Investigation of Supervised Dimensionality Reduction Methods for Phonetic Classification *

Heyun Huang
Department of Linguistics
Radboud University Nijmegen
Erasmuslaan 1, 6525 HT,
Nijmegen, the Netherlands
h.huang@let.ru.nl

Yang Liu
Department of Computing
The Hong Kong Polytechnic
University
Hong Kong, P.R. China
csygliu@comp.polyu.edu.hk

Lou Boves
Department of Linguistics
Radboud University Nijmegen
Erasmuslaan 1, 6525 HT,
Nijmegen, the Netherlands
l.boves@let.ru.nl

ABSTRACT

Automatic Speech Recognition (ASR) depends crucially on establishing acoustic models for speech units including phones. One disadvantage that lies in popular acoustic models is the lack of modeling speech continuity information. Stacking short-term features of consecutive frames may keep sufficient articulatory information. Unfortunately, the resultant high-dimensional feature space is still full of redundant information and also causes the curse of dimensionality for subsequent acoustic modeling. Motivated by this and some recent research [4, 15], our paper investigates the supervised dimensionality reduction methods to answer two research questions: whether local structures exist in the feature space formulated by stacking frames and whether the local structures help the acoustic modeling. Experimental results by TIMIT phonetic classification show that the assumed local structures do exist in the feature space and could be best described by nearest neighbor graphs.

Categories and Subject Descriptors

I.2.7 [Natural Language Processing]: Speech recognition and synthesis

General Terms

Design, Experimentation, Performance

Keywords

Automatic Speech Recognition, Phonetic Classification, Supervised Dimensionality Reduction, Local Structure, TIMIT

1. INTRODUCTION

*The research leading to these results has received funding from the [European Community's] Seventh Framework Programme [FP7/2007-2013] under grant agreement n°213850 - SCALE.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ICIMCS'11, August 5-7, 2011, Chengdu, Sichuan, China.
Copyright 2011 ACM 978-1-4503-0918-9/11/08 ...\$10.00.

Automatic Speech Recognition (ASR) [19, 17] is aiming at converting spoken words to text by machine, which has been widely researched for couples of decades. One of the most critical procedures is acoustical modeling of the basic linguistic units. Similar to human brain, ASR system always regards the phone as the basic unit and attempts to accurately model the posterior probabilities of each phone. In most of the ASR systems, Mel-Frequency Cepstral Coefficients (MFCC) have been proven as an effective model that characterizes the short-time speech signals. With these MFCCs, distributions of phones could be well trained for subsequent processing.

Current ASR systems compute MFCCs of short speech signals, which are always 25 milliseconds every 10 milliseconds. However, the procedure of speech production is always continuous. The underlying segmentation procedure prevents ASR systems from capturing the continuity information. To compensate the information loss, the differences and accelerators of MFCCs in successive frames are frequently used to model the continuity. They are heuristic in characterizing the time-evolution within a phone and also fail to represent the transitions between neighbored phones. As an alternative to these methods [13, 10, 12] capturing the speech trajectories, stacking consecutive frames of MFCCs becomes a possible approach to keep the continuity information of both within-phone and between-phone. If more frames are concatenated together, more useful information will be preserved at the cost of the increase of dimensionality. To be concrete, if we aim at representing the articulatory dynamics of syllables whose duration is 250 milliseconds on average, the consequent dimension of the MFCC stack will be larger than 300, which could be expected to preserve most useful information of phones.

However, the high-dimensional representation by stacking MFCCs is a double-edged sword, which probably exaggerates the pronunciation variability [16, 23] in two aspects. First, the phonetic length is flexible [18], which increases the difficulty of measuring similarities between two tokens with different lengths. Stacking more frames means allowing the larger variabilities of phonetic lengths. Second, the unpredictability of the neighbored phones, including their classes and lengths, will be more severe accompanied with concatenating frames. Meanwhile, the high-dimensional representation might result in the curse of dimensionality which detrimentally influences the subsequent classifiers [3], such as the k-Nearest Neighbor (kNN) classifier [22]. In summary, there is a great necessity of learning a compact and

discriminative representation from the MFCC stack.

