In: Speech Communication 26(4), pp. 259-282 (December 1998)
URL: http://www.dcs.shef.ac.uk/~miguel/papers/dim_red_epg.html

Dimensionality Reduction of Electropalatographic Data Using Latent Variable Models

Miguel Á. Carreira-Perpiñán Steve Renals
Dept. of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
{M.Carreira,S.Renals}@dcs.shef.ac.uk

June 24, 1998

Abstract

We consider the problem of obtaining a reduced dimension representation of electropalatographic (EPG) data. An unsupervised learning approach based on latent variable modelling is adopted, in which an underlying lower dimension representation is inferred directly from the data. Several latent variable models are investigated, including factor analysis and the generative topographic mapping (GTM). Experiments were carried out using a subset of the EUR-ACCOR database, and the results indicate that these automatic methods capture important, adaptive structure in the EPG data. Nonlinear latent variable modelling clearly outperforms the investigated linear models in terms of log-likelihood and reconstruction error and suggests a substantially smaller intrinsic dimensionality for the EPG data than that claimed by previous studies. A two-dimensional representation is produced with applications to speech therapy, language learning and articulatory dynamics.

Keywords: electropalatography (EPG), articulatory modelling, data reduction methods, dimensionality reduction, latent variable models, finite mixture distributions, mixture models, principal component analysis (PCA), factor analysis, mixtures of factor analysers, generalised topographic mapping (GTM), mixtures of multivariate Bernoulli distributions.

1 Introduction

The technique of electropalatography (EPG) is well established as a relatively noninvasive, conceptually simple and easy-to-use tool for the investigation of lingual activity in both normal and pathological speech. Qualitative and quantitative data about patterns of lingual contacts with the hard palate during continuous speech may be obtained using EPG, and the technique has been used in studies of descriptive phonetics, coarticulation, and diagnosis and treatment of disordered speech (Hardcastle et al., 1991a; Hardcastle et al., 1989). Typically, the subject wears an artificial palate moulded to fit the upper palate with a number of electrodes mounted on the surface to detect lingual contact (62 in the Reading EPG system (Hardcastle et al., 1991a)). The EPG signal is sampled at 100–200 Hz. Thus, for a given utterance, the sequence of raw EPG data consists of a stream of binary vectors with both spatial and temporal structure.

Potentially the tongue has many degrees of freedom owing to its lack of skeleton (Stone, 1991); however, a number of studies suggest that tongue movements in speech may be appropriately modelled using a few elementary articulatory parameters (e.g. (Nguyen et al., 1994; Nguyen et al., 1996) and references therein). Spatial redundancy arises due to the limited number of possible tongue configurations—only a tiny fraction of the potential $2^{62} \approx 4.6 \cdot 10^{18}$ EPG patterns can be produced by the articulatory system. Temporal redundancy arises due to the correlations between neighbouring frames: the position of the tongue (and thereby the EPG signal) changes slowly with time, compared to the acoustic signal. If the appropriate constraints were known, then it would be possible to represent the EPG signal using fewer dimensions.

In this paper we approach the problem of finding such a dimensionality reduction mapping, not from a physical point of view, but from a machine learning one. That is, we consider a data set consisting of a number of EPG patterns and try to learn a mapping into a low-dimensional space from the data, without any a priori assumption about it (other than a specific, but quite general, form) and using methods well-known in the statistics and pattern recognition fields. Our approach will be that of latent variable modelling, in which the data is assumed to have been generated stochastically from a small number of hidden variables. In common with most work on EPG data reduction we will concentrate on dimensionality reduction at the spatial level only, thus ignoring the constraints arising from coarticulation and other dynamics of the articulatory system.

The primary advantage of dimensionality reduction for EPG data is that it makes a sequence of EPG data more amenable to analysis, since direct manipulation of the raw EPG sequence is cumbersome due to its high dimensionality (Hardcastle et al., 1991b). A reduced-dimension representation in terms of a few variables can be inspected, plotted against time or against each other, etc. more easily. Additional advantages are the consequent reductions in storage space and data transmission rate—albeit of little relevance, since the EPG data has relatively low requirements for them.

The resultant dimensionality reduction may suggest a value for the intrinsic dimensionality of the EPG data, i.e., the number of degrees of freedom of the tongue-palate system in what concerns the generation of EPG patterns. It may also give some insight into the problem of the acoustic-to-articulatory mapping (Schroeter and Sondhi, 1994), in which values for articulatory parameters such as vocal tract area functions, lip positions and jaw dynamics are determined from the acoustic signal. Note that the EPG signal alone is an incomplete articulatory description, omitting such details as nasalisation and vocalisation. Hence, the mapping from phonemes to EPG patterns is not one-to-one since certain phonemes (e.g. /æ/ and /ɑ/) can produce the same EPG patterns.

The rest of the paper is organised as follows. Section 2 explains the advantages that unsupervised learning methods (in particular latent variable models) for EPG data have over fixed data reduction methods. Section 3 introduces the latent variable modelling framework. Section 4 describes the data set employed for the experiments. These experiments are described in sections 5, 6, 7 and 8. The paper is completed with a discussion of the results in section 9 and a conclusion in section 10.

2 Data reduction methods

Several techniques to extract features or other kind of condensed information from the EPG data have been developed (Hardcastle et al., 1991b; Jones and Hardcastle, 1995). These include the *totals* display, which monitors dynamic changes in different contact regions of the palate (e.g. in fig. 1, alveolar: rows 1–2; palatal: 3–5; velar: 6–8) as a function of time; the *centre of gravity* index, designed to locate the highest concentration of activated electrodes in a given EPG frame; or the *frequency of contact* with each electrode over a period of time, expressed as a percentage and usually calculated in a row-by-row basis.

The majority of these techniques are based on linear combinations of the EPG vector components. We refer to such linear combinations as *linear indices* and represent them using a D-dimensional basis vector $\mathbf{v}$. The score of a D-dimensional EPG pattern $\mathbf{t}$ with respect to index $\mathbf{v}$ is thus the projection of $\mathbf{t}$ on $\mathbf{v}$, i.e., $\mathbf{v}^T\mathbf{t} = \sum_{d=1}^{D} v_d t_d$. Observe that multiplying an index by a constant does not add any new information, so that $\mathbf{v}$, $2\mathbf{v}$ and $-2\mathbf{v}$ are all equivalent.

Many of these numerical indices have a fixed form, prescribed in advance. This form often reflects a phonetician's a priori beliefs about the relevance of different palate regions; for example, most work on EPG data reduction uses predetermined articulatory regions on the artificial palate (Byrd et al., 1995; Hardcastle et al., 1991b) via a linear combination of the individual contacts (see fig. 1). Alternatively, they may have been designed in some other way. For example, (Nguyen, 1995) applies the discrete cosine transform (DCT) to the EPG frames considered as images to obtain several linear indices, corresponding to low spatial frequencies. Even though this makes no a priori assumption about the data, it still has the disadvantage that this set of indices is fixed across data sets (because the indices are the basis vectors of the DCT, which for a given image size is fixed).

Ad hoc methods are not robust and will not perform well in situations where the speech deviates from the standard: impaired speakers, different speech styles or unusual accents. This makes desirable the use of automatic methods to extract structure from a data set without requiring any prior knowledge about it, such as factor analysis (which looks for linear correlations), neural networks (which can find more complex relationships), etc. Such techniques could also be useful in determining relevant articulatory regions empirically; this would allow for interspeaker differences in the details of electrode placement with respect to anatomical configurations or speaker-specific articulatory patterns (Byrd et al., 1995), rather than using fixed electrode locations for all speakers, as is customary. Thus, one of the goals of this paper is to find EPG cues in an empirical way, without referring to predetermined indices based on phonetic knowledge (such as those illustrated in fig. 1).

Adaptive dimensionality reduction for EPG data has been investigated using both linear systems and neural networks (Nguyen et al., 1996; Holst et al., 1995). Linear approaches have been based on a rotated principal components analysis (referred to as factor analysis in (Nguyen et al., 1996)). Two layer feed-forward neural network autoassociators used by (Nguyen et al., 1996) and (Holst et al., 1995) infer a set of basis vectors that spans a subspace similar[1] to that defined by the principal components (Bourlard and Kamp, 1988; Baldi and Hornik, 1989). These adaptive methods lack a natural interpretation as probability model (although one may be forced upon PCA (Tipping and Bishop, 1997)). In this paper we are concerned with methods that explicitly define a full probability model.

Graphically, we can represent both EPG patterns and indices as in figures 1 and 3: each small rectangle corresponds to one component of the D-dimensional vector (starting in the top left corner and ending in the

[1] If the neural network units have linear activation functions, then both spaces are exactly the same.

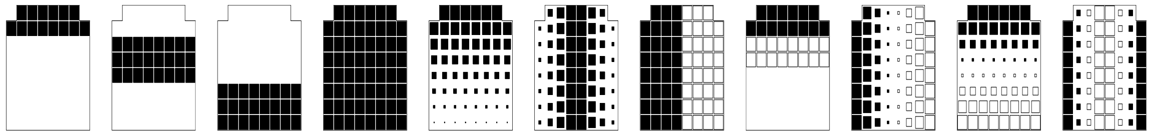

Figure 1: A selection of typical EPG data reduction indices reported in the literature (basis vectors of an orthogonal, linear projection): alveolar (ALV), palatal (PAL), velar (VEL), total of contacts (TOT), centre of gravity (COG), laterality (LAT), asymmetry (ASY), fricative (FRIC), left-right asymmetry (LRA), alveolar-palatal (APA) and lateral-median (LME).

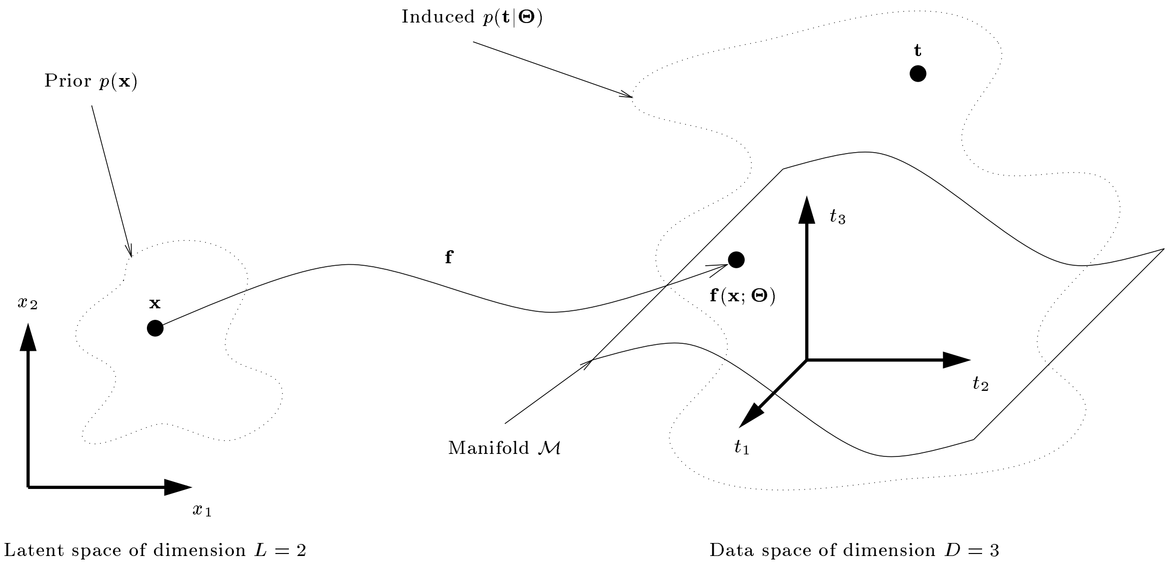

Figure 2: Schematic of a latent variable model with a 3-dimensional data space and a 2-dimensional latent space.

bottom right one), where $D = 62$. The colour of the rectangle, black or white, indicates the positive or negative sign of the component, and the size of the rectangle is proportional to the magnitude of the component. The shape of the figure (an 8×8 grid where the top corners are unused) follows that of the palate, the alveolar part being in the top and the velar part in the bottom. Note that in the conventional pictorial representation of EPG patterns, only binary data are allowed: each position is either empty or contains a full, black rectangle. Thus, we extend this representation to EPG indices by allowing any real value in each position. However, the reader should bear in mind that in an EPG pattern all components are either 0 or 1, and in an EPG index they can be any real number (positive or negative). Thus, fig. 1 depicts EPG indices while fig. 3 depicts EPG patterns.

