LIPREADING BY LOCALITY DISCRIMINANT GRAPH

Yun Fu [1,2,3*], *Xi Zhou* [1,2], *Ming Liu* [1,2], *Mark Hasegawa-Johnson* [1,2], *and Thomas S. Huang* [1,2]

[1]Beckman Institute, University of Illinois at Urbana-Champaign (UIUC), Urbana, IL 61801, USA
[2] Dept. of ECE, University of Illinois at Urbana-Champaign (UIUC), Urbana, IL 61801, USA
[3] Dept. of Statistics, University of Illinois at Urbana-Champaign (UIUC), Urbana, IL 61801, USA

ABSTRACT

The major problem in building a good lipreading system is to extract effective visual features from the enormous quantity of video sequences data. For appearance-based feature analysis in lipreading, classical methods, e.g. DCT, PCA and LDA, are usually applied to dimensionality reduction. We present a new pattern classification algorithm, called Locality Discriminant Graph (LDG), and develop a novel lipreading framework to successfully apply LDG to the problem. LDG takes the advantages of both manifold learning and Fisher criteria to seek the linear embedding which preserves the local neighborhood affinity within *same class* while discriminating the neighborhood among *different classes*. The LDG embedding is computed in closed-form and tuned by the only open parameter of k-NN number. Experiments on AVICAR corpus provide evidence that the graph-based pattern classification methods can outperform classical ones for lipreading.

Index Terms— Lipreading, graph embedding, discriminant analysis, audio-visual speech, discrete cosine transform.

1. INTRODUCTION

Over the past decades, numerous studies have demonstrated that the information of lip movement and speech signal are highly correlated and can be complementary for both human and machine perception [13, 14]. For example, some speech sounds are distinct in the visual domain, yet are easily confused in the audio domain. As a result, in the last a few years, lipreading has been attracted lots of research attentions for the improvement of the speech recognition performance, especially under noise environment [12].

The major problem in building a good lip-reading system is to extract effective visual features from the enormous quantity of data in video sequences. Many existing papers have introduced different algorithms for visual feature extraction. In principle, they can be categorized into 3 groups: appearance-based features, shape-based features, and combination of both [16]. It has been shown that the appearance-based features achieve better performances than the other categories [15]. For appearance-based technique, the image patches containing the entire speaker's mouth area are considered as the raw features for lipreading. However, the dimensionality of the mouth region, which is the number of pixels in the image patch, is usually too large to allow effective statistical modeling, by means of a prevalent Hidden Markov Model (HMM) [17]. Therefore, several classical linear transformation algorithms are adopted for dimensionality reduction [8], in which the most commonly applied transforms are the Discrete Cosine Transform (DCT) [18], Principal Components Analysis (PCA) [19] and Linear Discriminant Analysis (LDA) [20].

Recently, the manifold learning and dimensionality reduction in supervised manner have been the focus of considerable issues in image processing and pattern recognition. Unlike above traditional methods that consider Gaussian assumption and global feature space modeling, such new techniques, e.g. Locally Linear Embedding (LLE) [2], Laplacian Eigenmaps [4], Isomap [3], focus on preserving the local geodesic distances and neighborhood relationships which reflect the real geometry structure of the low-dimensional manifold without strongly depending on the data distribution. Moreover, the linearization form of these methods [9], e.g. Locality Preserving Projections (LPP) [5], and Locality Embedded Analysis (LEA) [1], are designed for more practical applications in pattern classification. Since lipreading needs data fusion and multicue audio-visual analysis [10], it is straightforward to introduce such manifold learning methods into this field. However, these linear algorithms mainly focus on preserving data localities and similarities in the manifold space so that discriminating power can not be guaranteed sufficiently high. As a result, the projected data points of different classes may still overlap after embedding. A few recent developed successful manifold learning methods [7, 6] take into account the Fisher criteria which explicitly aims at maximizing the discriminant capacity of the embedding.

