EXPLOITING LOCAL AND GLOBAL STRUCTURES FOR TIMIT PHONE CLASSIFICATION

Heyun Huang, Louis tenBosch, Jort Gemmeke, Bert Cranen, and Lou Boves

Department of Linguistics, Radboud University Nijmegen
Erasmaslaan 1, 6525, Nijmegen, the Netherlands
{h.huang, l.tenbosch, j.gemmeke, b.cranen, and l.boves}@let.ru.nl
web: http://lands.let.ru.nl

ABSTRACT

Using contextual information of phones is an effective way to improve the performance of phone classification tasks, but requires the use of dimensionality reduction. One of the disadvantages of Linear Discriminant Analysis (LDA), a popular dimensionality reduction method is that it is not able to account for local differences between the distributions of classes in the feature space. Newer methods, such as the Local Fisher Discriminant Analysis (LFDA), on the other hand, may overestimate the contribution of local distributions. In this paper, we propose to use a dimensionality reduction algorithm with an affinity matrix that allows finding the optimal trade-off between local and global information. Experiments on TIMIT show that both local and global information in the MFCC feature space are important for phone classification and that a substantial improvement can be achieved over both LDA and LFDA.

1. INTRODUCTION

MFCCs and PLP coefficients are powerful means to represent important local characteristics of speech signals. However, these features poorly represent temporal dynamics. The ballistic nature of the movements of the articulators implies that there is a substantial amount of continuity in the speech signal. The delta and delta-delta coefficients that are typically used, however, do not capture very well the properties of the trajectories that correspond to (demi-)syllables. Motivated by this insight, several attempts have been made to better capture this continuity. At least three different approaches have been proposed: (1) using features that combine time windows of different length (e.g. the TRAP features proposed in [1]), (2) modelling feature trajectories by polynomial regression (e.g. [2] [3]) and (3) by using sequences of feature vectors as units for subsequent modelling. In this paper, we will investigate the use of sequences of MFCC vectors for the classification of phone-like speech segments.

A sequence of feature vectors captures trajectories; thus, if a sufficiently long sequence were centered in a phone, it would contain all acoustic information that is available for the classification of this phone. However, this approach faces two obvious problems. First, capturing temporal dynamics implies capturing the different feature trajectories that characterise the transitions from and into the neighbouring phones. For the phone classification task this means that quite different trajectories must still be mapped onto a single phone class (lest one can succeed in modelling context-dependent phones). Second, stacking enough feature vectors, each consisting of more than 10 coefficients, to capture the transitions into and out of long phones, results in a high dimensional feature space that is sparsely populated, making it difficult to define effective distance measures.

In order to handle the so-called curse of dimensionality, several methods have been proposed for mapping a high-dimensional original feature space into a lower-dimensional space. In this paper, we focus on methods based on Fisher's Linear Discriminant Analysis (FDA). FDA aims to maximize the ratio of the between-class and within-class variance in the (reduced) feature space. Conventional FDA is a *global* dimensionality reduction method, because the distances between all pairs of observations are given equal weight. This is made particularly clear in the reformulation of the mathematics that avoids the use of class means and the global mean [4] [5] However, it is well known that some acoustic features are important for discriminating between some (classes of) phones, while they may be redundant for other (classes of) phones. Moreover, the relevance of a feature for the classification of a phone may depend on the segmental context of that phone. These considerations lead to the conclusion that we need to extend FDA in such a way that local structure can be brought to bear, in addition to the global distribution of the observations.

For that purpose a so-called Affinity Matrix has been introduced in the equations that define FDA (cf. Section 2). The graph-embedding dimensionality reduction framework [4] is a generalization of several dimensionality reductions methods that use an affinity matrix for a specific purpose. One of the first methods that used the affinity matrix to capture local structure information was Locality Preserving Projection (LPP) [6]. Local FDA (LFDA) extends conventional FDA by using a local kernel [5].

In [7] LFDA was applied to a Japanese speech recognition task. It was shown that the introduction of local structure improved the performance compared to the baseline FDA results. However, in general the relative importance of global and local information for subsequent classification tasks is not known in advance. In fact, in many practical problems local and global structures co-exist in the feature space. Therefore, one would want to have a dimensionality reduction algorithm that allows finding the optimal trade-off between global and local information. In this paper we present such a method.

