

GAUSSIAN PROCESS DYNAMICAL MODELS FOR NONPARAMETRIC SPEECH REPRESENTATION AND SYNTHESIS

Gustav Eje Henter^{1,*} Marcus R. Freen² W. Bastiaan Kleijn^{1,2}

¹School of Electrical Engineering, KTH – Royal Institute of Technology, Stockholm, Sweden

²School of Engineering and Computer Science, Victoria University of Wellington, New Zealand

ABSTRACT

We propose Gaussian process dynamical models (GPDMs) as a new paradigm in acoustic models of speech. By using multidimensional, continuous state-spaces, the technique can overcome familiar limitations of discrete-state, HMM-based speech models. The added dimensions allow the state to represent and describe more than just temporal structure as systematic differences in mean, rather than as mere correlations in a residual (as HMMs with dynamics or AR-HMMs do). Owing to their basis in Gaussian processes, the models avoid restrictive parametric or linearity assumptions on signal structure. We outline GPDM theory, and describe model setup and initialization schemes relevant to speech applications. Experiments demonstrate subjectively better quality of synthesized speech than from comparable HMMs. In addition, there is evidence for unsupervised discovery of salient speech structure.

Index Terms—acoustic models, stochastic models, non-parametric speech synthesis, sampling

1. INTRODUCTION

Hidden Markov models (HMMs) [1] constitute the dominant paradigm in model-based speech recognition and synthesis (e.g., [2]). HMMs are probabilistic, allowing them to deal with uncertainty in a principled manner, and strike an attractive balance between complexity and descriptive power: they avoid restrictive assumptions such as limited memory or linearity, but can still be trained efficiently on large databases.

Unfortunately, HMMs are not satisfactory stochastic representations of speech feature sequences [3]. Sampling from HMMs trained on speech acoustic data reveals several shortcomings of the model, in that durations are incorrect and the sound is warbly and unnatural. Contemporary model-based speech synthesis systems, HMM-based or not, therefore never sample from stochastic models for signal generation.

In this paper we introduce a new paradigm for nonparametric, nonlinear probabilistic modelling of speech based

on Gaussian process dynamical models (GPDMs). This approach has the potential to overcome all principal issues with HMMs and provide more realistic acoustic models. Like HMMs, they may later be used as building blocks which can be concatenated to form arbitrary speech utterances. In the remainder of the text, we motivate and describe GPDMs in the context of speech signal modelling, and present concrete results from a synthesis application.

2. BACKGROUND

We here explain the benefits of moving from Markov chains to continuous, multidimensional state-spaces, and introduce GPDMs as nonparametric dynamical models for speech.

2.1. Continuous, multidimensional state-spaces

Let $\underline{Y}=(Y_1 \dots Y_N)$ be a sequence of observations, here speech features, and let $\underline{X}=(X_1 \dots X_N)$ be the corresponding sequence of unobserved latent-state values. The features are typically continuous and real, $\mathbf{y}_t \in \mathbb{R}^D$, with D between 10 and 100. We consider \underline{Y} a $D \times N$ matrix-valued RV.

HMMs have a discrete state-space, $x_t \in \{1, \dots, M\} \forall t$. This is sufficient to model piecewise i.i.d. processes, but is not a good fit for speech since 1) HMMs have stepwise constant evolution, while speech mostly changes continuously, and 2) the implicit geometric state-duration distribution of the underlying Markov chain has much greater variance than natural speech sound durations. Dynamic features and hidden semi-Markov models [4] have been proposed to deal with issues 1) and 2) separately, respectively. Both shortcomings can however be addressed *simultaneously*, by considering a continuous state-space to represent incremental progress and intermediate sounds [5]. This is the approach explored here.

Typical speech HMMs use left-right Markov chains to encode long-range dependencies between features at different times, specifically the sequential order of sounds in an utterance. Other dependence-modelling is less structured. Short-range time-dependencies can be described as time-correlated deviations from the state-conditional feature mean, e.g., using dynamic features [3]. This enables gradual changes in expected value. Variation between comparable times in distinct

*This research is supported by the LISTA project. The project LISTA acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 256230.

realizations is usually only modelled as Gaussian deviations from a single state-conditional mean value.

