NOISE AWARE MANIFOLD LEARNING FOR ROBUST SPEECH RECOGNITION

Vikrant Singh Tomar, Richard C. Rose

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

ABSTRACT

This paper considers the application of discriminative manifold learning approaches in feature analysis for automatic speech recognition (ASR). The issue of manifold learning is addressed for feature space dimensionality reduction in domains involving noise corrupted speech. The locality preserving discriminant analysis (LPDA) approach to manifold learning is investigated. This class of techniques exploits the assumption that there is a structural relationship among data vectors which can be maintained by preserving the local relationships among the transformed data vectors. The paper presents a procedure for reducing the impact of varying acoustic conditions on manifold learning. Noise aware manifold learning (NAML) is described as an approach for exploiting estimated background characteristics to define the size of the local neighborhoods used for LPDA feature space transformations. It is shown that NAML significantly reduces the speech recognition WER in a noisy speech recognition task over LPDA, particularly at low signal-to-noise ratios.

Index Terms: Locality preserving discriminant analysis, graph embedding, dimensionality reduction, speech recognition

1. INTRODUCTION

Augmenting static features with dynamic spectral information has long been an important aspect of acoustic feature extraction for automatic speech recognition (ASR). One of the widely accepted techniques used for capturing this dynamic information is to combine multiple consecutive static feature vectors to form high dimensional super-vectors. These super vectors may represent on the order of 100 milliseconds of speech and have dimensionality as high as 200. This representation, however, introduces two issues. The first is high dimensionality of the super-vectors, and the second is high degree of correlation between feature vector components. These issues have inspired the use of feature space dimensionality reduction and discriminant analysis approaches in ASR.

In previous work [1], a new approach was presented for feature space transformation, termed 'locality preserving discriminant analysis' (LPDA) as applied to ASR. LPDA is a discriminative manifold learning technique. It not only attempts to preserve the underlying local sub-manifold based relationships of feature vectors but also maximizes a criterion related to the separability between classes of feature vectors. In [1], speech features derived from LPDA were reported to produce significant gains in ASR performance compared to features derived using other well-known feature space transformation techniques, such as linear discriminant analysis (LDA) [2] and locality preserving projections (LPP) [3, 4]. LDA discriminates between classes of feature vectors, but it does not take into account the local manifold based relationships of the data. LPP is an manifold based locality preservation technique, but it does not consider the inter-class discriminant structure of the data. Another example of discriminative feature space transformation techniques is heteroscedastic linear discriminant analysis (HLDA) [5] that takes into account heteroscedastic distribution of speech features in different classes. 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) with LDA [6, 7].

This work addresses the issue of performance degradation of manifold learning based feature space transformation approaches when applied to noise corrupted speech. Noise aware manifold learning (NAML) is presented as an approach which addresses the interaction between acoustic noise conditions and the structure of local neighborhoods used in manifold learning. The NAML framework is motivated by the fact that the shape and size of the local manifold structures are affected by the choice of the kernel scale parameter. The selection of this parameter has a crucial effect on the behavior of kernel, and consequently the performance of the features [4, 8]. Since the neighborhood structure is also affected by the presence of noise, there exists a significant interplay between the Gaussian kernel scale factor and acoustic noise. This work presents NAML as an effective way to apply manifold learning techniques to varying acoustic environments in ASR.

The relationship of this paper to previous work includes the application of manifold learning algorithms with some notion of discriminative power in other domains. Deng et al. [9] reported significant improvement in a face-recognition task while using a locality preserving discriminant technique. Chen et al. [10], and Yan et. al. [11] also reported gain in face-recognition accuracy using this class of techniques. The goal of this work is to exploit estimated background characteristics to select the size of the local neighborhoods used for LPDA feature space transformations. The discriminative manifold learning LPDA is chosen as an example manifold learning technique primarily because of the good performance obtained using LPDA in previous work [1]. However, the NAML framework can be extended to other manifold learning algorithms.