Concerning phonemic categories, it should be clear that, for practical reasons, no general labelling of the data was performed, as explained in section 4. However, we use the labels of fig. 3 (identified perceptually by a phonetician) with the specific purpose of assessing our results in sections 5 to 8.

3 Generative modelling using latent variables

In latent variable modelling the assumption is that the observed high-dimensional data is generated from an underlying low-dimensional process. The high dimensionality arises for several reasons, including stochastic variation and the measurement process. The objective is to learn the low dimensional generating process (defined by a small number of *latent variables* or hidden causes) along with a noise model, rather than directly learning a dimensionality reducing mapping.

We consider here the case where the latent variables are mapped by a fixed transformation into a higher-dimension observed space (measurement procedure) and noise is added there (stochastic variation). Such models are commonly known as latent variable models and have been used in a number of fields to explain a multivariate process in terms of a few independent variables (Bartholomew, 1987; Everitt, 1984).

Consider an observed sample (in *data space*) $\{\mathbf{t}_n\}_{n=1}^N \subset \mathbb{R}^D$ of N D-dimensional real vectors that has been generated by an unknown distribution. In latent variable modelling we assume that the distribution in data space $\mathbb{R}^D$ is actually due to a small number $L < D$ of latent variables acting in combination. We refer to this L-dimensional space as the *latent space*. Thus, a point $\mathbf{x}$ in latent space $\mathbb{R}^L$ is generated according to a prior distribution $p(\mathbf{x})$ and it is mapped onto data space $\mathbb{R}^D$ by a smooth mapping $\mathbf{f}$. Because $\mathbf{f}(\mathbb{R}^L)$ is an L-dimensional manifold in $\mathbb{R}^D$, in order to extend it to the whole D-dimensional data space we define a distribution $p(\mathbf{t}|\mathbf{x}) = p(\mathbf{t}|\mathbf{f}(\mathbf{x}))$ on $\mathbb{R}^D$, called the noise or error model. The prior in latent space $p(\mathbf{x})$, the smooth mapping $\mathbf{f}$ and the noise model $p(\mathbf{t}|\mathbf{x})$ are all equipped with parameters which we collectively call $\boldsymbol{\Theta}$. In latent variable modelling these parameters are optimised, typically to maximise the likelihood of the observed data given the parameters, $p(\mathbf{t}_n|\boldsymbol{\Theta})$. This optimisation is often carried out using an EM algorithm (Dempster et al., 1977). Figure 2 illustrates the idea of latent variable models.

The joint probability density function in the product space $\mathbb{R}^D \times \mathbb{R}^L$ is $p(\mathbf{t}, \mathbf{x})$ and integrating over latent space gives the marginal distribution in data space:

$$p(\mathbf{t}) = \int p(\mathbf{t}, \mathbf{x})\, d\mathbf{x} = \int p(\mathbf{t}|\mathbf{x}) p(\mathbf{x})\, d\mathbf{x}. \tag{1}$$

The log-likelihood of the parameters given the sample $\{\mathbf{t}_n\}_{n=1}^N$ is

$$l(\boldsymbol{\Theta}) = \log \prod_{n=1}^{N} p(\mathbf{t}_n|\boldsymbol{\Theta}) = \sum_{n=1}^{N} \log p(\mathbf{t}_n|\boldsymbol{\Theta}) \tag{2}$$

which is to be maximised under the maximum likelihood criterion for parameter estimation. This will provide with a set of values for the parameters, $\boldsymbol{\Theta}^* = \arg\max_{\boldsymbol{\Theta}} l(\boldsymbol{\Theta})$, corresponding to a (local) maximum of the log-likelihood. Note that the log-likelihood value $l(\boldsymbol{\Theta}^*)$ allows the comparison of any two latent variable models (or, in general, any two probability models), however different these may be.

Once the parameters $\boldsymbol{\Theta}$ are fixed, Bayes' theorem gives the posterior distribution in latent space given a data vector $\mathbf{t}$, i.e., the distribution of the probability that a point $\mathbf{x}$ in latent space was responsible for generating $\mathbf{t}$:

$$p(\mathbf{x}|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{t})} = \frac{p(\mathbf{t}|\mathbf{x})p(\mathbf{x})}{\int p(\mathbf{t}|\mathbf{x})p(\mathbf{x})\, d\mathbf{x}}. \tag{3}$$

Summarising this distribution in a single latent space point $\mathbf{x}^*$ results in a *reduced-dimension representative* of $\mathbf{t}$. This defines a corresponding mapping $\mathbf{F}$ from data space onto latent space, so that every data point $\mathbf{t}$ is assigned a representative in latent space, $\mathbf{x}^* = \mathbf{F}(\mathbf{t})$. Thus, it can be considered as an *inverse mapping* of $\mathbf{f}$. This mapping $\mathbf{F}$ will be most successful when the posterior distribution $p(\mathbf{x}|\mathbf{t})$ is unimodal and sharply peaked. Typical choices for $\mathbf{F}(\mathbf{t})$ are the mean, $\mathrm{E}\{\mathbf{x}|\mathbf{t}\}$, or the mode, $\arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{t})$.

Theoretical investigation of the conditions for smoothness of the inverse mapping $\mathbf{F}$ is under way. We can expect such smoothness to be related to the multimodality character of the posterior distribution of eq. (3) and to the discrete or continuous character of the latent space. For factor analysis and PCA the inverse mapping is smooth, but not for GTM in general. However, continuity of the inverse mapping is likely in practical situations where the posterior distribution is unimodal and sharply peaked for the majority of the training set data points.

Applying the mapping $\mathbf{f}$ to the reduced-dimension representative we obtain the reconstructed data vector $\mathbf{t}^* = \mathbf{f}(\mathbf{x}^*)$. The reconstruction error for that point $\mathbf{t}$ is defined in terms of some distance $\Delta(\mathbf{t}, \mathbf{t}^*)$ in data space and the average reconstruction error for the sample as $E_\Delta = \frac{1}{N}\sum_{n=1}^N \Delta(\mathbf{t}_n, \mathbf{t}_n^*)$. For example, the Euclidean distance provides with the usual mean squared error criterion $E_2 = \frac{1}{N}\sum_{n=1}^N \|\mathbf{t}_n - \mathbf{t}_n^*\|_2^2$.

3.1 Latent variable models

A latent variable model is specified by:

- The functional form of the prior in latent space $p(\mathbf{x})$;
- The smooth mapping $\mathbf{f} : \mathbb{R}^L \longrightarrow \mathbb{R}^D$ from latent space to data space;
- The noise model in data space $p(\mathbf{t}|\mathbf{x})$.

We have investigated two specific latent variable models: *factor analysis* and the *generative topographic mapping* (GTM). In both cases the parameters of the model are estimated using the EM algorithm. The E-step involves estimating the posterior distribution of the latent variables (using the current parameter values). The M-step is a maximisation of the log-likelihood function in which the posterior distribution estimated in the E-step is used to *fill in* the latent variables.

Both these choices of latent variable model implicitly assume that the observed data is continuous, while the EPG patterns are binary. In the experiments, we shall see that this does not present any problems. In section 7.2, we also fit a purely binary model to the data, a mixture of multivariate Bernoulli distributions.

3.1.1 Factor analysis

Factor analysis (Bartholomew, 1987; Everitt, 1984) uses a Gaussian distributed prior and noise model, and a linear mapping from data space to latent space. Specifically:

- The latent space prior $p(\mathbf{x})$ is unit normal:

$$p(\mathbf{x}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{4}$$

  The latent variables $\mathbf{x}$ are often referred to as the *factors*.

- The smooth mapping $\mathbf{f}$ is linear:

$$\mathbf{f}(\mathbf{x}) = \mathbf{\Lambda x} + \boldsymbol{\mu}. \tag{5}$$

  The columns of the $D \times L$ matrix $\mathbf{\Lambda}$ are referred to as the *factor loadings*.

- The data space noise model is normal centred in $\mathbf{f}(\mathbf{x})$ with diagonal covariance matrix $\mathbf{\Psi}$:

$$p(\mathbf{t}|\mathbf{x}) \sim \mathcal{N}(\mathbf{f}(\mathbf{x}), \mathbf{\Psi}). \tag{6}$$

  The D diagonal elements of $\mathbf{\Psi}$ are referred to as the *uniquenesses*.

The marginal distribution in data space is normal with a constrained covariance matrix:

$$p(\mathbf{t}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Lambda\Lambda}^T + \mathbf{\Psi}). \tag{7}$$

The posterior in latent space is also normal:

$$p(\mathbf{x}|\mathbf{t}) \sim \mathcal{N}\left(\mathbf{A}(\mathbf{t}-\boldsymbol{\mu}), (\mathbf{I} + \mathbf{\Lambda}^T\mathbf{\Psi}^{-1}\mathbf{\Lambda})^{-1}\right) \tag{8}$$

$$\mathbf{A} = \mathbf{\Lambda}^T(\mathbf{\Lambda\Lambda}^T + \mathbf{\Psi})^{-1} = (\mathbf{I} + \mathbf{\Lambda}^T\mathbf{\Psi}^{-1}\mathbf{\Lambda})^{-1}\mathbf{\Lambda}^T\mathbf{\Psi}^{-1}. \tag{9}$$

The reduced-dimension representative is taken as the posterior mean:

$$\mathbf{F}(\mathbf{t}) = \mathrm{E}\left\{\mathbf{x}|\mathbf{t}\right\} = \mathbf{A}(\mathbf{t}-\boldsymbol{\mu}). \tag{10}$$

Therefore, the inverse mapping $\mathbf{F}$ is smooth.

Note that orthogonal rotation of the factors ($\mathbf{\Lambda}' = \mathbf{\Lambda R}$ where $\mathbf{R}$ is an orthogonal matrix) does not alter the distribution in data space, $p(\mathbf{t})$. Thus, from all the factor loadings matrices $\mathbf{\Lambda}$, we are free to choose that which is easiest to interpret according to some criterion, e.g. by varimax rotation[2].

The parameters of a factor analysis model may be estimated using an EM algorithm (Rubin and Thayer, 1982):

E step: This requires computing the moments:

$$\mathrm{E}\left\{\mathbf{x}|\mathbf{t}_n\right\} = \mathbf{A}^{(\tau)}(\mathbf{t}_n - \boldsymbol{\mu})$$
$$\mathrm{E}\left\{\mathbf{x}\mathbf{x}^T|\mathbf{t}_n\right\} = \mathbf{I} - \mathbf{A}^{(\tau)}\mathbf{\Lambda}^{(\tau)} + \mathbf{A}^{(\tau)}(\mathbf{t}_n - \boldsymbol{\mu})(\mathbf{t}_n - \boldsymbol{\mu})^T(\mathbf{A}^{(\tau)})^T$$

for each data point $\mathbf{t}_n$ given the current parameter values $\mathbf{\Lambda}^{(\tau)}$ and $\mathbf{\Psi}^{(\tau)}$.

M step: This results in the following update equations for the factor loadings $\mathbf{\Lambda}$ and uniquenesses $\mathbf{\Psi}$:

$$\mathbf{\Lambda}^{(\tau+1)} = \left(\sum_{n=1}^{N} \mathbf{t}_n \,\mathrm{E}\left\{\mathbf{x}|\mathbf{t}_n\right\}^T\right)\left(\sum_{n=1}^{N} \mathrm{E}\left\{\mathbf{x}\mathbf{x}^T|\mathbf{t}_n\right\}^T\right)^{-1}$$
$$\mathbf{\Psi}^{(\tau+1)} = \frac{1}{N}\,\mathrm{diag}\left(\sum_{n=1}^{N} \mathbf{t}_n\mathbf{t}_n^T - \mathbf{\Lambda}^{(\tau+1)}\,\mathrm{E}\left\{\mathbf{x}|\mathbf{t}_n\right\}\mathbf{t}_n^T\right)$$

where the updated moments are used and the "diag" operator sets all the off-diagonal elements of a matrix to zero.

