Biomedical Signal Processing and Control xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Dimensionality reduction for visualization of normal and pathological speech data

J. Goddard [a,*], G. Schlotthauer [b], M.E. Torres [b,c,**], H.L. Rufiner [b,c]

[a] *Departamento de Ingeniería Eléctrica, Universidad Autónoma Metropolitana, Iztapalapa, México*
[b] *Facultad de Ingeniería, Universidad Nacional de Entre Ríos, CONICET, Oro Verde (E.R.), Argentina*
[c] *Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional de Litoral, CONICET, Santa Fe, Argentina*

ARTICLE INFO

Article history:
Received 13 June 2008
Received in revised form 18 November 2008
Accepted 14 January 2009
Available online xxx

Keywords:
Data visualization
Dimensionality reduction
Kernel methods
Pathological voice analysis

ABSTRACT

For an adequate analysis of pathological speech signals, a sizeable number of parameters is required, such as those related to jitter, shimmer and noise content. Often this kind of high-dimensional signal representation is difficult to understand, even for expert voice therapists and physicians. Data visualization of a high-dimensional dataset can provide a useful first step in its exploratory data analysis, facilitating an understanding about its underlying structure. In the present paper, eight dimensionality reduction techniques, both classical and recent, are compared on speech data containing normal and pathological speech. A qualitative analysis of their dimensionality reduction capabilities is presented. The transformed data are also quantitatively evaluated, using classifiers, and it is found that it may be advantageous to perform the classification process on the transformed data, rather than on the original. These qualitative and quantitative analyses allow us to conclude that a nonlinear, supervised method, called kernel local Fisher discriminant analysis is superior for dimensionality reduction in the actual context.

1. Introduction

Several feature extraction techniques have been proposed for pathological voice analysis and classification [1–5]. Most of them use measures that characterize different aspects of the voice signal, such as frequency perturbations and noise content. In these cases, a vector representation of the data is often chosen whose size impedes the data's visualization. Moreover, the 'curse' of the dimensionality is a well-known problem arising in classical pattern recognition. This situation is exacerbated when the data available are limited in quantity. Data visualization techniques can help alleviate this situation.

Data visualization can provide a practical tool in exploratory data analysis, which may enable us to gain a useful first insight into the information hidden in high-dimensional data. Techniques which transform a high-dimensional space into a space of fewer dimensions – often one, two or three – are collectively known as dimensionality reduction techniques. Dimensionality reduction techniques are not clustering tools, however, they can provide clues about clusters within the data, or they may be performed before applying a clustering or classification algorithm. Furthermore, the intrinsic structure of the data, such as its intrinsic dimensionality, may well be revealed after mapping the data to a lower dimensional space.

The most commonly used classical method for dimensionality reduction is perhaps principal component analysis (PCA), also known as the Karhunen–Loève transform, or singular value decomposition [6]. PCA performs an unsupervised linear mapping of the data to a lower dimensional space in such a way, that the variance of the data in the low-dimensional representation is maximized. A disadvantage of PCA is that the embedded subspace has to be linear. For example, if the data are located on a circle in a three-dimensional Euclidean space, $\mathbb{R}^3$, PCA will not be able to identify this structure. Another disadvantage is that PCA depends critically on the units in which the features are measured.

Sammon's mapping [7] is a classical nonlinear method, that performs a mapping such that the interpoint distances of the data are approximately preserved in the lower dimensional space.

In recent years, a number of other unsupervised visualization techniques have become available, and their application to data sets, such as those involving speech, is just being conducted [8–10]. Among these methods, those of kernel PCA (KPCA) [11], Gaussian process latent variable model (GP-LVM) [12] and local linear embedding (LLE) [13] are particularly relevant for our purposes in the present paper.

[*] Corresponding author.
[**] Corresponding author at: Facultad de Ingeniería, Universidad Nacional de Entre Ríos, CONICET, Oro Verde (E.R.), Argentina.
E-mail addresses: jgc@xanum.uam.mx (J. Goddard), metorres@santafe-conicet.gov.ar (M.E. Torres).

1746-8094/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bspc.2009.01.001

KPCA is a (usually) nonlinear extension of PCA using kernel methods. Kernel methods have been successfully applied in the fields of pattern analysis and pattern recognition [14], often providing better classification performance than other methods, and frequently playing a vital part in the nonlinear extension of classical algorithms. GP-LVM is a probabilistic, nonlinear, latent variable model that generalizes PCA. LLE, on the other hand, provides low-dimensional, neighborhood-preserving embeddings. This means that points which are 'close' to one another in a data space will also be close when mapped onto the low-dimensional space.

Often class information about the data in a given problem is known, and in this case, supervised dimensionality reduction techniques can be used. Fisher discriminant analysis (FDA) is one such classical supervised linear technique. However it tends to give undesired results if the samples in a class are multimodal. To overcome this drawback Sugiyama proposed the Local Fisher discriminant analysis (LFDA). This method maximizes between class separability and preserves within class local structure at the same time. FDA and LFDA are both linear methods, and the latter may be extended to kernel local Fisher discriminant analysis (KLFDA), a nonlinear dimensionality reduction technique, obtained by applying the kernel trick [15].

In this paper we extend a previous work [16]. Our aim here is to compare and discuss supervised and unsupervised dimensionality reduction techniques applied to normal and pathological speech data, which allow their visualization in lower dimensions. In Section 2, we present the real-world speech dataset employed and briefly describe the techniques which are applied to it. In Section 3, results are presented and discussed through qualitative and quantitative analysis. The conclusions are presented in Section 4.

2. Data and methods

In Section 2.1, we give a description of the speech data considered in the paper and the preprocessing applied to it in order to obtain a suitable vector representation. This is followed, in Section 2.2, by a brief review of the dimensionality reduction methods which will be compared here.

2.1. Normal and pathological data

The data used in this paper consisted of real voice samples of the sustained vowel /a/ for both normal and pathological voices. Among the pathological voice samples, we considered two voice disorders: dysphonia and paralysis. The voice samples were taken from the 'Disordered Voice Database', from the Massachusetts Eye and Ear Infirmary Voice and Speech Laboratory [17] and distributed by Kay Elemetrics [18]. The clinical information includes diagnostic information along with patient identification, age, sex, smoking status, and more.

In total there were 34 patients with dysphonic speech disorders, 61 with paralysis and a further 53 normal speakers. For each subject, a 14-features vector $\vec{x}$ was associated. These features were chosen using similar criteria to those in [2,3], namely: number and degree of voice breaks, fraction of locally unvoiced frames, three variables related to jitter (local, relative average perturbation and five-point period perturbation quotient), five related to shimmer (two local estimations, three-point amplitude perturbation, five-point amplitude perturbation and eleven-point amplitude perturbation) and three measures related to noise content. In previous work, this set of features provided the best performance in several classification tasks [2,3]. Moreover, a similar vector of features has been reported as the one providing the best results in [19], where different sets of characteristics and classification methods were compared. In what follows, for completeness, we include their definitions (cf. [2,3] for more details).

