Gaussian Process Dynamical Models for Phoneme Classification

Hyunsin Park *
Department of EE, KAIST
Daejeon, KOREA
hs.park@kaist.ac.kr

Chang D. Yoo
Department of EE, KAIST
Daejeon, KOREA
cdyoo@ee.kaist.ac.kr

1 Introduction

Currently, the hidden Markov model (HMM) is the predominant model studied and used for speech recognition. There has been undeniable progress in speech recognition through the study of HMM but the huge gap that exists between user's expectation and progress is also undeniable. There are essentially two limitation with the HMM: (1) The Markovian structure in HMM leads to limitation in what it can represent since the observations are conditionally independent given the states. The HMM can model only local dependency of speech by obtaining observations on a frame bases. However, the structure has been used since its computational benefit greatly outweighs its deficiency. (2) The state of the articulator represented as a discrete latent variables in HMM is another limitation that has been taken for granted with few reasons. Continuous states have been considered in the linear dynamic model (LDM) [2]. The limitations discussed above must be lifted for improving the speech recognition performance.

In this paper, Gaussian process dynamical model (GPDM) that can capture the nonlinearities of the data without overfitting is considered for phoneme classification. It can potentially overcome the limitation discussed above. The GPDM was proposed in [3] by augmenting Gaussian process latent variable model (GPLVM) with a latent dynamic model. In the GPDM, latent dynamics function and emission function are represented by Gaussian processes that make the model to be non-parametric then global dependency of speech can be considered. Moreover, since non-parametric models do not assume fixed model structure, they can be more flexible than parametric models. By assuming that the continuous dynamics and nonlinearity of the GPDM are well matched with the properties of the speech, the GPDM is used for speech modeling.

2 GPDM based classification

In the GPDM, the first order dynamic model in the latent space is represented by $\mathbf{x}_n = \mathbf{f}(\mathbf{x}_{n-1}) + \mathbf{u}_n$, and the latent-observation emission model by $\mathbf{y}_n = \mathbf{W}^{-1}(\mathbf{g}(\mathbf{x}_n) + \mathbf{v}_n)$, where $\mathbf{x}_n \in \Re^d$, $\mathbf{y}_n \in \Re^D$, $\mathbf{f}(\cdot) \in \Re^d$, $\mathbf{g}(\cdot) \in \Re^D$, $\mathbf{u}_n \in \Re^d$, $\mathbf{v}_n \in \Re^D$, and $\mathbf{W} \in \Re^{D \times D}$ denote latent source (trajectory, manifold), observation, latent dynamic function, emission function, latent dynamic model noise, emission model noise, and diagonal scaling factor, respectively. Here the ith elements of $\mathbf{f}$ and $\mathbf{g}$ follow Gaussian processes with zero-mean function and covariance functions k_f and k_g that are parametric interms of Σ_X and Σ_Y, respectively : $f_i \sim \mathcal{GP}(0, k_f(\mathbf{x}, \mathbf{x}'))$ and $g_i \sim \mathcal{GP}(0, k_g(\mathbf{x}, \mathbf{x}'))$. Both noise terms $\mathbf{u}_n$ and $\mathbf{v}_n$ are assumed Gaussian such that $\mathbf{u}_n \sim \mathcal{N}(\mathbf{0}, \sigma_u^2 \mathbf{I})$ and $\mathbf{v}_n \sim \mathcal{N}(\mathbf{0}, \sigma_v^2 \mathbf{I})$.

Let $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_N]^T$ and $\mathbf{Y} = [\mathbf{y}_1\ \mathbf{y}_2\ \cdots\ \mathbf{y}_N]^T$. Latent variable estimate $\hat{\mathbf{X}}$ and hyper-parameter estimate $\hat{\mathbf{\Theta}} = \{\hat{\Sigma}_X, \hat{\Sigma}_Y\}$ are obtained by MAP estimation as follows;

$$\{\hat{\mathbf{X}}, \hat{\mathbf{\Theta}}\} = \arg\max_{\mathbf{X}, \mathbf{\Theta}} \ln p(\mathbf{X}, \mathbf{\Theta} | \mathbf{Y}). \tag{1}$$

