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

{miguel,sjr}@dcs.shef.ac.uk   http://www.dcs.shef.ac.uk/~miguel

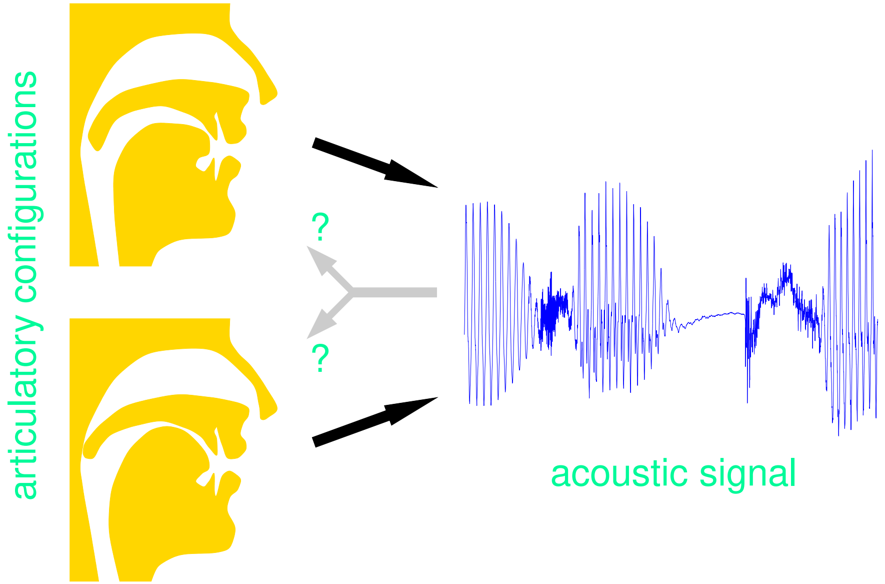
---

The 14th International Congress of Phonetic Sciences (ICPhS99)

San Francisco, USA, 1–7 August 1999

The acoustic-to-articulatory mapping problem

The positions, speeds, etc. of supraglottal mechanical elements of the vocal tract (tongue, jaw . . .) are called articulatory variables and are continuous functions of the time.



- ▷ A vocal tract configuration produces a unique acoustic signal. Thus the mapping articulatory $\longrightarrow$ acoustic is univalued (forward problem).
- ▷ An acoustic signal can be produced by different vocal tract configurations. The mapping acoustic $\longrightarrow$ articulatory is multivalued (inverse problem).

Approaches:

- ▷ Dynamic programming search in a large articulatory codebook: best performance, but very slow.
- ▷ Neural network (trained with the codebook): faster and more compact, but worse performance.
- ▷ Carefully prepared assembly of neural networks: approaches codebook performance and is fast.