Our paper is organized as follows. In Section 2 of this paper we briefly review the dimensionality reduction algorithm and a special view of capturing the local or global information. After addressing the importance of simultaneously exploring both local and global structures, we firstly introduce the theory of LFDA and then propose our extensions of the

affinity matrix in Section 3. In Section 5, experimental results are presented. Discussion and conclusion are presented in Section 6.

2. GLOBAL AND LOCAL DIMENSIONALITY REDUCTION

2.1 FDA: A Conventional Global Approach

In this paper, a speech segment in the form of a block of contiguous MFCC feature vectors is represented as a single high-dimensional vector $\mathbf{x}$ (by stacking consecutive frames). Suppose the classification problem can be stated in terms of J classes C_j, with each class C_j containing n_j of such vectors. The conventional Fisher Linear Discriminant Analysis (FDA) is based on two matrices. For each class C_j, the within-class scatter matrix $\mathbf{S}_j^{(w)}$ is defined as:

$$\mathbf{S}_j^{(w)} = \sum_i (\mathbf{x}_{ji} - \mu_j)(\mathbf{x}_{ji} - \mu_j)^t \tag{1}$$

in which μ_j denotes the mean of all vectors $\mathbf{x}_{ji}$, $1 \leq i \leq n_j$ in the class C_j, $1 \leq j \leq J$. The between-class scatter matrix $\mathbf{S}^{(b)}$ is defined as:

$$\mathbf{S}^{(b)} = \sum_j \frac{n_j}{n} (\mu_j - \mu)(\mu_j - \mu)^t \tag{2}$$

in which μ denotes the overall mean and n the total number of tokens.

Maximizing the ratio of the within-class scatter $\mathbf{S}^{(w)}$ and the between-class scatter $\mathbf{S}^{(b)}$ (the FDA criterion) requires finding the matrix $\mathbf{T}$ according to

$$\underset{\mathbf{T}}{\operatorname{argmax}} [tr((\mathbf{T}\mathbf{S}^{(w)}\mathbf{T})^{-1}\mathbf{T}\mathbf{S}^{(b)}\mathbf{T})] \tag{3}$$

Equations (1) and (2) which define the between-class and within-class scatter matrices can be rewritten to avoid the use of the means, and to emphasize the relation between pairs of vectors, in the following way:

$$\mathbf{S}^{(w)} = \frac{1}{2} \sum_{ij} A_{ij}^w (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^t \tag{4}$$

$$A_{ij}^w = \begin{cases} 1/n_c & \text{if } C_{\mathbf{x}_i} = C_{\mathbf{x}_j} \\ 0 & \text{if } C_{\mathbf{x}_i} \neq C_{\mathbf{x}_j} \end{cases}$$

$$\mathbf{S}^{(b)} = \frac{1}{2} \sum_{ij} A_{ij}^b (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^t \tag{5}$$

$$A_{ij}^b = \begin{cases} 1/n - 1/n_c & \text{if } C_{\mathbf{x}_i} = C_{\mathbf{x}_j} \\ 1/n & \text{if } C_{\mathbf{x}_i} \neq C_{\mathbf{x}_j} \end{cases}$$

where $C_{\mathbf{x}_i} = C_{\mathbf{x}_j}$ means that the observations $\mathbf{x}_i$ and $\mathbf{x}_j$ are members of the same class C. In the conventional definition of FDA, A_{ij}^w and A_{ij}^b are independent of the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. Therefore, FDA can be considered as a global dimensionality reduction method.

2.2 LFDA: Incorporating Local Information into FDA

In order to take local density information into account, the coefficients A_{ij} can be made dependent on the distance between two observations $\mathbf{x}_i$ and $\mathbf{x}_j$. In the Locality Preserving Projection-approach proposed in [6] the affinity matrix $\mathbf{A}$ is defined as:

$$A_{ij} = \exp(-||\mathbf{x}_i - \mathbf{x}_j||^2) \tag{6}$$

When applied to Eqs. (4) and (5) the net effect of Eq. (6) is that the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ is taken into account in the contribution to the between-class and within-class scatter matrix. We can go one step further. When using Eq. (6) all pairwise distances $||\mathbf{x}_i - \mathbf{x}_j||^2$ are weighted equally. However, this might not be optimal. To allow for different weights, a locality preserving scaling kernel can be adopted in LDA, as was done in [8] to obtain Local Fisher Discriminant Analysis (LFDA) by defining the weights as:

$$A_{ij} = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{\sigma_i \sigma_j}\right) \tag{7}$$