The rest of this paper is structured as follows. An introduction to LPDA is presented in Section 2. After describing the experimental setup in Section 3, Section 4.1 motivates the need for an NAML approach by analyzing the effect of noise level on the optimal choice of Gaussian kernel scale factor for manifold learning techniques. Noise aware extension of LPDA is presented in Section 4.2. Section 5 presents further evidence suggesting the impact of noise on manifold learning techniques. Finally, Section 6 concludes the paper.

2. LOCALITY PRESERVING DISCRIMINANT ANALYSIS

This section briefly introduces the LPDA formulation. Interested readers are referred to [1, 10, 11] for a more detailed discussion.

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

A generic supervised feature space dimensionality reduction problem can be defined as follows. Consider a training set of feature vectors, $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots \boldsymbol{x}_N\}$ in $\mathbb{R}^d$ such that all the vectors in $\boldsymbol{X}$ are labeled as belonging to one of a set of classes $c \in \{c_1, c_2, \cdots, c_{N_c}\}$, where N_c is the number of classes. The goal is to estimate the parameters of a projection matrix $\boldsymbol{P} \in \mathbb{R}^{d \times m}$, with $m \leq d$, to perform a constrained transformation of the features from a d-dimensional space onto an m-dimensional space.

LPDA attempts to maximize class separability, while preserving the local sub-manifold based relationships of the data vectors. Following the generic framework of graph embedding [11], LPDA acts by embedding the training dataset into two undirected weighted graphs, namely, the intrinsic graph $\mathcal{G}_i = \{\boldsymbol{X}, \boldsymbol{W}_i\}$ and the penalty graph $\mathcal{G}_p = \{\boldsymbol{X}, \boldsymbol{W}_p\}$. Here $\boldsymbol{X}$ represents the nodes of the graphs, which correspond to the vectors in the dataset. Therefore, $\boldsymbol{X}$ is same for both the intrinsic and penalty graphs. $\boldsymbol{W}_i$ and $\boldsymbol{W}_p \in \mathbb{R}^{N \times N}$ are the intrinsic and penalty affinity edge-weight matrices, respectively.

The affinity matrices characterize the statistical and geometrical similarities of the data points. The elements of the matrices are defined in terms of a Gaussian kernel as,

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

and

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

where ρ is the kernel scale parameter. $C(\boldsymbol{x}_i)$ denotes the class or label of vector $\boldsymbol{x}_i$. The function $I(\boldsymbol{x}_i, \boldsymbol{x}_j)$ indicates whether $\boldsymbol{x}_i$ lies in the near neighborhood of $\boldsymbol{x}_j$. In this work, a node $\boldsymbol{x}_i$ is connected to the 200 nearest neighbors belonging to the same class $C(\boldsymbol{x}_i)$ in the intrinsic graph, $\mathcal{G}_i$. Similarly, $\boldsymbol{x}_i$ is connected to the 200 closest neighbors *not* belonging to the class $C(\boldsymbol{x}_i)$ in the penalty graph, $\mathcal{G}_p$.

A scatter measure for a graph $\mathcal{G}$ over the target space vectors $\boldsymbol{y}_i \; \forall i = 1, 2, \cdots, N$, where $\boldsymbol{y}_i$ is obtained according to the projection $\boldsymbol{y}_i = \boldsymbol{P}^T \boldsymbol{x}_i$, can be defined by,

$$F(\boldsymbol{P}) = \sum_{i \neq j} ||\boldsymbol{y}_i - \boldsymbol{y}_j||^2 w_{ij} \tag{3a}$$

$$= 2\boldsymbol{P}^T \boldsymbol{X}(\boldsymbol{D} - \boldsymbol{W})\boldsymbol{X}^T \boldsymbol{P} \tag{3b}$$

where $\boldsymbol{D}$ is a diagonal matrix whose elements correspond to the column sum of the affinity matrix $\boldsymbol{W}$, *i.e.,* $\boldsymbol{D}_{ii} = \sum_j w_{ij}$. Depending on whether the goal is to preserve or eliminate the concerned graph structure, an optimal projection matrix $\boldsymbol{P}$ can be obtained by minimizing or maximizing the scatter in Eq. (3b).