In this paper, we present a new algorithm, called Locality Discriminant Graph (LDG), for feature extraction and representation in the application of lipreading. In supervised learning case, assuming there are generally two types of nearest neighbors of the same class and different class for each data

[*] *E-mail*: {yunfu2, xizhou2, mingliu1, jhasegaw, t-huang1}@uiuc.edu. This research was funded in Part by the U.S. Government VACE program, and in part by the NSF Grant CCF 04-26627. The views and conclusions are those of the authors, not of the US Government or its Agencies.

1-4244-1437-7/07/$20.00 ©2007 IEEE

point, the affinities of such neighborhood relations are modelled by locality linear reconstructions of NNs. Our proposed algorithm tries to seek the linear embedding which preserves the local neighborhood affinity within same class while discriminating the neighborhood affinity among different classes. Such LDG embedding is computed in closed-form and tuned by the only open parameter of k-NN number. The effectiveness and advantage of LDG are validated by applying it to the realistic lipreading scenario on AVICAR corpus, and comparing with several traditional techniques.

Our major contribution and conclusions are summarized in four points. (1) We present a new pattern classification algorithm LDG for discriminant subspace learning; (2) We develop a lipreading framework and successfully apply LDG algorithm to the problem; (3) This work demonstrates that the graph-based feature selection methods can outperform classical ones for lipreading. (4) To our best knowledge, this is a new try to apply graph-based methods to lipreading.

2. LIPREADING FRAMEWORK

Our lipreading system is structured through a flow diagram in Figure 1. First, AdaBoost-based face tracker and lip tracker are used to estimate the mouth region (Region of Interest, ROI). The entire image in the ROI creates the source high-dimensional feature space. Secondly, different dimensionality reduction algorithms, including classical linear transformation algorithms (DCT, PCA, LDA) and manifold learning algorithms (LPP, LEA, LDG) are used and compared for learning low dimensional feature space. Linear head position correction and model-based head position correction are operated on lip features for alleviating the influence by different head pose and unequal length of feature vector. Last, left-to-right and continuous density HMM is trained at word-level in the training database. Each HMM contains 7 states, while one or more Gaussian distributions with diagonal covariance matrix are associated with each state. The HMM-based recognizer was implemented using the hidden Markov model toolkit HTK version 3.1 [21].

3. LOCALITY DISCRIMINANT GRAPH

3.1. Discriminant Graph Modeling

The new algorithm, Locality Discriminant Graph (LDG), is based on the assumption that geometric relationship among the high dimensional data is described by the linear neighborhood reconstruction. It means that each data point can be represented by the linear combination of its k nearest neighbors. Considering the Fisher criteria, we define two types of k nearest neighbors for each data point: *same class* k-NN and *different class* $\widetilde{k}$-NN (k can be different from $\widetilde{k}$). Here the k-NN can be constructed by any kinds of distance metric

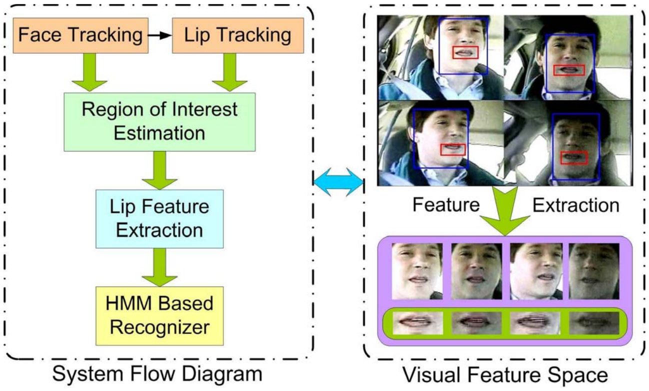


Fig. 1. Lipreading system structure.

(spatio-temporal, Euclidean, correlation) or searching strategies (kd-Tree, Hash).