2.1.1. Features

(1) Number of voice breaks. Any period of time, between two consecutive pitch detections, longer than 1.25 over a selected pitch floor (usually 75 Hz), is considered as a voice break.
(2) Degree of voice breaks. It is the total duration of the breaks in the voiced parts of the signal, divided by its total length. Silences at the beginning and at the end of the signal are not considered breaks.
(3) Fraction of locally unvoiced frames. This is the fraction of frames analyzed as unvoiced.
(4) Jitter or period perturbation quotient.
  (a) Jitter ratio (local) or jitt is defined as:

$$jitt = 1000\frac{(1/(n-1))\sum_{i=1}^{n-1} P_i - P_{i+1}}{(1/n)\sum_{i=1}^{n} P_i}, \tag{1}$$

  where P_i is the period of the ith cycle, in ms, and n is the number of periods in the sample.
  (b) Relative average perturbation (RAP) is defined as:

$$RAP = \frac{(1/(n-2))\sum_{i=2}^{n-1}|((P_{i-1}+P_i+P_{i+1})/3)-P_i|}{(1/n)\sum_{i=1}^{n} P_i}, \tag{2}$$

  where P_i and n are as above.
  (c) Five-point period perturbation quotient (ppq5) is defined as:

$$ppq5 = \frac{(1/(n-4))\sum_{i=3}^{n-2}|((\sum_{j=-2}^{2} P_{i+j})/3)-P_i|}{(1/n)\sum_{i=1}^{n} P_i}, \tag{3}$$

  where P_i and n are as above.
(5) Shimmer or amplitude perturbation quotient.
  (a) Local shimmer (shimm) is defined as:

$$shimm = \frac{(1/(n-1))\sum_{i=1}^{n-1}|A_i - A_{i+1}|}{(1/n)\sum_{i=1}^{n} A_i}, \tag{4}$$

  where A_i is the amplitude of the ith cycle and n is the number of periods in the sample.
  (b) Local shimmer (dB). It is defined as the previous one, but expressed in dB.
  (c) Three-point amplitude perturbation quotient (apq3) is defined as:

$$apq3 = \frac{(1/(n-2))\sum_{i=2}^{n-1}|((A_{i-1}+A_i+A_{i+1})/3)-A_i|}{(1/n)\sum_{i=1}^{n} A_i}, \tag{5}$$

  where A_i and n are as above.
  (d) Five-point amplitude perturbation quotient (apq5) is defined as:

$$apq5 = \frac{(1/(n-4))\sum_{i=3}^{n-2}|((\sum_{j=-2}^{2} A_{i+j})/3)-A_i|}{(1/n)\sum_{i=1}^{n} A_i}, \tag{6}$$

  where A_i and n are as above.
  (e) Eleven-point amplitude perturbation quotient (apq11) is defined as:

$$apq11 = \frac{(1/(n-10))\sum_{i=6}^{n-5}|((\sum_{j=-5}^{5} A_{i+j})/11)-A_i|}{(1/n)\sum_{i=1}^{n} A_i}, \tag{7}$$

  where A_i and n are as above.
(6) Noise-content parameters. These measures quantify the amount of glottal noise in the vowel waveform. In contrast to perturbation measures, they attempt to resolve the vowel waveform into signal and noise components, computing their

Please cite this article in press as: J. Goddard, et al., Dimensionality reduction for visualization of normal and pathological speech data, Biomed. Signal Process. Control (2009), doi:10.1016/j.bspc.2009.01.001

energies ratio, considering only the voiced parts of the signal. The three selected parameters are mean autocorrelation, mean harmonics-to-noise ratio (in dB) and mean noise-to-harmonics ratio [20].

2.2. Dimensionality reduction methods

Here we present a brief review of the dimensionality reduction methods which will be compared in the paper. The features vector will be denoted by $\vec{x} \in \mathbb{R}^P$ and the dimensionally reduced vector by $\vec{y} \in \mathbb{R}^Q$, with $Q < P = 14$. $\mathcal{X} = \{\vec{x}_1, \ldots, \vec{x}_N\}$ denotes the data set of features vectors used for each dimensionality reduction technique. The matrix with N-columns given by the features vectors in $\mathcal{X}$ will be notated as $\mathbf{X}$.

2.2.1. Principal component analysis

PCA is a dimensionality reduction method that attempts to efficiently represent the data by finding orthonormal axes which maximally decorrelate the data. The data are then projected onto these orthogonal axes. The principal components are precisely this set of Q orthonormal vectors, where Q is often chosen to be 2 or 3.

There are several equivalent ways to find the principal components, one being that of finding the Q eigenvectors $\vec{v}$ of the covariance matrix $\mathbf{C}$ of the data set, corresponding to the Q largest eigenvalues λ. If $\mathcal{X}$ is a zero mean data set in the Euclidean space $\mathbb{R}^P$, then the covariance matrix is given by:

$$\mathbf{C} = \frac{1}{n}\sum_{j=1}^{N} \vec{x}_j \vec{x}_j^T, \tag{8}$$

where $\vec{x}_j^T$ indicates the vector transpose and the corresponding eigenvalue equation is $\mathbf{C}\vec{v} = \lambda\vec{v}$.

PCA provides a linear mapping of the data onto the lower Q-dimensional space, but suffers from several problems previously mentioned in the introduction. In order to define a nonlinear extension of PCA, KPCA has been introduced.

2.2.2. Kernel principal component analysis

KPCA uses the notion of a kernel to modify the corresponding PCA algorithm. Generally, a positive semidefinite kernel k on $\mathcal{X} \times \mathcal{X}$ is defined as a real-valued function:

$$k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R},$$

such that:

(1) k is symmetric: $k(\vec{x}_i, \vec{x}_j) = k(\vec{x}_j, \vec{x}_i) \quad \forall\ i, j = 1, \ldots, N$.
(2) k is positive semidefinite, i.e.

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j k(\vec{x}_i, \vec{x}_j) \geq 0, \tag{9}$$

holds for $\forall\ \ \alpha_i,\ \ \alpha_j \in \mathbb{R}$ and for $n = 1, \ldots, N$.

It can be shown that given a kernel k, there exists a (Reproducing Kernel) Hilbert space $\mathcal{H}$ and a transformation $\phi : \mathcal{X} \rightarrow \mathcal{H}$ such that

$$k(\vec{x}_i, \vec{x}_j) = \langle \phi(\vec{x}_i), \phi(\vec{x}_j) \rangle, \tag{10}$$

holds. $\mathcal{H}$ is often referred to as the feature space and can be infinite-dimensional.