In LPDA, the goal is to penalize the structural properties of the penalty graph, while preserving the structure from the intrinsic graph. To this end, the ratio of the penalty graph scatter measure to that of the intrinsic graph is maximized to obtain the optimal projection matrix as,

$$\underset{\boldsymbol{P}}{\arg\max}\, \text{tr}\left((\boldsymbol{X}(\boldsymbol{D}_i - \boldsymbol{W}_i)\boldsymbol{X}^T\boldsymbol{P})^{-1}(\boldsymbol{P}^T\boldsymbol{X}(\boldsymbol{D}_p - \boldsymbol{W}_p)\boldsymbol{X}^T\boldsymbol{P})\right) \tag{4}$$

where the subscripts i and p signify ‘intrinsic’ and ‘penalty’ graphs, respectively [1,11]. Eq. (4) can be solved as a generalized eigenvalue problem,

$$(\boldsymbol{X}(\boldsymbol{D}_p - \boldsymbol{W}_p)\boldsymbol{X}^T)\boldsymbol{p}_{lpda}^j = \lambda_j(\boldsymbol{X}(\boldsymbol{D}_i - \boldsymbol{W}_i)\boldsymbol{X}^T)\boldsymbol{p}_{lpda}^j \tag{5}$$

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

3. TASK DOMAIN AND SYSTEM CONFIGURATION

The ASR performance experiments in this work are conducted on the European Telecommunications Standards Institute’s Aurora-2 speech in noise corpus. The training set consists of 8440 utterances collected from 55 male and 55 female speakers. The test dataset consists of a total of 4004 utterances artificially corrupted by three different noise types (subway, exhibition hall, and car) at SNRs ranging from 5 to 20 dB, and clean speech. The corpus was created by adding noise from multiple environments to connected digit utterances spoken in a quiet environment. As a result, the corpus represents a simulation of a speech in noise task, and one must be careful about generalizing these results to the wide range of actual speech in noise tasks.

The baseline features are configured as 39-dimensional Mel-frequency cepstrum coefficients (MFCC), consisting of 12 static coefficients, normalized log energy, and augmented by Δ-cepstrum, and $\Delta\Delta$-acceleration. The transformations, LDA, LPP and LPDA, are estimated using 117 dimensional super-vectors obtained by concatenating 9 vectors of MFCC augmented with log energy. For LDA and LPDA, classes are defined as the states of the continuous density hidden Markov models (CDHMM). The resultant projection matrix $\boldsymbol{P}$ is then used to perform discriminant feature space transformations on the 117-dimensional training and test vectors and project the features to a 39 dimensional space. During affinity calculations, a neighborhood size of $k = k_i = k_p = 200$ is taken for all the experiments.

The ASR system is configured using whole word CDHMM models with 16 states per word-model, plus 3 states for the silence model, and 1 state for the short pause model. There are 11 word-based CDHMM models, and a total of 180 states. Each state is modeled by a mixture of 3 Gaussians. STC transformations are applied prior to recognition to account for the correlation introduced to the transformed features by the LPDA, LPP and LDA projections, as described in [1, 6]. Finally, the ASR performance is reported in terms of %-word error rate (WER).

4. NOISE AWARE MANIFOLD LEARNING

This section presents noise aware manifold learning as applied to the LPDA approach described in Section 2. LPDA is selected as an example manifold learning technique primarily because of the good performance obtained using LPDA in previous work [1]. The need for a noise aware approach is motivated in Section 4.1 by quantifying the relationship between the Gaussian kernel size and the acoustic noise conditions. Noise aware LPDA (N-LPDA) is presented in Section 4.2 as an automatic procedure for choosing the kernel scale factor based on the estimated SNR level for various noise conditions.