in which σ_n (where n is i or j) is defined by

$$\sigma_i = ||\mathbf{x}_i - \mathbf{x}_l^L|| \tag{8}$$

In this expression, $\mathbf{x}_l^L$ denotes the l^{th} ranked neighbour in the list of L nearest neighbours in the same class as $\mathbf{x}_i$, ordered according to their distance to $\mathbf{x}_i$. This means that σ_i estimates the local density of the class around the observation $\mathbf{x}_i$ (the same holds for $\mathbf{x}_j$). This leads to the following definition of LFDA:

$$\mathbf{S}^{(w)} = \frac{1}{2} \sum_{ij} A_{ij}^w (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^t \tag{9}$$

$$A_{ij}^w = \begin{cases} A_{ij}/n_c & \text{if } C_{\mathbf{x}_i} = C_{\mathbf{x}_j} \\ 0 & \text{if } C_{\mathbf{x}_i} \neq C_{\mathbf{x}_j} \end{cases}$$

$$\mathbf{S}^{(b)} = \frac{1}{2} \sum_{ij} A_{ij}^b (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^t \tag{10}$$

$$A_{ij}^b = \begin{cases} A_{ij}(1/n - 1/n_c) & \text{if } C_{\mathbf{x}_i} = C_{\mathbf{x}_j} \\ A_{ij}/n & \text{if } C_{\mathbf{x}_i} \neq C_{\mathbf{x}_j} \end{cases}$$

Thus, the affinity matrix (7) in LFDA assigns a larger weight to observation pairs that are relatively close. At the same time, pairs of observations that are farther apart make smaller contributions to two scatter matrices.

3. GENERALIZED AFFINITY MATRIX

LFDA has proven itself as an effective way to exploit local and global structures simultaneously [5]. However, LFDA assigns a fixed weight to local and global information. In most practical situations the structure of the feature space is not known in advance. As a consequence, it is not known what the best trade-off is between local and global information. Conventional FDA is obtained by setting $A_{ij} = 1$ to when x_i and x_j are from the same class and $A_{ij} = 0$ when x_i and x_j are from different classes.

Thus, we can obtain a gradual transition between LFDA and FDA by introducing a parameter γ into the affinity matrix, which we apply as an exponent to the product $\sigma_i \sigma_j$. By doing so, we change Eq. (7) to:

$$A_{ij}(\gamma) = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{(\sigma_i \sigma_j)^\gamma}\right) \tag{11}$$

Eq. (11) can be regarded as a generalized affinity matrix in which γ determines the trade-off between local and global structure. Larger values of γ correspond to smaller weights of the affinity matrix. This results in smaller differences between the weights assigned to close and more distant pairs of observations, making global structure increasingly more important than local structure. More in particular it holds that with

- $\gamma \to +\infty, A_{ij}(\gamma) \to 1$, which is equivalent to the conventional FDA.
- $\gamma = 1$, the approach is equivalent to LFDA.
- $\gamma = 0, A_{ij}(\gamma)$ is equivalent to that of LPP [6].

In the following experiments we will investigate the trade-off between local and global structure as a function of γ in more detail.

4. EXPERIMENTAL SETUP

4.1 Feature extraction and classification task

In our phone classification experiments on TIMIT [9] we used the standard NIST training set (462 speakers, 3696 utterances, 139,852 phone tokens for training), development set (50 speakers, 400 utterances, 15,038 phone tokens) in line with the choice made by A. Halberstadt in [10], and the standard core test set (24 speakers, 192 utterances, 7195 phone tokens) for testing purpose. During modelling, the 64 TIMIT phone labels were reduced into 48 classes in line with [9]. Only when evaluating the models, we further mapped the labels into the commonly used 39 classes [9] to calculate the classification error rate. Glottal stops (q) are ignored both in training and testing.

The feature extraction is performed as follows. A short-time Fourier analysis is performed with a 25ms Hamming window and a 10ms window shift. For each frame, we compute 13 MFCC features: $c_0 - c_{12}$. For each phone, we extract 23 consecutive MFCC feature frames with the center frame aligned with the center frame of the phone as indicated by the manual labelling. This means that there is a context of 11 frames to the left and right of the center of the phone, which may or may not extend into neighbouring phones depending on the duration of the phone.

The MFCC features are reshaped to a single $13 \times 23 = 299$ dimensional feature vector. Motivated by the common tandem [11] of dimensionality reduction for classification, in our experiments, Principal Component Analysis (PCA) is first applied to map the 299-dimensional features into 150-dimensional feature vectors (these 150 explain 97% of the total variance). The resulting 150-dimensional feature vectors form the input for all the different LDA-based dimensionality reduction techniques that we compare in the remainder of this paper.