The most commonly used kernels are the polynomial and radial basis function kernels defined on $\mathbb{R}^P \times \mathbb{R}^P$ by:

$$k(\vec{x}_i, \vec{x}_j) = (\langle \vec{x}_i, \vec{x}_j \rangle + 1)^d, \tag{11}$$

$$k(\vec{x}_i, \vec{x}_j) = \exp\left(\frac{-\|\vec{x}_i - \vec{x}_j\|^2}{2\,\sigma^2}\right), \tag{12}$$

respectively, where $d = 1, 2, \ldots$ and $\sigma \in \mathbb{R}^+$. For these kernels the transformation ϕ is not explicitly defined, and the kernels are directly applied in the original data space. This is known as the 'kernel trick'.

For the kernels of Eqs. (11) and (12), it can be shown that KPCA is conceptually the same as performing standard PCA with the data set $\{\phi(\vec{x}_1), \ldots, \phi(\vec{x}_N)\}$ in the feature space $\mathcal{H}$ (with the above notation). Fortunately, the kernel trick, referred to above, can also be applied in this case, and the explicit use of ϕ avoided. Instead, the $N \times N$ kernel matrix $\mathbf{K} = \{k_{ij}\}$, with $k_{ij} = k(\vec{x}_i, \vec{x}_j)$, is introduced, and the equation:

$$\mathbf{K}\vec{v} = N\lambda\vec{v}, \tag{13}$$

is solved for $\lambda \in \mathbb{R}$ and $\vec{v} = (v_1, \ldots, v_N)^T \in \mathbb{R}^N$.

A projection of a new pattern $\vec{x}$ in data space onto the q-th principal component in feature space can be found using:

$$y_{\vec{x}}^q = \sum_{i=1}^{N} v_i^q\, k(\vec{x}, \vec{x}_i). \tag{14}$$

Observe that $y_{\vec{x}}^q$ represents the q-th component of the dimensionally reduced vector $\vec{y}$ associated with $\vec{x}$. In order to use KPCA, we have to choose a kernel function and, as in the case of PCA, decide on the number of dimensions on which to project.

2.2.3. Sammon's mapping

Sammon's mapping [7] is a classical method for producing a nonlinear mapping of a set of P-dimensional vectors to a lower-dimensional space, usually of 2 or 3 dimensions. The mapping is constructed such that the interpoint distances of the vectors are approximately preserved by the corresponding interpoint distances of their images in the lower-dimensional space; as such, Sammon's mapping is a form of multidimensional scaling. We denote by d_{ij} the distance between the vectors $\vec{x}_i$ and $\vec{x}_j$ (usually taken as the Euclidean distance measure). We wish to find N Q-dimensional vectors $\vec{y}_i$, $i = 1, \ldots, N$, with $Q = 2$ or 3, such that the error function E (often referred to as Sammon's stress) is minimized, where:

$$E(\mathbf{X}, \mathbf{Y}) = \frac{1}{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}} \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \frac{(d_{ij} - \delta_{ij})^2}{d_{ij}}, \tag{15}$$

and δ_{ij} denotes the Euclidean distance between the vectors $\vec{y}_i$ and $\vec{y}_j$. Sammon proposed a steepest descent procedure to search for a minimum of the error function, although it is computationally inefficient for large data sets. Another, well-known, disadvantage of the mapping is that it has to be recalculated when new points are added.

2.2.4. Gaussian process latent variable model

GP-LVM is a recent addition to dimensionality reduction techniques proposed by Lawrence [12,21]. GP-LVM is a probabilistic, nonlinear, latent variable model that generalizes PCA. GP-LVM implicitly learns a Gaussian process mapping between a low-dimensional latent space, $\mathcal{Y}$, and the usually high-dimensional data space, $\mathcal{X}$, in such a way that the points which are 'close' in latent space are mapped to points which are also 'close' in data space. GP-LVM can discover low dimensional manifolds in high-dimensional data with small data sets [21]. We may briefly describe the technique in terms of the following optimization problem [21]: Let $\mathbf{X}$ be the $N \times P$ matrix in which each row is a single training datum. Let $\mathbf{Y}$ denote the $N \times Q$ matrix whose rows represent the corresponding positions in latent space. Given a covariance function for the Gaussian process, $k(\vec{y}_i, \vec{y}_j)$, the likelihood of the data, given the latent positions, is

$$p(\mathbf{X}|\mathbf{Y},\vec{\beta}') = \frac{1}{(2\pi)^{PN/2}|\mathbf{K}_{\vec{\beta}'}|^{P/2}} \exp\left(-\frac{1}{2}tr(\mathbf{K}_{\vec{\beta}'}^{-1}\mathbf{XX}^T)\right), \qquad (16)$$

where $\mathbf{K}_{\vec{\beta}'}$ is known as the kernel matrix and $\vec{\beta}'$ denotes the kernel hyperparameters. The kernel matrix $\mathbf{K}_{\vec{\beta}'} = \{k_{ij}\}$ is defined by the covariance function, where $k_{ij} = k(\vec{y}_i, \vec{y}_j)$.

Learning in the GP-LVM consists of maximizing Eq. (16) with respect to $\mathbf{Y}$ and the $\vec{\beta}'$ components. Whilst this also tends to be computationally inefficient for large data sets, Lawrence proposes a practical algorithm for finding $\mathbf{Y}$ with GP-LVM. It should be noted that the $\mathbf{Y}$ found will not be unique.

2.2.5. Local linear embedding

LLE is an unsupervised learning algorithm that computes low-dimensional, neighborhood-preserving embeddings of high-dimensional inputs. LLE is performed in three steps. First, for each point in the data, it's κ nearest neighbors in the data are found (usually using Euclidean distance). Then, each point is approximated by convex combinations of it's κ nearest neighbors, to obtain a matrix of reconstruction weights, $\mathbf{W}$. Finally, low-dimensional embeddings (usually in a space of one or two-dimensions) are found such that the local convex representations are preserved. More precisely, for each vector $\vec{x}_i \in \mathcal{X}$ let $\mathcal{N}_i$ denote the set of indices of it's κ nearest neighbors. In order to find the reconstruction weights, $\mathbf{W} = \{w_{ij}\}$, the objective function:

$$E(\mathbf{W}) = \sum_i \left|\vec{x}_i - \sum_{j\in\mathcal{N}_i} w_{ij}\vec{x}_j\right|^2, \qquad (17)$$

has to be minimized, subject to $\sum_{j\in\mathcal{N}_i} w_{ij} = 1$.

