34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Biomedical Data Analysis by Supervised Manifold Learning

A. M. Álvarez-Meza[1], G. Daza-Santacoloma[1], and G. Castellanos-Domínguez[1]

Abstract— **Biomedical data analysis is usually carried out by assuming that the information structure embedded into the biomedical recordings is linear, but that statement actually does not corresponds to the real behavior of the extracted features. In order to improve the accuracy of an automatic system to diagnostic support, and to reduce the computational complexity of the employed classifiers, we propose a nonlinear dimensionality reduction methodology based on manifold learning with multiple kernel representations, which learns the underlying data structure of biomedical information. Moreover, our approach can be used as a tool that allows the specialist to do a visual analysis and interpretation about the studied variables describing the health condition. Obtained results show how our approach maps the original high dimensional features into an embedding space where simple and straightforward classification strategies achieve a suitable system performance.**

I. INTRODUCTION

The analysis of biomedical data is a challenge that mainly requires to discover the appropriated structure of the information embedded into the registers. Achieving a suitable analysis of the data allows to improve the performance of automatic systems to diagnostic support and simplify its implementation. Often, the embedded structure of the information is assumed as linear by many techniques, such as: principal component analysis (PCA), linear discriminant analysis (LDA), multidimensional scaling (MDS), etc. Nevertheless, the linearity assumption does not usually corresponds to the real behavior of the biomedical data. Indeed, the most common recordings describing the human health status (e.g. speech recordings, phonocardiogram, electrocardiograms, electroencephalograms, among others) are composed by several nonlinearly correlated variables lying in high dimensional spaces. In this regard, it is necessary to consider the analysis of the data by means of nonlinear dimensionality reduction (NLDR) techniques.

In this way, it is possible to represent in a low dimensional space (or embedding space) the high dimensional data, generally assuming that the input data are sampled from a smooth underlying manifold. The NLDR methods aim to obtain useful and compact representations of the information. Nonetheless, there are some limitations in the application of these kind of techniques when data lie in separated groups [4]. Above issue is faced during the design of automatic systems to diagnostic support. Indeed, most of the NLDR methods are not conceived to consider the class labels (e.g., control patients or pathological patients) of the data as extra information, which should enhance the low dimensional representation of the data, improving the system accuracy.

Several approaches relating supervised manifold learning have been presented. In [7] is reported a supervised Locally Linear Embedding (LLE) method, called α-LLE, which enforces the separation among classes according to a weighting parameter controlling the amount of class information that is incorporated. In [9], a local formulation of linear discriminant analysis is introduced. Next, in [6] the class labels are used to determine the neighbors of the training data so as to map overlapping high dimensional data into clusters in the embedded space. Mostly, these approaches either omit the preservation of the high dimensional local data structure in the embedding space (yielding overtraining), or they require the use of free parameters controlling the smoothing of the transformation to avoid the over fitting and to reduce the noise sensitivity. Furthermore, in [4] it is proposed a supervised version of LLE, termed C-LLE, which employs class labels as extra information to guide the procedure of dimensionality reduction allowing to figure out a suitable representation for each one of them. The amount of class label information incorporated in the embedding process is controlled by a tradeoff parameter, however, the upper bound of the tradeoff is not well defined, which can be problematic for an inexpert user and could increase the computational load of the algorithm.

Recently, some machine learning approaches have shown that using multiple kernels as similarity measure, instead of just one, can be useful to improve the feature extraction [5], [13]. In this sense, we propose a supervised NLDR method based on Laplacian Eigenmaps (LEM) algorithm [3] using multiple kernel representations (MKR). Our approach, which we named LEM-MKR, incorporates class label information while the local structure topology of the data is preserved during the mapping. The proposed LEM-MKR improves the accuracy of automatic systems to diagnostic support, reducing the computational complexity of the final classifiers, which could be useful to real-time implementations. Moreover, our approach also can be used as a tool that allows the specialist to do a visual analysis and interpretation about the studied variables of the health condition.

The remainder of this work is organized as follow. In section II the proposed LEM-MKR methodology is described. In section III the experimental conditions and results are presented. Finally, in sections IV and V we discuss and conclude about the attained results.

