VARIATIONAL INFERENCE AND LEARNING FOR SEGMENTAL SWITCHING STATE SPACE MODELS OF HIDDEN SPEECH DYNAMICS

Leo J. LEE *

University of Waterloo
Dept. of Electrical & Computer Engineering
Waterloo, ON, N2L 3G1, CANADA

Hagai ATTIAS and Li DENG

Microsoft Research
One Microsoft Way, Redmond
WA 98052-6399, USA

ABSTRACT

This paper describes novel and powerful variational EM algorithms for the segmental switching state space models used in speech applications, which are capable of capturing key internal (or hidden) dynamics of natural speech production. Hidden dynamic models (HDMs) have recently become a class of promising acoustic models to incorporate crucial speech-specific knowledge and overcome many inherent weaknesses of traditional HMMs. However, the lack of powerful and efficient statistical learning algorithms is one of the main obstacles preventing them from being well studied and widely used. Since exact inference and learning are intractable, a variational approach is taken to develop effective approximate algorithms. We have implemented the segmental constraint crucial for modeling speech dynamics and present algorithms for recovering hidden speech dynamics and discrete speech units from acoustic data only. The effectiveness of the algorithms developed are verified by experiments on simulation and Switchboard speech data.

1. INTRODUCTION

The goal of human speech production is to convey discrete linguistic symbols corresponding to the intended message, while the actual speech signal is produced by the continuous and smooth movement of the articulators with lots of temporal structures. This seemingly contradictory dual nature (discrete vs continuous) of speech can be amazingly utilized by human speech recognizers in a beneficial way to enhance the decoding of the underlying message from acoustic signals. However, so far this has been a serious challenge for acoustic modeling in both scientific research and practical applications. The conventional hidden Markov models (HMMs) used in the state-of-the-art speech technology, albeit putting enough emphasis on the symbolic nature of speech, have long been recognized to model the temporal dynamics very poorly, which result in some inherent weaknesses of the current speech technology built upon it. Efforts have since been made to improve the modeling of temporal dynamics and the ultimate goal is to turn the *coarticulation* behavior in natural speech from a *curse* (as in current speech technology) to a *blessing*. Currently there are two general trends in the speech research community to reach this goal: one is to extend upon HMM to better account for the temporal dynamics in acoustic signals directly [1, 2], the other is to use some kind of hidden dynamics, abstract or physically meaningful, to account for the temporal dynamics and subsequently map it to the acoustic domain [3, 4, 5]. The HMM extensions typically enjoy the benefit of being able to use the standard HMM training and test algorithms with some generalization, but have more models parameters and need more computation. The temporal dynamics at the surface acoustic level is also very noisy and difficult to extract. The hidden dynamics models (HDMs) are able to directly model the underlying dynamics with a parsimonious set of parameters and closer to the models developed in speech science, but they typically require the derivation of new training and test algorithms with various degrees of difficulty. This paper directly addresses such a problem for a class of HDMs to be described shortly. We focus on the posterior from which one computes sufficient statistics and use a new approximation as recently introduce in machine learning [6, 7] to approximate the intractable exact form.

The remainder of the paper is organized as follows: The model used in this work is described in Section 2, followed by some details of the algorithm development in Section 3. Section 4 shows the effectiveness of the algorithms by simulation and real speech data, followed by the conclusions and discussions of the future work in Section 5.

2. MODEL DESCRIPTION

The HDM is first proposed in a form of switching state-space models for speech applications. The state equation and observation equation are defined to be:

$$\mathbf{x}_n = \mathbf{A}_s \mathbf{x}_{n-1} + (\mathbf{I} - \mathbf{A}_s)\mathbf{u}_s + \mathbf{w}, \quad (1)$$

$$\mathbf{y}_n = \mathbf{C}_s \mathbf{x}_n + \mathbf{c}_s + \mathbf{v}, \quad (2)$$

where n and s are frame number and phone index respectively, $\mathbf{x}$ is the hidden dynamics and $\mathbf{y}$ is the acoustic feature vector (such as MFCC). The hidden dynamics is chosen to be the vocal-tract-resonances (VTRs) in this work, which are closely related to the smooth and target-oriented movement of the articulators. The state equation (1) is a linear dynamic equation with phone dependent system matrix $\mathbf{A}_s$ and target vector $\mathbf{u}_s$ and with build-in continuity constraints across the phone boundaries. The observation equation (2) represents a phone-dependent VTR-to-acoustic linear mapping. The choice of linear mapping

*The first author performed the work while at Microsoft Research.