The embeddings, $\{\vec{y}_1, \ldots, \vec{y}_N\}$, of the original data are obtained by minimizing the following objective function:

$$O(\mathcal{Y}) = \sum_i \left|\vec{y}_i - \sum_{j\in\mathcal{N}_i} w_{ij}\vec{y}_j\right|^2. \qquad (18)$$

An advantage of LLE is that it has few free parameters to set and a non-iterative solution, thus avoiding convergence to a local minimum.

Interesting relationships have recently been found between KPCA and LLE, as well as other well-known dimensionality reduction techniques cf. [22].

2.2.6. Fisher discriminant analysis

FDA is a linear, supervised dimensionality reduction method applied to a dataset, whose elements each have a class label assigned to them.

Let $\mathcal{C} = \{1, 2, \ldots, C\}$ denote the set of possible classes, and for each $\vec{x}_i \in \mathcal{X}$, let $c_i \in \mathcal{C}$ be its associated class label. Let N_c be the number of elements of $\mathcal{X}$ belonging to class c.

The *within-class scatter matrix* $\mathbf{S}^{(w)} \in \mathbb{R}^{N\times N}$ and the *between-class scatter matrix* $S^{(b)} \in \mathbb{R}^{N\times N}$ are defined by [15]:

$$\begin{aligned}\mathbf{S}^{(w)} &= \sum_{c=1}^{C}\sum_{i:c_i=c}(\vec{x}_i - \vec{\mu}_c)(\vec{x}_i - \vec{\mu}_c)^T,\\ \mathbf{S}^{(b)} &= \sum_{c=1}^{C} N_c(\vec{\mu}_c - \vec{\mu})(\vec{\mu}_c - \vec{\mu})^T,\end{aligned} \qquad (19)$$

where $\sum_{i:c_i=c}$ denotes the summation over i such that $c_i = c$, $\vec{\mu}_c$ is the mean of the samples in class c and $\vec{\mu}$ is the mean of all the samples. The FDA transformation matrix $\mathbf{T}_{FDA}$ is defined as follows:

$$\mathbf{T}_{FDA} \equiv \underset{\mathbf{T}\in\mathbb{R}^{P\times Q}}{argmax}[\mathrm{tr}((\mathbf{T}^T\mathbf{S}^{(w)}\mathbf{T})^{-1}\mathbf{T}^T\mathbf{S}^{(b)}\mathbf{T})]. \qquad (20)$$

In this way FDA seeks a transformation matrix $\mathbf{T}$ such that $\mathbf{S}^{(b)}$ is maximized whilst $\mathbf{S}^{(w)}$ is minimized. This problem is solved by means of a generalized eigenvalue problem.

2.2.7. Local Fisher discriminant analysis

FDA performs poorly if the samples in a class form several separate clusters (i.e. for multimodal classes). This undesired behavior is caused by the non-local nature of scatter matrices. So, we can reformulate FDA in a pairwise manner in order to include local information [15]. Define the *local* within-class scatter matrix

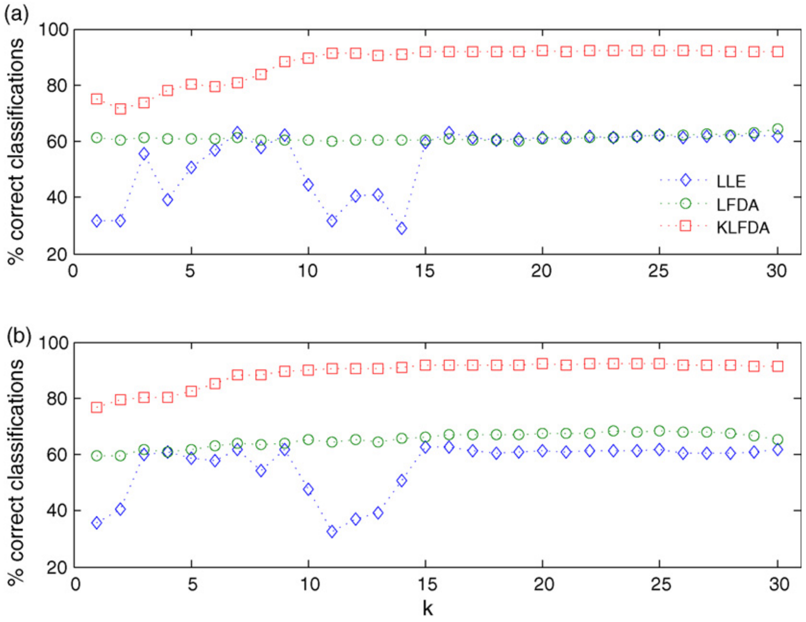

Fig. 1. Selection of the number of neighbors, κ. Average percentage of correct classifications obtained using 1000 realizations of a LDA classifier on the transformed data for $Q = 2$ (a) and $Q = 3$ (b).