In practice, these correlation-based approaches fail to capture important structure in speech variation, and do not produce realistic speech samples [6]. The sampled speech has a rapidly-varying, warbly quality to it due to the large magnitude of the noise-driven deviations from the feature mean. To obtain more pleasant-sounding output, speech synthesizers therefore generally avoid sampling, and only generate the most likely output sequence. This is known as *maximum likelihood parameter generation* (MLPG) [7].

The models considered here can use multidimensional state-spaces $\mathbf{x}_t \in \mathbb{R}^Q$ to represent structured differences between realizations. The added state-space dimensions may for instance track different pronunciations for the same utterance, e.g., stress-dependent pitch and formant evolution.

As both the continuous state-space and the extra dimensions give us flexibility to explain more empirically observed variation as *systematic* differences in mean, less variability will be attributed to residual, *random* variation. The estimated noise magnitude will thus decrease, making samples less warbly and more realistic. We now consider a specific model on such state spaces, based on Gaussian processes.

2.2. Gaussian process dynamical models

A dynamical model for \mathbf{Y} is defined by 1) an initial state distribution $f_{\mathbf{X}_1}(\mathbf{x}_1)$, 2) a stochastic mapping $f_{\mathbf{X}_{t+1}|\mathbf{X}_t}(\mathbf{x}_{t+1} | \mathbf{x}_t)$ describing state-space dynamics, and 3) a state-conditional observation distribution $f_{\mathbf{Y}_t|\mathbf{X}_t}(\mathbf{y}_t | \mathbf{x}_t)$. In speech, we think of \mathbf{x}_t as representing the state of the speaker—most importantly the sound currently being produced—while \mathbf{y}_t are the current acoustic features.

In a simple HMM describing a speech utterance or phone (for synthesis), $f_{\mathbf{X}_{t+1}|\mathbf{X}_t}$ is usually a left-right Markov chain. The output $f_{\mathbf{Y}_t|\mathbf{X}_t}$ is often Gaussian for synthesis tasks, but GMMs are common in recognition. In this paper, however, both $f_{\mathbf{X}_{t+1}|\mathbf{X}_t}$ and $f_{\mathbf{Y}_t|\mathbf{X}_t}$ will be modelled as continuous-valued densities, using stochastic regression based on Gaussian processes (GPs). (For a review of Gaussian processes please consult [8].) The resulting construction is known as a Gaussian process dynamical model, GPDM [9, 10].

For the output mapping $f_{\mathbf{Y}_t|\mathbf{X}_t}$, GPDMs use a technique known as Gaussian process latent-variable models (GP-LVMs) [11]. These assume the output is a product of Gaussian processes, one for each \mathbf{y}_t -dimension, with a shared covariance kernel $k_Y(\mathbf{x}, \mathbf{x}')$ that depends on latent variables \mathbf{X}_t . The processes are conditionally independent given \mathbf{x}_t , similar to assuming diagonal covariance matrices in conventional HMMs. The conditional output distribution becomes

$$f(\mathbf{y}|\mathbf{x}, \beta, \mathbf{w}) = \frac{1}{\sqrt{(2\pi)^{DN} |\mathbf{K}_Y(\mathbf{x}, \beta)|^D}} \cdot \prod_{d=1}^D w_d \exp\left(-\frac{1}{2} \mathbf{w}_d^T \mathbf{y}_d \mathbf{K}_Y^{-1}(\mathbf{x}, \beta) \mathbf{y}_d\right), \quad (1)$$

where the kernel matrix has entries $(\mathbf{K}_Y)_{t,t'} = k_Y(\mathbf{x}_t, \mathbf{x}_{t'} | \beta)$, β being a set of kernel hyperparameters. The scale factors w_d compensate for different variances in different output dimensions. The entries of \mathbf{X} are assumed Gaussian and i.i.d.

Using GP-LVMs for the output mapping essentially assumes that acoustic features \mathbf{y}_t have similar characteristics (mean and standard deviations) for similar speech states \mathbf{x}_t , e.g., the same phone being spoken, though the details depend on the chosen k_Y . This is similar in principle to HMMs, but is more flexible and does not assume that \mathbf{x}_t is quantized.