4.1. Importance of the Gaussian kernel scale factor

The Gaussian kernel scale factor, ρ in Eq. (1) and (2), governs the data relationships along the manifold, and in-turn plays an important role in the performance of the manifold learning algorithms [4, 8]. Most often this parameter is heuristically tuned to the given dataset. Too large a scale factor would tend to flatten the Gaussian kernel leading to a near-linear behavior, whereas a too small value would lack sufficient smoothing of the manifold, thus resulting in a kernel highly sensitive to noise. These claims are supported by experimental results presented in this section. For these experiments a multi-noise mixed training dataset was used to minimize the environmental mismatch between the training and testing conditions.

Table 1 demonstrates the dependence of LPDA approach on ρ values. ASR performance results using multi-noise mixed CDHMM training on the LPDA transformed features corresponding to two different values of ρ, $\rho_1 = 800, \rho_2 = 1000$, are given for three different noise types (Sub.=subway, Exh.=exhibition hall, and car). The next five columns of the table display the ASR %-WER performance for five different SNR levels (clean, 20 dB, 15 dB, 10 dB, and 5 dB).

Table 1. Comparison of LPDA ASR performance in terms of %-WER for two different values of ρ, *viz.,* $\rho_1 = 800, \rho_2 = 1000$. The best of the two 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
	1000	1.83	2.43	**3.29**	**5.25**	**11.82**
Exh.	800	**1.08**	**2.56**	**3.61**	6.79	16.17
	1000	1.38	2.56	3.72	**6.08**	**14.04**
Car	800	**1.73**	2.74	3.40	6.83	15.99
	1000	2.19	**2.27**	**3.02**	**5.04**	**15.33**

It can be observed from the results in Table 1 that $\rho = 800$ gives better performance in case of clean speech compared to $\rho = 1000$, however, performance falls faster as noise increases. Eventually $\rho = 1000$ produces better performance for high noise conditions. This trend is visible for all the noise types.

To conclude, an important finding that can be derived form these results is that the optimal value of the scale parameter is heavily influenced by the level of noise corruption in speech. Such dependence of the optimal choice of kernel scale factor on SNR level can be handled by multiple scale factors, each specific to a noise condition. An automatic scheme to achieve this is discussed next.

4.2. Noise aware LPDA

A noise aware LPDA scheme, which can automatically choose a noise-matched LPDA transformation for a given noise condition, is described as follows. First, a number of projection matrices are trained based on different values of ρ. Second, a heuristic technique, like cross-validation, is used to associate an optimal value of ρ – and hence a specific LPDA projection – to each SNR level of the data based on ASR performance. Note that an intermediate step of estimating the SNRs of the speech utterances is involved here. Recent research has produced a number of highly accurate SNR detection algorithms in speech domain [12, 13]. This work utilized a hybrid SNR detector based on the two algorithms [12, 13] to achieve average SNR detection accuracy of 85%. Then, separate CDHMM models are trained from the features obtained by using these different projection matrices. During testing, SNR is estimated for each testing utterance, followed by feature space transformation using the optimal projection matrix associated with the corresponding SNR level, and finally the corresponding model is used for recognition.

In this work, five different ρ's were chosen $\{800, 800|900, 900, 1000, 1000|3000\}$ based on ASR performance from a cross-validation experiment to train N-LPDA transformations. Here, the values in the format '$a|b$' refer to the two different scaling factors used for the intrinsic and penalty graph kernels, respectively. The results comparing performance of N-LPDA with that of three other LPDA transformations for different ρ values are shown in Table 2. Three separate tables are presented for three different noise types (Sub.=subway, Exh.=exhibition hall, and car). For each LPDA technique, ASR WER result are given for four different noise levels ranging from 20 to 5 db SNR for each noise type. The last column of the table shows ASR WER performance averaged over the four SNR levels.