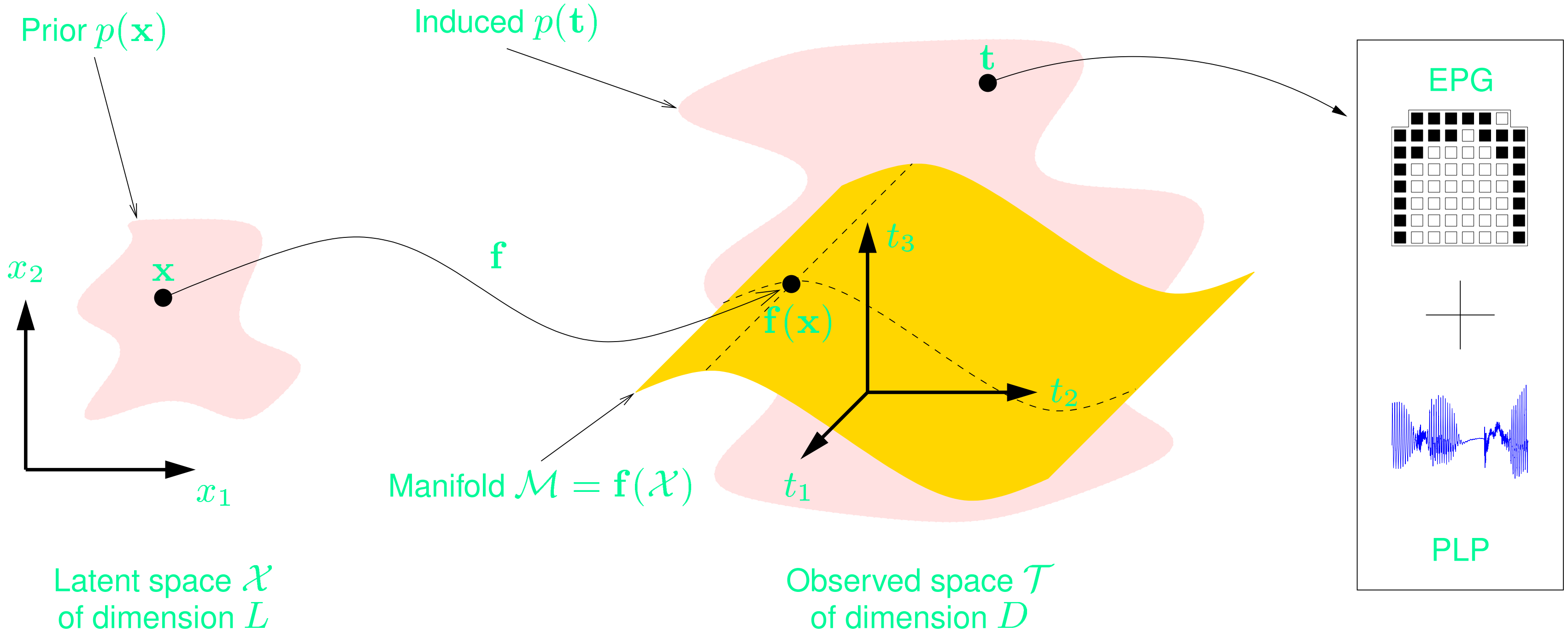
Joint density modelling of acoustic and articulatory variables

- Traditional mapping approximators (e.g. neural networks) cannot deal well with:
  - the stochastic character of real data
  - multivalued mappings (the nonuniqueness problem)

  . . . for which probabilistic models are more suitable.
- We propose:
  - a joint density model for the acoustic and articulatory variables
  - construction of the acoustic-to-articulatory mapping from a conditional distribution
  - use of continuity constraints (via dynamic programming) to solve the ambiguity.
- The correlations in the joint data (acoustic & articulatory) imply a low intrinsic dimensionality
- . . . which suggests using a latent variable model for the joint variables.

Latent variable models

Aim: to infer a low-dimensional representation of an observed, high-dimensional process.



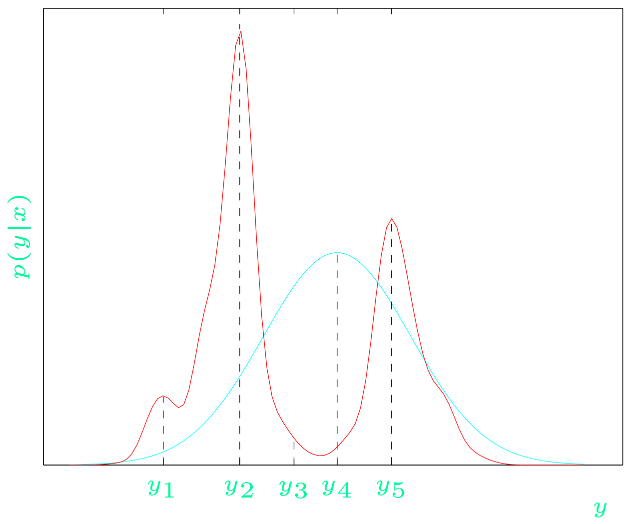
The data distribution in observed space $\mathcal{T}$ is modelled using a low-dimensional representation in latent space $\mathcal{X}$. In the figure the observed space consists of 62-dimensional EPG patterns plus 12-dimensional PLP coefficients, thus $D = 74$.

