In: Proc. of the 14th International Congress of Phonetic Sciences (ICPhS99),
San Francisco, USA, 1-7 August 1999.
URL: http://www.dcs.shef.ac.uk/~miguel/papers/icphs99.html

A LATENT VARIABLE MODELLING APPROACH TO THE ACOUSTIC-TO-ARTICULATORY MAPPING PROBLEM

Miguel Á. Carreira-Perpiñán and Steve Renals
Dept. of Computer Science, University of Sheffield, UK

ABSTRACT

We present a latent variable approach to the acoustic-to-articulatory mapping problem, where different vocal tract configurations can give rise to the same acoustics. In latent variable modelling, the combined acoustic and articulatory data are assumed to have been generated by an underlying low-dimensional process. A parametric probabilistic model is estimated and mappings are derived from the respective conditional distributions. This has the advantage over other methods, such as articulatory codebooks or neural networks, of directly addressing the nonuniqueness problem. We demonstrate our approach with electropalatographic and acoustic data from the ACCOR database.

1. INTRODUCTION

Recently, there has been a growing awareness in the speech recognition field that blind science based on acoustical information alone might not tend to more improvements of state of the art computational methods of speech production [5]. Blind science means here the use of complex generic statistical models (e.g., hidden Markov models or neural networks) that could be used for the description of any physical phenomenon because the strong assumptions that they make can usually be overcome by having a large number of parameters and of training data. While these methods can be applied successfully to a wide range of phenomena, there seems to be a limit on their speech recognition performance inasmuch as they are solely acoustic models.

Incorporating additional knowledge about the speech process can be done in several ways. For example, the hidden dynamic model of [8] adds some constraints derived from the mechanics of the articulatory system, in the form of a Kalman filter. In our approach, we simply use acoustic feature vectors augmented with articulatory components and let the latent variable model infer a hidden representation without enforcing any external constraints. Smoothness is an integral part of our model, which therefore does not require any ad-hoc module.

While this approach is very general and can be applied to speech recognition (e.g. by using the latent variables as features), in this paper we concentrate on its suitability to the acoustic-to-articulatory mapping problem.

2. INVERSE PROBLEMS

An inverse problem occurs when there is a one-to-many association between variables. Inverse problems pose great theoretical and computational difficulties. An example is the problem of the acoustic-to-articulatory mapping. It is well known that, while given a time sequence of vocal tract configurations there is a unique output acoustic signal, the converse is not true: multiple vocal tract configurations can produce a given acoustic signal [9]. A number of approaches have been applied with limited success, including articulatory codebooks and neural networks.

Probabilistic models offer a significant advantage over these methods in that it is possible to obtain many-to-one mappings between the variables being modelled. By training the model with both acoustic and articulatory data, it is then possible to compute conditional distributions of the acoustic variables given the articulatory ones and vice versa. A one-to-one mapping will result in a unimodal distribution, while a many-to-one mapping will result in a multimodal one. In both cases, the modes of the conditional distribution can be taken as the solution to inverting the mapping.

In this paper, the articulatory variables will be electropalatographic (EPG) frames, rather than the positions of the different articulators, due to the lack of a more appropriate data set. In this case, the mapping EPG $\rightarrow$ acoustic signal is many-to-one (since different phonemes may correspond to the same EPG frame).

3. GENERATIVE MODELLING USING LATENT VARIABLES

In latent variable modelling the assumption is that the observed high-dimensional data $\mathbf{t}$ is generated from an underlying low-dimensional process defined by a small number L of *latent variables* $\mathbf{x}$ [1]. The latent variables are mapped by a fixed transformation into a D-dimensional data space and noise is added there. The aim is to learn the low dimensional generating process along with a noise model, rather than directly learning a dimensionality reducing mapping. Note that the low-dimensional representation is abstract and may not necessarily be interpretable in terms of any physical variables. Latent variable models have been applied to data reduction of EPG data in [4].

A latent variable model is specified by a prior in latent space $p(\mathbf{x})$, a smooth mapping $\mathbf{f}$ from latent space to data space and a noise model in data space $p(\mathbf{t}|\mathbf{x})$. These three elements are equipped with parameters which we collectively call $\mathbf{\Theta}$. Integrating the joint probability density function $p(\mathbf{t},\mathbf{x})$ over latent space gives the marginal distribution in data space, $p(\mathbf{t})$. Given an observed sample in data space $\{\mathbf{t}_n\}_{n=1}^N$ of N D-dimensional real vectors that has been generated by an unknown distribution, a parameter estimate can be found by maximising the log-likelihood of the parameters $l(\mathbf{\Theta}) = \sum_{n=1}^N \log p(\mathbf{t}_n|\mathbf{\Theta})$, typically using an EM algorithm.