GP-LVMs were designed as probabilistic, local, nonlinear extensions of principal component analysis (PCA), and MAP estimation in a GP-LVM will therefore attempt to attribute as much as possible of the observed acoustic \mathbf{y} -variation as due to variations in the underlying speaker state \mathbf{X}_t .

GP-LVMs assume \mathbf{X} is i.i.d., and have no memory to account for context or to smooth estimated latent-space positions over time. GPDMs endow the GP-LVM states with simple, first-order autoregressive dynamics $f_{\Delta\mathbf{X}_t|\mathbf{X}_t}$, so that $\Delta\mathbf{X}_t = \mathbf{X}_{t+1} - \mathbf{X}_t$ is a stochastic function of \mathbf{X}_t . (Higher-order dynamics and other constructions are also possible [9].) Specifically, the next-step distributions for the $\Delta\mathbf{X}_t$ components are assumed to be given by separate Gaussian processes (with a shared kernel $k_X(\mathbf{x}, \mathbf{x}')$), conditionally independent of other dimensions and of \mathbf{Y} given \mathbf{x}_t . The joint probability distribution is more involved than for the GP-LVM, as the dynamics map a space onto itself. It can be written

$$f(\mathbf{x}|\alpha) = f_{\mathbf{X}_1}(\mathbf{x}_1) \frac{1}{\sqrt{(2\pi)^{Q(N-1)} |\mathbf{K}_X(\mathbf{x}, \alpha)|^Q}} \cdot \exp\left(-\frac{1}{2} \text{tr}(\Delta\mathbf{x} \mathbf{K}_X^{-1}(\mathbf{x}, \alpha) \Delta\mathbf{x}^T)\right) \quad (2)$$

where $\Delta\mathbf{x} = (\mathbf{x}_2 - \mathbf{x}_1, \dots, \mathbf{x}_N - \mathbf{x}_{N-1})$. This distribution is *not* Gaussian as \mathbf{K}_X depends on \mathbf{x} , and fair sampling requires Metropolis-type algorithms. An approximation called *mean prediction* [9] exists for sequentially generating latent-space trajectories of high likelihood, analogous to MLPG.

Using GPs to describe state dynamics represents an assumption that the state of the speaker, and thus the acoustic output, evolves similarly when the state is similar, quite like how HMMs work but without the discretization.

By endowing all hyperparameters with priors, a fully Bayesian nonparametric dynamical model is obtained.

2.3. Parameter estimation in GPDMs

GPDMs have a number of unobserved variables— α , β , \mathbf{W} , and \mathbf{X} —which have to be inferred from data. In principle, we would like to integrate out these unknowns. [10] shows convincingly that integrating out the latent trajectories \mathbf{X} (“two-stage MAP” in their [supplemental video](#)) produces highly realistic sampled data. However, this required stochastic integration with sampling from $f_{\mathbf{X}}$ as the integrals are intractable.

A faster alternative is to use MAP estimation by guessing some initial trajectories and then performing scaled con-

jugate gradient ascent (SCG). However, there are many local optima and MAP-estimated latent-space trajectories are typically noisy and random, as the GPDM tries to place as much variation as possible in the latent space. To get smooth dynamics, one may choose a fixed α with low noise, and only estimate the remaining unknowns, as we do here.

3. IMPLEMENTING GPDMs FOR SPEECH

GPDMs were first introduced to model motion capture data, and we are unaware of any prior applications to speech. The closest approximation familiar to us is speech-driven facial animation [12]. We here discuss specific issues in using GPDMs as speech acoustic models, and propose an initialization scheme for sequential signals such as speech utterances.

3.1. Feature representation

To create speech features suitable for GPDMs we employed the STRAIGHT system [13], widely used in speech synthesis. STRAIGHT generates three outputs: 1) an $F0$ contour, with pitch zero in unvoiced frames, 2) a filter spectrum, and 3) an aperiodicity spectrum. To get a more compact feature set, we represented the two spectra by 10 and 40 MFCCs, respectively, and downsampled the data to 100 fps. We also removed the mean of each component over the dataset.

The relative scale of the STRAIGHT outputs is arbitrary. Even though the scaling factors w in (1) can in principle compensate for different feature SNRs, we normalized all dimensions to have unit noise magnitude, as this is beneficial for PCA-based initialization schemes. Component SNRs were estimated by fitting a third-order AR-process to each dimension and looking at the standard deviation of the driving noise.

