A Correlational Discriminant Approach to Feature Extraction 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

A nonlinear discriminant analysis based approach to feature space dimensionality reduction in noise robust automatic speech recognition (ASR) is proposed. It utilizes a correlation based distance measure instead of the conventional Euclidean distance. The use of this 'correlation preserving discriminant analysis' (CPDA) procedure is motivated by evidence suggesting that correlation based cepstrum distance measures can be more robust than Euclidean based distances when speech is corrupted by noise. The performance of CPDA is evaluated in terms of the word error rate obtained by using CPDA derived features on a speech in noise task, and is compared to a number of Euclidean distance based approaches to feature space transformations, namely linear discriminant analysis (LDA), locality preserving projections (LPP), and locality preserving discriminant analysis (LPDA).

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

1. Introduction

An important aspect of feature extraction in automatic speech recognition (ASR) is to augment the static features, obtained from short-time speech segments, by the dynamic spectral information of speech. This augmentation can be achieved by concatenating multiple static feature vectors to form high-dimensional super vectors representing speech over 100 milliseconds and have dimensionality of as high as 200. This representation, however, introduces the problem of the high dimensionality of the super-vectors and the high degree of correlation between feature vector components.

These issues necessitate the application of feature space discrimination and dimensionality reduction techniques to ASR. A number of such techniques have been readily used in ASR, for example, linear discriminant analysis (LDA) [1, 2], and locality preserving projections (LPP) [3, 4]. LDA, being a purely discriminant approach, attempts to maximize separability among somehow defined classes of feature vectors. A shortcoming of LDA is its inability to preserve the local geometric relationships of the data-space, which is achieved by manifold learning approaches, such as, LPP. LPP treats the data-points to be residing on the surface of a lower dimensional manifold, and attempts to preserve the structure of this manifold in the transformed space. That being said, LPP itself lacks the power to exploit the discriminant structure of the data. Motivated by these facts, the authors introduced a new feature space transformation technique to ASR in the previous work [5]. The technique, termed 'locality preserving discriminant analysis' (LPDA), introduces discriminant component to manifold learning techniques. LPDA attempts to preserve within-class local manifold based relationships of the data points, while maximizing the separability between different classes.

However, the problem with all of these techniques, including LPDA, is that they are based on Euclidean distance metrics for characterizing the relationships between feature vectors. This work investigates a new discriminant approach, referred to as 'correlation preserving discriminant analysis' (CPDA), for feature space transformation. It builds upon LPDA by utilizing, instead of a Euclidean distance measure, a cosine-correlation based distance measure in describing the manifold domain relationships between data vectors. It is motivated by the fact that acoustic models based on Euclidean distance measures are highly susceptible to ambient noise. In particular, it has been shown that additive noise in the linear spectrum domain alters the norm of cepstrum domain feature vectors [6]. Another way to visualize this could be that noise causes the data points to scatter around their original position. As a result, the unseen test features may not obey the structure of the manifold learned from the training data during estimation of the feature space transformation. These factors contribute to misclassification during the feature space transformation. It has also been found that the angles between cepstrum vectors are comparatively more robust to noise [7]. This suggests a potential advantage to a discriminant transformation estimated using a correlation based objective function, especially in the context of noise robust ASR.

Fortunately, there has been a great deal of work on developing correlation based discriminant feature transformations in other application domains. Tang et. al. [8] utilized a cosine-correlation based technique for speaker clustering. Huang et. al. [9] demonstrated significant improvement in face-recognition when using a cosine-correlation distance measure as compared to an Euclidean distance measure. Ma and Lao [10] also reported gain in face-recognition accuracy when using a cosine-correlation based discriminant mechanism.

The rest of this paper will present an implementation of CPDA for feature space dimensionality reduction for acoustic modeling in ASR. Section 2 of this paper will briefly present the development of correlation preserving discriminant analysis (CPDA) approach and discuss its application to dimensionality reduction for robust ASR. An experimental study comparing the relative performance of CPDA with some Euclidean distance based techniques in terms of word error rate (WER) on the Aurora 2 speech in noise task will be presented in Section 3. Finally, Section 4 concludes the paper.

