Gaussian Process Dynamical Models for Non-parametric Speech Representation and Synthesis

Gustav Eje Henter[1], **Marcus R. Frean**[2], **and W. Bastiaan Kleijn**[1,2]

[1]*Sound and Image Processing Lab, School of Electrical Engineering,*
KTH – Royal Institute of Technology, Stockholm, Sweden
[2]*School of Engineering and Computer Science, Victoria University of Wellington, New Zealand*

Overview

We present a **new paradigm in speech acoustic models**:

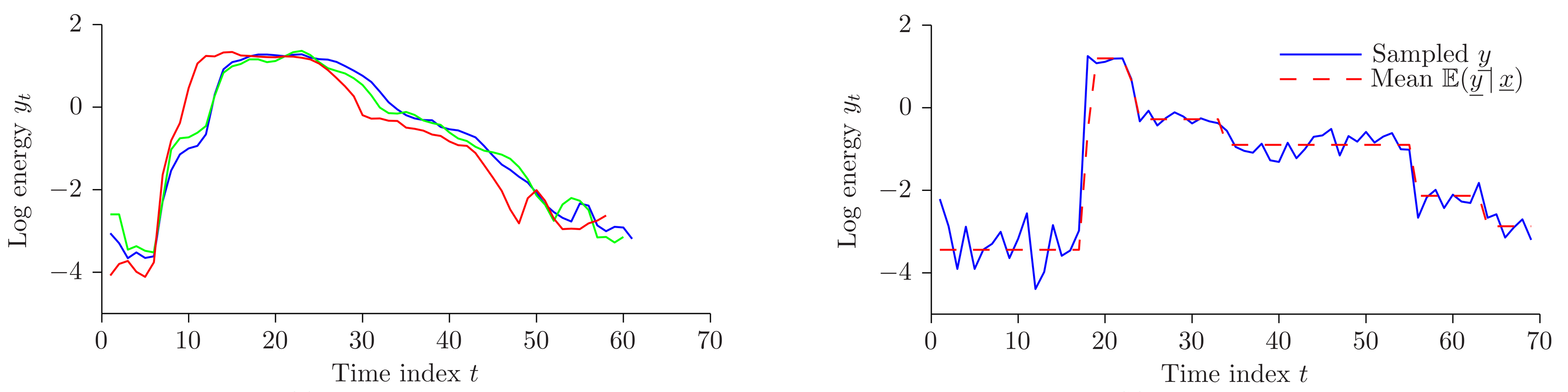
- Traditional H(S)MMs are not good models of speech
- Speakers are better represented by continuous, multidimensional state-spaces
- Nonparametric methods can discover the most salient speaker-state aspects
- We suggest using Gaussian process dynamical models (GPDMs)
- GPDMs generate more natural speech than HMMs in an experiment
- The multidimensional space can represent prosodic variation

Traditional HMMs Are Not Like Speech

HMM-based acoustic models do not sound like speech. Sample sequences:

1. Have **unnatural durations** (memoryless, geometric distribution)
   - Current solution: non-memoryless, semi-Markov models
2. Are **piecewise stationary** (constant), with discrete jumps
   - Current solution: add dynamic features
3. Are **unnaturally warbly**
   - All deviations from the mean contour are treated as noise
   - Current solution: only generate the most probable output (so-called MLPG)

Not sampling may hide the issues, but we are still not describing natural speech!



(a) Training data (b) HMM output

Figure 1: Samples from speech-trained HMMs are unnatural

What to Do

HMMs over-simplify reality. Speech and speakers are more complex than a single, no-skip left-right discrete-state HMM can describe.