The HTS system [2] uses a mixture of continuous and discrete distributions to represent voiced pitch or unvoiced excitations. This is unsuitable for GPs, which are designed for continuous data spaces. For simplicity, we have restricted ourselves to voiced speech in this initial work.

3.2. Covariance functions

Because the feature data mean has been removed it is appropriate to only consider zero-mean GPs. We then only need to specify k_X and k_Y to have a fully defined model.

For the dynamics, we chose a simple squared exponential (RBF) kernel with a noise term,

$$k_X(\mathbf{x}, \mathbf{x}') = \alpha_1 \exp\left(-\frac{\alpha_2}{2} \|\mathbf{x} - \mathbf{x}'\|^2\right) + \alpha_3^{-1} \delta_{\mathbf{x}, \mathbf{x}'}. \quad (3)$$

Linear and higher-order kernel terms are left as future work.

A similar RBF kernel with a noise term is an appealing choice also for k_Y , to model smooth output with some residual variation. Rapid changes and localized discontinuities such as plosives can be modelled with advanced kernels such as the Gaussian cdf Φ , though that has not been pursued here.

3.3. Advanced initialization

As the likelihood function for the latent \mathbf{x} has many local optima, the starting position $\underline{\mathbf{x}}^{(0)}$ in MAP is highly influential in determining the quality of the final model. We hence went to some lengths to compute a starting position that well expresses our expectations on process behaviour.

Initializing the latent-space variable trajectory by PCA, as in [10], ignores the time dimension of the data and produces a model where acoustically similar frames will evolve similarly regardless of utterance position, like in a (non-hidden) Markov chain. This is precisely what we strive to avoid.

As the most important variation in speech occurs along the time dimension, we initialized the first latent coordinate by the time from utterance start, as an indicator of progress through the sentence. Multiple training utterances were aligned by dynamic time warping. Remaining \mathbf{x} -dimensions were initialized by PCA, so that points at comparable times were spaced closer or farther according to acoustic similarity.

As the scale of the first latent dimension is arbitrary relative to other axes, it was rescaled according to $|\Delta x_1^{(0)}|_2 = \frac{1}{Q-1} \|\Delta \mathbf{x}_{2:Q}^{(0)}\|_2$, to have comparable RMS Δx -magnitude to the remaining dimensions. Finally, the mean was removed and all axes were rescaled equally so that $\|\underline{\mathbf{x}}^{(0)}\|_2 = DN$, to match the Gaussian prior mean and variance.

4. EXPERIMENTS

In order to assess the properties of GPDMs as stochastic models of speech, we performed several experiments with utterance synthesis and speech representation, and contrasted the results against comparable HMMs.

4.1. Speech synthesis

For synthesis applications, we are interested in the quality of samples and maximum probability output of our speech models. Two experiments were conducted on a data set containing the fully-voiced utterances ‘‘I’ll willingly marry Marilyn’’ and ‘‘our lawyer will allow your rule,’’ each spoken three times by a single, male speaker. GPDMs were trained with 1000 SCG iterations on the voiced frames of each of the two utterances, with dynamics hyperparameters fixed at 20 dB SNR to get smooth dynamic trajectories. Corresponding HMMs with 40 states per second were Baum-Welch-trained on the same data. Mean-predicted and sampled GPDM output was then compared in a listening test against the voiced sections of raw database utterances, speech resynthesized from training-data features, and MLPG and sampled output from the HMMs. In the test, eight subjects were asked to rate the resulting signal sources on a scale from 0 (completely unnatural) to 100 (completely natural) using a MUSHRA-like interface.

In the first experiment, GPDMs with one-dimensional state spaces were compared against HMMs without dynamic

Output type	Mean	St. dev.	t -test p -value
Deterministic	12	14	0.002
Sampled	9.7	15	0.014

Table 1. GPDM vs. HMM opinion score difference.