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

2. Correlation Preserving Discriminant Analysis

This section describes the correlation preserving discriminant approach to feature space dimensionality reduction. Section 2.1 presents the CPDA formulation which includes both graph embedding for characterizing the local relationships between data vectors and optimization of a non-linear objective function based on a correlation based measure of class separability. Section 2.2 presents a discussion of the issues associated with applying CPDA in ASR.

2.1. CPDA - Algorithm Formulation

The first step of training the coefficients of a CPDA transformation is to project the super-vectors described in Section 1 onto the surface of a unit hypersphere. This has the effect of discarding magnitude information while retaining the correlation based relationships between data vectors and is motivated in Section 2.2. Consider an ASR training dataset which is given by a matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots \boldsymbol{x}_N]^T$, where each row $\boldsymbol{x}_i \in \mathbb{R}^d$ represents a feature vector projected onto the surface of a d-dimensional unit hypersphere, and belongs to the class/label c_i. The goal in CPDA 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 class discrimination in the projected feature space while retaining the inherent data structure. For an arbitrary source space vector $\boldsymbol{x}_i$, the corresponding target space vector $\boldsymbol{y}_i$ is obtained according to the following transformation

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

CPDA follows the generalized framework of graph embedding for dimensionality reduction [11]. Graph embedding refers to representing the data space by one or more undirected connected graphs with the data vectors as the vertices of the graphs. An edge between any two nodes of the graphs can be represented by some measure of similarity or closeness between them. In CPDA, similarity between the nodes is represented by their cosine-correlation, as defined later. These graphs characterize the statistical and geometrical properties of the dataset, and can be represented by matrices, called the affinity matrices.

In CPDA, two undirected weighted graphs, the intrinsic graph $\mathcal{G}_i = \{\boldsymbol{X}, \boldsymbol{W}_i\}$, and the penalty graph $\mathcal{G}_p = \{\boldsymbol{X}, \boldsymbol{W}_p\}$ are defined. Here $\boldsymbol{X}$ – the nodes of the graph – represent the vectors of the dataset, and $\boldsymbol{W}_i$ and $\boldsymbol{W}_p \in \mathbb{R}^{N \times N}$ are the intrinsic and penalty affinity edge-weight matrices, respectively. The elements of the intrinsic and penalty graph weight matrices are defined as

$$w_{ij}^{intrinsic} = \begin{cases} exp(\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle - 1}{\rho}) & I(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 \\ 0 & I(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0 \end{cases} \tag{2}$$

and

$$w_{ij}^{penalty} = \begin{cases} exp(\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle - 1}{\rho}) & ; \quad I(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0 \\ 0 & ; \quad I(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 \end{cases} \tag{3}$$

where ρ is the kernel scale parameter. The function $I(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is defined as a logical <AND> of two indicator functions $e_c(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $e(\boldsymbol{x}_i, \boldsymbol{x}_j)$. The function $e_c(\boldsymbol{x}_i, \boldsymbol{x}_j)$ specifies whether $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ lie within the same class, and the function $e(\boldsymbol{x}_i, \boldsymbol{x}_j)$ indicates whether the two nodes are close in terms of cosine-correlation $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$. Since $\boldsymbol{x}_i$, and $\boldsymbol{x}_j$ are unit vectors, $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle = ||\boldsymbol{x}_i|| \cdot ||\boldsymbol{x}_j|| \cos\theta_{ij} = \cos\theta_{ij}$. Thus, the correlation between the two vectors simply denotes their angular affinity. The data vector $\boldsymbol{x}_j$ can be said to be close to $\boldsymbol{x}_i$ either if – in terms of correlation – it falls within the k-nearest neighbors of $\boldsymbol{x}_i$ or alternatively if $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \leq \epsilon$ (some threshold). In this work, the neighborhood of a data vector is defined by its k-nearest neighbors. In the intrinsic graph, a node $\boldsymbol{x}_i$ is connected to the k_i nearest neighbors, in correlation sense, belonging to the same class c_i. Contrastingly, in the penalty graph, a node $\boldsymbol{x}_i$ is connected to the k_p largest correlation neighbors *not* belonging to the class c_i. For the current task, the optimal values of k_i and k_p were empirically determined to be $k_i = k_p = 200$.