We consider the following latent variable models, for which EM algorithms are available:

Factor analysis [1], in which the mapping is linear, the prior in latent space is unit Gaussian and the noise model is diagonal Gaussian. The marginal in data space is then Gaussian with a constrained covariance matrix.

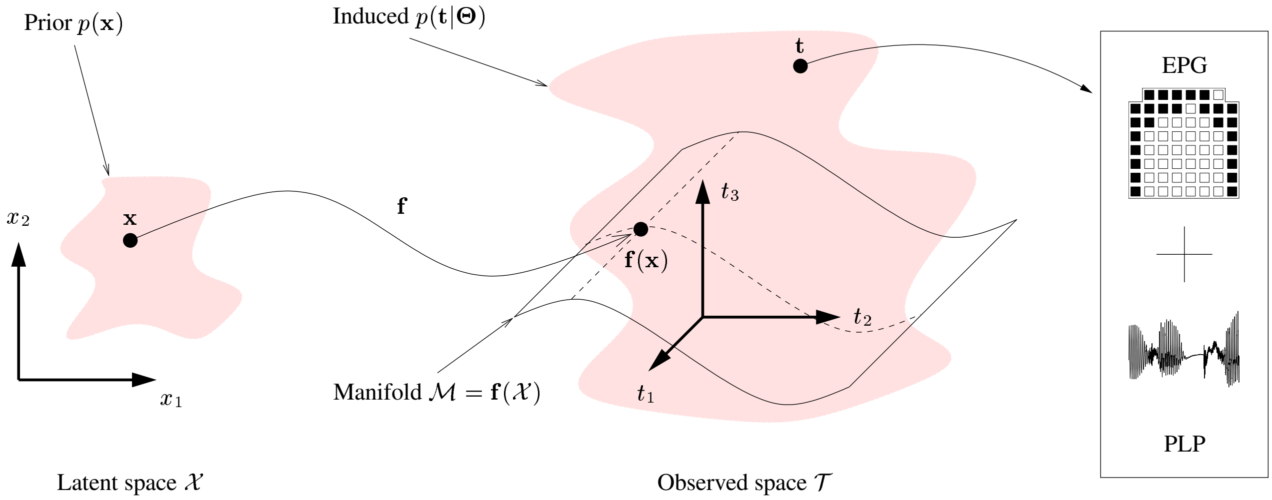

Figure 1: Schematic of a latent variable model where the observed data consists of EPG patterns and PLP coefficients.

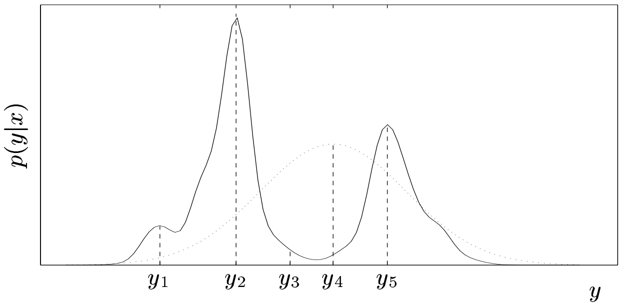

Figure 2: A unimodal (dotted line) and a multimodal conditional distribution (solid line). The vertical, dashed lines mark the modes and the means.

Principal component analysis (PCA), which can be considered a particular case of factor analysis with isotropic Gaussian noise model [10].

The generative topographic mapping (GTM) [2] is a nonlinear latent variable model, where the mapping is a generalised linear model, the prior in latent space is discrete uniform and the noise model is isotropic Gaussian. The marginal in data space is then a constrained mixture of Gaussians.