By comparing the average ASR WER in the last column of Table 2 for different LPDA techniques in different noise types, a clear observation can be drawn that N-LPDA gives improved average ASR performance compared to any single ρ choice. A close inspection of ASR WER for different SNR levels shows that N-LPDA provides improved overall performance in most cases. Thus, it is safe to say that the N-LPDA scheme reduces the dependency of neighborhood size on noise level for manifold learning schemes.

Table 2. ASR %-WER for mixed noise training and noisy testing on Aurora-2 speech corpus for LPDA technique with different ρ values. The best performance has been highlighted for each noise condition.

Noise	LPDA (ρ)	SNR (dB)				
		20	15	10	5	Avg.
Sub.	800	2.27	3.65	6.02	13.11	6,26
	1000	2.43	3.29	**5.25**	11.82	5.70
	1000|3000	**2.18**	3.29	5.28	11.73	5.62
	N-LPDA	**2.18**	**3.25**	**5.25**	**11.44**	**5.53**
Exh.	800	2.56	3.61	6.79	16.17	7.28
	1000	2.56	3.72	**6.08**	14.04	6.60
	1000|3000	**2.22**	3.64	6.66	**13.85**	6.59
	N-LPDA	2.28	**3.36**	**6.08**	**13.85**	**6.39**
Car	800	2.74	3.40	6.83	15.99	7.24
	1000	**2.27**	3.02	**5.04**	15.33	6.42
	1000|3000	2.30	**2.77**	5.19	12.73	**5.75**
	N-LPDA	2.36	2.92	**5.04**	**12.60**	**5.74**

To further demonstrate the effectiveness of N-LPDA, another set of experiments are performed where the ASR WER performance of N-LPDA and LPDA is compared to other well know feature space transformation techniques, namely linear discriminant analysis (LDA) and locality preserving projections (LPP). The results for this experiment are compared in Table 3 for the various noise conditions as described earlier. For each noise type, ASR WER are given for four different feature types, namely MFCC, LDA, LPP and N-LPDA (noise aware LPDA). It is apparent from the results in Table 3 that noise aware LPDA produces improved ASR performance compared to other feature extraction techniques for most noise conditions.

5. FURTHER EVIDENCE CHARACTERIZING THE IMPACT OF NOISE ON MANIFOLD LEARNING

The purpose of this section is to demonstrate the impact of mismatched environmental conditions on the LPDA manifold learning approach for estimating ASR feature space transformations. It is argued in Section 1 that manifold learning techniques benefit from the assumption that there is a structural relationship amongst data vectors which can be maintained by preserving the local relationships among the transformed data vectors. This suggests that if the presence of noise blurs these local relationships, the effectiveness of these techniques will be diminished. This phenomenon is examined under highly mismatched acoustic conditions by estimating LPDA transforms and training CDHMM models under clean conditions and evaluating ASR WER under a number of noisy conditions.

Table 3. ASR %-WER for mixed noise training and noisy testing on Aurora-2 speech corpus for LDA, LPP and N-LPDA. The best performance has been highlighted for each noise condition.

Noise	Feat.	SNR (dB)				
		Clean	20	15	10	5
Sub.	MFCC	1.76	2.99	4.0	6.21	11.89
	LDA	1.82	2.25	**2.93**	5.29	12.32
	LPP	1.66	2.33	3.50	5.71	13.26
	N-LPDA	**1.44**	**2.18**	3.25	**5.25**	**11.84**
Exh.	MFCC	1.89	3.34	3.83	6.64	**12.72**
	LDA	1.83	2.63	3.37	6.67	14.29
	LPP	1.76	2.56	4.23	8.55	16.91
	N-LPDA	**1.14**	**2.38**	**3.36**	**6.08**	13.85
Car	MFCC	1.99	2.77	3.36	5.45	**12.31**
	LDA	2.29	2.83	3.45	5.69	15.92
	LPP	1.88	2.71	3.61	6.08	14.97
	N-LPDA	**1.67**	**2.56**	**2.92**	**5.04**	13.60