4.2 Weighted k-Nearest Neighbour classification

As said, after the dimensionality reduction, a weighted k-Nearest Neighbour classifier [12] is used to classify the feature vectors into one of the 48 phone classes. This is done as follows. Let $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k$ be the k nearest neighbours of an observation $\mathbf{x}$. The weights of these neighbours are computed according to Eq. (12).

$$w_i = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}||^2}{\tau}\right), i = 1, 2, ..., k \qquad (12)$$

In this equation, k is the number of nearest neighbors while τ is a parameter that is able to control the contribution of each neighbor to the classifier. To determine the most likely class, the weights in each class are accumulated using Eq. (13) and the class label $\hat{C}$ associated with the largest sum is selected.

$$\hat{C} = \arg\max_c \sum_{\mathbf{x}_i \in C_c} w_i \qquad (13)$$

5. EXPERIMENTAL RESULTS AND ANALYSES

5.1 Experimental Results

To evaluate the proposed cascade algorithm PCA-(L)FDA-kNN, we use the development set and core test set described in the previous section to generate various experimental results. This was done as follows. First, reasonable values of the parameters γ, k and τ were sought by globally investigating their effect on both the development set and core test set. In preliminary experiments, it appeared that interesting effects took place within $k \in \{15, 16, \ldots, 60\}$, $\tau \in \{3.5, 3.625, \ldots, 6.875, 7\}$, and $\gamma \in \{0, 0.01, 0.02, \ldots, 1.5\}$. Fig.1 shows an example of our results. The figure describes the relationship between γ and classification accuracy with some specific setting of k and τ. The two curves represent the performance on development set and core test test. Four vertical lines indicate the values fof γ obtained in four different ways: according to the conventional LFDA (i.e. $\gamma = 1$, the magenta vertical line), the γ that is found by optimizing the performance on the development set (black line), the value of γ (the green vertical line) optimized on the core test set, and the specific setting $\gamma = 0$ that uses the affinity matrix in LPP (the cyan vertical line, we call it as "LPP" in the remainder for convenience). The third γ represents the situation in which tuning takes place on the evaluation test set ('oracle' performance).

The following three subsections will analyze the results to show the effectiveness of proposed method in different views. In all experiments, values of γ, k and τ will be given for the sake of comparison.

5.2 Analysis 1: Performance Evaluation in Development Set and Core Test Set

In our first analysis, all parameters are jointly optimized on the development set. The resulting optimized parameters are used for evaluation of the proposed method on the core test set. This optimization yields a performance of the locality-weighting algorithm of 74.45% accuracy on the core test set. For fair comparison, we compare this result with the PCA-FDA-kNN and the PCA-LFDA-kNN methods after the same type of optimization: parameter optimization on the development test set and scoring on the evaluation core test set. The PCA-FDA-kNN method yields 73.52% (column FDA in table. 1), while the PCA-LFDA-kNN method yields 74.00% (column LFDA). This table shows that weighting of global and local information by means of the γ parameter is useful for obtaining results beyond those of the LDA and the LFDA-based results.

5.3 Analysis 2: Robustness of Proposed Method

The first analysis shows the performance of the proposed locality-weighting algorithm compared to the LDA- and the

Table 1: Performance comparison of FDA, LFDA, and the proposed locality weighting method method on the core test set after optimizing the parameters on the development set.

Method	FDA	LPP	LFDA	Proposed Method
Accuracy	73.52	73.99	74.00	**74.45**

LFDA-based method after *jointly optimizing* the three parameters (k, τ, γ) on the development set. In the second analysis, we investigate whether this improvement is robust and holds for each (k, τ)-combination. To that end, model results were obtained for a range of different values of parameters k and τ. For each fixed combination of k and τ, γ has been optimized on the development set and used for testing on the core test set. This analysis is designed to show the performance comparison between the best-tuned γ on the development set and the original LFDA ($\gamma = 1$). These results can explicitly be explained in terms of Fig.1: for each "picture" (possible combination of parameters (k,τ)), we are interested in the difference between classification accuracy related to point A (for LFDA) / D (for LPP) and point B. It could be noted that the previous analysis part is interested in the "best picture" whose point F has the accuracy higher than that of any other "picture".
To show this, table 2 provides some basic statistical analysis of the performance *difference*. This table shows that we could achieve 0.4%, 0.88%, and 0.04% gain over LFDA in average, the best case, and the worst case respectively. The positive number in the *worst* case means that introducing γ into LFDA improves the classification performance for all (k, τ)-combinations. The mean gain (0.40) and its standard deviation (0.14) also proves that the gain is always moderate. Compared with LPP, although proposed method seldom performs worse, it gains a larger percentage than LFDA on average (0.58) and in the best case (1.14). In sum, proposed dimensionality reduction algorithm with γ outperforms both LFDA and LPP with kNN classifier.