1. The **state-space** should be **continuous**
   - We can be in-between sounds and key-frame states (solves 2 above)
   - Incremental progress between states can be remembered (solves 1 above)
2. The state-space should be **multidimensional**
   - Sentence position ("time") is just one aspect of speaker state
   - Overshoots, undershoots, prosody etc. now representable in state space
   - Meaningful variations are not treated as noise anymore (solves 3 above)

Follows industry trend from simple but exact towards advanced but approximated

(a) Discrete states (left-right HMM) (b) Continuous state-space

Figure 2: Comparison of one-dimensional state-spaces

Continuous State-Space Models

A dynamical model for $\boldsymbol{Y}_t$ with hidden (latent) state X_t is defined by:

1. An initial distribution $P(x_0)$
2. **Markovian state dynamics** $P(x_{t+1} \mid x_t)$
3. **State-dependent output** $P(\boldsymbol{y}_t \mid x_t)$
   – Usually assumed Gaussian, defined by means $\boldsymbol{\mu}_Y(x_t)$ and covariances $\boldsymbol{\Sigma}_Y(x_t)$

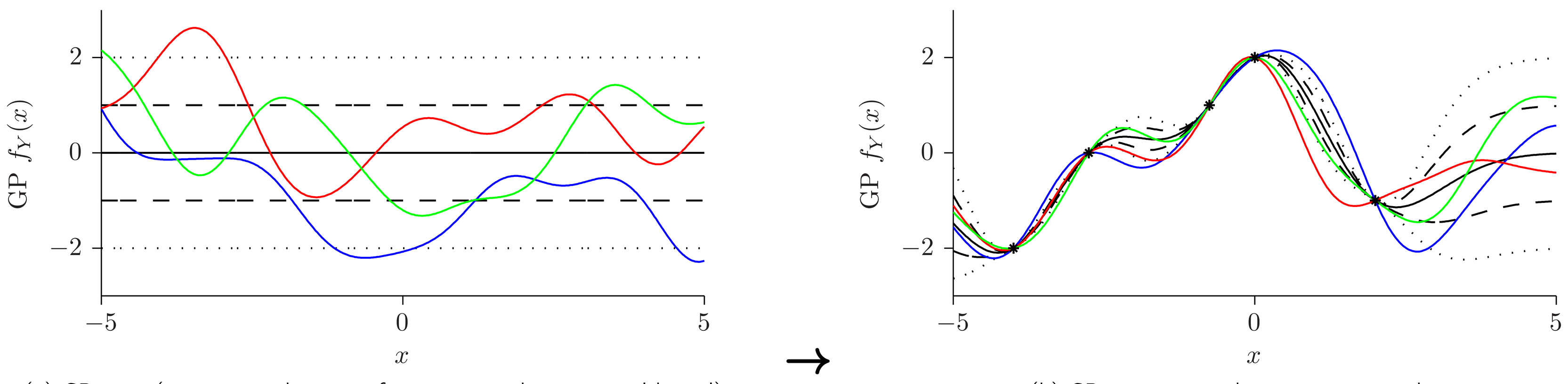
For discrete state-spaces $x_t \in \{1, \ldots, Q\}$, 1, 2, 3 can use general mappings.
For continuous state-spaces $\boldsymbol{x}_t \in \mathbb{R}^Q$, completely general mappings cannot be learned. We must make assumptions.

- Nonparametric assumptions are compelling
  – Similar states should evolve similarly (2) and generate similar output (3)
  – Let the model select the most salient aspects for the state-space to describe
  – Assume all distributions are Gaussian, for simplicity

→ This suggests basing our models on **Gaussian processes** (GPs), a Bayesian framework for nonparametric stochastic regression

Gaussian Processes in Brief

- GPs are like infinite-dimensional Gaussian vector distributions
  – Vector case: mean $i \in \mathbb{Z} \to \mu_i$, covariance $(i, j) \to \sum_{ij}$
  – GP case: mean $\boldsymbol{x} \in \mathbb{R}^Q \to \mu(\boldsymbol{x})$, covariance $(\boldsymbol{x}, \boldsymbol{x}') \to k(\boldsymbol{x}, \boldsymbol{x}')$
- Predictions are made through correlations with previous observations
- The covariance kernel $k(\cdot, \cdot)$ is a positive definite function
  – k expresses prior beliefs, e.g., that similar x-values have similar output
- GPs can be seen as priors over possible regression functions $f_Y(\boldsymbol{x})$



(a) GP prior (using smoothness-preferring squared exponential kernel) → (b) GP posterior – observations starred

Figure 3: Example Gaussian process – samples (color) and standard deviations (black)

Gaussian Process Dynamical Models

Dynamical models built from Gaussian processes are known as **GPDMs**.