Suppose we have the original data set $\mathcal{X} = \{\mathbf{x}_i : \mathbf{x}_i \in \mathbb{R}^D\}_{i=1}^n$. Considering supervised learning case, each sample pattern in the training set is associated with a category label l, so we have a label set $\mathcal{L} = \{l_i : l_i \in \mathbb{R}\}_{i=1}^n$. Denote the set of $\mathbf{x}_i$'s same class k nearest neighbors from set $\mathcal{X}$ as $\mathcal{X}_{\mathrm{s}}^{(i)} = \{\mathbf{x}_{\mathrm{s}(j)}^{(i)}\}_{j=1}^k$, satisfying $l_{\mathrm{s}(p)}^{(i)} = l_{\mathrm{s}(q)}^{(i)}$ for $p, q = 1, 2, \cdots, k$. In the same way, denote the set of $\mathbf{x}_i$'s different class $\widetilde{k}$ nearest neighbors from set $\mathcal{X}$ as $\mathcal{X}_{\mathrm{d}}^{(i)} = \{\mathbf{x}_{\mathrm{d}(j)}^{(i)}\}_{j=1}^{\widetilde{k}}$, satisfying $l_{\mathrm{d}(p)}^{(i)} \neq l_{\mathrm{d}(q)}^{(i)}$ for $p, q = 1, 2, \cdots, \widetilde{k}$. In the style of LLE modeling, the same class and different class reconstruction error are calculated by the following cost functions respectively.

$$\begin{cases} \varepsilon_{\mathrm{s}}(\mathbf{C}_{\mathrm{s}}) = \sum_{i=1}^n \left\|\mathbf{x}_i - \sum_{j=1}^k c_{\mathrm{s}(j)}^{(i)} \mathbf{x}_{\mathrm{s}(j)}^{(i)}\right\|^2 \\ \varepsilon_{\mathrm{d}}(\mathbf{C}_{\mathrm{d}}) = \sum_{i=1}^n \left\|\mathbf{x}_i - \sum_{j=1}^{\widetilde{k}} c_{\mathrm{d}(j)}^{(i)} \mathbf{x}_{\mathrm{d}(j)}^{(i)}\right\|^2 \end{cases} \tag{1}$$

where $\mathbf{C}_{\mathrm{s}}$ and $\mathbf{C}_{\mathrm{d}}$ are $n \times n$ matrices to encode the coefficients, and $\mathrm{s}(j), \mathrm{d}(j) = 1, 2, \cdots, n$. Define the reconstruction coefficient set $\mathcal{C}_{\mathrm{s}}^{(i)} = \{c_{\mathrm{s}(1)}^{(i)}, \cdots, c_{\mathrm{s}(k)}^{(i)}\}$ for same class case and $\mathcal{C}_{\mathrm{d}}^{(i)} = \{c_{\mathrm{d}(1)}^{(i)}, \cdots, c_{\mathrm{d}(\widetilde{k})}^{(i)}\}$ for different class case respectively, subject to the following constraints

$$\begin{cases} \sum_{j=1}^k c_{\mathrm{s}(j)}^{(i)} = 1, \text{ if } c_{\mathrm{s}(j)}^{(i)} \in \mathcal{C}_{\mathrm{s}}^{(i)}; \text{ else } c_{\mathrm{s}(j)}^{(i)} = 0. \\ \sum_{j=1}^{\widetilde{k}} c_{\mathrm{d}(j)}^{(i)} = 1, \text{ if } c_{\mathrm{d}(j)}^{(i)} \in \mathcal{C}_{\mathrm{d}}^{(i)}; \text{ else } c_{\mathrm{d}(j)}^{(i)} = 0. \end{cases} \tag{2}$$

We have $\mathbf{C}_{\mathrm{s}}[i, \mathrm{s}(j)] = c_{\mathrm{s}(j)}^{(i)}$, $\mathbf{C}_{\mathrm{d}}[i, \mathrm{d}(j)] = c_{\mathrm{d}(j)}^{(i)}$ and the other elements all 0. $\mathbf{C}_{\mathrm{s}}$ and $\mathbf{C}_{\mathrm{d}}$ are sparse matrices consisting of neighborhood characteristic of the original space.

