

Speech Recognition with a Generative Factor Analyzed Hidden Markov Model

Kaisheng Yao*, Kuldip K. Paliwal† and Te-Won Lee*

*Institute for Neural Computation, University of California at San Diego

†School of Microelectronic Engineering, Griffith University, Australia
kyao@ucsd.edu k.paliwal@griffith.edu.au tewon@ucsd.edu

Abstract

We present a generative factor analyzed hidden Markov model (GFA-HMM) for automatic speech recognition. In a traditional HMM, the observation vectors are represented by mixture of Gaussians (MoG) that are dependent on discrete-valued hidden state sequence. The GFA-HMM introduces a hierarchy of continuous-valued latent representation of observation vectors, where latent vectors in one level are acoustic-unit dependent and the latent vectors in a higher level are acoustic-unit independent. An expectation maximization (EM) algorithm is derived for maximum likelihood parameter estimation of the model. The GFA-HMM can achieve a much more compact representation of the intra-frame statistics of observation vectors than traditional HMM. We conducted an experiment to show that the GFA-HMM can achieve better performances over traditional HMM with the same amount of training data but much smaller number of model parameters.

1. Introduction

In the automatic speech recognition (ASR) problem, one is presented with multi-dimensional data with N dimension where it is assumed that the data is generated from acoustic sources that are modeled as discrete state q in a hidden Markov model (HMM). The transition of the states is assumed to encode the transition of the speech unit and the content of the uttered speech can be inferred by the well-known Viterbi algorithm. The task in speech modeling for ASR within the HMM framework is to obtain a compact and accurate model of the observations. However, this is a hard problem, since the observation vector is high dimensional and the elements in the observation vector contain second as well as higher order statistical information. Traditional approaches in modeling speech observations in an HMM make use of mixture of Gaussians (MoG) with usually a diagonal covariance matrix in each state, which implicitly models the intra-frame correlations.

Despite its pattern recognition appearance, the speech model in an HMM can be viewed in statistics as a latent representation. In particular, the discrete state q is the discrete latent representation of the speech unit and the discrete Gaussian index m in the MoG is the discrete latent representation of the density in that state. In this context, it is therefore natural to describe the N dimensional observation vector $\mathbf{y}(t)$ at time t as correlated in terms of a smaller set of L dimensional continuous-valued latent vector $\mathbf{x}(t)$. In this case, the most straightforward description of the continuous-valued latent representation of $\mathbf{y}(t)$ is given by the fol-

lowing linear model,

$$y_n(t) = \sum_{l=1}^L \Lambda_{nl} x_l(t) + v_n(t), \quad n = 1, \dots, N \quad (1)$$

where $y_n(t)$ denotes the n -th element in vector $\mathbf{y}(t)$ at time t . The $y_n(t)$ depends on linear combination of elements in $\mathbf{x}(t)$ with matrix $\Lambda = [\Lambda_{nl}]_{N \times L}$. The density of $\mathbf{y}(t)$ is also related to the N -dimensional noise $\mathbf{v}(t)$ with element $v_n(t)$. Note that the problem in Eq. (1) is general, since without certain constraints imposed on the model, the solution is non-trivial.

With appropriate constraint on $x_l(t)$ and $v_n(t)$, continuous-valued latent representation can be useful for modeling speech in a compact way. In [1], $\mathbf{x}(t)$ is distributed as $N(\mathbf{x}(t); \mathbf{0}, \mathbf{I})$, which explicitly model the correlations of elements in observation vectors. To model higher order statistics of the observation vectors explicitly, the latent vector $\mathbf{x}(t)$ has to be non-Gaussian.

In this paper, we present a novel speech model for automatic speech recognition. The key to our approach lies in the introduction of a hierarchical continuous-valued latent representation. Observation vector $\mathbf{y}(t)$ is correlated with an *acoustic unit dependent* continuous-valued latent vector $\mathbf{x}(t)$ whose density is dependent on the state $q(t)$ at time t . Since elements in vector $\mathbf{x}(t)$ depend on the same acoustic unit, the elements have correlations. The correlations can be compactly represented by another continuous-valued latent representation $\mathbf{z}(t)$, which is independent of the acoustic unit. In this paper, the $\mathbf{z}(t)$ is distributed as a standard Gaussian $N(\mathbf{z}(t); \mathbf{0}, \mathbf{I})$. Noise in the observation vector $\mathbf{y}(t)$ is modeled as a MoG and it depends on the state $q(t)$ at time t . This model is called a generative factor analyzed HMM (GFA-HMM), described schematically in Figure 1. The model gives a more compact representation of intra-frame statistics than the traditional HMM, and it shows improved performances over traditional HMM given the same amount of training data.

In the context of speech recognition, the originality of this new model can be viewed from the following point. Compared to the traditional HMM, which represents observation vectors as dependent solely on discrete states/mixtures, the GFA-HMM has another continuous-valued latent representation of the observation vectors. The latent vector in the continuous-valued latent representation can be simply modeled by a standard diagonal Gaussian $N(\mathbf{x}(t); \mathbf{0}, \mathbf{I})$, which reduces the model to HMMs with factorized covariance matrix [1]. When the continuous-valued latent vector $\mathbf{x}(t)$ are distributed as MoG, the model reduces to factor-analyzed HMM (FA-HMM) proposed in [2].