*Research carried out under the grants provided by "*Centro de Investigación e Innovación de Excelencia – ARTICA*" - COLCIENCIAS, a PhD. scholarship funded by Universidad Nacional de Colombia, and project 20201006594 funded by Universidad Nacional de Colombia and Universidad de Caldas

[1]All authors are with Signal Processing and Recognition Group, Universidad Nacional de Colombia, Campus La Nubia, km 7 via al Magdalena, Manizales-Colombia. {amalvarezme, gdazas, cgcastellanosd}@unal.edu.co

II. BACKGROUND

Laplacian Eigenmaps (LEM) is a NLDR technique based on preserving the intrinsic geometric structure of the manifold [3]. Let $\mathbf{X} \in \Re^{n \times p}$ the input data matrix with sample vectors $\mathbf{x}_i$ ($i = 1, \ldots, n$), LEM aims at providing a mapping to a low dimensional Euclidean space $\mathbf{Y} \in \Re^{n \times m}$, with row samples $\mathbf{y}_i$, being $m \ll p$. The algorithm apprises the following steps. First, an undirected weighted graph $G(V,E)$ is built; where V are the vertices and E the edges. In this case, there are n vertices, one for each $\mathbf{x}_i$. Nodes i and j are connected by $E_{ij} = 1$, if i is one of the k nearest neighbors of j (or viceversa), being measured by the Euclidean distance [3]. Second, a weight matrix $\mathbf{W} \in \Re^{n \times n}$ is calculated as $W_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$, if $E_{ij} = 1$, otherwise $W_{ij} = 0$, being $\kappa(\cdot,\cdot)$ a kernel function. After that, the $\mathbf{L} \in \Re^{n \times n}$ graph Laplacian matrix is given by $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D} \in \Re^{n \times n}$ is a diagonal matrix with $D_{ii} = \sum_j W_{ij}$. Finally, $\mathbf{Y}$ is calculated by minimizing

$$\sum_{ij} (\mathbf{y}_i - \mathbf{y}_j)^2 W_{ij}. \tag{1}$$

Note that above minimization implies a penalty if neighboring points $\mathbf{x}_i$ and $\mathbf{x}_j$ are mapped far apart. Equation (1) can be solved as the generalized eigenvalue problem $\mathbf{L}\mathbf{Y}_{:,s} = \lambda_s \mathbf{D}\mathbf{Y}_{:,s}$; where λ_s is the eigenvalue corresponding to the $\mathbf{Y}_{:,s}$ eigenvector, with $s = 1, \ldots, n$. First eigenvector is the unit vector with all equal components, while the remaining m eigenvectors form the embedded space. However, conventional NLDR approaches are not suitable for pattern recognition tasks, e.g., biomedical data analysis, because of the lacking of methods to incorporate label information in the mapping. A supervised NLDR mapping could enhance the data separability in further classification stages [4], [12].

Recently, some machine learning approaches have shown that using multiple kernels to infer the data similarity instead of just one (MKR), can be useful to improve the data interpretability [5], [13]. Thus, based on MKR we propose to incorporate the label information of the data into the LEM mapping. Given Z kernel functions, a combined kernel function can be computed as $\kappa_\xi(\mathbf{x}_i, \mathbf{x}_j) = \sum_{z=1}^{Z} \xi_z \kappa_z(\mathbf{x}_i, \mathbf{x}_i)$, subject to $\xi_z \geq 0$, and $\sum_{i=1}^{Z} \xi_z = 1$ ($\forall \xi_z \in \Re$). In this work, we propose to analyze the input data $\mathbf{X}$ considering both the local and class membership relationships among samples. Note that the optimization (1) is directly related to the weight matrix $\mathbf{W}$, which can be analyzed as a kernel matrix. Therefore, it is possible to employ different similarities measures in the LEM formulation by means of MKR, or as we called, a LEM-MKR approach. Hence, we consider two weight matrices, $\mathbf{W}_e$ and $\mathbf{W}_c$, in the LEM mapping process. The former is computed as

$$W_{e_{ij}} = \begin{cases} \kappa(\mathbf{x}_i, \mathbf{x}_j) & E_{ij} = 1 \\ 0 & E_{ij} = 0 \end{cases}, \tag{2}$$