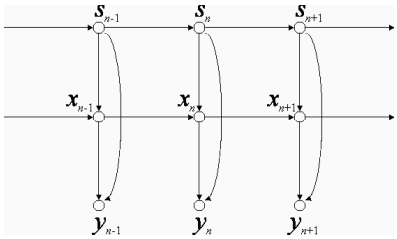

Fig. 1. HDM represented as a Bayesian network

is mainly due to the difficulty of algorithm development. The resulting algorithm can also be generalized to mixtures of linear mapping and piece-wise linear mapping within a phone. Gaussian white noises $\mathbf{w}_n$ and $\mathbf{v}_n$ are added to both the state and observation equations to make the model probabilistic. Similar models have been proposed and used by the authors previously [4, 8], where more insight and details are given.

To facilitate algorithm development, the HDM is also expressed in terms of probability distributions:

$$\begin{aligned} p(s_n = s \mid s_{n-1} = s') &= \pi_{ss'}, \\ p(\mathbf{x}_n \mid s_n = s, \mathbf{x}_{n-1}) &= \mathcal{N}(\mathbf{x}_n \mid \mathbf{A}_s x_{n-1} + \mathbf{a}_s, \mathbf{B}_s), \\ p(\mathbf{y}_n \mid s_n = s, \mathbf{x}_n) &= \mathcal{N}(\mathbf{y}_n \mid \mathbf{C}_s \mathbf{x}_n + \mathbf{c}_s, \mathbf{D}_s), \end{aligned} \quad (3)$$

where $\pi_{ss'}$ is the phone transition probability matrix, $\mathbf{a}_s = (\mathbf{I} - \mathbf{A}_s)\mathbf{u}_s$ and $\mathcal{N}$ denotes a Gaussian distribution with mean and precision matrix (inverse of the covariance matrix) as the parameters. The joint distribution over the entire time sequence is given by

$$p(\mathbf{y}_{1:N}, \mathbf{x}_{1:N}, s_{1:N}) = \prod_n p(\mathbf{y}_n|s_n, \mathbf{x}_n) p(\mathbf{x}_n|s_n, \mathbf{x}_{n-1}) p(s_n|s_{n-1}). \quad (4)$$

The conditional independence relations of the model can be seen more clearly from a graphic form (Bayesian network) as shown in Fig. 1.

3. MODEL INFERENCE AND LEARNING

Inference refers to the calculation of posterior distribution $p(s_{1:N}, \mathbf{x}_{1:N} \mid \mathbf{y}_{1:N})$ given all model parameters, while learning refers to the estimation of model parameters $\Theta = \{\mathbf{A}_{1:S}, \mathbf{a}_{1:S}, \mathbf{B}_{1:S}, \mathbf{C}_{1:S}, \mathbf{c}_{1:S}, \mathbf{D}_{1:S}\}$ given the complete distribution, usually in a maximum likelihood (ML) sense. Under the expectation-maximization (EM) framework, inference is the E step and learning is the M step. In our model, however, the posterior turns out to be a Gaussian mixture whose number of components is exponential in the number of states (or phones) and frames, and is therefore computationally intractable. Here we develop two approximations, called mixtures of Gaussian (MOG) and HMM posteriors respectively, based on *variational* techniques. The idea is to choose the approximate posterior $q(s_{1:N}, \mathbf{x}_{1:N} \mid \mathbf{y}_{1:N})$ with a sensible and tractable structure and optimize it by minimizing its Kullback-Liebler (KL) distance to the exact posterior. It turns out that this optimization can be performed efficiently without having to compute the exact (but intractable) posterior.