For the purpose of phonetic classification, we aim at learning a representation that enhances the similarity of tokens from the same class and dissimilarity of tokens from different classes. One way for realizing this idea is performing the Fisher Discriminant Analysis (FDA) [6], which is one of the most representative supervised dimensionality reduction algorithms. Generally speaking, FDA seeks an optimal linear combination of features which maximizes the between-class scatter and minimizes the within-class scatter simultaneously. However, an important assumption of FDA is that the data distribution of each class is an unimodal normal distribution, which makes that FDA is not capable of representing complex local variations and subtleties of the data structure like protrusions and concavities, which might exist in the vector space underlying the MFCC stack because of the co-articulation effects. In order to enhance the discriminant ability of FDA on complex data sets, several methods designed to describe the localities [11] have been proposed recently, including Local FDA (LFDA) [24, 21], Local Discriminant Embedding (LDE) [2], Marginal Fisher Analysis [26], Locality Sensitive Discriminant Analysis (LSDA) [1], and Average Neighborhood Margin Maximization (ANMM) [25]. The ideas inside these local methods are similar: they seek a linear projection that maximizes the separations among nearby points by constructing a neighborhood relationship graph.

In this paper, we review some of the aforementioned supervised dimensionality reduction methods and evaluate them by the performance comparison on the phonetic classification task. One major goal of our evaluation is to investigate whether the local structures exist in the feature space of MFCC stacks. In addition, for local methods, the approaches to construct the adjacency graph are then visited by comparing those algorithms with different ideas to put edges among tokens. With the optimal adjacency maps, several weighting methods to weight the edges will be investigated to show which idea is most suitable for describing the affinities among tokens.

The rest of the paper is organized as follows. Section 2 briefly introduces the phonetic classification task using MFCC stack as the basic unit. Section 3 reviews some representative local methods in the algorithmic level. Section 4 provides the experimental results and related analysis. The last section presents general discussion and draw our conclusion about the investigation.

2. THE PHONETIC CLASSIFICATION TASK

To recognize utterances or sentence like human brain, automatic speech recognition systems should first learn the basic linguistic unit. Phones have been widely acknowledged and applied as a suitable linguistic unit for acoustical modeling. In most of the ASR systems, words or sentences are always segmented into several short-time frames and the posterior probabilities of these frames are then estimated. Roughly speaking, the goal of acoustical modeling could be summarized as finding out a discriminative (or less overlapping) representation. The target of phonetic classification is to correctly classify newly-arrived phones by the classifier trained using training phones, which is inherently in accordance with the target of ASR systems. To be concrete, the common goal of phonetic classification and acoustical modeling in speech recognition is to separate acoustic

Figure 1: Flowchart of our phonetic classification scheme.

features from different classes of phones to the greatest extent. Therefore, although artificial, phonetic classification has been regarded as a common way to evaluate the effectiveness of acoustical models [8]. This paper will evaluate the acoustical models derived from supervised dimensionality reduction algorithms, whose result is probably instructive to the design of acoustical models for ASR systems.

2.1 The Classification Scheme

A common phonetic classification system contains feature extraction and subsequent classification modules. As mentioned above, the feature extraction module in our work could be described by stacking MFCC frames and then projecting them into a lower dimensional representation. The flowchart of our phonetic classifier is given in Fig. 1. It is worth mentioning that Principal Component Analysis (PCA) [14] is applied to reduce the dimensionality of original feature vector as a common tandem in discriminant analysis [5]. Different supervised dimensionality reduction methods are performed after PCA.

The subsequent classifier is the Weighted k-Nearest Neighbor (WkNN) classifier [22], which classifies the reduced low-dimensional feature vectors into one of the given phonetic classes. Suppose that $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k$ are the k nearest neighbors of a test vector $\vec{x}$. The weights of these neighbors are calculated by Eq. 1:

$$w_i = \exp\left(-\frac{\|\vec{x}_i - \vec{x}\|^2}{\tau}\right), i = 1, 2, \dots, k \quad (1)$$

The label of test vector is assigned by the class whose accumulated weights are the biggest. Selection of parameters k and τ will be discussed in the experimental result part.