considering traditional LEM assumptions. Now, let $\mathbf{c} \in \Re^{n \times 1}$ a class label vector with $c_i \in \{1, \ldots, C\}$, being C the number of classes in $\mathbf{X}$, $\mathbf{W}_c$ can be expressed as

$$W_{c_{ij}} = \delta(c_i - c_j) \left[\kappa(\mathbf{x}_i, \mathbf{x}_j) + \kappa(\mathbf{x}_i, \boldsymbol{\mu}_i) + \kappa\left(\mathbf{x}_j, \boldsymbol{\mu}_j\right) \right] \tag{3}$$

It is important to note that the function $\delta(c_i - c_j)$ in (3) penalizes the non-class memberships, where $\delta(c_i - c_j) = 1$, if and only if $c_i = c_j$, otherwise, it is equal to zero value. Besides, $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ are the average vectors of such class from $\mathbf{x}_i$ and $\mathbf{x}_j$ belong. The terms $\kappa(\mathbf{x}_i, \boldsymbol{\mu}_i)$ and $\kappa\left(\mathbf{x}_j, \boldsymbol{\mu}_j\right)$ aim to reveal the similarity of each sample with the average intra-manifold structure of each class. Moreover, $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ considers the similarity between samples of the same class. Inspired by MKR, equations (2) and (3) can be used to computed the weight matrix $\mathbf{W}_T$ as

$$\mathbf{W}_T = \xi_e \mathbf{W}_e + \xi_c \mathbf{W}_c, \tag{4}$$

subject to $\xi_e + \xi_c = 1$. From equation (4), the combined Laplacian matrix is calculated as $\mathbf{L}_T = \mathbf{D}_T - \mathbf{W}_T$, where $\mathbf{D}_T \in \Re^{n \times n}$ is a diagonal matrix with $D_{T_{ii}} = \sum_j W_{T_{ji}}$. Therefore, a new NLDR objective function can be written as $\sum_{ij} (\mathbf{y}_i - \mathbf{y}_j)^2 W_{T_{ij}}$, which can be solved as a generalized eigenvalue problem, fixing ξ_e and ξ_c.

The ξ_e and ξ_c parameters in (4) give a tradeoff between the local appearance and the class label similarities retained in $\mathbf{Y}$. If $\xi_e = 0$ ($\xi_c = 1$), we have the original mapping of LEM. As ξ_c increases, then ξ_e decreases due to the constraint $\xi_e + \xi_c = 1$. Therefore, for a given pair of points (ξ_e, ξ_c), we can infer both, the local and the class label representation errors as $\varepsilon_e = \sum_{ij} (\mathbf{y}_i - \mathbf{y}_j)^2 W_{e_{ij}}$, and $\varepsilon_c = \sum_{ij} (\mathbf{y}_i - \mathbf{y}_j)^2 W_{c_{ij}}$, respectively. Looking for the simultaneous minimization of both errors, we employ the parametric plot ε_e versus ε_c, as a tool to study the behavior of these quantities, with all them normalized between 0 and 1. Based on the L-curve criteria for Tikhonov regularization, the point with maximum curvature results to be a good choice for ξ_e and ξ_c. Note that ε_e determines if the underlying data structure is not well preserved in $\mathbf{Y}$, while ε_c establishes the quality of the separability among classes.

On the other hand, even when NLDR algorithms provide an embedding for a fixed dataset, it is necessary to generalize their results to new locations in the input space. Therefore, given the embedding space $\mathbf{Y}$, a new sample $\mathbf{x}_{\text{new}}$ can be mapped by the minimization of $\|\mathbf{x}_{\text{new}} - \sum_{r=1}^{k} v_r \boldsymbol{\eta}_r\|^2$, where $\sum_{r=1}^{k} v_r = 1$, being $\boldsymbol{\eta}_r$ one of the k nearest neighbors of $\mathbf{x}_{\text{new}}$ in $\mathbf{X}$ (see [14] for details). In Fig. 1 a comparison between traditional PCA projection and LEM-MKR is presented for a synthetic nonlinear structured data [4].