It is necessary to say a few words about our previous approaches and other related work in the literature before presenting the current one. Most of our previous algorithms are developed under the assumption of hard phone boundaries which are either known or estimated separately by some heuristic methods [9], and the intractable exact posterior is approximated by a single Gaussian. This is also true for most of the work in a broad range of literatures for switching state space models. In contrast, the current approach uses soft phone assignments that are estimated under an unified EM framework as in [6, 7], but unlike [6, 7], our approximation doesn't factorize s from $\mathbf{x}$ and results in a multimodal posterior over $\mathbf{x}$ instead of a unimodal one, which is justifiably more suitable for speech applications.

3.1. MOG posterior

Under this approximation q is restricted to be:

$$q(s_{1:N}, \mathbf{x}_{1:N}) = \prod_n q(\mathbf{x}_n \mid s_n) q(s_n), \quad (5)$$

where from now on the dependence of the q's on the data $\mathbf{y}$ is omitted but always implied.

Minimizing the KL divergence between q and p is equivalent to maximizing the following functional $\mathcal{F}$,

$$\begin{aligned} \mathcal{F}[q] &= \sum_{s_{1:N}} \int d\mathbf{x}_{1:N}\, q(s_{1:N}, \mathbf{x}_{1:N}) \cdot \\ &\quad [\log p(\mathbf{y}_{1:N}, \mathbf{x}_{1:N}, s_{1:N}) - \log q(s_{1:N}, \mathbf{x}_{1:N})], \end{aligned} \quad (6)$$

which is also a lower bound of the likelihood function and will be subsequently used as the objective function in the learning (M) step.

By taking *calculus of variation* to optimize $\mathcal{F}$ w.r.t. $q(\mathbf{x}_n|s_n)$ and $q(s_n)$, it turns out that each component $q(\mathbf{x}_n|s_n)$ follows a Gaussian distribution, i.e.,

$$q(\mathbf{x}_n \mid s_n = s) = \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\rho}_{s,n}, \boldsymbol{\Gamma}_{s,n}), \quad (7)$$

and the parameters $\boldsymbol{\rho}_{s,n}$ and $\boldsymbol{\Gamma}_{s,n}$ are given by

$$\boldsymbol{\Gamma}_{s,n} = \mathbf{C}_s^T \mathbf{D}_s \mathbf{C}_s + \mathbf{B}_s + \sum_{s'} \gamma_{s',n+1} \mathbf{A}_{s'}^T \mathbf{B}_{s'} \mathbf{A}_{s'}, \quad (8)$$

$$\begin{aligned} \boldsymbol{\Gamma}_{s,n}\boldsymbol{\rho}_{s,n} &= \mathbf{B}_s(\mathbf{A}_s \sum_{s'} \gamma_{s',n-1}\boldsymbol{\rho}_{s',n-1} + \mathbf{a}_s) \\ &+ \sum_{s'} \gamma_{s',n+1} \mathbf{A}_{s'}^T \mathbf{B}_{s'}(\boldsymbol{\rho}_{s',n+1} - \mathbf{a}_{s'}) \\ &+ \mathbf{C}_s^T \mathbf{D}_s(\mathbf{y}_n - \mathbf{c}_s), \end{aligned} \quad (9)$$

where $\gamma_{s,n} = q(s_n = s)$ and is computed from

$$\begin{aligned} \log \gamma_{s,n} &= f_1(\boldsymbol{\rho}_{s,n}, \boldsymbol{\Gamma}_{s,n}, \Theta) + f_2(\boldsymbol{\rho}_{s',n-1}, \boldsymbol{\Gamma}_{s',n-1}, \Theta) \\ &+ f_3(\boldsymbol{\rho}_{s',n+1}, \boldsymbol{\Gamma}_{s',n+1}, \Theta). \end{aligned} \quad (10)$$

f's denote linear functions whose expressions are too lengthy to be written down here. Eq. (8) and (9) are coupled linear equations given model parameters Θ and γ's and can be solved efficiently by sparse matrix techniques. Eq. (10) is a nonlinear equation by itself and has to be solved by iteration. Eq. (8), (9) and (10) constitutes the inference or E step of the algorithm and have to be solved iteratively all together after some proper initializations.