The location parameter $\boldsymbol{\mu}$ is estimated by the sample mean, and does not take part in the EM algorithm.

[2]Varimax rotation (Kaiser, 1958) finds an orthogonal rotation of the factors such that, for each new factor, the loadings are either very large or very small (in absolute value). The resulting rotated matrix $\mathbf{\Lambda}'$ has many values clamped to (almost) 0, that is, each factor involves only a few of the original variables. This simplifies factor interpretation.

3.1.2 Principal component analysis

Principal component analysis (PCA) is a factor analysis in which the uniquenesses are constrained to be equal; i.e., $\mathbf{\Psi}$ is isotropic, $\mathbf{\Psi} = \sigma^2\mathbf{I}$. PCA is not usually constructed as a probability model, and is usually inferred by minimising a least squares reconstruction error.

We refer to a principal components or factor analysis having order L when the first L principal components or factors are extracted. For a fixed data set we may regard PCA as an *additive* technique, insofar a PCA of order L produces the same principal components as a PCA of order $L-1$ plus a new, additional principal component. However, PCA followed by (varimax) rotation does not have this property anymore.

The property of additivity does not necessarily hold for factor analysis: that is, the factors found by a factor analysis of order L are, in general, all different from those found by a factor analysis of order $L-1$. That means that one can only talk about the joint collection of L factors (or the linear subspace spanned by them). However, as reported in section 5, for the EPG data set we used, we found that this property of *factor additivity* holds with good approximation: after varimax rotation, the factor loadings obtained in a factor analysis of order $L-1$ appear again with little modification in a factor analysis of order L. This has an important consequence: the factors can be ordered by a *relevance* criterion, in the same way that the principal components can be ordered by the proportion of directional variance they explain. So, a factor analysis of order 1 produces the first factor; a factor analysis of order 2 finds that same factor plus a second one, etc. In this situation, it is possible to refer individually to each factor as having an independent status.

3.1.3 GTM

The generative topographic mapping (GTM) (Bishop et al., 1998) is a nonlinear latent variable model which has been proposed as a principled alternative to self-organising feature maps (Kohonen, 1995). Specifically:

- The prior in latent space, $p(\mathbf{x})$, is discrete uniform, assigning nonzero probability only to the points $\{\mathbf{x}_k\}_{k=1}^K$, usually arranged in a regular grid:

$$p(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K}\delta(\mathbf{x}-\mathbf{x}_k). \tag{11}$$

  This discrete prior can be seen as a Monte Carlo sampling of a continuous, uniform prior.

- The mapping $\mathbf{f}$ is a generalised linear model:

$$\mathbf{f}(\mathbf{x}) = \mathbf{W}\boldsymbol{\phi}(\mathbf{x}), \tag{12}$$

  where $\mathbf{W}$ is a $D \times F$ matrix and $\boldsymbol{\phi}$ an $F \times 1$ vector of fixed basis functions.

- The noise model $p(\mathbf{t}|\mathbf{x})$ is an isotropic normal centred in $\mathbf{f}(\mathbf{x})$:

$$p(\mathbf{t}|\mathbf{x}) \sim \mathcal{N}(\mathbf{f}(\mathbf{x}), \sigma^2\mathbf{I}). \tag{13}$$

The marginal distribution in data space is a constrained mixture of Gaussians:

$$p(\mathbf{t}) = \frac{1}{K}\sum_{k=1}^{K} p(\mathbf{t}|\mathbf{x}_k), \tag{14}$$

and the posterior in latent space is discrete:

$$p(\mathbf{x}_k|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{x}_k)}{\sum_{i=1}^{K} p(\mathbf{t}|\mathbf{x}_i)}. \tag{15}$$

The reduced-dimension representative can be taken as the posterior mean or the posterior mode, which can be quite different from each other if the posterior distribution is multimodal. The inverse mapping $\mathbf{F}$ is not continuous in general, although it will be approximately continuous if the posterior distribution (15) is unimodal and sharply peaked for most points in data space.

The parameters of a GTM model may be estimated using the EM algorithm:

E step: This requires computing the *responsibility* $R_{nk} = p(\mathbf{x}_k|\mathbf{t}_n)$ of each latent space point $\mathbf{x}_k$ having generated point $\mathbf{t}_n$ using eqs. (11)-(15) with the current parameter values $\mathbf{W}^{(\tau)}$ and $\sigma^{(\tau)}$.

M step: This results in the following update equations for the parameters $\mathbf{W}$ and σ, respectively:

$$\mathbf{\Phi}^T \mathbf{G}^{(\tau)} \mathbf{\Phi} (\mathbf{W}^{(\tau+1)})^T = \mathbf{\Phi}^T (\mathbf{R}^{(\tau)})^T \mathbf{T}$$

$$(\sigma^{(\tau+1)})^2 = \frac{1}{ND} \sum_{k=1}^{K} \sum_{n=1}^{N} R_{nk}^{(\tau)} \left\| \mathbf{f}(\mathbf{x}_k) - \mathbf{t}_n \right\|^2$$

where $\mathbf{\Phi} = (\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_K)^T$, $\mathbf{T} = (\mathbf{t}_1, \dots, \mathbf{t}_N)^T$, $\mathbf{R}$ is an $N \times K$ matrix with elements R_{nk} and $\mathbf{G}$ is a $K \times K$ diagonal matrix with elements $G_{kk} = \sum_{n=1}^{N} R_{nk}$. Solving for $\mathbf{W}$ requires (pseudo-)inverting the matrix $\mathbf{\Phi}^T \mathbf{G} \mathbf{\Phi}$ at each iteration.

3.2 Finite mixtures of latent variable models

Finite mixtures (Everitt and Hand, 1981) of latent variable models can be constructed in the usual way as

$$p(\mathbf{t}) = \sum_{m=1}^{M} p(m) p(\mathbf{t}|m) \tag{16}$$

where:

- $p(\mathbf{t}|m)$, $m = 1, \dots, M$ are latent variable models not necessarily based on latent spaces of the same dimension, i.e.,

  $$p(\mathbf{t}|m) = \int p(\mathbf{t}, \mathbf{x}|m)\, d\mathbf{x} = \int p(\mathbf{t}|\mathbf{x}, m) p(\mathbf{x}|m)\, d\mathbf{x}$$

  where $p(\mathbf{x}|m)$ is the prior distribution in the latent space of the mth component, $p(\mathbf{t}|\mathbf{x}, m)$ its noise model and $\mathbf{f}_m : \mathbb{R}^{L_m} \longrightarrow \mathbb{R}^D$ its mapping from latent space into data space.

- $p(m)$ are the mixing proportions.

The joint density is $p(\mathbf{t}, \mathbf{x}, m) = p(\mathbf{t}|\mathbf{x}, m) p(\mathbf{x}|m) p(m)$ and the finite mixture distribution can be expressed as the marginalisation of $p(\mathbf{t}, \mathbf{x}, m)$ over $\mathbf{x}$ and m:

$$p(\mathbf{t}) = \sum_{m=1}^{M} \int p(\mathbf{t}, \mathbf{x}, m)\, d\mathbf{x} = \sum_{m=1}^{M} p(m) \int p(\mathbf{t}|\mathbf{x}, m) p(\mathbf{x}|m)\, d\mathbf{x} = \sum_{m=1}^{M} p(m) p(\mathbf{t}|m).$$

Once the model is formulated and the functional forms of each latent variable model are fixed, estimation of the parameters can be done by maximum likelihood. As usual with mixture models, the mixing proportions are taken as parameters $p(m) = \pi_m$ and included in the estimation process; observe that one parameter π_m is not free due to the constraint $\sum_{m=1}^{M} \pi_m = 1$.

Maximum likelihood estimation can be conveniently accomplished with an EM algorithm:

E step: This requires computing the *responsibility* $R_{nm} = p(m|\mathbf{t}_n)$ of each component m having generated point $\mathbf{t}_n$ using Bayes' theorem:

$$R_{nm} = \frac{p(\mathbf{t}_n|m) p(m)}{p(\mathbf{t}_n)} = \frac{\pi_m p(\mathbf{t}_n|m)}{\sum_{m=1}^{M} \pi_m p(\mathbf{t}_n|m)}$$

where $p(\mathbf{t}|m)$ is given by the latent variable model with the parameter values of the current iteration.

M step: This results in several update equations for the parameters. The update equations for the mixing proportions are independent of the type of latent variable model used:

$$\pi_m^{(\tau+1)} = \frac{1}{N} \sum_{n=1}^{N} R_{nm}^{(\tau)} \pi_m^{(\tau)}.$$

The equations for the rest of the parameters (from the individual latent variable models) depend on the specific functional form of $p(\mathbf{t}|\mathbf{x}, m)$ and $p(\mathbf{x}|m)$, but often they are averages of the usual statistics weighted by the responsibilities. EM algorithms are available to train mixtures of factor analysers (Ghahramani and Hinton, 1996) and of principal component analysers (Tipping and Bishop, 1997).

A reduced-dimension representative can be obtained as the reduced-dimension representative of the mixture component with the highest responsibility: $\mathbf{x}^* = \mathbf{F}(\mathbf{t}) = \mathbf{x}^*_{m^*}$ such that $m^* = \arg\max_m p(m|\mathbf{t})$. A

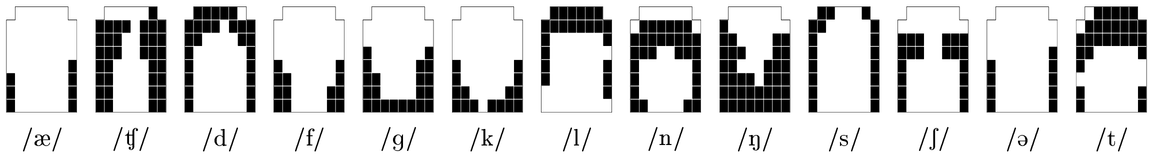

Figure 3: Representative EPGs for the typical stable phase of different phonemes.

reconstructed vector in data space can then be obtained as the reconstructed vector of component m^*: $\mathbf{t}^* = \mathbf{f}_{m^*}(\mathbf{F}_{m^*}(\mathbf{t}))$. Alternatively, one could average the reconstructed vectors as

$$\mathbf{t}^* = \sum_{m=1}^{M} p(m|\mathbf{t})\mathbf{f}_m(\mathbf{F}_m(\mathbf{t})) = \sum_{m=1}^{M} p(m|\mathbf{t})\mathbf{t}_m^*. \tag{17}$$

In particular, mixtures of factor analysers can be constructed where each factor analyser of L factors is characterised by two kinds of parameters: the mean vector $\boldsymbol{\mu}_m$ and the loadings matrix $\mathbf{\Lambda}_m$ (containing L loading vectors), in addition to the mixing proportion π_m.

3.3 Mixtures of multivariate Bernoulli distributions

We also modelled the EPG data sets with a mixture of multivariate Bernoulli distributions, where each component m is modelled as a multivariate Bernoulli distribution with parameters $\mathbf{p}_m = (p_{m1}, \dots, p_{mD})^T$:

$$p(\mathbf{t}|m) = \prod_{d=1}^{D} p_{md}^{t_d}(1 - p_{md})^{1-t_d} = \prod_{d=1}^{D} p(t_d|m). \tag{18}$$

Note that this is not a latent variable model because it does not have the form of eq. (1): there is no underlying latent space. Although one could consider the variable m, which indexes the components, as a latent variable (this is the principle of *latent class analysis* (Everitt and Hand, 1981)), we are interested in possibly multidimensional latent spaces.

While all the previously commented models were designed for continuous variables, disregarding the binary nature of the EPG data, this is a purely binary model. Also note that, while a multivariate Bernoulli distribution considers each component of the $\mathbf{t}$ vector independently ($p(\mathbf{t}|m)$ being a product distribution), this independence is broken in the mixture model, which thus can account for correlations between variables.