Latent variable models (cont.)

- Latent space prior $p(\mathbf{x})$, mapping $\mathbf{f}(\mathbf{x})$, noise model in observed space $p(\mathbf{t}|\mathbf{x})$: parameters $\boldsymbol{\Theta}$.
- Marginalisation in latent space (often difficult): $p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{x})p(\mathbf{x})\,d\mathbf{x}$.
- Maximum likelihood parameter estimation from sample $\{\mathbf{t}_n\}_{n=1}^N$, usually via an EM algorithm: $\arg\max_{\boldsymbol{\Theta}} l(\boldsymbol{\Theta}) = \sum_{n=1}^N \log p(\mathbf{t}_n|\boldsymbol{\Theta})$.
- Dimensionality reduction mapping via posterior distribution in latent space: $p(\mathbf{x}|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{t})}$.
- The latent dimension L must be fixed in advance—could use model selection.

Latent variable model	Prior in latent space $p(\mathbf{x})$	Mapping $\mathbf{f}$	Noise model $p(\mathbf{t}|\mathbf{x})$	Density in observed space $p(\mathbf{t})$
Factor analysis (FA)	$\mathcal{N}(\mathbf{0},\mathbf{I})$	linear	diagonal normal	constrained Gaussian
Principal component analysis (PCA)	$\mathcal{N}(\mathbf{0},\mathbf{I})$	linear	spherical normal	constrained Gaussian
Generative topographic mapping (GTM)	discrete uniform	generalised linear model	spherical normal	constrained Gaussian mixture

Deriving a functional relationship from a conditional distribution



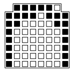
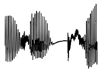
- Consider a conditional distribution $p(y|x = x_0)$.
- In principle, each solution of $y = f^{-1}(x_0)$ should correspond to a mode of $p(y|x = x_0)$.
- If the entropy is low (i.e., the distribution is very informative), we can pick representative points of $p(y|x)$:
  - Single choice (univalued mapping): pick the mean.
  - Multiple choice (multivalued mapping): pick all the modes.
- The correct mode at each point is selected using a continuity constraint minimised by dynamic programming.
- For unimodal, symmetric distributions (e.g. factor analysis) nothing of this matters!

Our problem: prediction of PLP coefficients and EPG patterns

Two characteristics of the EPG variables:

- They are an incomplete representation of the vocal tract.
- They are binary—but temporal continuity as proximity between consecutive frames still applies.