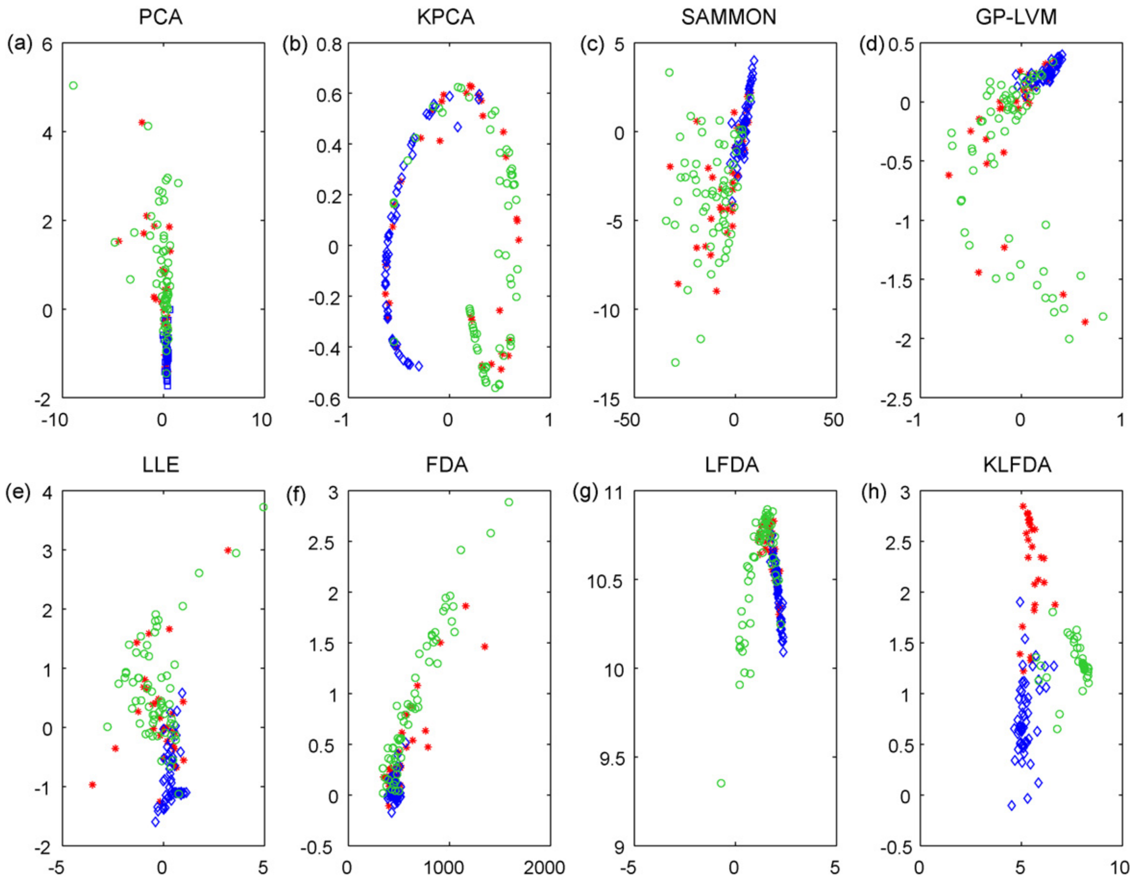

Fig. 2. Reduction to dimension $Q = 2$. (a) PCA, (b) KPCA, (c) Sammon, (d) GP-LVM, (e) LLE, (f) FDA, (g) LFDA and (h) KLFDA applied to the voice data: normal (diamonds), dysphonia (stars) and paralysis (circles).

$\tilde{\mathbf{S}}^{(w)}$ and the *local* between-class scatter matrix $\tilde{\mathbf{S}}^{(b)}$ by:

$$\tilde{\mathbf{S}}^{(w)} = \frac{1}{2}\sum_{i,j=1}^{N}\tilde{w}_{ij}^{(w)}(\vec{x}_i-\vec{x}_j)(\vec{x}_i-\vec{x}_j)^T,$$
$$\tilde{\mathbf{S}}^{(b)} = \frac{1}{2}\sum_{i,j=1}^{N}\tilde{w}_{ij}^{(b)}(\vec{x}_i-\vec{x}_j)(\vec{x}_i-\vec{x}_j)^T, \qquad (21)$$

where

$$\tilde{w}_{ij}^{(w)} \equiv \begin{cases} a_{ij}/N_c & \text{if } c_i = c_j = c, \\ 0 & \text{if } c_i \neq c_j, \end{cases}$$
$$\tilde{w}_{ij}^{(b)} \equiv \begin{cases} a_{ij}(1/N - 1/N_c) & \text{if } c_i = c_j = c, \\ 1/N & \text{if } c_i \neq c_j, \end{cases}$$

and $\mathbf{A}$ is an *affinity matrix*, that is an $N \times N$ matrix whose element a_{ij} corresponds to the affinity between $\vec{x}_i$ and $\vec{x}_j$. We assume that $a_{ij} \in [0, 1]$, with a_{ij} large if $\vec{x}_i$ and $\vec{x}_j$ are 'close' and a_{ij} is small if $\vec{x}_i$ and $\vec{x}_j$ are 'far apart'. This means that, sample pairs in the same class which are far apart, have less influence on $\tilde{\mathbf{S}}^{(w)}$ and $\tilde{\mathbf{S}}^{(b)}$. In this work the value of each a_{ij} was taken as:

$$a_{ij} = \exp\left(\frac{-\|\vec{x}_i - \vec{x}_j\|^2}{\sigma^2}\right). \qquad (22)$$

Then we obtain the LFDA transformation matrix $\mathbf{T}_{LFDA}$ in an analogous way to that of $\mathbf{T}_{FDA}$, but replacing $\mathbf{S}^{(w)}$ and $\mathbf{S}^{(w)}$ with $\tilde{\mathbf{S}}^{(w)}$ and $\tilde{\mathbf{S}}^{(b)}$, respectively.

2.2.8. *Kernel local Fisher discriminant analysis*

As previously mentioned, LFDA can be extended to nonlinear dimensionality reduction. By using the kernel trick, a regularized version of the LFDA generalized eigenvalue problem can be expressed as [15]:

$$\mathbf{K}\tilde{\mathbf{L}}^{(b)}\mathbf{K}\tilde{\vec{v}} = \tilde{\lambda}(\mathbf{K}\tilde{\mathbf{L}}^{(w)}\mathbf{K} + \epsilon\mathbf{I}_N)\tilde{\vec{v}}, \qquad (23)$$

where $\mathbf{K}$ is the kernel matrix with elements $k_{ij} = k(\vec{x}_i, \vec{x}_j)$, $\tilde{\mathbf{L}}^{(b)}$ and $\tilde{\mathbf{L}}^{(w)}$ are the so called graph-Laplacian matrices of dimension $N \times N$, ϵ is a small constant and $\mathbf{I}_N$ is the identity matrix. For further details see Appendix C in [15].

Let $\{\tilde{\vec{v}}_k\}_{k=1}^{N}$ be the generalized eigenvectors associated with the generalized eigenvalues $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \ldots \geq \tilde{\lambda}_N$ of Eq.(23). Then the embedded image of $\phi(\vec{x})$ in $\mathcal{H}$ is given by:

$$\left(\sqrt{\tilde{\lambda}_1}\vec{v}_1 | \sqrt{\tilde{\lambda}_2}\vec{v}_2 | \ldots | \sqrt{\tilde{\lambda}_q}\vec{v}_q\right)^T \begin{pmatrix} k(\vec{x}_1, \vec{x}) \\ k(\vec{x}_2, \vec{x}) \\ \vdots \\ k(\vec{x}_N, \vec{x}) \end{pmatrix}, \qquad (24)$$

where $(\vec{\xi}_1 | \cdots | \vec{\xi}_N)$ stands for a matrix with column vectors $\vec{\xi}_i \in \mathbb{R}^N$. This kernelized variant of LFDA is called KLFDA.