In the transformed space, a generalized scatter measure for a graph $\mathcal{G}$ can be given by

$$S = \sum_{i \neq j} ||f(\boldsymbol{x}_i, \boldsymbol{P}) - f(\boldsymbol{x}_j, \boldsymbol{P})||^2 w_{ij} \tag{4}$$

where $f(\boldsymbol{x}_i, \boldsymbol{P}) = \frac{\boldsymbol{P}^T \boldsymbol{x}_i}{||\boldsymbol{P}^T \boldsymbol{x}_i||}$ is the transformed unit vector corresponding to the vector $\boldsymbol{x}_i$ in the source space. Note that for $||\boldsymbol{P}^T \boldsymbol{x}_i|| = 0$, $f(\boldsymbol{x}_i, \boldsymbol{P}) = 0$. Eq. (4) can be simplified as

$$S = 2 \sum_{i \neq j} \left(1 - \frac{f_{ij}}{f_i f_j}\right) w_{ij} \tag{5}$$

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 $f_{uv} = \boldsymbol{x}_u^T \boldsymbol{P} \boldsymbol{P}^T \boldsymbol{x}_v$.

The goal is to penalize the properties inherent to the penalty graph (maximize its scatter S_p), while at the same time preserve the properties inherent to the intrinsic graph (minimize its scatter S_i). Thus, the following function can be defined as a measure of class separability and graph-preservation [9].

$$F(\boldsymbol{P}) \;=\; S_p - S_i \;=\; 2 \sum_{i \neq j} \left(1 - \frac{f_{ij}}{f_i f_j}\right) \cdot w_{ij}^{p-i} \tag{6}$$

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

$$\boldsymbol{P}_{CPDA} = \arg\max_{\boldsymbol{P}} F(\boldsymbol{P}). \tag{7}$$

To optimize expression (6), the gradient ascent rule can be utilized as follows. $P := P + \alpha \nabla_P F$, with

$$\nabla_P F \;=\; 2 \sum_{i \neq j} \left[\frac{f_{ij} \boldsymbol{x}_i \boldsymbol{x}_i^T}{f_i^3 f_j} + \frac{f_{ij} \boldsymbol{x}_j \boldsymbol{x}_j^T}{f_i f_j^3} - \frac{\boldsymbol{x}_i \boldsymbol{x}_j^T + \boldsymbol{x}_j \boldsymbol{x}_i^T}{f_i f_j} \right] \boldsymbol{P} \cdot w_{ij}^{p-i} \tag{8}$$

where α is the gradient scaling factor, and $\nabla_P F$ represents the gradient of the expression (6) with respect to $\boldsymbol{P}$.

Unfortunately, $F(\boldsymbol{P})$ is a non-linear and non-convex function, thus the optimization does not have a closed-form solution. In particular, for a high-dimensional data-space, the gradient ascent might converge to a local maxima. Therefore, a good initialization is critical for achieving a sufficient optimization. To this end, the mapping function can be approximated to that of a linear transformation rooted in graph-embedding by neglecting the normalization term, *i.e.*, setting $f(\boldsymbol{x}_i, \boldsymbol{P}) = \boldsymbol{P}^T \boldsymbol{x}_i$. A closed-form solution for a good initial projection matrix is then achieved by solving the eigenvalue problem [5, 11]

$$(\boldsymbol{X}(\boldsymbol{D}_p - \boldsymbol{W}_p)\boldsymbol{X}^T)\boldsymbol{p}_j = \lambda_j (\boldsymbol{X}(\boldsymbol{D}_i - \boldsymbol{W}_i)\boldsymbol{X}^T)\boldsymbol{p}_j, \tag{9}$$

where $\boldsymbol{D}$ is a diagonal matrix whose elements correspond to the row sum of the matrix $\boldsymbol{W}$. The subscript i and p signifies 'intrinsic' and 'penalty' matrices respectively. The vector $\boldsymbol{p}_j$ indicates the j^{th} column of the transformation matrix $\boldsymbol{P}_{CPDA}$.

2.2. Discussion

It is important to motivate the use of a correlational discriminant approach for estimating feature space transformations. There are two main factors that contribute to the strength of CPDA in feature analysis for ASR. The first is that it utilizes a cosine correlation distance for defining relationships between the feature vectors in the manifold space. The second is that it combines graph embedding based within class manifold learning with inter-class discrimination. These factors are discussed in more details in the subsequent paragraphs.