- Assume different dimensions are independent given x_t
  – Like assuming diagonal covariance matrices
- Output mapping $\boldsymbol{Y}_t(\boldsymbol{x}_t, \boldsymbol{\beta})$ and dynamic mapping $\Delta \boldsymbol{X}_t(\boldsymbol{x}_t, \boldsymbol{\alpha})$
  – Different kernels k_Y, k_X, with shapes governed by hyperparameters $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$
- Given a state-sequence $\underline{\boldsymbol{x}}$, the output distribution $f_{\underline{\boldsymbol{Y}} \mid \underline{\boldsymbol{X}}}(\underline{\boldsymbol{y}} \mid \underline{\boldsymbol{x}}, \boldsymbol{\beta})$ is Gaussian
  – The covariance matrix $\boldsymbol{K}_Y(\underline{\boldsymbol{x}}, \boldsymbol{\beta})$ is a function of $\underline{\boldsymbol{x}}$
- The state-sequence distribution is **non-Gaussian** since $\boldsymbol{K}$ depends on $\underline{\boldsymbol{x}}$ itself

$$f_{\underline{\boldsymbol{X}}}(\underline{\boldsymbol{x}} \mid \boldsymbol{\alpha}) \propto \left|\boldsymbol{K}_X^{-1}(\underline{\boldsymbol{x}}, \boldsymbol{\alpha})\right| \exp\left(-\frac{1}{2}\sum_{q=1}^{Q} \underline{\Delta x_q}^T \boldsymbol{K}_X^{-1}(\underline{\boldsymbol{x}}, \boldsymbol{\alpha}) \underline{\Delta x_q}\right)$$

– Non-Gaussianity makes sampling and parameter estimation challenging
– Currently available approximations (e.g., MAP) introduce quality loss
– Results from motion capture show natural samples are possible

Experiments

- Feature extraction (pitch + cepstra), synthesis using STRAIGHT at 100 fps
- k_Y, k_X squared-exponential covariance kernels with white noise terms
  – Not suitable for discontinuous data, so fully voiced utterances were used

1. **Synthesis experiment**
   - $Q = 1$ dim. GPDMs vs. many-state left-right no-skip HMMs
   - Data: three examples of each utterance
   - Subjects rated signal naturalness in a MUSHRA-like test
   - High-probability GPDM output rated better than HMM MLPG ($p = 0.0017$)
   - GPDM samples rated better than HMM samples ($p = 0.014$)
   - High-probability output still much more natural than sampling
2. **Representation experiment**
   - Data: six examples of an utterance, but with two different stress patterns
   - A $Q = 3$ dim. GPDM separates the two prosodic variations in latent space
   - Within each group (colored in Figure 4) there is a common representation

The Future

- GPDMs provide a powerful framework, to which HMM tricks can be adapted
- GPDM computational effort can be made tractable through approximations
- With improved parameter estimation, GPDMs can perform better still
- Extension to arbitrary speech synthesis is underway

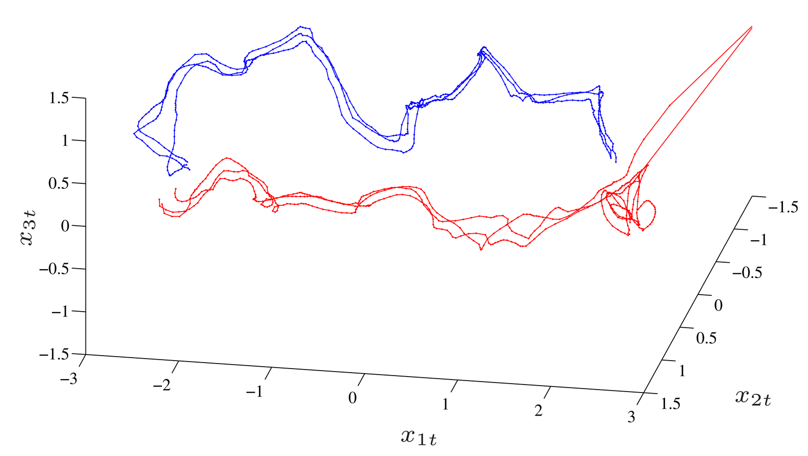


Figure 4: Learned latent-space trajectories, color coded by stress pattern

LISTA