From the above formulation, we have two cost functions with $\mathcal{X}$, $\mathcal{X}_{\mathrm{s}}^{(i)}$, $\mathcal{X}_{\mathrm{d}}^{(i)}$ known and $\mathbf{C}_{\mathrm{s}}$, $\mathbf{C}_{\mathrm{d}}$ unknown. For graph modeling, we can calculate the two kinds of coefficients with each data point centered in local by solving

$$\begin{cases} \mathcal{C}_{\mathrm{s}}^{(i)} = \arg\min(\varepsilon_{\mathrm{s}}(\mathbf{C}_{\mathrm{s}})). \\ \mathcal{C}_{\mathrm{d}}^{(i)} = \arg\min(\varepsilon_{\mathrm{d}}(\mathbf{C}_{\mathrm{d}})). \end{cases} \tag{3}$$

Then we can fill in the elements of matrices $\mathbf{C}_{\mathrm{s}}$ and $\mathbf{C}_{\mathrm{d}}$ with above results. Notice that here we already have two parameters, k and $\widetilde{k}$, to tune for better localized graph modeling.

3.2. Subspace Learning

Our goal is to find a linear projection $P \in \mathbb{R}^{D \times d}$ of our desired subspace, which preserves the local neighborhood affinity within same class while discriminating the neighborhood affinity among different classes at the same time. Hence we adopt the similar cost function as Equation 1 to represent the model in the embedded subspace. The objective function for modeling P is formulated as follows

$$\begin{cases} \varepsilon_{\mathrm{s}}(\mathbf{P}) = \sum_{i=1}^{n} \left\| \mathbf{P}^T \mathbf{x}_i - \sum_{j=1}^{k} c_{\mathrm{s}(j)}^{(i)} \mathbf{P}^T \mathbf{x}_{\mathrm{s}(j)}^{(i)} \right\|^2 \\ \varepsilon_{\mathrm{d}}(\mathbf{P}) = \sum_{i=1}^{n} \left\| \mathbf{P}^T \mathbf{x}_i - \sum_{j=1}^{\widetilde{k}} c_{\mathrm{d}(j)}^{(i)} \mathbf{P}^T \mathbf{x}_{\mathrm{d}(j)}^{(i)} \right\|^2 \end{cases} \tag{4}$$

Notice that we have above $\mathbf{C}_{\mathrm{s}}$ and $\mathbf{C}_{\mathrm{d}}$ and $\mathcal{X}$ known but matrix $\mathbf{P}$ unknown this time. After embedding, we want to keep the class relation reflected by the known labels. That is to say, in the low-dimensional subspace, we want to preserve neighboring points close if they have the same label, while prevent points of other classes from entering the neighborhood. We finally have the constrained optimization problem

$$\mathbf{P} = \max(\varepsilon_{\mathrm{d}}(\mathbf{P})), \text{ subject to } \varepsilon_{\mathrm{s}}(\mathbf{P}) = t. \tag{5}$$

where t is a small constant, such as 1. This equation can be solved by Lagrange Optimization.

3.3. The LDG Algorithm

The proposed LDG algorithm in matrix forms is summarized as follows. Due to space limit, some details are omitted.