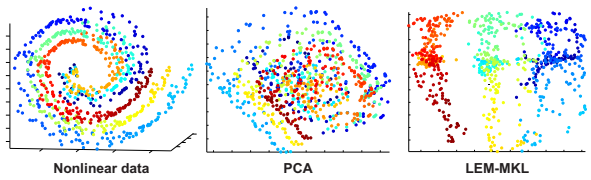


Fig. 1. Triple swiss-roll ($n = 1000$, $p = 3$, $C = 3$, $m = 2$). Traditional PCA projection does not unfold the underlying structure of the input data, and it overlaps samples of different classes. LEM-MKR projection conserves the local structure of each swiss-roll while separates, as well as possible, samples from different classes.

III. EXPERIMENTS

We aim to demonstrate the advantages of our approach as a tool to outperform diagnosis support systems for automatic detection of diseases. A general scheme of such kind of system can be summarized as in Fig. 2. First, a preprocessing stage is used to prepare the raw data for further analysis. Then, some features are estimated according to the studied phenomenon. In most of the cases, it is difficult to directly interpret the obtained information, due to the complexity and the large amount of obtained features. Before a classifier can be applied with a reasonable hope of generalization, a small number of useful features will have to be extracted.

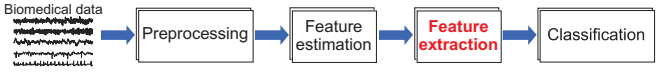


Fig. 2. Diagnosis support system scheme for automatic detection of diseases.

LEM-MKR is tested as feature extraction stage, using a linear kernel to find the local and the class membership relationships among samples in (2) and (3). We compare the performance of LEM-MKR against some linear and nonlinear unsupervised feature extraction methods: PCA, LLE, LEM [3], [14], and against the supervised NLDR techniques: α-LLE and C-LLE [4], [12]. The dimension of the embedding space m is fixed looking for a 95% of expected local variability [7]. The k value for the NLDR algorithms is chosen according to [10], in which an specific number of neighbors for each sample is computed. The parameter α in α-LLE is selected from the set $\{0, 0.2, \ldots, 1\}$ according to the training error. Two straightforward classifiers are tested: linear discriminant classifier (**ldc**), and k-nearest neighbors classifier (**knnc**). A 10 folds cross-validation scheme is employed to determine the performance of the system. The number of neighbors for **knnc** is optimized with respect to the leave-one-out error of the training set.

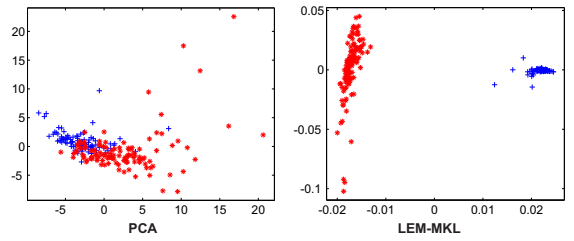
Four biomedical databases are tested. The former contains voice records of children with and without Cleft Lift and Palate (CLP). The main goal is to detect and characterize the presence of hypernasality in the speech registers. This dataset is provided by *Signal Processing and Recognition Group-SPRG* of the Universidad Nacional de Colombia, Manizales. The database holds 266 voice records from children between 5 and 15 years old, who uttered the Spanish vowels. There are 110 children labeled by a phoniatry expert as healthy, and 156 labeled as hypernasal. We employ the preprocessing and feature estimation stages of the voice records according to [11], leading 147 acoustic features for each sample.

The second database is the phonocardiographic database (PCG), also provided by SPRG, which is composed by 35 adult subjects (16 normals and 19 with murmur). Eight recordings were taken from each patient, corresponding to the four traditional focuses of auscultation (mitral, tricuspid, aortic and pulmonary areas) in the phase of post-expiratory and post-inspiratory apnea. The signals were digitized at 44.1kHz with 16-*bits* per sample. Furthermore, in order to select beats without artifacts and another type of noise that can degrade the performance of the algorithms, a visual and audible inspection was carried out by cardiologists. Thus, 548 individual beats were extracted, 274 for each class, using an R-peak detector. The recordings are characterized by means of a time-frequency representation, particularly, the spectrogram obtained by a Short-Time Fourier Transform (38 frequencies and 480 instants of time, see [2] for details).