Model learning involves taking derivatives of $\mathcal{F}$ w.r.t. all the model parameters and setting them to zero. This results in a set of linear equations which can be solved easily. Since this step is standard as to all EM approaches with no special difficulties, the detail equations are omitted.

3.2. HMM posterior

Under this approximation q is taken to be

$$q(s_{1:N}, \mathbf{x}_{1:N}) = \prod_{n=1}^{N} q(\mathbf{x}_n \mid s_n) \cdot \prod_{n=2}^{N} q(s_n \mid s_{n-1}) \cdot q(s_1). \quad (11)$$

First we define two posterior transition probabilities:

$$\begin{aligned} \eta_{s's,n} &= q(s_n = s \mid s_{n-1} = s'), \\ \bar{\eta}_{s's,n} &= q(s_n = s \mid s_{n+1} = s') = \frac{\eta_{s's,n+1}\gamma_{s,n}}{\gamma_{s',n+1}}, \end{aligned} \quad (12)$$

where γ is the same as in the previous section. It turns out that each $q(\mathbf{x}_n|s_n)$ is again a Gaussian distribution, and $\boldsymbol{\rho}_{s,n}$ and $\boldsymbol{\Gamma}_{s,n}$ are given by coupled linear equations having the same form as (8) and (9), except that the γ's are replaced by η's and $\bar{\eta}$'s. These equations can again be solved by sparse matrix techniques. The γ's and η's themselves can be solved by the following efficient backward-forward procedure given the model parameters and all the $\boldsymbol{\rho}$'s and $\boldsymbol{\Gamma}$'s.

1. Initialize: $z_{s,N+1} = 1$ for all s.
2. Backward pass: for $n = N, \cdots, 2$

$$\begin{aligned} z_{s,n} &= \sum_{s'} \exp(f_{ss',n}) z_{s',n+1} \ , \\ \eta_{ss',n} &= \frac{1}{z_{s,n}} \exp(f_{ss',n}) z_{s',n+1} \ . \end{aligned} \quad (13)$$

3. For $n = 1$:

$$\begin{aligned} z_1 &= \sum_{s} \exp(f_{s,1}) z_{s,2} \ , \\ \gamma_{s,1} &= \frac{1}{z_1} \exp(f_{s,1}) z_{s,2} \ . \end{aligned} \quad (14)$$

4. Forward pass: for $n = 2, \cdots, N$

$$\gamma_{s,n} = \sum_{s'} \eta_{s's,n} \gamma_{s',n-1} \ . \quad (15)$$

Again, f's are functions of the $\boldsymbol{\rho}$'s, $\boldsymbol{\Gamma}$'s and model parameters whose expressions are too lengthy to be given here. Also remember that the complete E step still has to iterate between the calculation of $q(\mathbf{x}_n \mid s_n)$ and $q(s_n \mid s_{n-1})$. The model learning is quite similar to the MOG case and the detail equations are omitted.