3.3.1. Calculate Coefficient Matrices.

Define the $k \times k$ local Gram matrix $\mathbf{G}_i$ for each $\mathbf{x}_i$ as $\mathbf{G}_i[p,q] = \left(\mathbf{x}_i - \mathbf{x}_{\mathrm{s}(p)}^{(i)}\right)^T \left(\mathbf{x}_i - \mathbf{x}_{\mathrm{s}(q)}^{(i)}\right)$, we have $\mathbf{C}_{\mathrm{s}}^{(i)} = \frac{\mathbf{G}_i^{-1}\mathbf{1}}{\mathbf{1}^T \mathbf{G}_i^{-1} \mathbf{1}}$. Then $\mathbf{C}_{\mathrm{s}}[i, \mathrm{s}(j)] = \mathbf{C}_{\mathrm{s}}^{(i)}(j)$. In the same way, define the $\widetilde{k} \times \widetilde{k}$ local Gram matrix $\widetilde{\mathbf{G}}_i$ for each $\mathbf{x}_i$ as $\widetilde{\mathbf{G}}_i[p,q] = \left(\mathbf{x}_i - \mathbf{x}_{\mathrm{d}(p)}^{(i)}\right)^T \left(\mathbf{x}_i - \mathbf{x}_{\mathrm{d}(q)}^{(i)}\right)$, we have $\mathbf{C}_{\mathrm{d}}^{(i)} = \frac{\widetilde{\mathbf{G}}_i^{-1}\mathbf{1}}{\mathbf{1}^T \widetilde{\mathbf{G}}_i^{-1} \mathbf{1}}$. Then $\mathbf{C}_{\mathrm{d}}[i, \mathrm{d}(j)] = \mathbf{C}_{\mathrm{d}}^{(i)}(j)$.

3.3.2. Calculate Projection Matrices.

After we get $\mathbf{C}_{\mathrm{s}}$ and $\mathbf{C}_{\mathrm{d}}$, we can calculate projection matrix $\mathbf{P}$ by solving the following eigenvalue problem

$$\mathbf{X}\left(\mathbf{D}_{\mathrm{d}} - \mathbf{C}_{\mathrm{d}}\right)^T \left(\mathbf{D}_{\mathrm{d}} - \mathbf{C}_{\mathrm{d}}\right) \mathbf{X}^T \mathbf{P} = \Lambda \mathbf{X}\left(\mathbf{D}_{\mathrm{s}} - \mathbf{C}_{\mathrm{s}}\right)^T \left(\mathbf{D}_{\mathrm{s}} - \mathbf{C}_{\mathrm{s}}\right) \mathbf{X}^T \mathbf{P}. \tag{6}$$

where $\mathbf{D}_{\mathrm{d}}[i,i] = \sum_{j=1}^{n} \mathbf{C}_{\mathrm{d}}[i,j]$ and $\mathbf{D}_{\mathrm{d}}[i,i] = 0$ for $\forall\ i \neq j$, also $\mathbf{D}_{\mathrm{s}}[i,i] = \sum_{j=1}^{n} \mathbf{C}_{\mathrm{s}}[i,j]$ and $\mathbf{D}_{\mathrm{s}}[i,i] = 0$ for $\forall\ i \neq j$.

As a result, we have $[\mathbf{p}_1\ \ \mathbf{p}_2\ \ \cdots\ \ \mathbf{p}_d]$, the generalized eigenvectors that correspond to the d largest eigenvalues in Equation 6. We can instead solve the following eigen decomposition equation with Singular Value Decomposition (SVD)

$$\mathbf{A}\mathbf{P} = \Lambda \mathbf{P} \tag{7}$$

where $\mathbf{A} = \left[\mathbf{X}\left(\mathbf{D}_{\mathrm{s}} - \mathbf{C}_{\mathrm{s}}\right)^T \left(\mathbf{D}_{\mathrm{s}} - \mathbf{C}_{\mathrm{s}}\right) \mathbf{X}^T \mathbf{P}\right]^{-1} \left[\mathbf{X}\left(\mathbf{D}_{\mathrm{d}} - \mathbf{C}_{\mathrm{d}}\right)^T \left(\mathbf{D}_{\mathrm{d}} - \mathbf{C}_{\mathrm{d}}\right) \mathbf{X}^T \mathbf{P}\right]$ and $\mathbf{P}$ denotes the learned subspace.