2.2 Speech Corpus for Evaluation: TIMIT

TIMIT [7], the well-known speech corpus recorded in Texas Instruments (TI) and transcribed in Massachusetts Institute of Technology (MIT), will be used for performance evaluation as a common way [17, 4, 20, 9] in the speech community. There are 61 different (labeled) phones in TIMIT that cover all possible phones of American English. To evaluate the acoustical models effectively, the authors in [17] suggested folding 61 phone labels into 48 phone labels, which is a common setting and is adopted in our paper. To train the classifier, the standard NIST training set is used in this experiment, which includes 462 speakers, 3,696 utterances,

Figure 2: Examples of vowel's MFCC stack: 23 frames $\times$ 12 MFCCs . Eight examples are shown for each class: /aa/, /ae/, /ey/, /eh/, /ih/, and /iy/.

Figure 3: Examples of plosive's MFCC stack: 23 frames $\times$ 12 MFCCs . Eight examples are shown for each class: /k/, /t/, /p/, /g/, /d/, and /b/.

and 139,852 phones. Meanwhile, they act as the templates of subsequent weighted kNN classifier. Some parameters in the dimensionality reduction and kNN methods are tuned using the development set proposed in [8], in which there are 50 speakers, 400 utterances, and 15,038 phones. The standard test set, called core set in TIMIT, contains 24 speakers, 192 utterances, and 7,196 phones, which is used to evaluate the performance of different algorithms. To investigate characteristics of dimensionality reduction algorithms, the remaining utterances from standard NIST test set (43,137 phones) are used in corresponding experiments.

To characterize each phone, the feature space is generated as follows: A short-time Fourier analysis is performed every 10 ms using a 25 ms Hamming window. These Fourier coefficients are then transformed to 12 MFCCs plus an energy coefficient. To build a contextual representation of a phone, 23 frames are concatenated and the center frame of the stack is that of the phone. Consequently, every phonetic segment is represented by a 299(= 13 $\times$ 23)-dimensional feature block.

To visualize the MFCC stacks, Fig.2 and Fig.3 show some examples of six types of vowels and plosives, respectively. PCA is then performed on the 299-dimensional vectors to generate 150-dimensional vectors which preserve 97% of the energy of original data. As possible keys to the black box in Fig.1, those supervised dimensionality reduction algorithms are performed on the PCA-based reduced vectors whose output vectors are then evaluated by the subsequent classifier.

3. REVIEW OF SUPERVISED DIMENSIONALITY REDUCTION METHODS

Given a training set with n points $\vec{x}_i \in \mathbb{R}^D$ ($i = 1, 2, \dots, n$), each data point is assigned by a label $c_{x_i} = c \in \{1, 2, \dots, C\}$. Meanwhile, the number of data points in class c is denoted by n_c . As the most widely-used supervised dimensionality reduction method, Fisher Discriminant Analysis (FDA) [6] aims at exploring the projection matrix $\mathbf{W} \in \mathbb{R}^{D \times d}$ ($d \leq D$) such that the low-dimensional representations $\vec{z}_i \in \mathbb{R}^d$ obtained by $\vec{z}_i = \mathbf{W}^T \vec{x}_i$ have the most discriminative power. Concretely speaking, FDA maximizes the so-call Fisher Criterion described by the following formulation:

$$\arg \max_{\mathbf{W}} \frac{\text{tr}(\mathbf{W}^T \mathbf{S}^b \mathbf{W})}{\text{tr}(\mathbf{W}^T \mathbf{S}^w \mathbf{W})} \quad (2)$$

where $\mathbf{S}^b$ and $\mathbf{S}^w$ denote the within-class and between-class scatter matrices, respectively. Actually, these two matrices can be reformulated by accumulating the pairwise scatter matrices, i.e., $(x_i - x_j)(x_i - x_j)^T$, as follows [24, 26]:

$$\mathbf{S}^w = \frac{1}{2} \sum_{ij} A_{ij}^w (x_i - x_j)(x_i - x_j)^T \quad (3)$$

$$\mathbf{S}^b = \frac{1}{2} \sum_{ij} A_{ij}^b (x_i - x_j)(x_i - x_j)^T \quad (4)$$

$$A_{ij}^b = \begin{cases} 1/n - 1/n_c & \text{if } c_{x_i} = c_{x_j} = c \\ 1/n & \text{if } c_{x_i} \neq c_{x_j} \end{cases}$$

where A_{ij}^w and A_{ij}^b are called affinity matrices or adjacency graphs [11]. From the definition in Eq. 3 and Eq. 4, we can see that the edges are imposed between all data pairs and the weights are independent of the corresponding distances. Therefore, FDA is not able to capture the local information.