There are two observations made in this section to support the hypothesis that environmental mismatch may affect manifold learning techniques to a greater extent than other feature types, for example, unaltered MFCC features. First, the effect of noise on ASR WER is evaluated when both MFCC features and LPDA features are applied to the speech in noise task described in Section 3. It is observed that the WER is far higher for the LPDA transformed features than that observed when no feature space transformation is performed. Second, the effects of noise are measured again after transforming the HMM covariance matrices to reduce the acoustic model mismatch with respect to the noisy test data. Surprisingly, it is found that, after transforming the model parameters, the gain in ASR performance observed for the LPDA transformed features is significantly higher than the that for the MFCC features.

The experimental results supporting these observations are displayed in Table 4. The table contains ASR WERs for test conditions corresponding to a range of SNRs in the subway noise condition. The two major rows in Table 4 represent the two different features being evaluated, namely untransformed MFCC features and the LPDA transformed features. For each feature set, the percent WER is displayed with respect to SNR level when no acoustic adaptation is performed, referred to in Table 4 as "None", and when unsupervised regression based covariance adaptation is performed on the HMM model during recognition, referred to in the table as "Cov.".

It is clear from observing the "None" labeled rows of Table 4 that there is a significant increase in WER for both the untransformed and LPDA transformed features as the testing conditions become increasingly mismatched with respect to the clean training conditions. However, it is also clear that the increase in WER is far greater for the case of the LPDA features. This suggests that imposing the structural constraints associated with manifold learning is actually increasing the confusability of the data when corrupted by additive noise.

It is well known that the presence of noise introduces distortions in the covariance structure of the data [14]. These distortions result in changes in the probability densities of noisy speech features resulting in a mismatch with respect to model distributions trained under clean conditions. For the particular case of manifold learning, it is also true that the unseen test features may not obey the structure of the manifold learned from clean training features during LPDA estimation. This results in a higher degradation in ASR performance when manifold learning techniques are employed for feature space transformation.

These observations concerning the impact of noise on the covariance of the data and the impact of the mismatched covariance on the assumed underlying manifold structure of the data suggests that some form of environmental compensation should reduce the effects of noise on the LPDA transformed features. The rows of Table 4 labeled as having "Cov." adaptation display the WERs obtained when applying unsupervised regression based covariance adaptation to transforming Gaussian mixture covariance matrices in the CDHMM model. A multiple pass adaptation scenario is used where maximum likelihood linear regression (MLLR) based covariance transforms were estimated from all test utterances corresponding to a given noise level [15]. While this scenario is not consistent with the scenario used to obtain the "no adaptation" results shown in Table 4, it serves to demonstrate the added impact mismatched conditions have on the LPDA manifold learning approach.

Table 4. ASR %-WER for clean training and subway noise testing on Aurora-2 speech corpus for MFCC, LPP and LPDA features, with and without environmental compensation.

Features	Adapt.	20 dB	15 dB	10 dB	5 dB
MFCC	None	2.52	6.97	24.01	54.19
	Cov.	2.67	5.59	16.49	44.83
LPDA	None	7.80	18.94	40.71	61.20
	Cov.	1.96	4.30	13.75	39.89

It is clear from the "Cov." results in the table that WERs for both LPDA transformed and untransformed features are significantly reduced at almost all SNR levels. Part of this reduction in both cases is due to the fact that all utterances from a given SNR level are used to estimate the regression based adaptation matrix. Note that this is an expected and well known phenomenon. However, the WER reductions for the case of LPDA transformed features are remarkable. In fact, while LPDA WERs are considerably higher than the WERs for the untransformed features with uncompensated models, the LPDA WERs are considerably lower than those obtained for the untransformed features when covariance normalization is applied.

