both the LDA and LPP feature transformations is defined to be a HMM state. There are $P = 1562$ clustered states in the HMM system described above, and state labels are assigned to concatenated input vectors through Viterbi segmentation with the baseline MFCC based HMM models.

Table 1 displays the WERs for the baseline MFCC-based system and multiple systems implemented using LDA and LPP feature transformations evaluated on all four RM test sets. The input data vectors for all LDA and LPP based systems in Table 1 consist of 9 concatenated static feature vectors. Results are displayed for the cases where LDA and LPP transformations are applied independently and also for the cases where LDA and LPP transformations are followed by a maximum likelihood linear transformation (MLLT) [14, 12]. MLLT is a well known technique for obtaining a data projection that has the effect of the maximizing the likelihood of the projected data under a diagonal covariance assumption.

It is clear from Table 1 that the relative performance of all transformation techniques varies considerably across the different RM test sets. It can also be seen that the average performance obtained using the LPP transformation, displayed in column 6, represents a relative improvement of 6.8% with respect to the average baseline WER of 4.4%. After applying the MLLT transformation, the combined LPP+MLLT transformation resulted in relative WER reduction of 16.1% relative to the baseline system and a 6.1% relative

reduction in WER relative to the LDA+MLLT system.

Table 1. *WER obtained on RM test sets using Ba seline MFCC features, LDA features, and LPP features*

Feature Set	RM Test Set				
	Feb89	Oct89	Feb91	Sep92	Ave.
baseline	2.85	3.76	4.03	6.96	4.40
LDA	3.12	3.17	3.90	6.92	4.28
LDA+MLLT	2.73	3.32	3.34	6.33	3.93
LPP	3.16	3.73	3.34	6.14	4.10
LPP+MLLT	3.01	3.06	2.90	5.78	**3.69**

Figure 2 describes the effect of varying the number of concatenated static feature frames that are input to the LPP and LDA transformations. The average WERs are displayed for the systems as the number of concatenated frames are varied from 5 to 11. Figure 2 shows that the effect on WER observed when reducing the number of concatenated input frames below nine frames is far less for LPP than that for LDA.

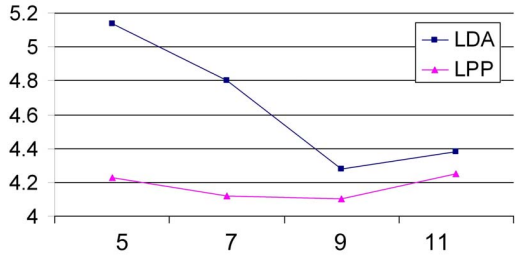

Fig. 2. *Average WER obtained using LPP and LDA transformations on data vectors consisting of from 5 to 11 concatenated frames.*

Table 2 describes the effect of varying the heat kernel factor ρ in Equation 7. The average WERs are displayed for values of ρ ranging from 1 to 1000 and for input data vectors consisting of 7 and 9 concatenated frames. It is clear from the table that the selection of ρ has a significant effect on WER. This is especially true for the higher dimensional input feature space. Currently, obtaining an optimum value for ρ for a specific task must be done empirically.

Table 2. *Average WER obtained using transformed features with different ρ on the RM test set.*

Frms.$\backslash\rho$	1.0	2.0	5.0	1000.0
7	5.67	**4.12**	4.48	4.98
9	7.55	4.42	**4.10**	5.02

5. SUMMARY AND CONCLUSIONS

In this paper, a new manifold based dimensionality reduction algorithm, LPP, was applied to feature extraction in ASR. LPP is fundamentally different from existing feature dimensionality reduction approaches that have been applied to ASR. This is because it exploits the assumption that speech is embedded in a nonlinear manifold space. The advantage of LPP with respect to other manifold based methods is that it is applied in the form of a linear transformation matrix and can be easily employed to unseen data. A class based LPP was implemented to handle the large amounts of training data in ASR, and huge memory and computation requirements were avoided. Finally, in experiments performed on the RM corpus, LPP has shown to provide a 16% reduction in WER with respect to a baseline MFCC based ASR system.

6. REFERENCES

[1] K. Beulen and H. Ney, "Experiments with linear feature extraction in speech recognition," in *European Conference on Speech Communication and Technology*, Madrid, Spain, 1995, pp. 1415–1418.

[2] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition," *Speech Communication*, vol. 26, pp. 283–297, 1998.

[3] J. W. Hung and L. S. Lee, "Optimization of temporal filters for constructing robust features in speech recognition," *IEEE Transactions on Audio, Speech and language Processing*, vol. 14, no. 3, pp. 808 – 832, 2006.

[4] V. Jain and L. Saul, "Exploratory analysis and visualization of speech and music by locally linear embedding," in *IEEE Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2004, pp. 984–987.

[5] A. Jansen and P. Niyogi, "Intrinsic fourier analysis on the manifold of speech sounds," in *IEEE Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, 2006, pp. 241–244.

[6] A. Errity and J. McKenna, "An investigation of manifold learning for speech analysis," in *Proceedings of the International Conference on Spoken Language Processing*, Pittsburgh, USA, 2006, pp. 2506–2509.

[7] J.B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[8] S.T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[9] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems 14*, 2001.

[10] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems 16*, 2003.

[11] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, 2005.

[12] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *IEEE Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, 2000, pp. 1129–1132.

[13] P. Somervuo, B. Chen, and Q. Zhu, "Feature transformations and combinations for improving ASR performance," in *European Conference on Speech Communication and Technology*, Geneva, Switzerland, 2003, pp. 477–480.

[14] R. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *IEEE Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Seattle, USA, 1998, pp. 661–664.