The idea of preserving local information using affinity matrix for linear dimensionality reduction was first proposed in [11]. To extend this idea into pattern classification tasks and provide more general solution, Yan et al. [26] proposed a graph embedding framework, which generalizes most of the local linear dimensionality reduction methods. According to this framework, the crucial procedure is to build a graph that considers the neighborhood relationship and endeavors to preserve it by a linear projection for both within-class and between-class scatters. Concretely, the decisive step is to determine A_{ij}^w and A_{ij}^b in which Laplacian graphs are required. Those methods [24, 2, 26], to be reviewed in this section, differ inherently in designing A_{ij}^w and A_{ij}^b from two aspects: constructing the adjacency graph and choosing the weights [11].

3.1 Local Fisher Discriminant Analysis

Similar to FDA, LFDA [24] uses a full adjacency map. In other words, there are edges on all pairs between data points in both A_{ij}^w and A_{ij}^b . The differences lie in the weights of edges of A_{ij}^w . FDA treats each edge equivalently important within a class and thus the weights of edges from the same class are equal. LFDA incorporates the locality idea into FDA by weighting the edges according to the corresponding distances:

$$S^w = \frac{1}{2} \sum_{ij} A_{ij}^w (x_i - x_j)(x_i - x_j)^T \quad (5)$$

$$A_{ij}^w = \begin{cases} a_{ij}/n_c & \text{if } c_{x_i} = c_{x_j} = c \\ 0 & \text{if } c_{x_i} \neq c_{x_j} \end{cases}$$

$$S^b = \frac{1}{2} \sum_{ij} A_{ij}^b (x_i - x_j)(x_i - x_j)^T \quad (6)$$

$$A_{ij}^b = \begin{cases} a_{ij}(1/n - 1/n_c) & \text{if } c_{x_i} = c_{x_j} = c \\ 1/n & \text{if } c_{x_i} \neq c_{x_j} \end{cases}$$

in which a_{ij} is decided by the scaling kernel [27]:

$$a_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma_i \sigma_j}\right) \quad (7)$$

where σ_p ($p = i$ or j) is defined by:

$$\sigma_p = \|x_p - x_p^L\| \quad (8)$$

x_p^L is the L^{th} nearest neighbor of x_p from the same class.

3.2 Local Discriminant Embedding / Marginal Fisher Analysis

The two methods, LDE [2] and MFA [26], employ the similar idea to construct the adjacency graph: each point is only connected with its k nearest neighbors in both within-class and between-class scatters. For each data point x_i , its k_1/k_2 nearest neighbors from the same class and other classes consist of two subsets $N_w(x_i)$ and $N_b(x_j)$:

$$N_w(x_i) = \{x_i^j \mid \|x_i^j - x_i\| < \|x_i^j - x_i^{k_1}\|, c_{x_i^j} = c_{x_i}\} \quad (9)$$

$$N_b(x_i) = \{x_i^j \mid \|x_i^j - x_i\| < \|x_i^j - x_i^{k_2}\|, c_{x_i^j} \neq c_{x_i}\} \quad (10)$$

Therefore, the definition of LDE/MFA can be described as follows:

$$S^w = \frac{1}{2} \sum_{ij} A_{ij}^w (x_i - x_j)(x_i - x_j)^T \quad (11)$$

$$A_{ij}^w = \begin{cases} a_{ij} & \text{if } x_i \in N_w(x_j) \text{ or } x_j \in N_w(x_i) \\ 0 & \text{otherwise} \end{cases}$$

$$S^b = \frac{1}{2} \sum_{ij} A_{ij}^b (x_i - x_j)(x_i - x_j)^T \quad (12)$$

$$A_{ij}^b = \begin{cases} 0 & \text{otherwise } c_{x_i} = c_{x_j} = c \\ b_{ij} & \text{if } x_i \in N_b(x_j) \text{ or } x_j \in N_b(x_i) \end{cases}$$

where a_{ij} and b_{ij} represent the weights of the edge connecting x_i and x_j . Two typical weighting functions are used in [2, 26, 11]:

- Single-Minded: $a_{ij} = b_{ij} = 1$
- Heat-Kernel:

$$a_{ij} = \exp(-\|x_i - x_j\|^2/t_a) \quad (13)$$

$$b_{ij} = \exp(-\|x_i - x_j\|^2/t_b) \quad (14)$$

Table 1: Performance Comparison between FDA and Local FDA on the core test set (upper row) and the development set (lower row) with best tuned parameters on the development set and the core test set.