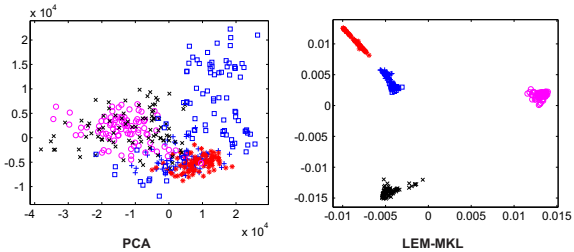
The third database is the Oxford Parkinson's disease detection (PARKINSON), which is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). The data target is to discriminate healthy people from those with PD. The dataset contains 195 voice recording, which are preprocessed and characterized as in [8], obtaining 23 acoustic features for each sample.

Finally, an EPILEPSY database is tested, which contains EEG signals of 29 patients with medically intractable focal epilepsies. They were recorded by the Department of Epileptology of the University of Bonn, by means of intracranially implanted electrodes [1]. The database comprises five sets (denoted as Z,O,N,F,S) composed of 100 single channel EEG segments, which were selected and extracted after visual inspection from continuous multichannel EEG to avoid artifacts (e.g. muscular activity or eye movements). All EEG signals were recorded with an acquisition system of 128 channels, using average common reference. Data was digitized at 173.61 Hz, and time–frequency and time–varying decomposition methods are used as feature extraction stage.

Table I presents the dataset properties and the classification accuracy for all the studied feature extraction methods. Moreover, the PCA and LEM-MKR 2D projections are shown for CLP and EPILEPSY (Figures 3(a) and 3(b)).



(a) CLP



(b) EPILEPSY

Fig. 3. Datasets visualization using PCA and LEM-MKR.

IV. DISCUSSION

From Figures 3(a) and 3(b), it is possible to observe that the low dimensional space found by PCA overlaps the

TABLE I

CLASSIFICATION RESULTS (AVERAGE ACCURACY ± STANDARD DEVIATION FOR 10-FOLD CROSS VALIDATION)

Dataset	Classifier	Without DR	PCA	LLE	LEM	α - LLE	C - LLE	LEM - MKR
CLP	**ldc**	75.63±06.49	83.66±08.64	83.55±08.66	84.87±04.89	63.41±15.70	**91.16±03.71**	**92.83±04.91**
$n = 238,\ p = 147,\ C = 2,\ m = 3$	**knnc**	87.39±06.62	81.97±07.30	85.27±07.74	82.77±05.82	66.38±09.31	**91.63±04.81**	**90.33±05.06**
PCG	**ldc**	Diverges	89.59±02.88	84.51±05.00	82.85±05.76	**94.35±04.87**	92.69±03.98	92.13±04.28
$n = 548,\ p = 18240,\ C = 2,\ m = 15$	**knnc**	Diverges	**96.53±01.61**	88.15±04.27	83.38±03.34	92.90±04.12	93.99±03.63	93.59±04.36
PARKINSONS	**ldc**	77.44±06.48	75.89±04.27	77.84±06.31	74.35±03.29	**89.76±05.31**	86.73±04.72	**88.29±05.90**
$n = 197,\ p = 23,\ C = 2,\ m = 4$	**knnc**	87.04±07.01	84.11±06.88	78.89±08.73	73.52±09.63	**88.18±05.97**	86.23±04.60	85.74±07.13
EPILEPSY	**ldc**	20.00±00.00	67.20±04.24	63.60±05.87	66.00±06.04	83.60±04.88	85.60±04.40	**90.80±05.18**
$n = 500,\ p = 2052,\ C = 5,\ m = 5$	**knnc**	**94.00±02.83**	74.20±05.53	79.60±04.70	72.60±06.11	83.00±09.81	90.20±03.19	92.60±04.90

observations, because it does not consider neither the local relationships nor the class membership similarities among samples. Hence, PCA is not able to unfold the underlying data structure, and its embedding is not suitable to separate different classes. Otherwise, LEM-MKR (our approach) preserves the local geometry of the original space, keeping away samples of different classes. The above statement can be explained by the tradeoff between the local similarity and the class membership matrices, which allows to unfold the main structure of the data in a space with lower dimension m than the original input space dimension p (see Table I).