The primary motivation for CPDA, over LPDA [5] and others, is the evidence suggesting that adding noise to clean speech results in distortion in the magnitude of cepstrum feature vectors but has a relatively small effect on the correlation between cepstrum vectors [6, 7]. Correlation based distance measures have also been implemented during recognition in continuous density hidden Markov models (CDHMM) based ASR systems and were found to achieve lower WERs on speech in noise tasks than the standard Euclidean based measure [7]. This implies that the CPDA feature space transformation may exhibit the same level of robustness when compared to the techniques based on Euclidean distance metrics.

There are several other general statements that can be made with reference to the superiority of the CPDA transform over conventional techniques like LDA and LPP. Due to the fact that both LPDA and CPDA are built upon the same framework of graph-embedding, the following can be considered as their common advantages. The most important aspect of CPDA is its ability to maximize the discrimination among various classes while preserving within-class local manifold based relationships. The conventional approaches could achieve one or the other. The second advantage of CPDA is that it makes no assumption about the distribution of input data, whereas the techniques like LDA expects the class conditional distribution of the data to be Gaussian. Third, unlike LPP, which preserves the manifold based structure of the entire dataset without considering distribution of the data among various classes, CPDA only preserves within-class manifold based relationships.

One well known drawback associated with all manifold learning approaches to feature space transformation is high computational complexity [10]. Both LPDA and CPDA, and even LPP, suffer from this issue. In addition, compared to LPDA and LPP, CPDA has slightly higher computational complexity because of its requirement to achieve the optimal objective function through gradient ascent. This represents a definite disadvantage of CPDA transformations when applied to the generally large speech processing tasks. For example, as compared to the other application domains described in [8–10], this work addressed a far larger task involving 180 states and a corpus consisting of 1.4 million 117-dimensional feature vectors.

Finally, correlation among the projected feature space is an issue common to all feature space transformations including LDA, LPP, LPDA and CPDA. This conflicts with the diagonal Gaussian covariance property generally assumed in CDHMM based ASR systems. To this end, either maximum likelihood linear transformation (MLLT) [2] or semi-tied covariances (STC) [12] should be applied in the transformed feature space to maximize the likelihood of the data under the constraint that feature vectors be uncorrelated.

3. Experimental Study

This section describes the experimental study performed to evaluate the ASR performance obtained using the CPDA approach to feature space transformations. Performance is reported as the word error rate (WER) obtained for the Aurora 2 speech in noise task domain. Section 3.1 describes the Aurora 2 task domain and the baseline ASR system configuration. Section 3.2 provides the experimental results across a range of noise types and signal-to-noise ratios. For comparison, ASR WER for LDA, LPP, and LPDA approaches are also provided.

3.1. Task domain and system configuration

All systems described Section 3.2 are evaluated on the European Telecommunications Standards Institute (ETSI) Aurora-2 speech corpus. The corpus represent a simulated speech in noise task created by artificially adding noise from multiple environments to connected digit utterances spoken in a quiet environment. For this reason, one must be careful about generalizing the results presented here to the wide range of actual speech in noise tasks. The training dataset consists of 8440 multi-noise mixed utterances collected from 55 male and 55 female speakers. The test results reported in Section 3.2 were evaluated over 4004 utterances corrupted by four different noise types at SNRs ranging from 5 to 20 dB.

For baseline comparison, 39-dimensional MFCC features resulting from the concatenation of 12 static coefficients, normalized log energy, Δ-cepstrum, and $\Delta\Delta$-acceleration were used. The input features for feature transformations were the 117-dimensional super-vectors consisting of 9 concatenated vectors of MFCC coefficients augmented by log energy. For discriminant analysis, the super-vectors were associated with classes corresponding to the CDHMM states. The resultant projection matrix $\boldsymbol{P}$ was then used to project the 117-dimensional training and test vectors to a 39 dimensional space.

The ASR system was 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 were 11 word-based CDHMM models, and a total of 180 states. Each state was modeled by a mixture of 3 Gaussians. STC transforms [12] were estimated to account for the correlation introduced to the features by the transformation approaches as described in Section 2.2.