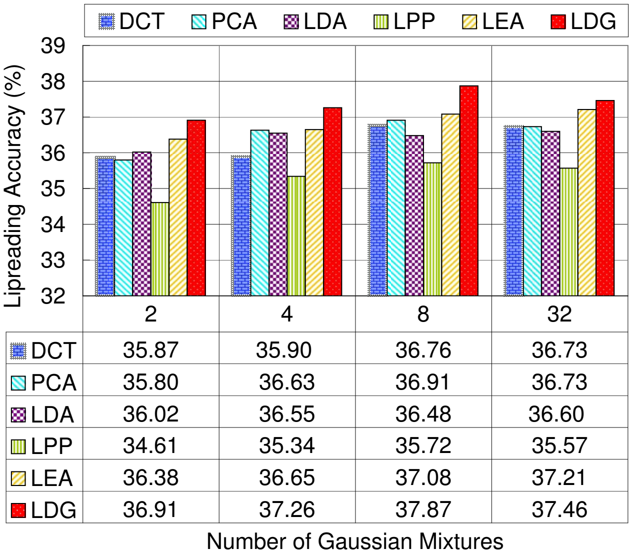
3.3.3. Calculate Data Projection.

Any new input data $\mathbf{X}_{\mathrm{new}}$ can be represented by the new coordinates $\mathbf{Y}_{\mathrm{new}} = \mathbf{P}^T \mathbf{X}_{\mathrm{new}}$.

4. EXPERIMENTS

We demonstrate the effectiveness of our algorithm and compare it with other state-of-the-art methods using the UIUC AVICAR corpus [11], shown in Figure 1. The UIUC AVICAR corpus data are recorded in a real car environment using a multi-sensory array, consisting of eight microphones on the sun visor and four video cameras on the dashboard. The four cameras are fixed in different locations to take four views of images synchronously. The training set has 851 utterances from 21 talkers, and the testing set has 490 utterances from 13 different talkers *not* in the training set. Both male and female talkers from various language backgrounds are contained in the training and testing sets. The script of each utterance is the connected digits. We train the multi-talker whole-word HMMs for all digits in the training set.

The entire pixels in the ROIs of the four sub-images in each frame create the high-dimensional source feature space.



	2	4	8	32
DCT	35.87	35.90	36.76	36.73
PCA	35.80	36.63	36.91	36.73
LDA	36.02	36.55	36.48	36.60
LPP	34.61	35.34	35.72	35.57
LEA	36.38	36.65	37.08	37.21
LDG	36.91	37.26	37.87	37.46

Fig. 2. Comparison of the highest lipreading accuracy (percentage) of all the 6 methods under different number of Gaussian mixtures per HMM state.

Table 1. Number of feature dimension corresponding to the highest lipreading accuracy in Figure 2.

	Gaussian Mixtures			
Methods	**2**	**4**	**8**	**32**
DCT	49	37	49	44
PCA	45	42	47	41
LDA	48	48	49	49
LPP	49	41	48	38
LEA	42	40	44	33
LDG	41	41	41	50

We first compress the pixels in each of the four sub-images to the corresponding upper-left 584 DCT coefficients. After stacking them together, we have in total 2,336 dimensions for the combined feature vector. Then we compress this DCT feature further to the first 100 PCA coefficients. Finally, different subspace learning methods, DCT, PCA, LDA, LPP, LEA and LDG, are applied to represent the lipreading feature in $1 \sim 50$ dimensions after dimensionality reduction. Figure 2 shows the comparison of the highest lipreading accuracy (percentage) of all the 6 methods under different number (2,4,8,32) of Gaussian mixtures per HMM state. Table 1 shows the number of feature dimension corresponding to the highest lipreading accuracy in Figure 2. We can observe that the graph-based manifold learning and dimensionality reduction methods, e.g. LEA and LDG, outperform the classical linear subspace learning methods, e.g. DCT, PCA and LDA, in all the cases under different number of Gaussian components. The computational complexity of these methods are all comparable to each other, but LDG consistently achieves the highest lipreading accuracy in all the cases with comparable best dimension numbers to that of the other 5 methods after dimensionality reduction. It is due to the database complexity and the pre-compression on the original feature by DCT that cause the marginal improvement by LDG for this database.

5. CONCLUSION

We have presented a new pattern classification algorithm, LDG, and developed a novel lipreading framework to adopt LDG as the feature selection method. LDG has been validated to be effective for feature representation and dimensionality reduction. Experimental results also demonstrate that, with comparable number of dimension for feature representation, the graph-based pattern classification methods, e.g. LEA and LDG, can outperform classical ones, e.g. DCT, PCA and LDA, for appearance-based feature analysis in lipreading.