Thus, it can be concluded that the direct impact of noise on the manifold learning procedure described in Section 2 occurs through distortions in the local neighborhoods for the manifold learning algorithm. These local neighborhoods are defined by the affinity matrices and the associated Gaussian kernels. The N-LPDA approach presented in Section 4 directly deals with this issues by considering the relationship between the size of the Gaussian kernels and the noise SNR levels.

6. CONCLUSION

This paper has investigated the effect of acoustic noise conditions on manifold learning approaches for feature space transformations in CDHMM based ASR. It was found that the structural constraints associated with manifold learning approaches result in transformed features that are more sensitive to mismatch in acoustic conditions than untransformed MFCC features. It was also shown that environment dependent performance degradation can be traced to the choice of the size of the local neighborhood used for defining local affinity matrices in manifold learning. These observations led to a multi-model approach for improving the robustness of manifold learning based feature space transformations, referred to here as noise aware manifold learning (NAML). This involves automatic selection from a set of noise-matched LPDA transformations to find a transform that best matches the estimated noise conditions associated with a given utterance. The approach was shown to provide reduced WER across a range of acoustic conditions with respect to LDA and LPP based feature space transformations.

7. REFERENCES

[1] Vikrant Singh Tomar and Richard C. Rose, “Application of A Locality Preserving Discriminant Analysis Approach to ASR,” in *International Conference on Information Science, Signal Processing, and their Applications (ISSPA)*, Montreal, QC, Canada, 2012.

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

[3] Xiaofei He and Partha Niyogi, “Locality preserving projections,” in *ICASSP: IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[4] Yun Tang and Richard Rose, “A study of using locality preserving projections for feature extraction in speech recognition,” in *ICASSP: IEEE Intl. Conf. on Acoustics, Speech, and Signal Pro.*, 2008.

[5] Nagendra Kumar, *Investigation of silicon-auditory models and generalization of linear discriminant analysis for improved speech recognition*, Ph.D. thesis, Johns Hopkins University, Baltimore, MD, 1997.

[6] Mark Gales, “Semi-tied covariance matrices for hidden Markov models,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272 – 281, May 1999.

[7] M. J. F. Gales, “Maximum likelihood multiple subspace projections for hidden Markov models,” vol. 10, no. 2, pp. 37–47, 2002.

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

[9] Deng Cai, Xiaofei He, and et al., “Locality sensitive discriminant analysis,” in *Intl. Joint Conference on Artifical Intelligence*, 2007.

[10] H.-T. Chen, H.-W. Chang, and T.-L. Liu, “Local discriminant embedding and its variants,” in *IEEE Conference on Computer Vision and Patter Recognition*, 2005, vol. 1, pp. 846–853.

[11] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Quang Yang, and Stephen Lin, “Graph embedding and extensions: A generalized framework for dimensionality reduction,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40 – 51, Jan 2007.

[12] Chanwoo Kim and Richard M Stern, “Robust Signal-to-Noise Ratio Estimation Based on Waveform Amplitude Distribution Analysis,” *Interspeech 2008*, pp. 2598–2601, 2008.

[13] M. Vondrasek and P. Pollak, “Methods for speech SNR estimation: Evaluation tool and analysis of VAD dependency,” *Radioengineering*, vol. 14, no. 1, pp. 7, 2005.

[14] Angel De La Torre, Jose C Segura, Carmen Benitez, and Javier Ramirez, “Speech recognition under noise conditions: Compensation methods,” in *Robust Speech Recognition and Understanding*, Michael Grimm and Kristian Kroschel, Eds., number June, chapter 25, pp. 439 – 460. InTech Education and Publishing, 2007.

[15] M. J. F. Gales and P. C. Woodland, “Mean and variance adaptation within the MLLR framework,” *Computer Speech and Language*, pp. 249–264, 1996.