Gaussian Process Dynamical Models for Nonparametric Speech Representation and Synthesis

Gustav Eje Henter\textsuperscript{1}, Marcus R. Frean\textsuperscript{2}, and W. Bastiaan Kleijn\textsuperscript{1,2}

\textsuperscript{1}Sound and Image Processing Lab, School of Electrical Engineering, KTH – Royal Institute of Technology, Stockholm, Sweden
\textsuperscript{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!

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

Continuous State-Space Models

A dynamical model for $Y_t$ 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 $P(y_t \mid x_t)$
   - Usually assumed Gaussian, defined by means $\mu_1(x)$ and covariances $\Sigma_1(x)$
   - For discrete state-spaces $x_t \in \{1, \ldots, Q\}$, 1, 2, 3 can use general mappings.
   - For continuous state-spaces $x_t \in \mathbb{R}^n$, 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 $i, j \to \sum_{k} k_{i,j,k}$
  - GP case: mean $x \in \mathbb{R}^D \to \mu(x)$, covariance $x, x' \to k(x, x')$
- Predictions are made through correlations with previous observations
- The covariance kernel $k(x, \cdot)$ 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 $f(x)$

Experiments

- Feature extraction (pitch + cepstra), synthesis using STRAIGHT at 100 fps
- $k_1, k_2$ 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