3. Results and discussion

The eight methods described in the previous section were applied to the voice samples introduced in Section 2.1, transforming them, in different experiments, to easily visualized vectors in Euclidean spaces with two and three dimensions. All

Please cite this article in press as: J. Goddard, et al., Dimensionality reduction for visualization of normal and pathological speech data, Biomed. Signal Process. Control (2009), doi:10.1016/j.bspc.2009.01.001

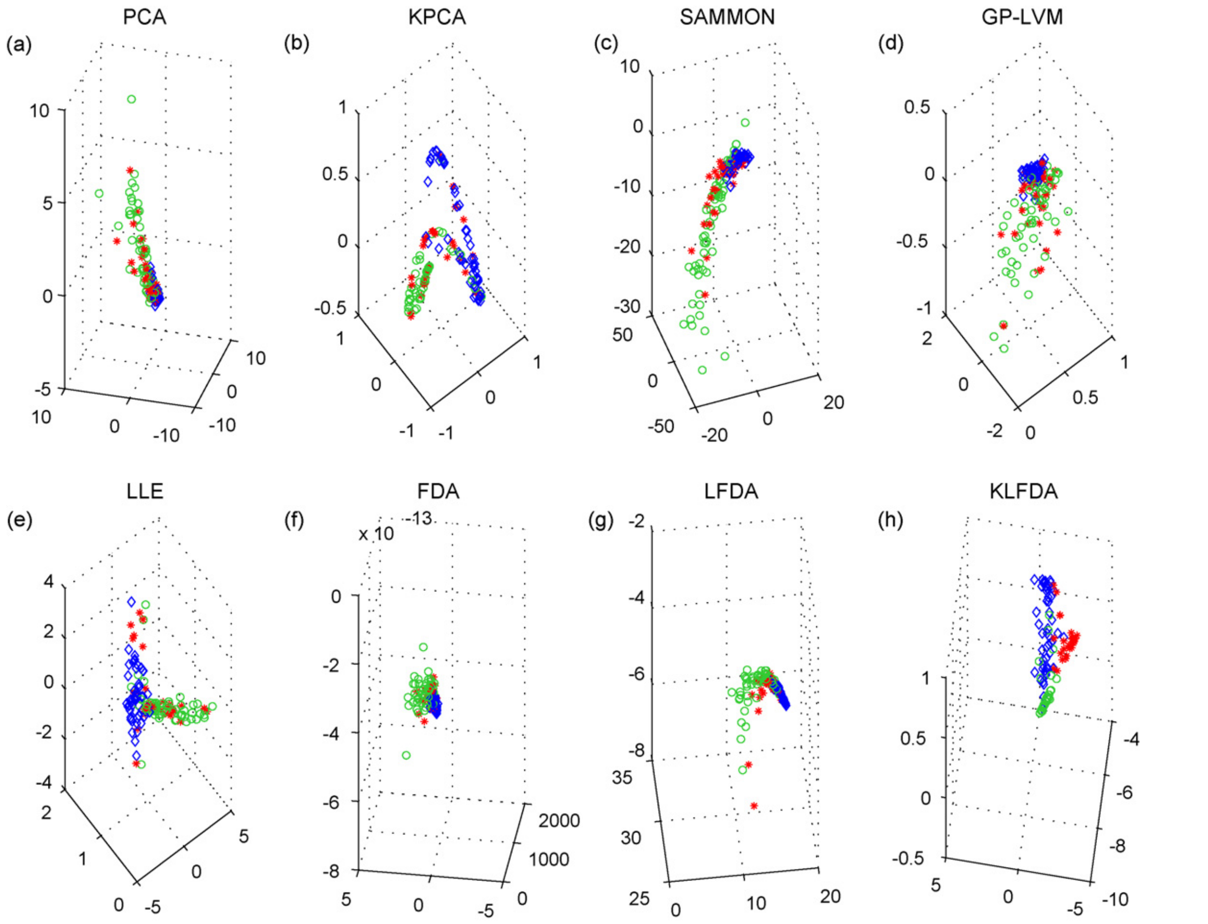

Fig. 3. Reduction to dimension $Q = 3$. (a) PCA, (b) KPCA, (c) Sammon, (d) GP-LVM, (e) LLE, (f) FDA, (g) LFDA and (h) KLFDA applied to the voice data: normal (diamonds), dysphonia (stars) and paralysis (circles).

the experiments were conducted using Matlab and publicly available implementations of KPCA,[1] GP-LVM,[2] LLE,[3] LFDA and KLFDA.[4] In KPCA and KLFDA different kernels were tested, and we present the best results, which were obtained using a radial basis function kernel given in Eq. (12). Also, given its better performance in preliminary tests, a multi-layer perceptron was selected for the kernel in GP-LVM. For each method the parameters have been selected *ad hoc* as the ones that provide the best discrimination.

In order to select the number of neighbors, κ, to be used in the dimensionality reduction, we applied multiple linear discriminant analysis (LDA) [23] on the mapped data, with $Q = 2$ and $Q = 3$, respectively, for a number of κ neighbors which varied from 1 to 30. In this way we fitted a multivariate normal density function to each group, with a pooled estimate of covariance. The transformed data were randomly split 1000 times into training and test sets, with sizes of 67 and 33% of the full dataset, respectively. Fig. 1 shows the average percentage of correct classifications obtained for each test set. For both values of Q we can see that the LLE method is unstable up to $\kappa = 15$. Moreover, for $\kappa > 15$, no noticeable effect on the performance of these methods is observed. A similar analysis was conducted for selecting the parameter σ in Eq. (12) for KPCA, yielding $\sigma = 4.6$ for $Q = 2$ and $\sigma = 2.5$ for $Q = 3$. For KLFDA the same kernel was chosen with $\sigma = 1.949$ for both values of Q. For LLE, LFDA and KLFDA the number of chosen neighbors was 7, 33 and 23, respectively.

Fig. 2 shows the data mapped into a two-dimensional Euclidean space by each of the eight dimensionality reduction techniques. In the case of PCA for two dimensions, the amount of variance explained was 97.69%. However, as we can appreciate in Fig. 2(a), PCA did not provide an appropriate visual discrimination between the three different voice groups. Also, whilst KPCA (Fig. 2(b)), LLE (Fig. 2(e)), and LFDA (Fig. 2(g)) suggest a separation between normal and pathological voices, these methods do not seem to distinguish between the two voice disorders. Further, neither Sammon (Fig. 2(c)), GP-LVM (Fig. 2(b)) nor FDA (Fig. 2(f)) provide a clear clustering of the data. Finally, we can appreciate in Fig. 2(h) that KLFDA is the only one that provides a visual discrimination of the data into the three classes. This result is consistent with Fig. 1(a).

Though one might expect better results from the supervised methods (FDA, LFDA and KLFDA), due to the additional information available, this is not always the case. It is important to observe that even if KPCA and LLE are unsupervised methods, their transformed data can visually distinguish between normal (diamonds) and pathological (stars and circles) data. This is not the case for the supervised method of FDA (cf. Fig. 2(d)), probably because the distributions of the classes are multimodal.