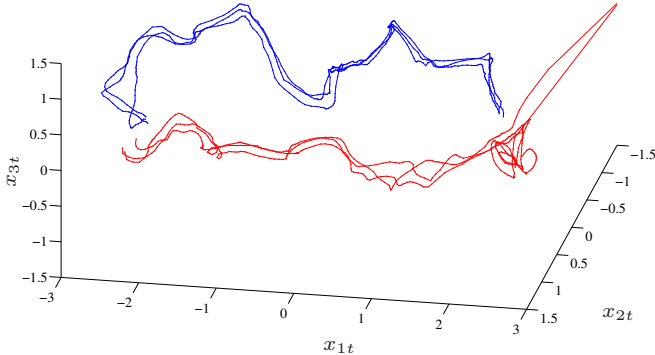


Fig. 1. Latent-space trajectories separated in 3D.

features. Thus no model could pass any information between frames beyond the current utterance position. In this scenario, subjects judged GPDM output as more natural than that of HMMs, both in the case of sampling and MLPG (table 1). The differences are significant according to paired t -tests. GPDM duration modelling, in particular, is noticeably better.

In the second experiment, GPDMs with $Q=3$ were compared against HMMs with delta and delta-delta features. Both these models can pass additional information between frames. With latent trajectories initialized from training data trajectories and then optimized to attain maximum probability, listeners judged mean GPDM output as significantly better than MLPG from the HMMs ($p=$=$$

4.2. Speech representation

In a final experiment, we explored how the additional state-space dimensions in GPDMs can represent multimodal distributions and speech variability. For this, we used the utterance “our lawyer will allow your rule,” thrice pronounced with the stress on “lawyer,” and thrice more stressing the word “allow.”

If not spotted and handled properly, such data inconsistency can degrade the quality of traditional speech synthesis systems. However, a GPDM with $Q=3$ trained on this data correctly separates the two prosodic variations in the latent space, as seen in figure 1, and can represent both varieties simultaneously. Note that this structure was not imposed beforehand (as is typically necessary to model the situation with an HMM), but was recovered automatically from the data.

5. CONCLUSIONS AND FUTURE WORK

We have described how models with continuous, multidimensional state-spaces can avoid the shortcomings of traditional,

discrete-state hidden Markov models of speech. Furthermore, Gaussian process dynamical models possess these advantages and can model speech without restrictive parametric assumptions. The advantages of GPDMs, including automatic structure discovery, are affirmed by experimental evidence.

Further efforts are needed, particularly for the training stage, to realize the full potential of the models and to apply them as building blocks in arbitrary speech synthesis. Work is presently underway to address these limitations.

6. REFERENCES

- [1] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
- [2] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, “The HMM-based speech synthesis system version 2.0,” in *Proc ISCA SSW6*, 2007, vol. 6, pp. 294–299.
- [3] H. Zen, K. Tokuda, and T. Kitamura, “Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences,” *Comput Speech Lang*, vol. 21, no. 1, pp. 153–173, 2007.
- [4] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Hidden semi-Markov model based speech synthesis,” in *Proc ICSLP 2004*, 2004, pp. 1393–1396.
- [5] G. E. Henter and W. B. Kleijn, “Intermediate-state HMMs to capture continuously-changing signal features,” in *Proc Interspeech 2011*, 2011, vol. 12, pp. 1817–1820.
- [6] M. Shannon, H. Zen, and W. Byrne, “The effect of using normalized models in statistical speech synthesis,” in *Proc Interspeech 2011*, 2011, vol. 12.
- [7] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in *Proc ICASSP 2000*, 2000, pp. 1315–1318.
- [8] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.
- [9] J. M. Wang, D. J. Fleet, and A. Hertzmann, “Gaussian process dynamical models,” in *Proc NIPS 2005*, 2006, vol. 18, pp. 1441–1448.
- [10] J. M. Wang, D. J. Fleet, and A. Hertzmann, “Gaussian process dynamical models for human motion,” *IEEE T Pattern Anal*, vol. 30, no. 2, pp. 283–298, Feb. 2008.
- [11] N. D. Lawrence, “The Gaussian process latent variable model,” Tech. Rep. CS-06-03, The University of Sheffield, Department of Computer Science, 2006.
- [12] S. Deena, S. Hou, and A. Galata, “Visual speech synthesis by modelling coarticulation dynamics using a non-parametric switching state-space model,” in *Proc ICMI-MLMI*, 2010, vol. 12.
- [13] H. Kawahara, “STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds,” *Acoust Sci & Tech*, vol. 27, no. 6, pp. 349–353, 2006.