4 Data set description

We have used a subset of the EUR-ACCOR database (Marchal and Hardcastle, 1993). This database, designed for the cross-language study of coarticulation, contains different kinds of measurements (electropalatography, pneumotachography, laryngography and some EMA, as well as the original acoustic signal) for utterances, isolated words and nonsense items spoken by several subjects in seven different European languages (French, English, German, Italian, Catalan, Swedish and Irish Gaelic) and varying speech styles (slow, fast, etc.). The EPG data consists of 62-bit frames sampled at 200 Hz.

We have selected EPG frames corresponding to the English language, in which each of 6 different subjects recorded 14 different utterances. The data set was divided into 6 parts, one for each speaker, and the data of each speaker was split into a training and a test set: utterances 1, 3–5 and 9–14 were put into the training set and utterances 2 and 6–8 into the test set (see appendix B). The criterion followed was to have a balance between the number of different EPG frames (i.e., excluding repeated patterns) between the test and training set: roughly 80% for training and 20% for test (depending on each speaker).

All the data used were unlabelled. Labelling each EPG frame with the phoneme it corresponds to is a time-consuming process, requiring manual annotation of the EPG sequence in relation to the acoustic signal (and perhaps other signals) by a phonetician; automatic annotation is possible using a speech recognition program but prone to errors. This is another reason to perform unsupervised learning of the EPG data, so that articulatory categories can be extracted directly from the data in the form of factors, prototypes or, more generally, a latent-variable space.

5 Results using factor analysis and principal component analysis

For all the experiments we report, maximum likelihood factor analysis was performed using an EM algorithm (Rubin and Thayer, 1982) with random starting points. Once the relative increase in log-likelihood was

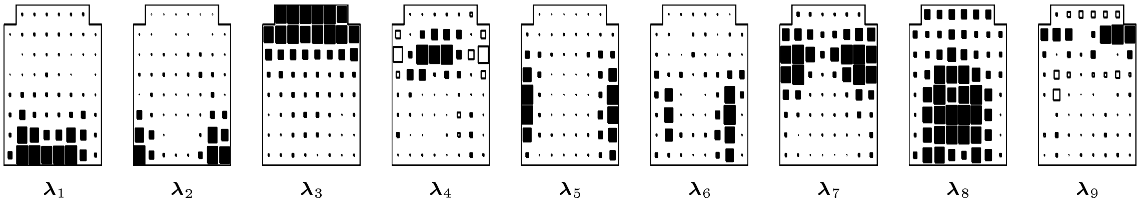

Figure 4: Factors for speaker RK after varimax rotation.

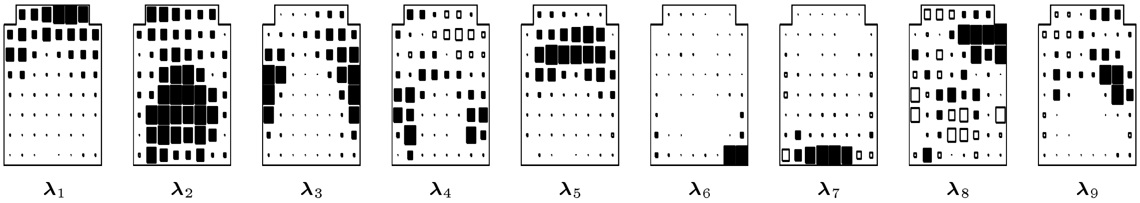

Figure 5: Factors for speaker HD after varimax rotation.

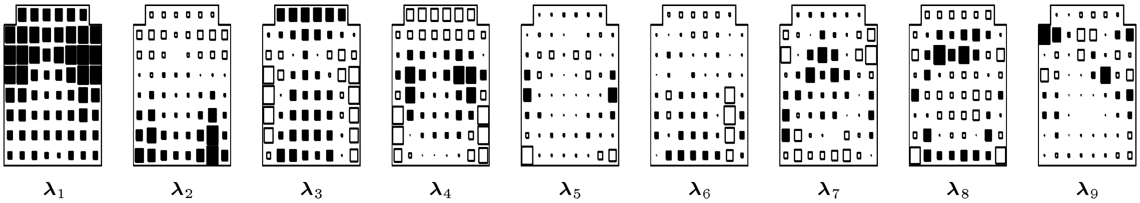

Figure 6: Principal components for speaker RK.

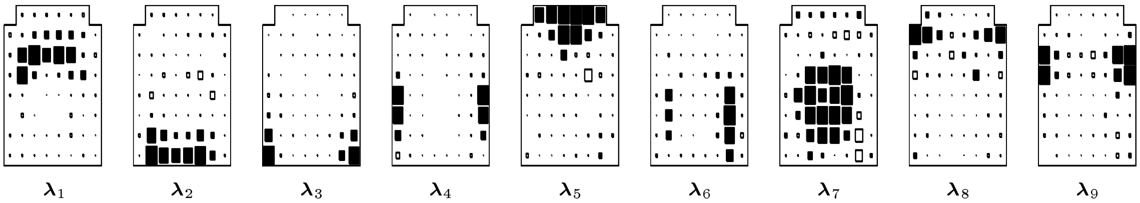

Figure 7: Principal components for speaker RK after varimax rotation.

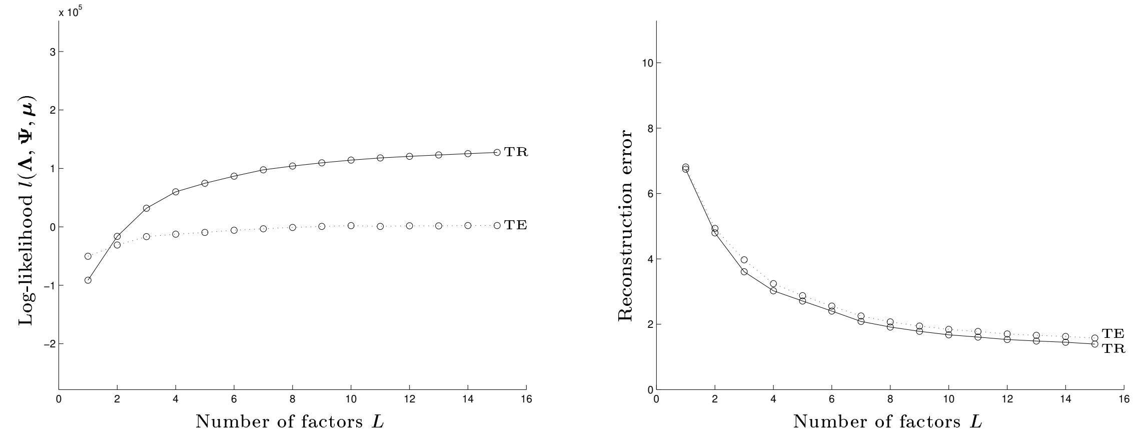

Figure 8: Log-likelihood (left) and reconstruction error (right) of the factor analysis model for speaker RK. TR and TE refer to training and test set, respectively.

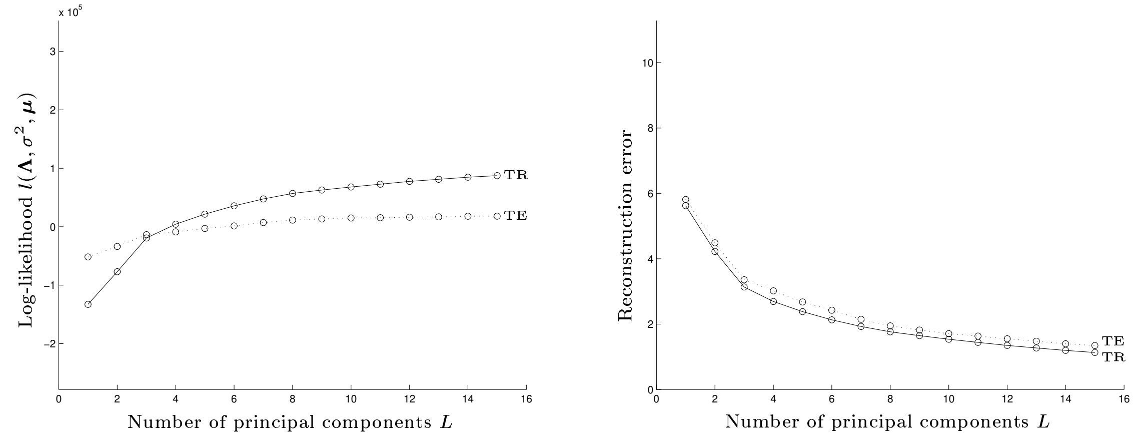

Figure 9: Log-likelihood (left) and reconstruction error (right) of the PCA model for speaker RK. TR and TE refer to training and test set, respectively.

smaller than 10^{-7}, the iterative procedure was stopped and varimax rotation was applied to the factors.

Figure 4 shows the factor loadings obtained for a factor analysis of order 9, ordered by relevance for speaker RK. Considering the loading vectors as linear indices, it is apparent that many of the indices of figure 1 may be associated to the factor loadings or to linear combinations of these loadings (e.g. ALV, PAL and VEL with factors 3, 7 and 1, respectively).

This association may be regarded as an empirical justification of these phonetic indices, since they account for a large proportion of the correlation between variables, as well as possessing a straightforward interpretation. However, if we compare figure 4 with figure 5 (corresponding to speaker HD) we see that, although some factors are approximately equal for both speakers (e.g. factor 8 in RK is factor 2 in HD), there are significant differences, revealing different articulation trends. For example, speaker HD tends to produce EPG patterns which are much less symmetrical than those of RK. Such variations are difficult to track with fixed indices; e.g. the asymmetry index will give no extra information for a speaker that does not produce any asymmetric patterns.

We also applied principal component analysis to the same training sets as factor analysis. The first 9 principal components for speaker RK are shown in figure 6. We note that most of the basis vectors contain both positive and negative values, thus lacking a simple interpretation. After a varimax rotation (figure 7) most basis vectors contain a majority of positive values, with several being similar to some of the factors of fig. 4. This is a consequence of the uniquenesses matrix $\mathbf{\Psi}$ being relatively isotropic (a multiple of the identity matrix), in which case factor analysis is equivalent to PCA. However, in a general situation the two methods give different results.

The factor analysis and PCA representations may be evaluated in terms of log-likelihood and reconstruction error. The left side of figures 8 and 9 show the log-likelihood of the factor analysis and PCA models for the training and test data sets of speaker RK and for different numbers of factors or principal components, respectively. In both models the log-likelihood value was computed as:

$$l(\boldsymbol{\Theta}) = -\frac{N}{2}\left(D\log 2\pi + \log|\boldsymbol{\Sigma}(\boldsymbol{\Theta})| + \operatorname{tr}\left(\mathbf{S}\boldsymbol{\Sigma}(\boldsymbol{\Theta})^{-1}\right)\right) \tag{19}$$

where $\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{t}_n - \bar{\mathbf{t}})(\mathbf{t}_n - \bar{\mathbf{t}})^T$ is the sample covariance matrix of the data set, with $\bar{\mathbf{t}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{t}_n$ being the sample mean. $\boldsymbol{\Sigma}(\boldsymbol{\Theta})$ is the covariance matrix under the model specified with parameters $\boldsymbol{\Theta}$, i.e., $\boldsymbol{\Sigma}(\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\mu}) = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}$ for factor analysis and $\boldsymbol{\Sigma}(\boldsymbol{\Lambda}, \sigma^2, \boldsymbol{\mu}) = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \sigma^2\mathbf{I}$ for PCA, according to the probabilistic PCA model of (Tipping and Bishop, 1997). Note that factor analysis is systematically better than PCA, as the theory predicts, because PCA is a particular case of factor analysis in which the uniquenesses are forced to be equal. That is, factor analysis is a more complete model than PCA. Not shown in the figure is the fact that the maximum achievable log-likelihood is identical for PCA (using the limit of 62 principal components) and factor analysis (using 51 factors, the identifiability limit[3]).