In Fig. 3, we display the data mapped into a three-dimensional Euclidean space using the same dimensionality reduction techniques as in the previous figure. We can appreciate that no further information can be extracted from visual inspection. In the case of PCA (Fig. 3(a)), even though the explained variance was 98.78%, the

[1] http://www.cs.unimaas.nl/l.vandermaaten.
[2] http://www.dcs.shef.ac.uk/~neil/fgplvm.
[3] http://www.cs.toronto.edu/~roweis/lle/code.html.
[4] http://sugiyama-www.cs.titech.ac.jp/~sugi/software/LFDA.

Please cite this article in press as: J. Goddard, et al., Dimensionality reduction for visualization of normal and pathological speech data, Biomed. Signal Process. Control (2009), doi:10.1016/j.bspc.2009.01.001

Table 1
Classification with 3-NN. The results of Tukey's multiple comparison test based on the non-parametric Kruskal–Wallis test. Original 14-dimensional vector *vs*. data transformed to two (2d) and three (3d) dimensions. The averaged percentages of correct classifications are presented in the second column. In the third column, those methods not significantly different at a significance level $\alpha = 0.05$, are shown.

	Correct. Class. (%)	Not S.D. from
Original	63.46	GP-LVM 2d, LFDA 2d, PCA 3d, Sam 3d, LFDA 3d
PCA 2d	59.38	LLE 2d, KPCA 3d
KPCA 2d	60.91	Sam 2d, LLE 2d
Sam 2d	61.84	KPCA 2d, GP-LVM 2d, PCA 3d, LFDA 3d
GP-LVM 2d	62.46	Original, Sam 2d, GP-LVM 2d, PCA 3d, Sam 3d, LFDA 3d
LLE 2d	60.17	PCA 2d, KPCA 2d
FDA 2d	47.65	FDA 3d
LFDA 2d	64.39	Original, Sam 3d
KLFDA 2d	***88.39***	KLFDA 3d
PCA 3d	62.83	Original, Sam 2d, GP-LVM 2d, Sam 3d, LFDA 3d
KPCA 3d	58.38	PCA 2d, LLE 3d
Sam 3d	63.36	Original, GP-LVM 2d, LFDA 2d, PCA 3d, LFDA 3d
GP-LVM 3d	66.04	–
LLE 3d	57.57	KPCA 3d
FDA 3d	47.65	FDA 2d
LFDA 3d	62.91	Original, Sam 2d, GP-LVM 2d, PCA 3d, Sam 3d
KLFDA 3d	88.36	KLFDA 2d

method cannot differentiate between the voice groups. It is worth observing that the image of the data under KPCA appears to approximate a one-dimensional curve for $Q = 2$ (Fig. 3(b)), and it appears to correspond to the folding of the one in Fig. 3(b). As in the two-dimensional case, KLFDA is the method that provides the best visual discrimination of the three classes.

In order to obtain a quantitative comparison between the eight methods, two simple classifiers were applied to the transformed data: κ-nearest neighbor (κ-NN) with $\kappa = 3$ and LDA. This value for κ was chosen after performing the test, varying its value from 1 to 15. The data were randomly split into training (67%) and test (33%) sets as above. The mean classification results for the test sets are presented in Tables 1 and 2 for κ-NN and LDA, respectively. In each Table, the first row corresponds to the classification of the original 14-dimensional data; in the first column we present the method's name, followed by '2d' or '3d', depending on the reduced data dimension. Finally, in the second column, the averaged percentage of correct classifications – over the 1000 randomly split test sets – is displayed for each reduction technique. The highest results are highlighted in bold. The last column indicates the methods that were not significantly different from the method in that row; significance was decided using Tukey's multiple comparison test based on the non-parametric Kruskal–Wallis test (KWT) [24], at a significance level of $\alpha = 0.05$. KWT has been selected, given that it does not assume a normal distribution of the population, which is in agreement with the obtained results.

Table 2
LDA classification. The results of Tukey's multiple comparison test based on the non-parametric Kruskal–Wallis test. Original 14-dimensional vector *vs*. data transformed to two (2d) and three (3d) dimensions. The averaged percentages of correct classifications are presented in the second column. In the third column, those methods not significantly different at a significance level $\alpha = 0.05$, are shown.

	Correct. Class. (%)	Not S.D. from
Original	64.16	PCA 2d, LLE 2d
PCA 2d	64.89	Original, LFDA 3d
KPCA 2d	69.23	KPCA 3d
Sam 2d	61.96	GP-LVM 2d, LLE 2d, PCA 3d, Sam 3d, LLE 3d
GP-LVM 2d	61.17	Sam 2d, Sam 3d, LLE 3d
LLE 2d	63.06	Original, Sam 2d, PCA 3d, Sam 3d, LLE 3d
FDA 2d	67.46	LFDA 2d, GP-LVM 3d
LFDA 2d	66.77	FDA 2d, GP-LVM 3d, LFDA 3d
KLFDA 2d	***92.26***	KLFDA 3d
PCA 3d	62.81	Sam 2d, Sam 3d, LLE 2d, LLE 3d
KPCA 3d	68.68	KPCA 2d
Sam 3d	62.03	Sam2d, GP-LVM 2d, LLE 2d, PCA 3d, LLE 3d
GP-LVM 3d	66.99	FDA 2d, LFDA 2d, LFDA 3d
LLE 3d	62.33	Sam 2d, GP-LVM 2d, LLE 2d, PCA 3d, Sam 3d, GP-LVM 3d
FDA 3d	47.65	–
LFDA 3d	65.96	LFDA 2d, GP-LVM 3d
KLFDA 3d	92.22	KLFDA 2d

As can be observed in the second column of both tables, the best performance was provided by KLFDA (in Table 1, KLFDA-2d: 88.39% and KLFDA-3d: 88.36%; in Table 2: KLFDA-2d: 92.26%, KLFDA-3d: 92.22%), with no significant differences between the dimensions (2d and 3d). This is in agreement with the visual, qualitative results, already discussed in previous Figures.

In the case of κ-NN classification (Table 1) we can appreciate that LFDA 2d, KLFDA 2d, GP-LVM 3d and KLFDA 3d provided higher correct classifications than using the original 14-dimensional feature vectors. However, only GP-LVM 3d and KLFDA (2d and 3d) are significantly different from the other methods. We can therefore conclude that they have better capabilities for the discrimination tasks considered here, in agreement with the visualization of speech data previously shown in Figs. 2 and 3.

On the other hand, in the case of LDA classification, KPCA (2d and 3d), GP-LVM 3d, FDA 2d, LFDA (2d and 3d) and KLFDA (2d and 3d) have a higher percentage of correct classifications than that for the original feature vectors. It is important to observe that PCA 2d and LLE 2d are not significantly different from the classification obtained using the original feature vectors.