3.1. Regression in latent variable models

Once the latent variable model has been trained using data from the observed space (that is, vectors consisting of both acoustic and articulatory information), we have a probabilistic model $p(\mathbf{x}, \mathbf{t})$ for all the variables of interest. For simplicity of notation, we omit the dependence on the parameters and the model. Using the standard operations of marginalisation and conditioning, it is possible to obtain the distributions of any variable(s) with respect to any other variable(s). For example, to find the distribution in latent space, we compute the posterior distribution of the latent variables with respect to the observed ones,

$$p(\mathbf{x}|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{t})},$$

which leads to dimensionality reduction and has been investigated in [4]. Here, we construct conditional distributions of the form $p(\mathbf{t}_{\mathcal{J}}|\mathbf{t}_{\mathcal{I}})$ where $\mathcal{I}, \mathcal{J} \in \{1, \dots, D\}$ are sets of indices and D is the dimensionality of the observed space. For example, if $\mathcal{I} = \{1, 7, 3\}$, then $\mathbf{t}_{\mathcal{I}} = (t_1 t_7 t_3)$. From a conditional distribution $p(\mathbf{t}_{\mathcal{J}}|\mathbf{t}_{\mathcal{I}})$ it is possible to construct a functional relationship $\mathbf{t}_{\mathcal{J}} = \mathbf{f}(\mathbf{t}_{\mathcal{I}})$ provided that the entropy of this conditional distribution is low. That is, given $\mathbf{t}_{\mathcal{I}}$, only a small region of the space of $\mathbf{t}_{\mathcal{J}}$ should have nonnegligible probability mass: $p(\mathbf{t}_{\mathcal{J}}|\mathbf{t}_{\mathcal{I}})$ is sharply peaked. To derive a functional relationship $y = f(x)$ from a conditional distribution $p(y|x)$, one can take a point that conveniently summarises the information contained in $p(y|x)$, e.g., the mean or the mode(s). If $p(y|x)$ is unimodal (like the dotted curve in fig. 2), the mean will usually be near the mode (value y_4). But if $p(y|x)$ is multimodal (like the solid curve in fig. 2), then each mode is potentially a valid solution (values y_1, y_2, y_5), while the mean (value y_3) may be a misleading estimate if it lies in a low-probability area (however, on the average it may have a lower reconstruction error, as we show below). This is the case with *inverse problems*, such as the acoustic-to-articulatory mapping, where one-to-many mappings may appear. The advantage of probabilistic methods is that in principle it is possible to find all possible values of y as a function of x (for a given value of x) by locating all the modes of $p(y|x)$. For the latent variable models investigated here, the distribution $p(\mathbf{t})$ in observed space is either Gaussian (factor analysis, PCA) or a mixture of isotropic Gaussians (GTM), and so $p(\mathbf{t}_{\mathcal{J}}|\mathbf{t}_{\mathcal{I}})$ is again Gaussian or a Gaussian mixture, respectively. The Gaussian case offers no problem as the mean coincides with the mode and the distribution is unimodal. For Gaussian mixtures, it is possible to find all the modes efficiently using gradient ascent combined with quadratic optimisation starting from each centroid [3]. Spurious modes may be discarded if their probability is lower than a suitable threshold. Additionally, it is possible to obtain error bars (i.e., a confidence interval) at each mode by locally approximating the density function by a normal distribution. However, if the dimensionality of $\mathbf{t}_{\mathcal{J}}$ is high, the error bars become very wide due to the curse of the dimensionality.

4. PREDICTION OF PLP COEFFICIENTS AND EPG PATTERNS

To demonstrate the ability for regression of latent variable models, we trained three different models using a small subset of the ACCOR database [7]. This database, designed for the cross-language study of coarticulation, contains electropalatography and acoustic measurements (among other measurements) for utterances in different European languages and varying speech styles (slow, fast, etc.). We selected the utterance "Put your hat on the hatrack and your coat in the cupboard" for speaker FG and computed from its acoustic waveform 12th-order PLP coefficients [6] plus the log-energy at 200 Hz. The EPG data consists of 62-bit frames sampled at 200 Hz, which we consider as 62-dimensional vectors of real numbers, with components indexed from 1 (top left) to 62 (bottom right). Thus, the resulting sequence consisted of over 600 75-dimensional real vectors, and was split into a training (80%) and a test (20%) set. All the data used were unlabelled.

The models trained were factor analysis (FA) with 9 factors (= dimensionality of latent space), principal component analysis (PCA) with 9 principal components (= dimensionality of latent space) and the generative topographic mapping (GTM) with a latent space of dimension 2 and a 20×20 grid.