3.2. Results

Table 1 compares the recognition performance of CPDA with LDA, LPP, and LPDA for four noise types and four SNRs ranging from 5dB to 20dB. Results pertaining to clean testing have been omitted for brevity. Four separate tables are displayed, one for each noise type (subway, car, exhibition hall, and airport) where each of these tables contains ASR WER for five different systems.

For each of these tables, the first row displays the baseline ASR WER obtained using mixed condition HMM training when no feature transformation is performed. The second row, labeled "LDA", corresponds to application of the LDA discriminant projection matrix to the concatenated MFCC feature vectors as described in the previous section. The third row, labeled "LPP" corresponds to the features obtained as a result of LPP approach. The fourth row "LPDA" corresponds to features obtained by applying the LPDA transformation to the con-

Table 1: 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**
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
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**
Airport	Baseline	3.42	4.88	8.49	16.58
	LDA	3.18	4.11	7.72	15.65
	LPP	4.35	6.95	10.38	21.15
	LPDA	**3.10**	4.09	7.49	15.09
	CPDA	3.67	**4.02**	**7.43**	**12.56**

catenated 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, an STC transformation is performed to minimize the impact of the data correlation resulting from the application of feature space transformations.

The results in Table 1 demonstrate the clear advantage of the LPDA family of approaches (LPDA and CPDA) over the conventional techniques such as LDA, and LPP. Furthermore, it can also be observed that CPDA outperforms all other techniques, including LPDA, in most cases. More importantly, the performance gain from LPDA to CPDA is relative to the level of noise corruption in the test features. Despite not producing the best results for the 20 dB (low noise) case, CPDA produces lowest WER for high noise cases. The most noteworthy result here is that of all the techniques given in the table CPDA exhibit the slowest drop in ASR performance with the increase in noise level. This demonstrate an improvement in ASR performance in high noise conditions when using a cosine-correlation based discriminant analysis for feature dimensionality reduction as compared to conventional Euclidean distance based techniques.

One crucial factor affecting the performance of a manifold learning approach is the shape and size of the neighborhood in the manifold space [4, 5, 13]. Typically, a nonlinear kernel is used to map the feature data onto the manifold space. The geometrical structure of the target mapping is then governed by the kernel scale parameter, ρ, as defined in Section 2.1. This parameter is generally empirically determined for a given task. For the results presented in Table 1, the optimal values of ρ have been carefully estimated for the LPP and LPDA experiments from development data. However, it was not practical in this work to perform the same level of optimization for ρ in the case of CPDA. This might explain the inferiority of CPDA performance to that of LPDA in certain noise conditions in Table 1.

4. Conclusion

This paper presents a discriminant analysis algorithm for feature space dimensionality reduction which utilizes a cosine-correlation based distance measure instead of the Euclidean distance based measures. The algorithm further utilizes the generalized framework of graph-embedding to eliminate dependency on the data distribution. When compared to the traditional Euclidean distance based techniques, the presented algorithm promises better performance for speech recognition in noisy environments. It is supported by the fact that cosine-correlation based distance is more robust to additive noise.

5. References

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

[2] G. Saon, M. P. an R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *ICASSP: IEEE Intl. Conf. on Acoustics, Speech, and Signal Pro.*, 2000.

[3] X. He and P. Niyogi, "Locality preserving projections," in *ICASSP: IEEE Intl. Conf. on Acoustics, Speech, and Signal Pro.*, 2003.

[4] Y. Tang and R. 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] V. S. Tomar and R. C. Rose, "Application of A Locality Preserving Discriminant Analysis Approach to ASR," To appear in *International Conference on Information Science, Signal Processing, and their Applications (ISSPA)*, Montreal, QC, Canada, 2012. [Online]. Available: http://www.ece.mcgill.ca/~vtomar/Publications/ISSPA12-LPDA.pdf

[6] D. Mansour and B. H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition," *IEEE Trans. on of Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1659 – 1671, Nov. 1989.

[7] B. A. Carlson and M. A. Clements, "Application of a weighted projection measure for robust hidden markov model based speech recognition," in *ICASSP: IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1991, pp. 921–924.

[8] H. Tang, S. M. Chu, and T. S. Huang, "Spherical discriminant analysis in semi-supervised speaker clustering," in *NAACL HLT*, Boulder, Colorado, USA, June 2009, pp. 57 – 60.

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

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

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

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

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