Test Set	Original	FDA	LFDA
Core Set	68.08	73.47	74.00
Development	68.85	74.16	74.44

4. RESULTS AND ANALYSIS

This section reports the performance of different supervised dimensionality reduction methods. The performance is evaluated by the classification accuracy after cross-validation. Each dimensionality reduction method is first performed on the development set. Its parameters (like k_1, k_2, t_a, t_b for LDE with heat kernel) and other two parameters (k, τ) in kNN are optimized on this set. To be concrete, these parameters are optimized by grid search. Preliminary experiments show the their interesting ranges for kNN: $15 < k < 40, 3.5 < \tau < 6.5$. For the parameters in the dimensionality reduction methods, the ranges will be introduced in the corresponding experiments. The tuning methods will adopt uniform sampling in the ranges and then searching for the optimal values in the development set, which will then be applied in the core test set. Meanwhile, the roles of development and core test sets will be changed. Furthermore, due to the inherent mismatch between development set and core test set, we will also show the classification accuracy with slightly perturbed values of those parameters. For fair comparison, the reduced dimension is 47, i.e., the number of classes minus 1, which was suggested in [1].

4.1 Substantiating Local Structure in TIMIT

Compared with FDA, the local FDA (LFDA) can be regarded as a natural way to incorporate the local structure. The close data pairs from the same class are regarded more important than the distant data pairs reflected via the value a_{ij} in Eq. 5. Thus, the first experiment is to compare the performance of FDA [6] and local FDA [24]. We use the default setting of LFDA in [24], which means the parameter L in Eq. 8 is set as 7 [27]. Meanwhile, the classification on the original space is implemented to show the benefit obtained by dimensionality reduction. Table 1 shows the results. A significant gain by FDA is achieved (5.39% on core test set and 5.31% on development set) upon the original space. We may also observe that the classification accuracy of LFDA on both development and core test sets is higher than that of FDA, which proves the existence of local structures to some extent.

4.2 Investigation of Graph Construction

LFDA considers the local structure of feature space of MFCC stack. However, the distant data pairs still make some contributions to the within-class scatter matrix. In order to prove this viewpoint explicitly, the definition of LFDA in Eq. 5 and Eq. 6 could be regarded as a special case of LDE/MFA by setting k_1 and k_2 as:

$$k_1 = n_c \text{ if } c_{x_i} = c \quad (15)$$

$$k_2 = \sum_i n_i - n_c \text{ if } c_{x_i} = c \quad (16)$$

Therefore, it is meaningful to investigate k_1 and k_2 that de-

Table 2: Performance Comparison between FDA and LDE.

Test Set	FDA	LDE/MFA
Core Set	73.47	75.69
Development	74.16	77.10

Figure 4: The relationship between classification accuracy and two k -parameters: an interesting diagonal pattern could be observed.

termine the adjacency graphs. To remove the effects of the weights on the edges, the single-minded kernel is adopted. Therefore, the cross-validation is involved with four parameters: (k_1, k_2, k, τ) . For the comparison with FDA, the small values of k_1 and k_2 are visited. After the preliminary trial, we choose the grid as $4 \leq k_1 \leq 20$ (at the interval of 1) and $5 \leq k_2 \leq 100$ (at the interval of 5).

The results are given in Table 2. The highest accuracy on the development set is achieved when $k_1 = 12$ and $k_2 = 55$ on the grid. A significant improvement over FDA could be observed: 2.22% on core test set and 2.61% on the development set, which further substantiates the existence of local structures.

After reporting the classification accuracy, a more interesting issue might be the relationship between the performance and parameters (k_1, k_2) , which is shown in Fig. 4. We use all tokens from TIMIT testing set. The darkness of red increases with the improvement of classification accuracy, which motivates us to draw several conclusions as follows:

- The diagonal pattern in the figure explicitly reveals that the underlying relationship between k_1 and k_2 is approximately **linear**, especially for small k_1 and k_2 .
- The pervasive red shows the robustness of the idea inside LDE/MFA because the red in Fig. 4 means a higher classification accuracy ($> 74.5\%$) than FDA.