In fig. 3 we used GTM to reconstruct parts of the EPG frame given other parts of it. Note how the reconstructed pattern is slightly different when the left half is given than when the right half is given, revealing asymmetry in the tongue movement. When the bottom half is given, the distribution for the top half happens to be multimodal, with several patterns (corresponding to open vowels and alveolars) becoming possible. Determining which one of these patterns is the true one would require additional information, such as the previous EPG frame in the sequence or the PLP coefficients in this frame.

Tables 1 and 2 show results for the reconstruction error of the EPG frame given the PLP coefficients and vice versa. For the linear-normal models (FA and PCA), the conditional distribution is always normal and so the only point to consider to reconstruct the vector is the conditional mean (equal to the conditional mode). For GTM, we tried three possibilities: the conditional mean; the global conditional mode (g-mode), i.e., the mode with highest probability; and the conditional mode closest to the vector to be reconstructed (c-mode). The c-mode gives a lower bound for the error using any kind of mode, but it is unknown a priori. We use it here to see how good the modes can perform over the mean.

The average error $E_\delta = \frac{1}{N}\sum_{n=1}^{N} \delta(\hat{\mathbf{t}}_n - \mathbf{t}_n)$, where $\mathbf{t}_n$ is the true vector and $\hat{\mathbf{t}}_n$ the reconstructed one, was computed for two different distances: the quadratic norm $\delta(\mathbf{t}) = \|\mathbf{t}\|_2^2 = \sum_{d=1}^{D} t_d^2$ and the maximum norm $\delta(\mathbf{t}) = \|\mathbf{t}\|_\infty = \max_{d=1,\dots,D} |t_d|$.

Regarding method comparison, the tables show that GTM attains the smallest error of any kind in almost all cases, while PCA has the highest error, at not much distance from FA. The reason for PCA performing worse than FA may be the fact that PCA assumes an isotropic error noise, but our variables have different ranges and variances (e.g. the EPG variables are in $[0,1]$ while the log-energy is in $[1,5]$ approximately). GTM also uses an isotropic error noise (although it is possible to extend it to a general diagonal noise model) but it compensates by having a nonlinear mapping. Therefore, the use of a nonlinear mapping is of paramount importance, as a nonlinear model with a latent space of dimension 2 can outperform linear methods with latent dimensionality of 9.

Regarding the use of the mean or a mode, the average reconstruction errors for GTM show that the mean performs better than the global mode in most cases but worse than the closest mode in all cases. We employed a fourth strategy, not shown in the tables, where the mean is used if the conditional distribution is unimodal and the global mode if it is multimodal. This gave an error virtually equal to that of the global mode, which indicates that the divergence occurs in multimodal conditional distributions. We conclude that the best predictor is one of the modes, but not necessarily the mode with the highest probability. Without any additional information it is not possible to tell a priori which mode is the best.

5. DISCUSSION

We have presented a class of probabilistic models which is able, by constructing conditional distributions and deriving from them functional relationships, to perform dimensionality reduction and multivariate regression for one-to-many mappings. We have showed its applicability to inverse problems, such as the acoustic-to-articulatory mapping. Both acoustic and articulatory data are used for training, but only the acoustics are necessary for testing.

A regression can be seen as missing data imputation, where given the present values $\mathbf{t}_{\mathcal{I}}$, the missing values $\mathbf{t}_{\mathcal{J}}$ are to be filled in using the knowledge of the distribution $p(\mathbf{t}_{\mathcal{J}}|\mathbf{t}_{\mathcal{I}})$. Thus, the same formalism applies to missing data imputation.

Our model does not currently contemplate the dynamic character of speech, which is necessary to predict the articulatory variables from the acoustic ones. However, continuity constraints may be applied in the latent space, so that of all the conditional modes only those that give a continuous trajectory in latent space are selected.

A potential problem of this approach is that the dimension of the latent space has to be fixed in advance, although an optimal one could be found by cross-validation. An additional problem of methods that sample the latent space, like GTM, is that their computational cost grows exponentially with the dimension of the latent space.

6. REFERENCES

[1] David J. Bartholomew. *Latent Variable Models and Factor Analysis*. Charles Griffin & Company Ltd., London, 1987.

[2] Christopher M. Bishop, Markus Svensén, and Christopher K. I. Williams. GTM: The generative topographic mapping. *Neural Computation*, 10(1):215–234, January 1998.