3.3. Speech specific issues

There are a number of important issues to be solved before the above algorithms can be applied to speech, and they are discussed here:

1. Parameter initialization: It is important to initialize the parameters appropriately for an iterative local optimization procedure such as EM. Our HDM enjoys the benefit of being closely related to speech-specific knowledge and some key parameters, especially the phone targets, can be reliably initialized from a formant synthesizer. Due to the small number of total parameters, others can be easily initialized by a small amount of hand-labeled VTR data.
2. Segmental constraint: The probabilistic form of the HDM allows phone transitions to occur at each frame, which is undesirable for speech. In training, we construct a series of time-varying transition matrices $\pi_{ss'}$ based on the given phonetic transcript (or one created from a lexicon if only word transcripts are given) and some initial segmentation to impose the segmental constraint and force the discrete-state component of the model to be consistent with the phonetic transcript. Such an approach also greatly reduces the number of possible phones s that have to be summed up at each time step, including (8)-(10), (13)-(15) and the calculation of all the f's. The segmental constraint in recognition is discussed in 4.
3. Hidden dynamics recovery: It is both informative (especially for debugging) and desirable to recover the hidden VTR, and it is calculated by:

$$\hat{x}_n = \sum_s \gamma_{s,n} \boldsymbol{\rho}_{s,n} \quad (16)$$

for both the MOG and HMM posterior assumptions.
4. Recognition strategy: Here we seek the most likely phone sequence given a sequence of observation. For the MOG case, this is simply accomplished by choosing the maximum γ at each frame; while for the HMM posterior we need to perform Viterbi decoding by using γ and η, e.g., the initialization and induction equation for the scoring are:

$$V_1(s) = \gamma_{s,1}, \ V_n(s') = \max_{1 \le s \le S} [V_{n-1}(s) \eta_{ss',n}] \gamma_{s',n}. \quad (17)$$

It is highly desirable to incorporate segmental (or minimal duration) constraint and language weighting in the recognition stage and this is implemented by Viterbi decoding with modified transition matrices for both cases (in MOG the transition matrix is created from scratch while in HMM the changes are merged into η). Such a strategy allows HDM to be used in phone recognition directly *without* resorting to an N-best list provided by HMM.

4. EXPERIMENTAL RESULTS

The results presented in this section are obtained by running the variational EM algorithm with MOG posterior. The correctness of implementation and effectiveness of the algorithm is first verified by simulation data. An example is shown in Fig. 2 and 3. Fig. 2 shows one of the training tokens (10 in total) with three dynamic regimes (or phones). Only the observation $\mathbf{y}$ is passed to the variational EM algorithm and the model parameters are initialized to be away from the true ones. After the algorithm converges, it learns the parameters quite well, e.g., the true and estimated parameters for the state equation are

$$\begin{aligned} \mathbf{A} &= [0.9\ 0.85\ 0.95], \quad \hat{\mathbf{A}} = [0.8922\ 0.7212\ 0.8623], \\ \mathbf{u} &= [2.0\ 2.5\ 1.8], \quad \hat{\mathbf{u}} = [2.0617\ 2.4011\ 1.8316]. \end{aligned}$$

Fig. 3 shows the hidden dynamics recovery for a test sequence, and the underlying phone sequence is also recognized perfectly for this simple example.

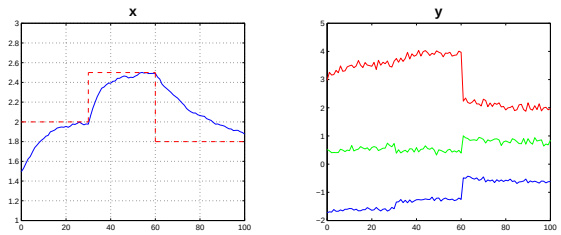

Fig. 2. Simulation data for training.

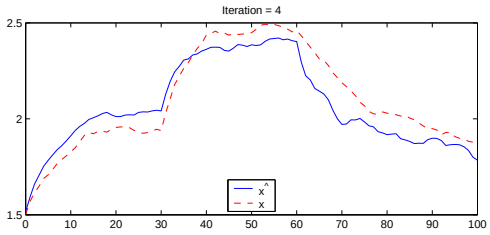

Fig. 3. Hidden dynamics recovery for simulation data. Broken red line: true; solid blue line: estimated.