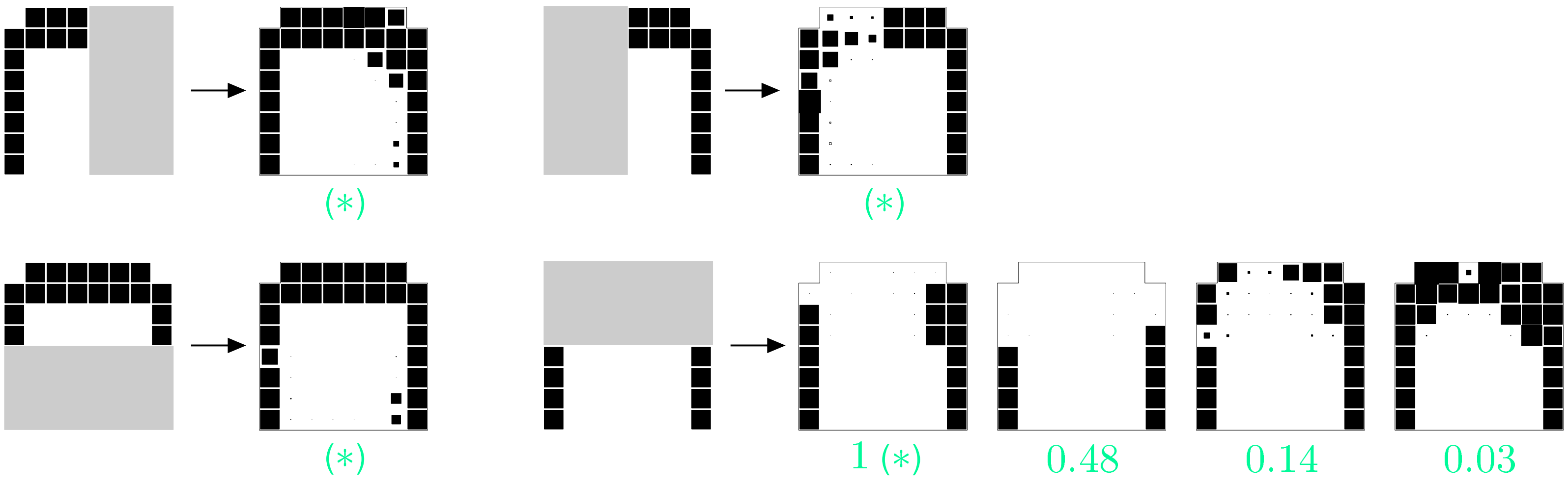
The dataset

- 62-dimensional EPG patterns  plus 12-dimensional PLP coefficients  sampled at 200 Hz. Both mappings, EPG $\longrightarrow$ PLP and PLP $\longrightarrow$ EPG, are one-to-many.
- 10 utterances from the ACCOR database for speaker RK were used for training (over 7500 74-dimensional vectors). 4 utterances were used for testing.
- The silence intervals at the beginning and end (but not inside) of each utterance were removed.

The models

- Factor analysis with a latent space of dimension $L = 9$ (total: 825 parameters).
- GTM with a latent space of dimension $L = 2$, a 30×30 latent grid and a 7×7 RBF grid (total: 3751 parameters).

Reconstruction of single EPG frames



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.

Reconstruction results: average squared error

Average squared error $\frac{1}{N}\sum_{n=1}^{N}\left\|\mathbf{t}_n - \hat{\mathbf{t}}_n\right\|^2$ for utterance "We tore down the outbuildings" with $N = 579$ points, without start and end silence intervals and without gain.

Pattern of missing data	Factor analysis	MLP	GTM				
			`mean`	`gmode`	`rmode`	`cmode`	`dpmode`
EPG $\rightarrow$ PLP	0.3185	0.3063	0.2843	0.3040	0.2983	0.2399	0.2821
PLP $\rightarrow$ EPG	5.7824	5.6305	5.5576	7.8321	9.2272	2.4032	8.4136
75% missing	2.9885		3.1947	3.6480	4.0173	2.2315	2.4717
50% missing	1.7364		1.8605	1.9608	2.1005	1.5297	1.6100
25% missing	0.8000		0.8151	0.8498	0.8832	0.6948	0.7178

Problem type	Low error $\longrightarrow$ High error
EPG $\rightarrow$ PLP	$\mathtt{cmode} \approx \text{MLP} < \mathtt{dpmode} \lesssim \mathtt{mean} \lesssim \mathtt{gmode} \lesssim \text{FA} < \mathtt{rmode}$
PLP $\rightarrow$ EPG	$\mathtt{cmode} \ll \text{MLP} < \mathtt{mean} \lesssim \text{FA} < \mathtt{dpmode} < \mathtt{gmode} < \mathtt{rmode}$
General	$\mathtt{cmode} < \mathtt{dpmode} \ll \mathtt{mean} \lesssim \text{FA} \approx \mathtt{gmode} < \mathtt{rmode}$

Summary

- Probabilistic models (in particular latent variable models) estimate mappings where:
  - the data is noisy
  - nonuniqueness exists (inverse problems).
- Advantages:
  - Physical knowledge of the problem (e.g. forward mapping) not required: just joint data.
  - Applicable to varying patterns of missing data: the observed variables are treated symmetrically, unlike methods based in function approximators.
  - Insensitive to time warping.
- Disadvantages:
  - Sensitivity to the smoothness of the density model.
  - High computational cost at reconstruction time.
  - Difficulty of density estimation in high dimensions.
- We don't model the temporal evolution of the data (unlike HMMs or Kalman filters).