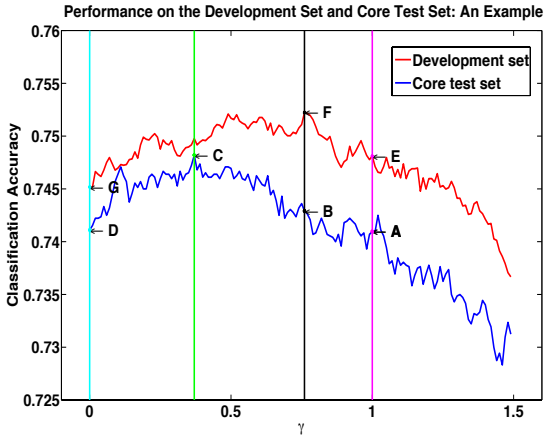

Figure 1: An example of performance evaluation on the development and core test set. Points A, B, C, and D show the classification accuracy of LFDA ($\gamma = 1$), the γ optimized on the development set, the optimal γ optimized on the core test set, and LPP ($\gamma = 0$) on the core test set. Points E, F, and G show the classification accuracy of LFDA ($\gamma = 1$), the γ optimized on the development set, and LPP ($\gamma = 0$) on the development set. These points are noted for easy understanding of our experiments

Table 2: Statistical analysis on the performance gain on core test set of the best γ of development set(%). The method column means the one compared with proposed method.

Method	Avg. Gain (std)	Max. Gain	Min. Gain
LFDA	0.40(0.15)	0.88	0.05
LPP	0.58(0.22)	1.14	-0.19

5.4 Analysis 3: Effect of γ – Exploring the Performance Upperbound in an Oracle Test

In the previous analyses, the parameter performance was based on optimization on a separate development set. Practically, in many pattern recognition problems, the development set and (core) test set are artificially defined and hardly match with each other. This also holds for the the sets at hand in TIMIT. To that end, it is useful to explore how well the algorithm performs in an 'oracle' setting, after tuning on the test set itself.
Here, the optimum γ is obtained by optimization on the core test set. Referring to Fig.1, the differences between the classification accuracy associated to Point A/D and C of all possible settings of (k,τ) is presented in Table 3. This table shows the ceiling of the performance improvement (in percentage) between the LFDA/LPP-based method and the best-tuned proposed locality-weighting method.

Table 3: Statistical analysis on the performance gain when γ(%) is optimized on the test set. The method column means the one compared with proposed method.

Method	Avg. Gain (std)	Max. Gain	Min. Gain
LFDA	0.70(0.13)	1.11	0.19
LPP	0.87(0.19)	1.39	0.19

6. DISCUSSIONS AND CONCLUSIONS

The conventional Fisher Discriminant Analysis (LDA) optimizes the between-class scatter compared to the average within-class scatter. This approach treats each pair of observations x_i with the same weight, without using the fact that many data sets have a different local structure.

In this paper we present our idea to explore the local and global structural information of TIMIT. This is done by first generalizing the conventional LDA by weighting all the between-pair distances. The novel affinity matrix acts as a generalized form of original LFDA [5]) algorithm in order to optimally exploit the local manifold structure together with the global information that we assume is present in the acoustic space in which we represented the TIMIT phonemes that we want to classify. The eventual equations are presented in eqs. and , in which the affinity matrices have been generalized.

Our LDA baseline uses a feature representation that is based on the use of a block of consecutive MFCC feature vectors. In the experiments, we have used a block size of 23 frames, which are 10 ms spaced apart. This means that phones are represented in a way that does not critically rely on information about phone boundaries. It also implies that, if the blocks are long enough, information from neighboring phones leaks into the block representation. This information may be of help in the classification of the phones but also harmful due to the unpredictability of the neighboring phones' labels.