Similar experiments have been performed for a small amount of speech data from the Switchboard database. Fig. 4 shows the hidden dynamics (VTR) recovery for one of the five training sentences used, and the same is shown for a short test sentence in Fig. 5. By applying simple minimum duration constraint and adjusting the variance level of silence (also modeled as a phone but it needs some special treatment since it doesn't really fit into the state equation of HDM), the phone sequence is recognized perfectly for this simple task.

5. CONCLUSIONS AND FUTURE WORK

We have described in this paper the variational technique for inference and parameter estimation for the segmental switching state space models designed for parsimonious representation of hidden speech dynamics and the resulting context-dependent behavior in speech acoustics. Two types of approximate posterior PDFs are assumed, MOG and HMM, and the respective inference algorithms are presented in detail. We address several specific issues related to speech modeling, including the way in which the segmental nature of the speech dynamics is taken into account. We have conducted comprehensive simulation experiments verifying the effectiveness of the algorithms. Preliminary experiments on recovering hidden speech dynamics (VTR frequencies) and on recovering discrete-state speech units (phone decoding) using acoustic speech data alone demonstrate the potential of the variational approach developed in this study for conversational speech recognition. The algorithms developed in this work also make important contributions to the machine learning community in general.

Our future work will focus on large-scale speech recognition tasks, as well as on improving various components of the segmental switching state space model and the related variational approximation techniques.

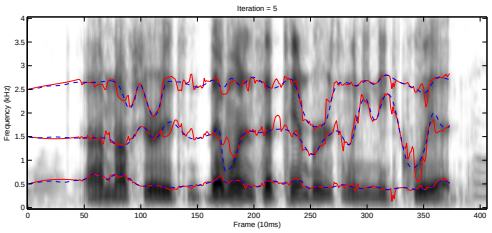

Fig. 4. VTR recovery in the training data. Broken blue line: hand-labeled VTR; solid red line: estimated VTR.

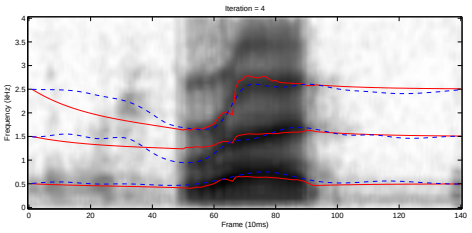

Fig. 5. VTR recovery in the test data. Broken blue line: hand-labeled VTR; solid red line: estimated VTR.

6. REFERENCES

[1] M. Ostendorf, V. V. Digalakis, and O. A. Kimball, "From HMMs to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 360–378, 1996.

[2] K. Reinhard and M. Niranjan, "Diphone subspace mixture trajectory models for HMM complementation," *Speech Communication*, vol. 38, pp. 237–265, 2002.

[3] H. B. Richards and J. S. Bridle, "The HDM: A segmental hidden dynamic model of coarticulation," in *Proc. ICASSP*, Phoenix, 1999, pp. 357–360.

[4] L. Deng and J. Z. Ma, "Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics," *J. Acous. Soc. Am.*, vol. 108, no. 6, pp. 3036–3048, 2000.

[5] Y. Gao, R. Bakis, J. Huang, and B. Xiang, "Multistage coarticulation model combining articulatory, formant and cepstral features," in *Proc. ICSLP*, Beijing, 2000, pp. 25–28.

[6] V. Pavlovic, B. Frey, and T. Huang, "Variational learning in mixed-state dynamic graphical models," in *Proc. UAI*, Stockholm, 1999, pp. 522–530.

[7] Z. Ghahramani and G. E. Hinton, "Variational learning for switching state-space models," *Neural Computation*, vol. 12, pp. 831–864, 2000.

[8] L. J. Lee, P. Fieguth, and L. Deng, "A functional articulatory dynamic model for speech production," in *Proc. ICASSP*, Salt Lake City, 2001, pp. 797–800.

[9] J. Z. Ma and L. Deng, "A path-stack algorithm for optimizing dynamic regimes in a statistical hidden dynamic model of speech," *Computer, Speech and Language*, vol. 14, pp. 101–114, 2000.