6. REFERENCES

[1] Y. Fu and T.S. Huang, *Locally Linear Embedded Eigenspace Analysis*, www.ifp.uiuc.edu/∼yunfu2/papers/LEA-Yun05.pdf.

[2] S.T. Roweis and L.K. Saul, “Nonlinear Dimensionality Reduction by Locally Linear Embedding,” *Science*, vol. 290, pp. 2323-2326, 2000.

[3] J.B. Tenenbaum, V.de Silva and J.C. Langford, “A Global Geometric Framework for Nonlinear Dimensionality Reduction,” *Science*, vol. 290, pp. 2319-2323, 2000.

[4] M. Belkin and P. Niyogi, “Laplacian Eigenmaps for Dimensionality Reduction and Data Representation,” *Neural Computation*, vol. 15, no. 6, pp. 1373-1396, 2003.

[5] X.F. He and P. Niyogi, “Locality Preserving Projections,” *Proc. of NIPS’03*, 2003.

[6] Q.B. You, N.N. Zheng, S.Y. Du, Y. Wu, “Neighborhood Discriminant Projection for Face Recognition,” *IEEE Conf. on ICPR’06*, pp. 532-535, 2006.

[7] H.-T. Chen, H.-W. Chang and T.-L. Liu, “Local Discriminant Embedding and Its Variants,” *IEEE Conf. on CVPR’05*, pp. 846-853, 2005.

[8] A.M. Martinez and A.C. Kak, “PCA versus LDA,” *IEEE Trans. on PAMI*, vol. 23, no. 2, pp. 228-233, 2001.

[9] S.C. Yan, D. Xu, B.Y. Zhang and H.J. Zhang, “Graph Embedding: A General Framework for Dimensionality Reduction,” *IEEE Conf. on CVPR’05*, pp. 830-837, 2005.

[10] S. Lafon, Y.Keller, and R.R. Coifman, “Data Fusion and Multicue Data Matching by Diffusion Maps,” *IEEE Trans. on PAMI*, vol. 28, no. 11, pp. 1784-1797, 2006.

[11] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T.S. Huang, “AVICAR: An Audiovisual Speech Corpus in a Car Environment,” *ICSLP’04*, pp. 2489-2492, 2004.

[12] I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox, and R. Harvey, “Extraction of Visual Features for Lipreading,” *IEEE Trans. on PAMI*, vol. 24, no. 2. pp. 198-213, 2002.

[13] N.P. Erber, “Auditory-Visual Perception of Speech,” *J. Speech & Hearing Disorders*, vol. 40, pp. 481-492, 1975.

[14] E.D. Petajan, *Automatic Lipreading to Enhance Speech Recognition*, Ph.D thesis, UIUC, 1984.

[15] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, “Recent Advances in the Automatic Recognition of Audio-Visual Speech,” *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306-1326, 2003.

[16] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari and J. Zhou, “Audio-Visual Speech Recognition,” Technical Report, Workshop 2000, CLSP, Johns Hopkins University, 2000.

[17] S.E. Levinson, L.R. Rabiner, and M.M. Sondhi, “An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition,” *J. The Bell System Technical*, vol. 62, no. 4, pp. 1035-1074, 1983.

[18] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, “Dynamic Bayesian networks for audio-visual speech recognition,” *EURASIP J. Appl. Signal Process*, no.11, pp.1274-1288, 2002.

[19] G. Potamianos, H. P. Graf, and E. Cosatto, “An Image Transform Approach for HMM based Automatic Lipreading,” *IEEE Conf. on ICIP’98*, pp. 173-177, 1998.

[20] P. Duchnowski, U. Meier, and A. Waibel, “See Me, Hear Me: Integrating Automatic Speech Recognition and Lip-Reading,” *ICSLP’94*, pp. 547-550, 1994.

[21] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland, “The HTK Book,” Cambridge University, 1996.