The right side of figures 8 and 9 shows the reconstruction error of the factor analysis and PCA models, respectively. For PCA, the projection on the L principal components was used; for factor analysis, the projection using (10) was used. The reconstruction error was defined in the usual way:

$$E_2 = \frac{1}{N}\sum_{n=1}^{N}\|\mathbf{t}_n - (\boldsymbol{\Lambda}\mathbf{A}(\mathbf{t} - \boldsymbol{\mu}) + \boldsymbol{\mu})\|_2^2 \tag{20}$$

where for PCA, $\boldsymbol{\Lambda} = \mathbf{A}^T$ contains the first L principal components and for factor analysis, $\boldsymbol{\Lambda}$ contains the factor loadings and $\mathbf{A}$ is defined in (9). According to the squared reconstruction error criterion PCA is systematically better than factor analysis; note that PCA is optimal among linear mappings according to this criterion on the training set. Further the directional variance (along each component or factor) decreases monotonically in PCA, while in factor analysis it does not—although it has a tendency to decrease with the relevance order.

6 Results using GTM

A factor analysis goodness-of-fit test[4] (valid in the asymptotic limit of large samples) is available for the null hypothesis that "the data sample can be explained with L factors" (Everitt, 1984), i.e., that the covariance of the observed variables can be accounted for with L factors. For our training sets (of sizes well above $N = 5\,000$ in all cases) and at a significance level of 5%, this hypothesis was rejected for all $L < 45$. A similar test is available for PCA and the null hypothesis was rejected for all values of L at the same significance level. This suggests that linear methods are not powerful enough and are able to extract only a fraction of the information contained in the EPG data.

Nonlinear mappings are accommodated by the latent variable framework, but arbitrary mappings and probability distributions present insurmountable mathematical and computational difficulties, particularly in the analytical evaluation of integral (1) (Carreira-Perpiñán, 1997). Therefore the actual choice is limited. The generative topographic mapping (GTM) (Bishop et al., 1998) is a latent variable model well suited to visualisation, since the latent space is effectively limited to two dimensions[5]. We trained GTM models with the following parameters: 20×20 grid in two-dimensional latent space ($K = 400$ points), scaled to the $[-1, 1] \times [-1, 1]$ square, and $\sqrt{F} \times \sqrt{F}$ grid in the same square of F Gaussian basis functions of width equal to the separation between basis functions centres; $\sqrt{F}$ varied from 3 to 14. Training was achieved by an EM algorithm, stopping when the relative increment in log-likelihood was smaller than 10^{-4}; the starting point was not random, but derived from the first principal components of the data set, as described in (Bishop et al., 1998).

Figure 10 gives the log-likelihood (left) and the reconstruction error (right) for the GTM models (speaker RK). We can see that:

- While the log-likelihood for the training set increases monotonically with increasing F, that of the test set reaches a maximum at around $F = 49$ and starts decreasing again. This means that overfitting is occurring for $F > 49$: the model is learning too well the training set but cannot generalise to the test set properly. The phenomenon of overfitting did not appear in the previous models, probably due to the small number of parameters used by them (although the log-likelihood curves for the test set do flatten if many factors or principal components are used). Note that the reconstruction error continues decreasing monotonically past $F = 49$.

[3] In order that the factor analysis model be identifiable, the number of degrees of freedom $\frac{1}{2}((D-L)^2 - (D+L))$ must be positive (Bartholomew, 1987; Everitt, 1984), which gives an upper bound for the number of factors to be extracted.

[4] This statistical test should be considered only as qualitative evidence, since we follow a Bayesian approach, under which hypothesis tests are not generally valid (Berger, 1985, pages 145–157). Also, this test requires normality of the population, which does not hold for our EPG data sets.

[5] Although theoretically it may use a latent space of any dimension L, in practice one uses $L \leq 2$, since the computational complexity is approximately exponential in L due to the curse of the dimensionality.

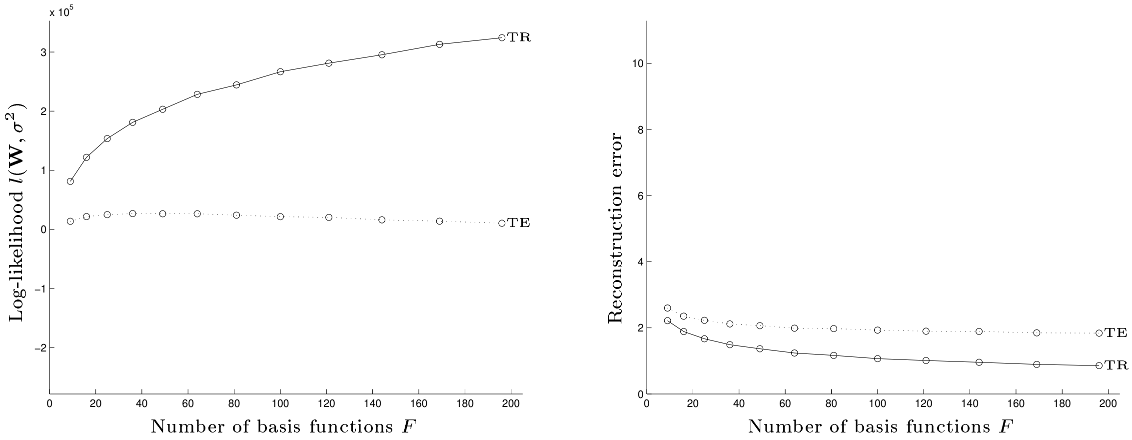

Figure 10: Log-likelihood (left) and reconstruction error (right) of the GTM model for speaker RK. TR and TE refer to training and test set, respectively.

- The log-likelihood of the GTM model (even with a relatively small number of basis functions, to avoid overfitting) is better than the highest possibly attainable log-likelihood of both PCA and factor analysis (i.e., using any number of principal components or factors). The reconstruction error using as reduced-dimension representative the mode of the posterior distribution[6] is very small, comparable to that of a PCA with 10 principal components or more (i.e., a latent space of dimension $L = 10$).

We remark that GTM is using a latent space of only dimension 2, while the curves for factor analysis and PCA correspond to latent spaces of higher dimension.

GTM is then a much better generative model for the EPG data set than PCA or factor analysis, due to the fact that the mapping between latent and data space used by GTM is an expansion in radial basis functions, which have been proven to be able to approximate any smooth function to arbitrary accuracy given enough basis functions (Park and Sandberg, 1993). However, the number of parameters of a GTM model is typically much higher than that of the other models we have analysed.

As we mentioned in section 3, the reduced-dimension representative for a data point $\mathbf{t}$ is obtained as a particularly informative point in latent space according to the posterior distribution for $\mathbf{t}$. For factor analysis and PCA this posterior distribution is normal (more or less broad, depending on the covariance matrix) and there is no doubt as to what reduced-dimension representative to select, because the normal is unimodal: the mean (equal to the mode). This is not necessarily the case for GTM, though: in principle, the posterior distribution for a given data point can be multimodal, which makes difficult selecting a single reduced-dimension representative. However, once our GTM model was trained, we found that the posterior distribution was unimodal and sharply peaked for over 90% of the data points. Thus, almost every EPG pattern can be associated to a unique point in two-dimensional latent space. Clearly, the coordinates of the reduced-dimension representative could be considered as (nonlinear) EPG data reduction indices.

The two-dimensional GTM model is slow to train using large data sets—around 60 times slower than real time on this problem using a Sun Ultra-170 workstation. However, for a trained model, evaluating the posterior distribution in latent space (and hence the reduced-dimension representative) is fast.

7 Results using mixture models

7.1 Mixtures of factor analysers

We modelled the same training sets with a mixture of factor analysers. Again, a maximum likelihood estimate for the parameters of the model was found via an EM algorithm (Ghahramani and Hinton, 1996), with random starting point and stopping when the relative increment in log-likelihood was smaller than 10^{-4}. The parameters are now the mixing proportions π_m of the mixture (that correspond to the prior probabilities of the different mixture components) and the means $\boldsymbol{\mu}_m$ and factor loadings $\boldsymbol{\Lambda}_m$ of each factor analyser.

Figure 11 shows the results for a mixture with 4 components, each of them a factor analysis of first order. The means coincide with some of the typical EPG patterns of fig. 3 (and with some of the prototypes found by the mixture of multivariate Bernoulli distributions of the next section). The corresponding loading vectors coincide with or are approximately a linear combination of several of the factor loadings of fig. 4. This implies that the mixture does not find new factors, but that it places different factors in different locations of data

[6] Using as reduced-dimension representative the mean of the posterior distribution gave a slightly larger error in all cases.

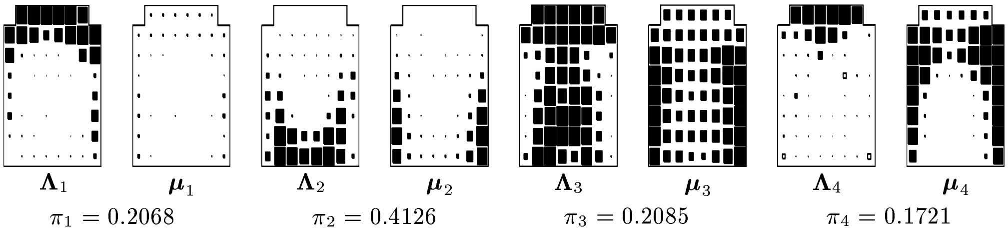

Figure 11: Means $\boldsymbol{\mu}_m$ and factor loadings $\boldsymbol{\Lambda}_m$ for a mixture of $M = 4$ factor analysers, each of $L = 1$ factor (for speaker RK). Below each pair $(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m)$ is the mixing proportion π_m of component m.

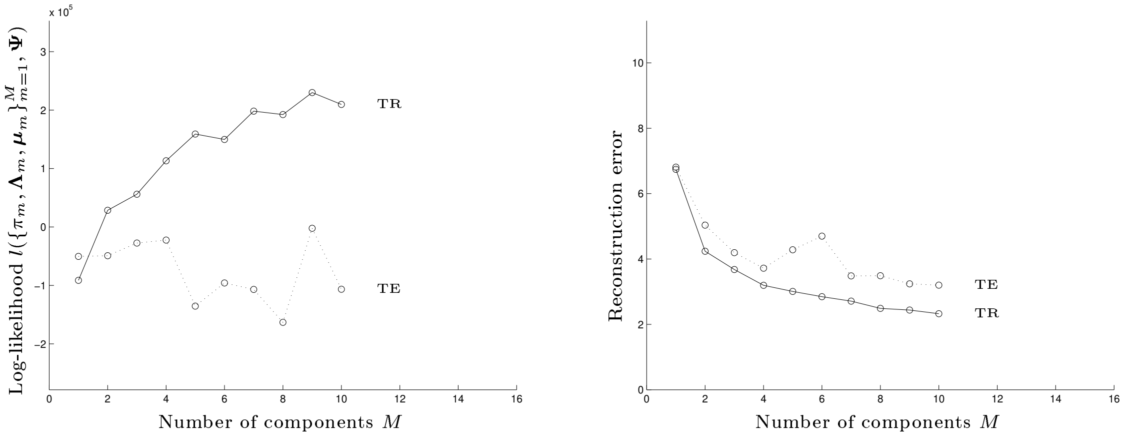

Figure 12: Log-likelihood (left) and reconstruction error (right) of the mixture of factor analysers model for speaker RK. TR and TE refer to training and test set, respectively.

space, adapting to the local behaviour of the correlations. For example, component number 2 is located in the area of vowels like /æ/ and is associated to a factor stressing velar patterns with some left-right asymmetry.

In all our experiments, a first-order factor analysis was used for each component (which made varimax rotation unnecessary), varying only the number of components in the mixture. The reason for not using higher order factor analysers is that the estimation algorithm systematically tends to a singularity of the likelihood surface, where some of the uniquenesses tend to zero and the $\boldsymbol{\Psi}$ matrix is not positive definite. This is a well-known problem in the literature of factor analysis and finite mixture distributions (Bartholomew, 1987; Everitt and Hand, 1981), called *Heywood case*.