In regard to the attained classification results presented in Table I, our approach presents, in most of the cases, a suitable classification accuracy, for both kind of classifiers **ldc** and **knnc**. Therefore, LEM-MKR allows to identify the nonlinear structure of the input data, finding an embedding space where a classifier with a simple decision boundary (i.e., **ldc**) can be used. Indeed, it is important to emphasize that the proposed tradeoff selection avoids the need of a manual tuning, finding a tradeoff that compensates both the intrinsic geometry conservation and the separability margin.

Moreover, it can be noticed how traditional PCA, and unsupervised NLDR algorithms, LLE and LEM, do not attain reliable classification performances. On the other hand, the α-LLE algorithm, overall, seems to obtain a high classification performance, when the technique is used in conjunction with a nonlinear boundary classifier. Nonetheless, as the number of classes grows or when the underlying data structure is more complex, the classification accuracy strongly diminishes, because of the induced overtraining. Finally, C-LLE methodology seems to be a good alternative for classifications tasks. However, the lower performance in some databases in comparison with our proposal (i.e., PARKINSONS and EPILEPSY), can be explained by the lack of the tuning of the required free parameter, which is not well defined in the original formulation [4].

V. CONCLUSIONS

A feature extraction methodology was proposed to learn the underlying data structure of biomedical information. For such purpose, a manifold learning framework based on LEM is enhanced by MKR, in order to learn both the local data structure and the class label relationships among samples. In addition, we established an scheme for automatic selection of the required free parameters, based on a tradeoff between the contribution of the local and the class membership relationships among observations. Proposed LEM-MKR minimizes, as well as possible, the local structure and the class separability representation errors in the embedding space. Our approach was tested as feature extraction stage to develop diagnosis support systems for automatic detection of diseases from biomedical data. Attained results showed how our methodology unfold the data main structure, mapping the original high dimensional space (original estimated features) to an embedding space where simple and straightforward classification strategies can be used to obtain a suitable system performance. Moreover, our approach is also a tool to visually analyze the studied phenomenon by the specialist. As future work, it should be interesting to test our methodology using other kind of similarity/dissimilarity measures, according to the application of interest.

REFERENCES

[1] R. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. Elger, "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state," *Phys. Rev. E*, vol. 64, pp. 71–86, 2001.

[2] L. D. Avendano, G. C. Domínguez, and J. I. G. Llorente, "Feature extraction from parametric time frequency representations for heart murmur detection," in *EMBC*, 2010.

[3] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[4] G. Daza, G. Castellanos, and J. C. Principe, "Locally linear embedding based on correntropy measure for visualization and classification," *Neurocomputing*, vol. 80, no. 0, pp. 19 – 30, 2012.

[5] M. Gonen and E. Alpaydin, "Localized multiple kernel regression," in *ICPR*, 2010.

[6] Q. Jiang, M. Jia, J. Hu, and F. Xu, "Machinery fault diagnosis using supervised manifold learning," *Mechanical Systems and Signal Processing*, vol. 23, pp. 2301–2311, 2009.

[7] O. Kouropteva, O. Okun, and M. Pietikäinen, "Supervised locally linear embedding algorithm for pattern recognition," in *IbPRIA*, 2003.

[8] M. Little, P. McSharry, S. Roberts, D. Costello, and I. Moro, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *BioMedical Engineering*, vol. 6, pp. 1–19, 2007.

[9] M. Loog and D. de Ridder, "Local discriminant analysis," in *ICPR*, 2006.

[10] A. Álvarez, J. Valencia, G. Daza, and G. Castellanos, "Global and local choice of the number of nearest neighbors in locally linear embedding," *Pattern Recognition Letters*, vol. 32, no. 16, pp. 2171 – 2177, 2011.

[11] J. Orozco, S. Murillo, A. Álvarez, J. Arias, E. Delgado, J. Vargas, and G. Castellanos, "Automatic selection of acoustic and non-linear dynamic features in voice," in *INTERSPEECH*, 2011.

[12] M. Pillati and C. Viroli, "Supervised locally linear embedding for classification: An application to gene expression data analysis," in *CLADAG*, 2005.

[13] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[14] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.