4.3 Comprehensive Investigation with Weighted Edges

After investigating the ways to construct adjacency maps with single-minded kernel, we apply other common kernels to weight the edges: heat kernel and adaptive kernel. The

Table 3: Performance Comparison Among Different Graphs and Weights on Core Test Set.

	All Connected Graph	kNN Graph
Single-Minded	73.47	75.63
Heat Kernel	74.60	75.54
Adaptive Kernel	74.77	75.52

Table 4: Performance Comparison Among Different Graphs and Weights on Development Set

	All Connected Graph	kNN Graph
Single-Minded	74.16	77.10
Heat Kernel	74.98	77.13
Adaptive Kernel	75.20	77.26

adaptive kernel is a natural extension from Eq. 7:

$$a_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{(\sigma_i \sigma_j)^{\gamma_a}}\right) \quad (17)$$

$$b_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{(\sigma_i \sigma_j)^{\gamma_b}}\right) \quad (18)$$

Cross-validation here means jointly optimizing (k, τ, t_a, t_b) for heat kernel and $(k, \tau, \gamma_a, \gamma_b)$ for adaptive kernel. Besides the single-minded kernel, two ideas to construct the adjacency graphs will be combined with three kernels and the resultant dimensionality reduction methods will be evaluated on development set and core test set. It is worth mentioning that FDA uses single-minded kernel and all-connected graph. The idea of LFDA is identical with the combination by the all-connected graph and a special adaptive kernel ($\gamma_a = 1, \gamma_b = +\infty$). The results are given in Table 3 and Table 4. The best performance on development set and core test set are achieved by the kNN graph with single-minded kernel and adaptive kernel, respectively. Besides the summary on the best performance, there are some additional interesting observations from the tables:

- The kNN graph is apparently better than the all-connected graph with all weight functions on both development set and core test set, which means that the locality information is better modeled by kNN graph than the weighting functions that is dependent on the pair-wise distances.
- Comparing three kinds of kernels, we may draw different conclusions for two graphs. The heat kernel and adaptive kernel are both superior to the single-minded kernel on the all-connected graph, whereas the performance of three kernels are comparative on the kNN graph. The similar performance of three kernels on the kNN graph imply that it might be more meaningful to decide which data pairs should be connected than to allocate the importance of these connections.

5. DISCUSSION AND CONCLUSION

This paper investigated several representative supervised dimensionality reduction methods from the perspective of adjacency graph construction and applied them to the TIMIT phonetic classification task. First of all, the phonetic classification task was introduced with the theoretical necessity of

reducing the dimension of high-dimensional MFCC stacks. The reason of stacking MFCC frames is that the speech continuity information might be contained to the most extent. Motivated by the co-articulatory effects, we intend to investigate the existence of the local structure in the original feature space formulated by MFCC stacks and whether they can help the classification. In this paper, we explored above issues using the supervised dimensionality reduction methods, which generally compose three steps: constructing adjacency graphs, weighting the edge, and finding the projection matrix. This paper investigated two ways to construct adjacency graphs (all-connected graph and kNN graph) and three common kernels to impose the weights on the edges (single-minded kernel, heat kernel, and adaptive kernel). Four representative supervised dimensionality reduction methods are included in our study: FDA [6], LFDA [24], LDE [2], and MFA [26].

The first experiment is designed to show the improvement achieved by applying FDA to the feature space generated by PCA (keeping 97% energy) on MFCC stacks. The performance improvement indicates the necessity and effectiveness of supervised dimensionality reduction. To explore whether there are some local structure in the PCA-based feature space, a heuristic local method, LFDA, was compared with FDA. The moderate improvement proves the existence of local structures to some extent. As both FDA and LFDA implicitly adopt all-connected graphs, the subsequent experiment aimed at evaluating the kNN graph with small k_1, k_2 values and single-minded kernel. The classification accuracy generated by kNN graphs is much higher than all-connected graphs, FDA. We then used all testing tokens from TIMIT to investigate the relationship between classification accuracy and parameters k_1 and k_2 . These combinations leading high classification accuracy appear as an interesting diagonal pattern, which means that there must be some implicit relationship between k_1 and k_2 . Meanwhile, for most of the combinations, their results are better than FDA, which shows the robustness of the kNN graph. The last experiment comprehensively reported the results under different adjacency graphs and weighting functions.