As before, the left side of figure 12 shows the log-likelihood of the mixture of factor analysers model for all data sets of speaker RK, and the right side the reconstruction error (M is the number of components in the mixture). The log-likelihood of the mixture is always superior to that of simple factor analysis (the latter being a more restricted maximum likelihood model), and the reconstruction error is always smaller as well, although there is not much difference. The mixtures considered do not result in a significantly improved performance over factor analysis.

7.2 Mixtures of multivariate Bernoulli distributions

We trained a mixture of multivariate Bernoulli distributions with an EM algorithm (Everitt and Hand, 1981; Wolfe, 1970), using random starting values for the parameters and stopping when the relative increment in log-likelihood was smaller than 10^{-4}. In this case, the only parameters to be estimated (apart from the mixture proportions) are the Bernoulli probabilities $\mathbf{p}_m = (p_{m1}, \dots, p_{mD})^T$ of each component, which can be considered as the prototype vectors for that component, since the mean of a Bernoulli distribution coincides with its $\mathbf{p}$ value. Observe that, since each Bernoulli probability (of the corresponding electrode being activated) is in the range $[0, 1]$, the prototypes obtained are interpretable as EPGs without the necessity of any transformation, such as varimax rotation.

Again, we observed that for this data set the mixtures also have approximately the property of additivity mentioned in section 3.1.2, and so we can order the prototypes obtained. A mixture with $M = 9$ components produced the prototypes shown in fig. 13, in decreasing order of relevance from left to right. Those prototypes are easily recognisable as some of the typical EPG patterns of fig. 3, e.g. $\mathbf{p}_2$ with /ʃ/, $\mathbf{p}_5$ with /l/ or $\mathbf{p}_6$ with /g/. Thus, the mixture may be doing a good job at estimating the density of the distribution of the EPG data

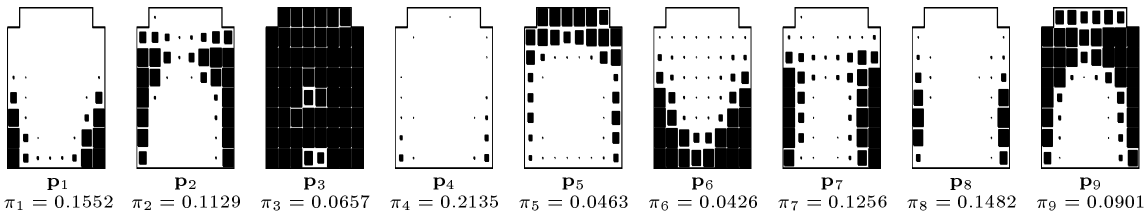

Figure 13: Prototypes for a mixture of multivariate Bernoulli distributions (for speaker RK).

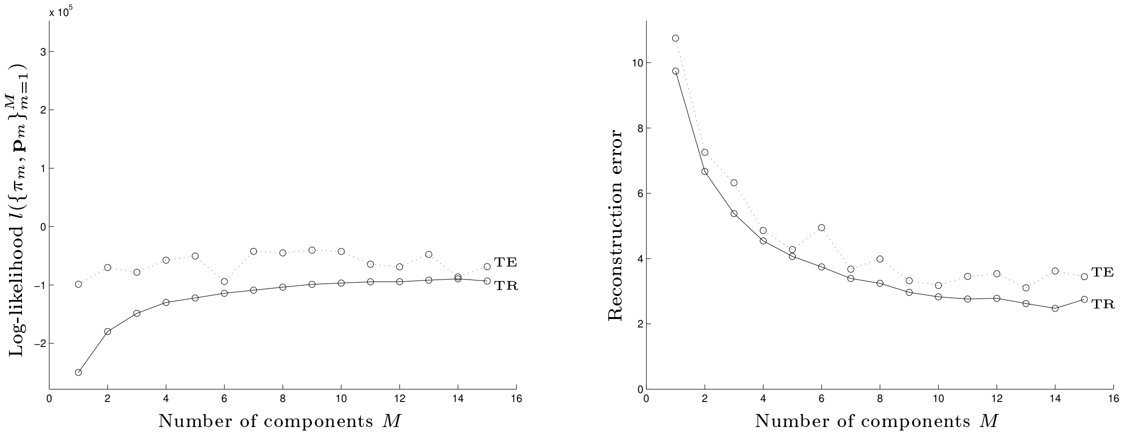

Figure 14: Log-likelihood (left) and reconstruction error (right) of the mixture of multivariate Bernoulli distributions model for speaker RK. TR and TE refer to training and test set, respectively.

in $D = 62$ dimensions, but—as noted in section 3.2—it does not achieve dimensionality reduction. The most it can do is to assign each data vector to its most likely component (in the sense of highest posterior probability) and reconstruct it as the prototype ($\mathbf{p}_m$ vector) of that component. This is essentially vector quantisation.

Figure 14 (left) shows the log-likelihood for speaker RK (where M is the number of components used). The addition of new components after the first 4 or 5 does not increase much the log-likelihood. Figure 14 (right) shows the reconstruction error, which is quite large compared to most of the other methods, confirming that this model is not good for vector reconstruction.

Also, note that the phenomenon of overfitting appears here again: the log-likelihood for the test set reaches a plateau at about $M = 4$.

EM estimation of a mixture of multivariate Bernoulli distributions will always give a positive likelihood from almost every starting point (Carreira-Perpiñán and Renals, 1998). In particular, boundary values (where some p_{md} can be 0 or 1) will not give rise to a likelihood equal to 0 (and a log-likelihood equal to $-\infty$) for points in the training set. However, this is not necessarily the case for other data sets. This is an undesirable feature arising from the Bernoulli distribution itself: a Bernoulli distribution of parameter $p = 0$ (1) can never generate a 1 (0). The other models studied (based in the normal distribution) assign to any point in the domain a strictly positive probability, however small this may be. Thus, for each speaker, a few points from the test set (usually less than 2%) were removed in a number of occasions in order to obtain a finite log-likelihood in figure 14.

8 Two-dimensional visualisation of EPGs

We consider now the issue of graphically representing a set of EPG frames, possibly a sequence obtained during an utterance. When there are more than two indices, the usual way to do this is by plotting the variation of all the indices with the time in an X-Y plot (called *trajectory* or *contact profile*) where time goes in abscissas and the index or indices of interest in ordinates (Byrd et al., 1995; Hardcastle et al., 1991b). However, when there are only two indices a more natural representation is possible: both index values are used as the X-Y coordinates in a plane, with points consecutive in time being linked by an arrow. If the two indices are powerful enough to adequately discriminate between different articulatory classes, the two-dimensional representation will be a *map* of articulatory regions.

There are statistical methods specifically suited to find the best projection of a multidimensional data set in the sense that it maximises a certain criterion. For example, discrimination between classes (Fisher's linear discriminant), departure from normality (projection pursuit) or a stress measure (multidimensional scaling); see (Huber, 1985; Mardia et al., 1979). These methods do not propose a model for the high-dimensional data. In contrast, our latent variable models are trained not find an optimal low-dimensional projection but to model best the data according to the maximum likelihood criterion. Nevertheless, the two-dimensional representation found by them (in particular by GTM) turns out to be good[7], as one might intuitively expect.

Figure 15 shows such a representation using pairs of factors for speaker RK. In the left picture, factors 1 and 2 were used, while in the right one 3 and 4 were used. The labels correspond to the typical EPG patterns of figure 3, while the trajectory corresponds to the highlighted fragment of the following utterance: “I prefer **Kant to** Hobbes for a good bedtime book,” with phonemic transcription /aɪ prɪ'fɜ 'kænt tə 'hɒbz fər ə 'gʊd 'bedtaɪm 'bʊk/. In each picture, the trajectory was obtained by projecting onto two-dimensional latent space each of the EPG frames of the utterance fragment and linking consecutive ones by a straight line. For visualisation purposes, we are free to choose the best pair from the factors extracted. In fact, factors 3-4 give a poor view of the data while factors 1-2 give a better one.

Figure 16 shows, for the case of two-dimensional GTM with $F = 49$ basis functions, the two-dimensional latent space with the same typical EPG patterns and utterance fragment as in fig. 15. Now, points in two-dimensional latent space which correspond to consecutive EPG frames in the utterance have been labelled with a correlative number to show more clearly the path followed by the reduced-dimension representative. The GTM map presents a much clearer layout of the articulatory classes than the best pair of factors did.

These two-dimensional maps obtained by latent variable models can be useful to analyse articulatory dynamics. Since the inverse mapping $\mathbf{F}$ defined in section 3 is continuous for factor analysis and approximately continuous for GTM, and since the EPG signal is assumed to be a continuous function of the time (due to the mechanical constraints of the tongue-palate system), discontinuities in the latent space trajectory must be due to abrupt jumps in the sampled EPG sequence in data space. This is the case for the transition from /æ/ to /nt/ in the utterance mentioned (fig. 17). The reason is that the EPG sampling frequency (equal to 100 Hz), while generally appropriate, is too low in situations where the tongue moves very fast and this results in missing EPG frames. The latent variable two-dimensional maps are able to detect this situation. Note, however, that the concept of continuity is not well defined because:

- The EPG frames are binary. Still, for a high enough sample frequency, one can expect most neighbouring frames to be very close (in both the Euclidean and Hamming distance sense).
- The latent space in GTM is discrete. However, owing to the fact that for most data points the posterior distribution was unimodal and sharply peaked, examination of the EPG sequence of the utterance showed that nearby points in data space were most times projected on neighbouring points in latent space.

Two-dimensional maps are also useful in speech therapy, speech assessment or language learning applications. For example, in speech therapy, most systems (such as the IBM SpeechViewer (Pratt et al., 1993)), provide real-time visual and auditory feedback on a range of speech parameters. Auditory feedback is available as normal or reduced-speed playback. Visual feedback includes cartoon graphics, intended for the speaker, and technical displays, useful for the therapist to monitor the details of the speaker's performance or to show aspects of speech patterning. The latter include displays over time of the fundamental frequency or the amplitude, 3D spectrograms, etc. If articulatory regions are drawn on the two-dimensional latent space, the speaker will receive instantaneous feedback in an intuitive way about the correctness of the articulatory gesture that he is trying to produce, a difficult task to achieve with the raw data. See (Hardcastle et al., 1991a) for a specific account of the uses of electropalatography (using both raw and reduced data) in speech therapy.

9 Discussion

9.1 Method comparison

Figures 18 and 19 show the log-likelihood and reconstruction error, respectively, for the training and test sets of speaker RK and all the models studied. The curves for other speakers were also very similar[8]. Removing repeated EPG patterns from the data set (keeping at most one instance of every EPG pattern) did not affect considerably either the curves or the prototypes or factors found (the mixing proportions do get affected, being related to the prior distribution of the components).

[7] We realise that our comparisons here are of qualitative value only. However, making them quantitative would take us too far away, as it would require defining an index of *goodness of two-dimensional representation* and comparing our latent variable models with some of the other methods. A further obstacle is that our data are unlabelled.

[8] The ranking of the methods was virtually the same for all our six speakers. A Friedman test showed significant n-wise differences (at a significance level of 1%) between the effect of the methods. Although this test does not require normality of the population (being a nonparametric test), again this evidence should be only qualitative under the Bayesian viewpoint.