In the three LDA-, LPP- and LFDA-based methods, k, τ are relevant parameters, while the locality-weighting method introduces a novel parameter, γ. We performed experiments by tuning three parameters (γ for the proposed weighting method and k, τ for the back-end kNN classifier), all on the TIMIT development set and core test set. Table 1 shows the cross-validation results of four methods (proposed method, LFDA, LPP, and baseline LDA). From the classification accuracy, one concludes that three methods with locality information, namely the LFDA-based method, LPP-based method and the proposed weighting method, outperform the conventional FDA. This means that the local structure can indeed be effectively exploited in the feature space. Meanwhile, the proposed method outperforms the other two methods: moderate gains (0.45% and 0.46%) could be obtained. In the second analysis, the robustness of proposed method was proven by investigating the performance gain by cross-validation of the parameter γ only, for fixed combinations of k and τ. Here, multiple kNN classifiers were applied, showing that the performance could be improved (nearly) for all combinations of k and τ over LFDA and LPP. This shows that weighting local and global structures in the LDA/LFDA-front-end makes sense, also in the case of a less optimal kNN back-end. The third analysis showed the upper (ceiling) limit of the gain we could achieve, by means of an oracle experiment in which the test set itself was used for optimizing the model parameters. The significant gain (1.11% and 1.39% in maximum for LFDA and LPP) obtained in this case also proves the potential usefulness of the proposed weighting method.

In summary, we firstly substantiated the co-existence of local and global structure in the feature space of TIMIT. Furthermore, a generalized affinity matrix with the parameter γ provided us a better way to explore these structures in terms of dimensionality reduction algorithms. The effectiveness of our approach was proven by TIMIT phone classification by outperforming original LFDA and LPP.

It is interesting to note that our optimal setting of LDA+kNN outperformed the kNN classifier used in [13] when it was combined with MFCC features (74.45% versus 73.96%). However, their kNN classifier outperformed our system when it was used with Boosted Maximum Mutual Information features (78.38%). Although their usage of the phonetic boundary information in TIMIT might account for our degradation to their performance to some extent, we still plan to investigate whether it is possible to integrate prior information about the most likely phone class in our extended version of FLDA.

7. ACKNOWLEDGEMENT

The research leading to these results has received funding from the [European Community's] Seventh Framework Programme [FP7/2007-2013] under grant agreement $n^{\circ}213850$ - SCALE. The research of Jort F. Gemmeke was supported by the Dutch-Flemish STEVIN project MIDAS and by IWT project ALADIN.

REFERENCES

[1] H. Hermansky and P. Jain, "Band-independent speech-events categories for TRAP based ASR," in *Proc. Eurospeech 2003*, 2003, pp. 1013 – 1016.

[2] H.Gish and K.Ng, "Parametric trajectory models for speech recognition," in *Proceedings of ICASSP-93*, 1996, pp. 466–469.

[3] Y. Han, J. de Veth, and L. Boves, "Trajectory clustering for solving the trajectory folding problem in automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15:4, pp. 1425 – 1434, 2007.

[4] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong jiang Zhang, Qiang Yang, Senior Member, and Stephen Lin, "Graph embedding and extension: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 29, pp. 40–51, 2007.

[5] Masashi Sugiyama and Sam Roweis, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis," *Journal of Machine Learning Research*, vol. 8, pp. 1027–1061, 2007.

[6] Xiaofei He and Partha Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems 16*, Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, Eds. MIT Press, Cambridge, MA, 2004.

[7] M. Sakai, N. Kitaoka, and K. Takeda, "Feature transformation based on discriminant analysis preserving local structure for speech recognition," in *Proceedings ICASSP-2009*, 2009, pp. 3813 – 3816.

[8] Lihi Zelnik-manor and Pietro Perona, "Self-tuning spectral clustering," in *Advances in Neural Information Processing Systems 17*. 2004, pp. 1601–1608, MIT Press.

[9] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hmms," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[10] Andrew K. Halberstadt, "Heterogeneous acoustic measurements and multiple classifiers for speech recognition," *Ph.D Thesis, MIT*, 1998.

[11] S. Fidler, D. Skocaj, and A. Leonardis, "Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling," *IEEE Trans. on Pattern Analysis and Machine Intelligence,*, vol. 28, pp. 337–350, 2006.

[12] S.Tan, "Neighbor-weighted k-nearest neighbor for unbalanced text corpus," *Expert Systems with Applications*, vol. 28, pp. 667–671, 2005.

[13] T. N. Sainath, A. Carmi, D. Kanevsky, and B. Ramabhadran, "Bayesian compressive sensing for phonetic classification," in *Proceedings ICASSP10*, 2010, pp. 4370–4373.