The step-by-step experiments show the effectiveness of local supervised dimensionality reduction methods. Compared with the features in the original space and projected by FDA, the local method with kNN-graph (that was adopted by LDE [2] and MFA [26]) achieved 7.55%/8.41% and 2.16%/3.10% accuracy improvement on TIMIT core test/development set. The encouraging result not only substantiates the existence of the local structure in MFCC stacks but also implies that modeling the co-articulatory effects with the idea of locality is highly promising.

6. REFERENCES

- [1] D. Cai, X. He, K. Zhou, J. Han, and H. Bao. Locality sensitive discriminant analysis. In *Proc. 20th IJCAI*, pages 708–713, 2007.
- [2] H.-T. Chen, H.-W. Chang, and T.-L. Liu. Local discriminant embedding and its variants. In *Proc. CVPR*, pages 846–853, 2005.
- [3] R. Duda, P. Hart, and D. Stork. Pattern classification. 2001.
- [4] A. Errity, J. McKenna, and B. Kirkpatrick. Manifold learning-based feature transformation for phone classification. In *Advances in Nonlinear Speech Processing*, volume 4885, pages 132–141. 2007.
- [5] S. Fidler, D. Skocaj, and A. Leonardis. Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28:337–350, 2006.
- [6] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, 7:179–188, 1936.
- [7] J. S. Garofolo. *Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database*. National Institute of Standards and Technology (NIST), Gaithersburg, MD, 1988.
- [8] A. K. Halberstadt. Heterogeneous acoustic measurements and multiple classifiers for speech recognition. *Ph.D. Thesis, MIT*, 1998.
- [9] A. K. Halberstadt and J. R. Glass. Heterogeneous acoustic measurements for phonetic classification.
- [10] 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*, 15:4:1425 – 1434, 2007.
- [11] X. He and P. Niyogi. Locality preserving projections. In *NIPS 16*. 2004.
- [12] H. Hermansky and P. Jain. Band-independent speech-events categories for TRAP based ASR. In *Proc. Eurospeech*, pages 1013 – 1016, 2003.
- [13] H. Gish and K. Ng. Parametric trajectory models for speech recognition. In *Proc. ICASSP*, pages 466–469, 1996.
- [14] H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Edu. Psychol.*, 24:417–441, 498–520, 1933.
- [15] A. Jansen and P. Niyogi. Intrinsic fourier analysis on the manifold of speech sounds. In *in Proc. ICASSP*, pages 241–244, 2006.
- [16] D. Jurafsky, W. Ward, J. Zhang, K. Herold, X. Yu, and S. Zhang. What kind of pronunciation variation is hard for triphones to model? In *Proc. ICASSP*, pages 577–580, 2001.
- [17] K. F. Lee and H. W. Hon. Speaker-independent phone recognition using hmms. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(11):1641–1648, 1989.
- [18] L. C. W. Pols, X. Wang, and L. F. M. ten Bosch. Modelling of phone duration (using the timit database) and its potential benefit for asr. *Speech Communications.*, 19:161–176, August 1996.
- [19] L. Rabiner and B.-H. Juang. *Fundamentals of speech recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.
- [20] T. N. Sainath, A. Carmi, D. Kanevsky, and B. Ramabhadran. Bayesian compressive sensing for phonetic classification. In *Proc. ICASSP*, pages 4370–4373, 2010.
- [21] M. Sakai, N. Kitaoka, and K. Takeda. Feature transformation based on discriminant analysis preserving local structure for speech recognition. In *Proc. ICASSP*, pages 3813 – 3816, 2009.
- [22] S. Tan. Neighbor-weighted k-nearest neighbor for unbalanced text corpus. *Expert Systems with Applications*, 28:667–671, 2005.
- [23] H. Strik and C. Cucchiari. Modeling pronunciation variation for asr: A survey of the literature. *Speech Communications.*, 29:225–246, November 1999.
- [24] M. Sugiyama and S. Roweis. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *JMLR*, 8:1027–1061, 2007.
- [25] F. Wang and C. Zhang. Feature extraction by maximizing the average neighborhood margin. In *Proc. CVPR*, 2007.
- [26] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. Graph embedding and extension: A general framework for dimensionality reduction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29:40–51, 2007.
- [27] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *NIPS 17*, pages 1601–1608, 2004.