[3] Miguel Á. Carreira-Perpiñán. Mode-finding in Gaussian mixtures. Technical Report CS–99–03, Dept. of Computer Science, University of Sheffield, March 1999.

[4] Miguel Á. Carreira-Perpiñán and Steve Renals. Dimensionality reduction of electropalatographic data using latent variable models. *Speech Communication*, 26(4):259–282, December 1998.

[5] Li Deng. A dynamic, feature-based approach to the interface between phology and phonetics for speech modeling and recognition. *Speech Communication*, 24(4):299–323, July 1998.

[6] Hynek Hermansky. Perceptual linear predictive (PLP) analysis of speech. *J. Acoustic Soc. Amer.*, 87(4):1738–1752, April 1990.

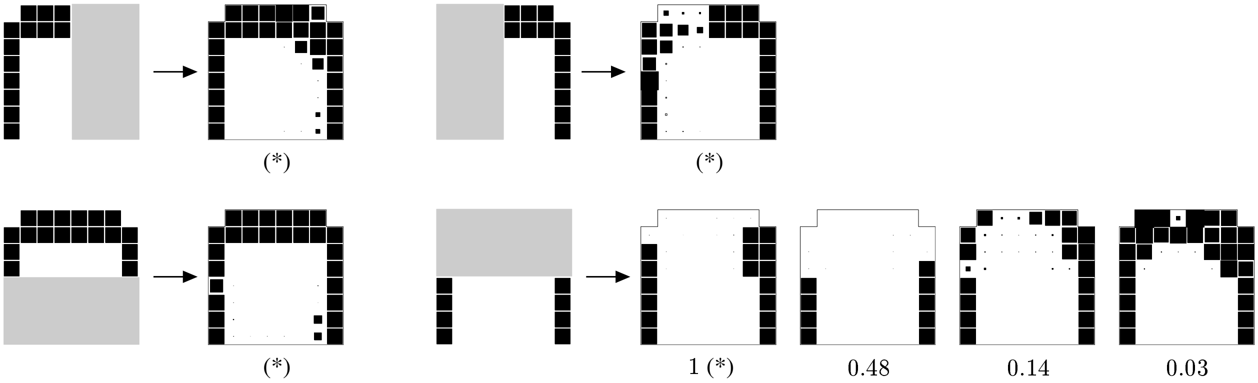

Figure 3: Use of the conditional distribution modes to predict, or reconstruct, variables in observed space. Here, we use the GTM model to compute the distribution of the EPG part greyed out (the unknown values) conditional on the EPG part which is not greyed out (the known values). The modes are given to the right of the arrow, labelled with their normalised probability if there is more than one mode. In all four cases, the mean (marked *) coincided approximately with one of the modes. Note the left-right asymmetry.

	Average quadratic error $\|\cdot\|_2^2$					Average maximum error $\|\cdot\|_\infty$				
	Factor analysis	PCA	GTM mean	GTM g-mode	GTM c-mode	Factor analysis	PCA	GTM mean	GTM g-mode	GTM c-mode
Training set	3.7635	3.8099	2.2736	2.8681	0.9466	0.7418	0.7497	0.5922	0.6155	0.3849
Test set	3.5060	3.5396	2.7667	3.5012	1.4830	0.7210	0.7283	0.6603	0.7043	0.4442

Table 1: Average reconstruction error of the EPG patterns given the PLP coefficients.

	Average quadratic error $\|\cdot\|_2^2$					Average maximum error $\|\cdot\|_\infty$				
	Factor analysis	PCA	GTM mean	GTM g-mode	GTM c-mode	Factor analysis	PCA	GTM mean	GTM g-mode	GTM c-mode
Training set	0.8870	1.2113	0.5777	0.6218	0.4230	0.6519	0.7532	0.4154	0.4085	0.3103
Test set	0.8632	1.2383	0.7967	0.9059	0.6147	0.6364	0.7314	0.5065	0.5324	0.3971

Table 2: Average reconstruction error of the PLP coefficients given the EPG patterns.

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

[8] Hywel B. Richards and John S. Bridle. The HDM: A segmental hidden dynamic model of coarticulation. In *Proc. ICASSP '99*, Phoenix, Arizona, USA, May 1999.

[9] Juergen Schroeter and Man Mohan Sondhi. Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Trans. Speech and Audio Process.*, 2(1):133–150, January 1994.

[10] Michael E. Tipping and Christopher M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2):443–482, February 1999.