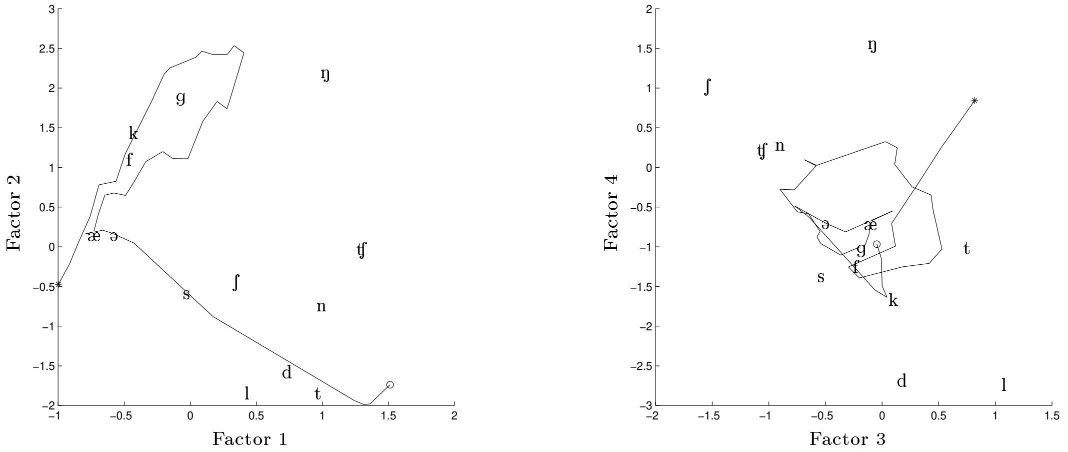

Figure 15: Two-dimensional plot of the trajectory of the utterance fragment “I prefer **Kant to** Hobbes for a good bedtime book” using factor analysis for speaker RK (left: latent space of factors 1 and 2; right: latent space of factors 3 and 4). The start and end points are marked as $*$ and $\circ$, respectively. The phonemes are located on the figure by projecting the prototypes of figure 3 using equation (10).

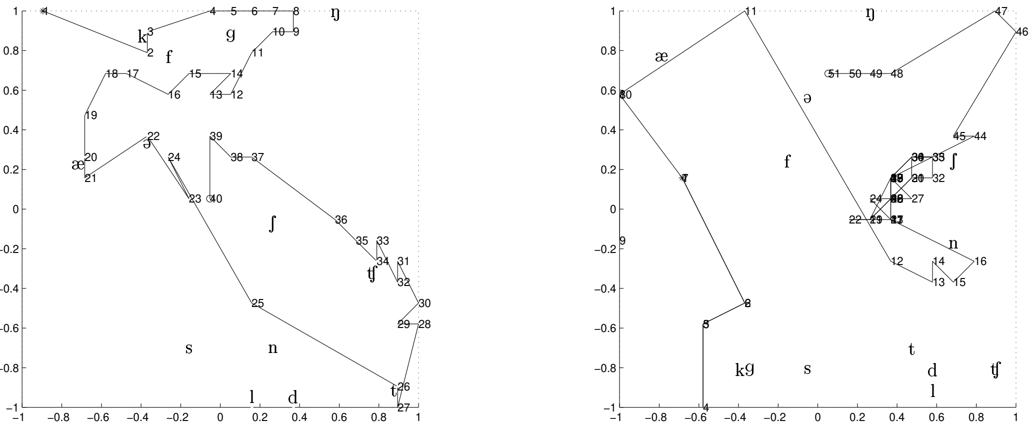

Figure 16: Two-dimensional plot of the trajectory of the utterance fragment “I prefer **Kant to** Hobbes for a good bedtime book” using GTM with $F = 49$ basis functions and a 20×20 latent space grid (left: speaker RK; right: speaker HD). The start and end points are marked as $*$ and $\circ$, respectively. Points are numbered correlatively. The phonemes are located on the figure by projecting the indices of figure 3 using the posterior mean of the GTM model.

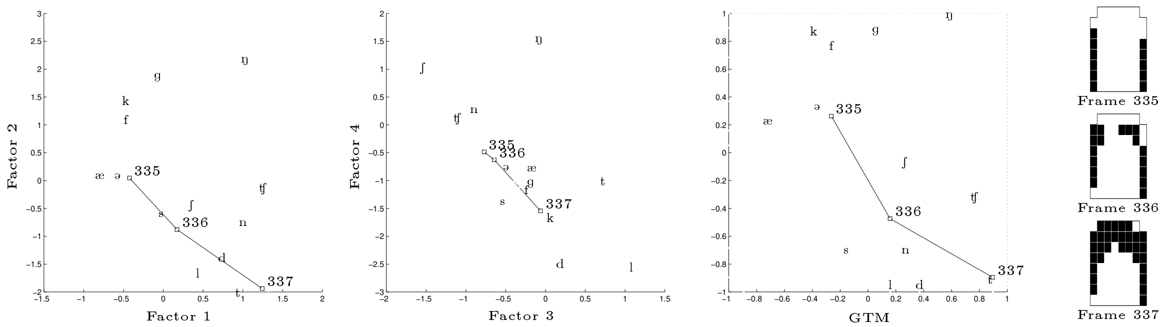

Figure 17: Discontinuities in the latent space due to discontinuities in the EPG sequence of the utterance fragment “I prefer **Kant to** Hobbes for a good bedtime book.” For speaker RK, the transition from /æ/ to /nt/ occurs from frame 335 to frame 336. The right part of the figure shows the three consecutive EPG frames. The plots on the left part show the corresponding trajectory in the two-dimensional map of factor analysis (factors 1 vs. 2, left; factors 3 vs. 4, centre) and GTM (right).

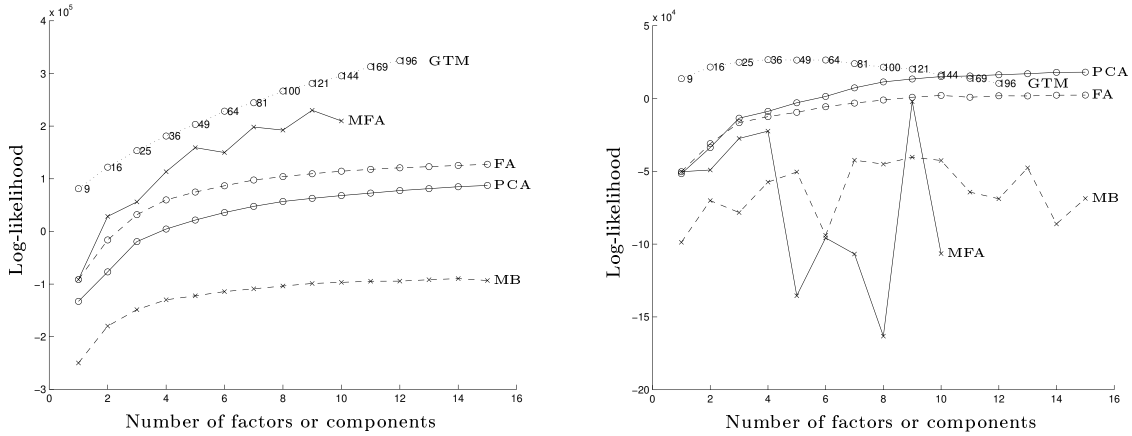

Figure 18: Comparison between methods in terms of log-likelihood for speaker RK (left: training set; right: test set): factor analysis (FA), principal component analysis (PCA), generative topographic mapping (GTM), mixtures of factor analysers (MFA) and mixtures of multivariate Bernoulli distributions (MB). Note that the x axis refers to the order of the factor analysis or principal component analysis, the number of mixture components in the case of mixture models and the square root of the number of basis functions in the case of GTM.

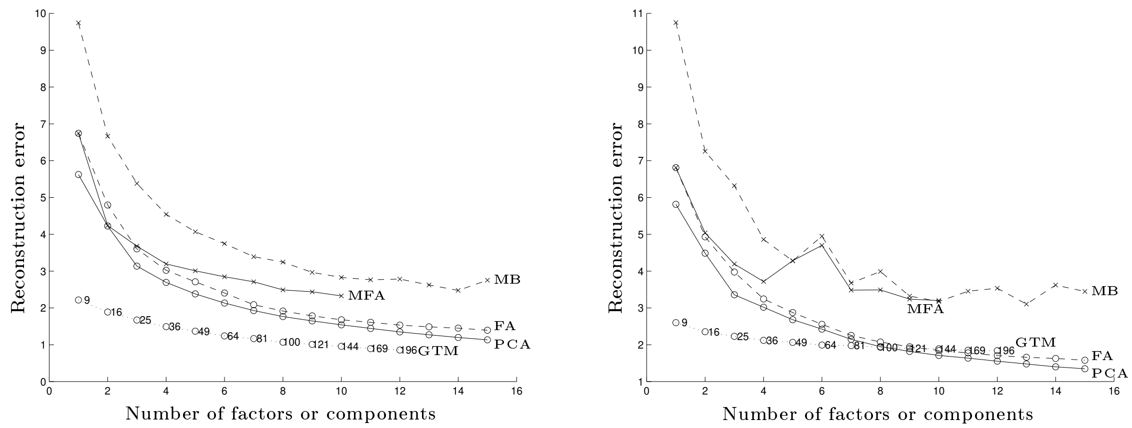

Figure 19: Comparison between methods in terms of reconstruction error for speaker RK (left: training set; right: test set): factor analysis (FA), principal component analysis (PCA), generative topographic mapping (GTM), mixtures of first order factor analysers (MFA) and mixtures of multivariate Bernoulli distributions (MB). Note that the x axis refers to the order of the factor analysis or principal components analysis, the number of mixture components in the case of mixture models and the square root of the number of basis functions in the case of GTM.

Our primary criterion is the log-likelihood in the test set, since it measures the generalisation power of the generative model. Reconstruction error is of secondary importance from a modelling perspective, since it treats equally the true data and the noise—although a good model in terms of likelihood will usually have a low reconstruction error as well.

For the cases studied, overfitting in the log-likelihood can appear if the number of parameters of the model is large enough (e.g. GTM in the right side of fig. 18), but the reconstruction error presents a steady decrease in both the training and test sets for any number of parameters (fig. 19).

Linear models (factor analysis and PCA) The training set curves confirm the theory: factor analysis outperforms PCA in log-likelihood (PCA being a particular case of factor analysis) while PCA outperforms factor analysis in reconstruction error (since PCA is the linear transformation with minimal squared reconstruction error). Interestingly, PCA outperforms factor analysis in log-likelihood in the test set for speakers RK and HD. We can conclude that for our data set the PCA model is almost as good as the factor analysis one. Both methods are computationally straightforward.

Mixtures of latent variable models have the advantage of providing with different latent variable models in different data space regions, thus adapting to local structure of the data. This often is a good modelling assumption, in that data are usually due to a number of different, concurrent processes. The mixtures of factor analysers, using a latent space of only one dimension, have a performance similar to that of factor analysis with several factors. This makes them promising for higher dimensions of the latent space. However, their practical application is limited as long as an algorithm which efficiently avoids singularities in the log-likelihood surface is not available. One could perhaps overcome this by constraining the domain of some parameters, although the question arises of how much to constrain them. More generally, one could use regularisation techniques (Girosi et al., 1996) in the form of a prior on the parameters. Not all mixtures are affected by this problem, either because of the nature of the likelihood surface (e.g. mixtures of multivariate Bernoulli distributions (Carreira-Perpiñán and Renals, 1998)) or because a direct algorithm to find maxima of it is available (e.g. mixtures of PCAs (Tipping and Bishop, 1997)).

Mixtures of multivariate Bernoulli distributions fare worse than all the other methods in all categories despite being the only model purely for binary data. Computationally they are also more costly than linear models. Their use for EPG data modelling is thus discouraged.

The generative topographic mapping (GTM), a nonlinear latent variable model, appears to be the best model in all categories. Taking the point of $F = 49$ radial basis functions (which gives the maximum likelihood in the test set), it clearly outperforms all the other models in log-likelihood and also reconstruction error (PCA needs a latent space of more than $L = 10$ dimensions to achieve a comparable reconstruction error). GTM achieves this with a latent space of only $L = 2$ dimensions, while the other methods were tested up to $L = 15$ (except the mixtures of factor analysers). The major drawback of GTM is its computational complexity, exponential in the dimension of the latent space, which limits it to $L = 2$ dimensions in practice. Nevertheless, two-dimensional visualisation can be of interest in many speech applications, as we noted in section 8.

9.2 Model validity

Validity of a priori indices The results from factor analysis show that:

- The similarity between the factors found and some of the linear data reduction indices proposed in the literature (Hardcastle et al., 1991b; Jones and Hardcastle, 1995), which have been designed using a priori assumptions, gives empirical support to the validity of the latter: such indices effectively capture a good deal of linear correlation in the data.
- However, the fact that the factors differ according to the EPG database employed indicates that fixed indices will not perform well on all databases (corresponding to different speakers, different speech styles, etc.). In order to capture as much structure as possible from a given database it is necessary to let the data speak for itself.