After comparing the results obtained using the unsupervised techniques, we can observe that for both of the classification methods, GP-LVM 3d improves on the classification in the original space. Nevertheless, the superiority of KLFDA 2d and 3d is evident. Again, all these quantitative results confirm the intuition gained by visual inspection of the previous figures.

4. Conclusions

In the present paper, we have compared the visualization capabilities of eight dimensionality reduction techniques, when they are applied to pathological and normal speech data. The pathological data included two different speech disorders: paralysis and dysphonia.

Supervised and unsupervised methods have been considered: PCA, KPCA, Sammon, GP-LVM, LLE, FDA, LFDA and KLFDA. The data were mapped by these methods to both two- and three-dimensional Euclidean spaces. Qualitative and quantitative analyses were conducted. Whilst dimensionality reduction was performed using all the data, classification was applied to 1000 randomly chosen training and test sets.

The qualitative analyses of the visual results indicate that discrimination using two or three dimensions is similar. However, almost all the methods could only distinguish two groups. An exception was provided by the KLFDA method, which gave a clear visual separation of the three classes: normal, paralysis and dysphonia.

A quantitative study was also performed using two different classifiers. The results indicate that the KLFDA reduction method also provides the best percentage of correct classification, in both

Please cite this article in press as: J. Goddard, et al., Dimensionality reduction for visualization of normal and pathological speech data, Biomed. Signal Process. Control (2009), doi:10.1016/j.bspc.2009.01.001

cases. In fact, there is a difference of over 28% between these latter results and those obtained with the original 14 dimensional data.

In the case of a patient with paralysis or a dysphonic voice, this type of low dimensional data visualization could help a voice therapist or physician to better understand a patient's condition by considering the relative localization of the patients voice samples in a bidimensional space. Furthermore, the patients' evolution during a given therapy could also be visually monitored.

Acknowledgements

This work was carried out with the financial support of UNER, UNL (Project CAI+D No 012-72) and CONICET (Argentina), ANPCyT, as well as CONACYT (Mexico) and SECYT (Argentina) under Project ME/PA03-EXI/031, and PROMEP-SEP (Mexico).

References

[1] R.J. Baken, R.F. Orlikoff, Clinical Measurement of Speech and Voice, 2nd ed., Singular Thompson Learning, 2000.

[2] G. Schlotthauer, M.E. Torres, M.C. Jackson-Menaldi, Automatic classification of dysphonic voices, WSEAS Trans. Signal Process. 2 (9) (2006) 1260–1267.

[3] G. Schlotthauer, M.E. Torres, M.C. Jackson-Menaldi, A pattern recognition approach to spasmodic dysphonia and muscle tension dysphonia automatic classification, J. Voice, in press.

[4] N. Sáenz-Lechón, J.I. Godino-Llorente, V. Osma-Ruiz, P. Gómez-Vilda, Methodological issues in the development of automatic systems for voice pathology detection, Biomed. Signal Process. Control 1 (2006) 120–128.

[5] M.E. Torres, L.G. Gamero, H.L. Rufiner, C. Martínez, D.H. Milone, G. Schlotthauer, Study of complexity in normal and pathological speech signals, in: Proc. of the 25th Annu. Int. Conf. of the IEEE Eng. in Med. and Biol. Soc., 2003, 2339–2342.

[6] I.T. Jolliffe, Principal Component Analysis, volume XXIX of Springer Series in Statistics, 2nd ed., Springer, 2002.

[7] J.W. Sammon, A Non-Linear Mapping for Data Structure Analysis, IEEE Trans. Comput. C-18(5) (1969) 401–409.

[8] M.A. Carreira-Perpinan, Continuous latent variable models for dimensionality reduction and sequential data reconstruction, PhD thesis, University of Sheffield, UK, 2001.

[9] V. Jain, L.K. Saul, Exploratory analysis and visualization of speech and music by locally linear embedding, in: Proc. Int. Conf. Acoust. Speech Signal Process, vol. 3, Canada, (2004), pp. 984–987.

[10] A. Kocsor, L. Tóth, Kernel-based feature extraction with a speech technology application, IEEE Trans. Signal Process. 52 (8) (2004) 2250–2263.

[11] B. Schölkopf, A. Smola, K.-R. Müller, Kernel Principal Component Analysis, Chapter Advances in Kernel Methods—Support Vector Learning, MIT Press, Cambridge, MA, 1998, pp. 327–352.

[12] N.D. Lawrence, Probabilistic non-linear principal component analysis with Gaussian process latent variable models, J. Mach. Learn. Res. 6 (2005) 1783–1816.

[13] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.

[14] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge U.P., 2004.

[15] M. Sugiyama, Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis, J. Mach. Learn. Res. 8 (2007) 1027–1061.

[16] J. Goddard, F. Martínez, G. Schlotthauer, M.E. Torres, H.L. Rufiner, Visualization of normal and pathological speech data. In: C. Manfredi (Ed.), Proc. of Models and Anal. of Vocal Emissions for Biomed. Appl.: 5th Int. Workshop, Italy, pp. 32–36, 2007.

[17] Massachusetts Eye, Voice Ear Infirmary, and Speech Lab. Disorder voice database version 1.03, 1994.

[18] Kay Elemetrics Corporation, Disordered Voice Database Model 4337, MEEIVS Lab. Boston, MA, 1994.

[19] A. Gelzinis, A. Verikas, M. Bacauskiene, Automated speech analysis applied to laryngeal disease categorization, Comput. Methods Prog. Biomed. 91 (2008) 36–47.

[20] P. Boersma, Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, in: Institute of Phonetic Sciences, University of Amsterdam, Proc., vol. 17, 1993, 97–110.

[21] N.D. Lawrence, Gaussian process models for visualisation of high dimensional data, in: Advances in Neural Information Processing Systems (NIPS), MIT Press, Cambridge, MA, 2004.

[22] J. Ham, D.D. Lee, S. Mika, B. Schölkopf, A kernel view of the dimensionality reduction of manifolds, in: R. Greiner, D. Schuurmans (Eds.), Proc. of the 21st. Int. Conf. on Mach. Learn., 2004, 369–376.

[23] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification. Stat. for Eng. and Inf. Sci., 2nd ed., Wiley Interscience, 2001.

[24] W.H. Kruskal, W.A. Wallis, Use of ranks in one-criterion variance analysis, J. Am. Stat. Assoc. 47 (1952) 583–621.

Please cite this article in press as: J. Goddard, et al., Dimensionality reduction for visualization of normal and pathological speech data, Biomed. Signal Process. Control (2009), doi:10.1016/j.bspc.2009.01.001