*This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the Human Resources Development Program for Convergence Robot Specialists support program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2011-C7000-1001-0001)

In the GPDM application in [3], all observations are concatenated within a segments for training. This allow long term dependency among the observations. However, as the number of segments increases, the computational complexity of parameter estimation polynomially increases. Therefore, in this paper, an approximation of the log posterior with Q segments is used for training as follows;

$$\ln p\left(\mathbf{X}^{(1:Q)}, \mathbf{\Theta} | \mathbf{Y}^{(1:Q)}\right) \approx \sum_{q=1}^{Q} \ln p(\mathbf{Y}^{(q)} | \mathbf{X}^{(q)}, \mathbf{\Theta}) + \sum_{q=1}^{Q} \ln p(\mathbf{X}^{(q)} | \mathbf{\Theta}) + \ln p(\mathbf{\Theta}) \tag{2}$$

This is similar to the partially independent training conditional (PITC) approximation in [4] without inducing variables.

For phoneme classification with learned phoneme GPDMs, the MAP decoding scheme is adopted. Let Θ_l be the estimated parameter set of l-th phoneme GPDM. Given an observation $\mathbf{Y}_{te}$, phoneme classification result $\hat{l}$ is obtained by

$$\hat{l} \quad = \quad \arg\max_{l} \left(\max_{\mathbf{X}} \ln p(\mathbf{X}, \Theta_l | \mathbf{Y}_{te}) \right). \tag{3}$$

3 Experiments

The proposed GPDM for phoneme classification is evaluated on the TIMIT database. In the experiments, the whole TRAIN set of the TIMIT database is used for training the phonemes using the GPDMs. The standard phoneme label set that consists of 48 labels for the TIMIT database is used for training. The core TEST set of the TIMIT database that consists of 192 sentences is used for testing. Final classification results are obtained by allowing some phoneme confusions that make the size of label set to be 39 and are same with other phoneme classification literatures. As observations, 39 dimensional MFCCs (13 static coefficients, Δ, and $\Delta\Delta$.) are used. In our GPDM, each dimension of observation shares the hyper-parameters and correlations between dimensions are not considered and observation are decorrelated using PCA.

Table 1: Phoneme classification results [%]
(1-mix: single Gaussian, 8-mix: mixture of 8 Gaussians, Lin: linear)

HMM		LDM	GPDM			
1-mix	8-mix		Lin-Lin	Lin-RBF	RBF-Lin	RBF-RBF
61.1	70.7	63.4	63.4	54.7	56.2	50.2

Table 1 shows the results of phoneme classification experiments for comparing the GPDM with the LDM and the HMM. In the HMM, 3 state left-to-right structure with a mixture of diagonal Gaussians at each state is adopted. In the LDM, full covariance Gaussians are used for initial state and noise distribution. The parameters of LDM and HMM are estimated by EM algorithm. In the case of GPDM, hyper-parameters and latent variables are estimated by SCG algorithm. For the kernel functions of the latent dynamic model and the emission model, RBF kernel or linear kernel are used. The dimensions of the latent spaces of the LDM and the GPDM are set to 2. Among the results of (dynamics kernel - emission kernel) GPDMs in the table, the best result is obtained by using linear kernels for both the dynamics and emission functions. The *Lin-Lin* GPDM shows better performance than the 1-mix HMM and similar with the LDM, but worse than the 8-mix HMM.

References

[1] F. Jelinek, "Continuous speech recognition by statistical methods," Proceedings of the IEEE, Vol.64, pp.532-556, 1976.

[2] J. Frankel and S. King, "Speech Recognition Using Linear Dynamic Models," IEEE Trans. Audio, Speech, and Language Processing, Vol.15, pp.246-256, 2007.

[3] J.M. Wang, D.J. Fleet, and A. Hertzmann, "Gaussian Process Dynamical Models for Human Motion, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.30, pp.283-298, 2008.

[4] J. Quiñonero-Candela and C.E. Rasmussen, "A Unifying View of Sparse Approximate Gaussian Process Regression," Journal of Machine Learning Research, Vol.6, pp.1939-1959, 2005.