Model identifiability A class of models is identifiable if there exist no two different sets of values for the model parameters (apart from permutations and rotations) that produce exactly the same probability distribution. Interpreting models from an unidentifiable class may thus be difficult (Bartholomew, 1987; Everitt, 1984; Everitt and Hand, 1981). The following results are known:

- PCA is always identifiable except in the case of equality of eigenvalues of the sample covariance matrix.
- Factor analysis followed by (varimax) rotation is also identifiable.
- While mixtures of multivariate Bernoulli distributions are unidentifiable for all dimensions (Gyllenberg et al., 1994), estimation of the mixture parameters from different, random starting points can still produce meaningful results (Carreira-Perpiñán and Renals, 1998).
- To our knowledge, no identifiability results are known for either mixtures of factor analysers or GTM.

Model additivity In section 3.1.2 we introduced the concept of model additivity for PCA and factor analysis, and we found that a similar phenomenon was observed for mixtures of multivariate Bernoulli distributions. There we claimed that often—whether the model is identifiable (e.g. factor analysis) or not (e.g. mixtures of multivariate Bernoulli distributions)—the factors or prototypes found in a model of order L are very approximately those of a model of order $L-1$ plus a new one. Here we add some more facts, based in our experience with the data sets analysed in this paper:

- Sometimes one does not obtain a new, stable factor or prototype but an intermediate one which will unfold into two new, stable ones in higher-order models.
- Past a certain order the models begin to produce repeated versions of some of the prototypes (typically the more relevant ones). This indicates that that order is probably too high and that a smaller one should be tried.
- While standard PCA is additive (the principal components being the eigenvectors of the sample covariance matrix ordered by the magnitude of the associated eigenvalue), PCA followed by varimax rotation is not. The components change considerably between adjacent orders.

In our data sets the phenomenon of additivity holds, but we do not have a theoretical basis for a general situation.

9.3 Intrinsic dimensionality of the EPG data

All the models described need to know in advance either the dimension of the latent space L or the number of components in the mixture M. This is a common disadvantage of most dimensionality reduction methods and is directly related to the determination of the intrinsic dimensionality of the EPG data. That is, the number of degrees of freedom of the articulatory system that generates the EPG patterns: the tongue and the hard palate. This is a matter currently being investigated (e.g. see (Stone, 1991; Nguyen et al., 1994; Nguyen et al., 1996) and references therein); a dimension of 5 to 10 is usually suggested. Alternatively, one can try to infer this intrinsic dimension directly from the data set by standard statistical techniques of model selection.

In the models analysed in this paper, one can often get a rough idea of what the appropriate number of latent variables or components is by examining the factors or prototypes found. For example, if in a mixture of multivariate Bernoulli distributions several prototypes look very similar (e.g. see fig. 13), it is obvious that one should try a mixture with fewer components even if the log-likelihood decreases a bit. Also, plotting the log-likelihood in the training and test sets versus the number of latent variables or components can help to avoid overfitting by selecting L or M such that the log-likelihood with the test set is maximal.

The fact that the dimension of the latent space or the number of components found in any of these ways may be optimal with respect to the model does not guarantee that it actually matches the intrinsic dimensionality of the EPG data, because the model may be bad for this data. Nevertheless, the two-dimensional representation produced by GTM and its performance in log-likelihood and reconstruction error (compared to that of the other methods, which use more latent variables) suggest that the intrinsic dimensionality of the EPG data may be substantially smaller than that claimed by other studies.

Determining the intrinsic dimensionality of the EPG data would be a significant step towards the solution of the acoustic-to-articulatory mapping and the construction of a realistic mechanical model of the vocal tract. However, a purely phenomenological, or black-box, approach—where all hope for finding a simple model is abandoned—may be more fruitful from a practical point of view. That is, a low-dimensional model of high complexity (with a high number of parameters) may offer no help to understand the physical behaviour of the articulatory system yet be able to mimic it to a desired level of accuracy.

10 Conclusion

We have shown that latent variable models, owing to their ability to produce a reduced-dimension representation of a data set, can be useful to present the relatively high-dimensional information contained in the EPG sequence in a way which is easier to understand and handle, e.g.:

- The loading vectors and principal component vectors of factor analysis and PCA, respectively, can be used as conventional EPG data reduction indices, with the advantage of adapting better to the database under consideration.
- A two-dimensional latent space representation can be used for the analysis of EPG sequences in several applications, such as speech therapy, language learning and articulatory dynamics.

This is a significant advantage over general probability models and mixtures of them, which are suitable for classification or clustering but not for dimensionality reduction.

The superior performance of a two-dimensional GTM model over the other methods indicates that the EPG data may be appropriately accounted for with a very small number of degrees of freedom as long as a nonlinear, complex model is used.

Acknowledgments

This work was supported by a scholarship from the Spanish Ministry of Education and Science, by ESPRIT Long Term Research Project SPRACH (20077) and by an award from the Nuffield Foundation. The authors acknowledge Markus Svensén and Zoubin Ghahramani for the Matlab implementations of GTM and the mixtures of factor analysers, respectively, and Alan Wrench for providing them with the ACCOR data. Matlab implementations of principal component analysis, factor analysis (and varimax rotation) and mixtures of multivariate Bernoulli distributions are available from the first author at URL

`http://www.dcs.shef.ac.uk/~miguel/research/epg.html`.

A Notation

We denote scalars in italics (m) and vector and matrix quantities in boldface ($\mathbf{x}$, $\mathbf{\Lambda}$). $\mathbf{I}$ is the identity matrix. $\delta(\mathbf{x})$ is the Dirac delta function. $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ is the multivariate normal distribution of mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$. $\|\mathbf{x}\|_2^2 = \sum_{i=1}^{L} x_i^2$ is the squared Euclidean, or sum of squares, or L_2 norm. $\mathrm{E}\{\mathbf{x}\} = \int \mathbf{x}p(\mathbf{x})\,d\mathbf{x}$ is the mean of a (continuous) distribution.

B Sentences from the EUR-ACCOR database employed in the experiments, with phonemic transcription

Detailed annotations of some of the utterances can be found in (Nicolaidis and Hardcastle, 1994).

1. The hostess should always wear clean gloves.
2. When the song's the thing, the cast can't sing.
3. All the keys have been handed out, Bill.
4. I prefer Kant to Hobbes for a good bedtime book.
   /aɪ prɪ'fɜ 'kænt tə 'hɒbz fər ə 'gʊd 'bedtaɪm 'bʊk/
5. Fred can go, Susan can't go, and Linda is uncertain.
6. We tore down the outbuildings.
7. Bella ran past the school, then came towards us.
8. The cold front is expected to pass this way on Sunday.
9. The catalogue lists just his own brand of tyre.
10. Put your hat on the hatrack and your coat in the cupboard.
11. She climbed up to see behind the clock.
12. It's a good thing he's not just fooling around.
13. Stats by such a student will surely be trashed after tests.
14. The lads showed us a strategy for trimming his moustache surreptitiously.

References

Baldi, P. and Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2:53–58.

Bartholomew, D. J. (1987). *Latent Variable Models and Factor Analysis.* Charles Griffin & Company Ltd., London.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer Series in Statistics. Springer-Verlag, Berlin, second edition.

Bishop, C. M., Svensén, M., and Williams, C. K. I. (1998). GTM: The generative topographic mapping. *Neural Computation*, 10(1):215–234.

Bourlard, H. and Kamp, Y. (1988). Autoassociation by the multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, 59:291–294.

Byrd, D., Flemming, E., Mueller, C. A., and Tan, C. C. (1995). Using regions and indices in EPG data reduction. *Journal of Speech and Hearing Research*, 38(4):821–827.

Carreira-Perpiñán, M. Á. (1997). Density networks for dimension reduction of continuous data: Analytical solutions. Technical Report CS–97–09, Dept. of Computer Science, University of Sheffield.

Carreira-Perpiñán, M. Á. and Renals, S. J. (1998). Practical identifiability of finite mixtures of multivariate Bernoulli distributions. Submitted to *Neural Computation.*

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39(1):1–38.

Everitt, B. S. (1984). *An Introduction to Latent Variable Models.* Monographs on Statistics and Applied Probability. Chapman & Hall, London, New York.

Everitt, B. S. and Hand, D. J. (1981). *Finite Mixture Distributions.* Monographs on Statistics and Applied Probability. Chapman & Hall, London, New York.

Ghahramani, Z. and Hinton, G. E. (1996). The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, University of Toronto.

Girosi, F., Jones, M., and Poggio, T. (1996). Regularization theory and neural networks architectures. *Neural Computation*, 7:219–269.

Gyllenberg, M., Koski, T., Reilink, E., and Verlaan, M. (1994). Non-uniqueness in probabilistic numerical identification of bacteria. *J. Appl. Prob.*, 31:542–548.

Hardcastle, W. J., Gibbon, F. E., and Jones, W. (1991a). Visual display of tongue-palate contact: Electropalatography in the assessment and remediation of speech disorders. *Brit. J. of Disorders of Communication*, 26:41–74.

Hardcastle, W. J., Gibbon, F. E., and Nicolaidis, K. (1991b). EPG data reduction methods and their implications for studies of lingual coarticulation. *J. of Phonetics*, 19:251–266.

Hardcastle, W. J., Jones, W., Knight, C., Trudgeon, A., and Calder, G. (1989). New developments in electropalatography: A state-of-the-art report. *J. Clinical Linguistics and Phonetics*, 3:1–38.

Holst, T., Warren, P., and Nolan, F. (1995). Categorising [s], [ʃ] and intermediate electropalographic patterns: Neural networks and other approaches. *European Journal of Disorders of Communication*, 30(2):161–174.

Huber, P. J. (1985). Projection pursuit. *Annals of Statistics*, 13(2):435–475 (with comments, pp. 475–525).

Jones, W. and Hardcastle, W. J. (1995). New developments in EPG3 software. *European Journal of Disorders of Communication*, 30(2):183–192.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200.

Kohonen, T. K. (1995). *Self-Organizing Maps.* Number 30 in Springer Series in Information Sciences. Springer-Verlag, Berlin.

Marchal, A. and Hardcastle, W. J. (1993). ACCOR: Instrumentation and database for the cross-language study of coarticulation. *Language and Speech*, 36(2, 3):137–153.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis.* Probability and Mathematical Statistics Series. Academic Press, New York.

Nguyen, N. (1995). EPG bidimensional data reduction. *European Journal of Disorders of Communication*, 30:175–182.

Nguyen, N., Hoole, P., and Marchal, A. (1994). Regenerating the spectral shape of [s] and [ʃ] from a limited set of articulatory parameters. *J. Acoustic Soc. Amer.*, 96(1):33–39.

Nguyen, N., Marchal, A., and Content, A. (1996). Modeling tongue-palate contact patterns in the production of speech. *J. of Phonetics*, 24:77–97.

Nicolaidis, K. and Hardcastle, W. J. (1994). Articulatory-acoustic analysis of selected English sentences from the EUR-ACCOR corpus. Technical report, SPHERE (Human capital and mobility program).

Park, J. and Sandberg, I. W. (1993). Approximation and radial-basis-function networks. *Neural Computation*, 5(2):305–316.

Pratt, S., Heintzelman, A. T., and Ensrud, D. S. (1993). The efficacy of using the IBM Speech Viewer vowel accuracy module to treat young children with hearing impairment. *Journal of Speech and Hearing Research*, 29:99–105.

Rubin, D. B. and Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76.

Schroeter, J. and Sondhi, M. M. (1994). Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Trans. Speech and Audio Process.*, 2(1):133–150.

Stone, M. (1991). Toward a model of three-dimensional tongue movement. *J. of Phonetics*, 19:309–320.

Tipping, M. E. and Bishop, C. M. (1997). Mixtures of principal component analysers. In *Proceedings of the IEE Fifth International Conference on Artificial Neural Networks*. London:IEE.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, 5:329–350.