Part of the work was carried out when the first author was with ATR Spoken Language Translation Research Laboratories, Kyoto, Japan.

Note that the model is not only dynamic but also non-linear. It is dynamic, because the latent representation is dependent on state in HMM. It is non-linear, because, as shown in Eq. (3), the densities of noise $\zeta_q(t)$ in the continuous-valued latent vector $\mathbf{x}(t)$ are mixture of Gaussians with diagonal covariance matrix \mathbf{V}_{qj} that might be different for each j .

Remark on the number of free parameters: Referring to Figure 1, the number of free parameters (NoFP) in GFA-HMM can be calculated separately for each of the latent representations. In particular, for Eq. (3), NoFP is $S \times K \times L + 2 \times S \times M^x \times L$, where $M^x = \max\{M_q^x : q = 1, \dots, S\}$. For Eq. (7), NoFP is $S \times N \times L + 2 \times S \times M^y \times N$, where $M^y = \max\{M_q^y : q = 1, \dots, S\}$. As a whole, the NoFP for GFA-HMM is given as,

$$\text{NoFP}_{\text{GFA-HMM}} = S \times L \times (K + N) + 2 \times S \times (M^x \times L + M^y \times N)$$

Note that the traditional HMM has NoFP as $2 \times S \times M^y \times N$. The seemingly more complex formula of NoFP for GFA-HMM does not mean that the GFA-HMM requires more free parameters to achieve a performance comparable to traditional HMM. On the contrary, a fair comparison should be based on the comparable number of free parameters. The GFA-HMM imposes explicit continuous-valued latent representation of observation vectors, which lack in traditional HMMs. This structural information may make GFA-HMM be more compact over traditional HMM in the sense that the GFA-HMM can achieve better performance than traditional HMM with the number of free parameters that is less than in the traditional HMM. We will verify this statement through experiments in Section 4.

3. Maximum likelihood parameter estimation of the GFA-HMM

Fig. 1 has shown that the only observable variable is $\mathbf{y}(t)$. Other variables are hidden. EM algorithm is applied to derive maximum likelihood estimation of the GFA-HMM parameters, which are collectively denoted by

$$\Theta = (a_{qp}, \mathbf{C}_q, c_{qj}, \xi_{qj}, \mathbf{V}_{qj}, \mathbf{\Lambda}_q, \pi_{qm}, \mu_{qm}, \mathbf{\Sigma}_{qm}) \quad (14)$$

The estimation process is iterated between calculation of posterior statistics and updating of parameters.

3.1. Posterior statistics

Traditional HMM has a discrete-valued latent representation of observations, i.e., state q and Gaussian mixture index m . The GFA-HMM also has a hierarchy of continuous-valued latent representation by $\mathbf{x}(t)$ and $\mathbf{z}(t)$. Accordingly, posterior statistics are calculated on both discrete and continuous latent variables.

Estimation formulae for posterior statistics of discrete sequences $Q(T)$, $M(T)$ and $J(T)$ are similar to the traditional HMM. Denote the posterior probability of being in state q at time t given observation sequence $\mathbf{y}(T)$ and model parameter Θ , $p(q|\mathbf{y}(T), \Theta)$, as $\gamma_q(t)$. With the likelihood in Eq. (11), it can be obtained by the forward-backward algorithm as in traditional HMM, i.e.,

$$\gamma_q(t) = \frac{\alpha_q(t)\beta_q(t)}{\sum_{i=1}^S \alpha_i(t)\beta_i(t)} \quad (15)$$

where $\alpha_q(t) = p(\mathbf{y}(1), \dots, \mathbf{y}(t), q(t) = q|\Theta)$ accounts for the probability of the partial observation sequence $(\mathbf{y}(1), \dots, \mathbf{y}(t))$ and state q at time t given model parameter Θ , while $\beta_i(t) =$

$p(\mathbf{y}(t+1), \dots, \mathbf{y}(T)|q(t) = i, \Theta)$ is the probability of the partial observation sequence $(\mathbf{y}(t+1), \dots, \mathbf{y}(T))$ given state i at time t and model parameter Θ . The above formula is similar for $\gamma_{qmj}(t)$ and $\gamma_{qm}(t)$.

Regarding the posterior distribution of the continuous-valued latent vector $\mathbf{x}(t)$ given observation $\mathbf{y}(t)$, state q , mixture m and j at time t , by Bayes rule, it is given as,

$$p(\mathbf{x}(t)|\mathbf{y}(t), q, m, j, \Theta) = \frac{p(\mathbf{y}(t)|\mathbf{x}(t), q, m, \Theta)p(\mathbf{x}(t)|q, j, \Theta)}{p(\mathbf{y}(t)|q, m, j, \Theta)}$$

As shown in Eq. (4), Eq. (8) and Eq. (9), each component in the right side of the above equation is Gaussian. Accordingly, the posterior distribution is Gaussian as well. It can be verified that the posterior distribution, $p(\mathbf{x}(t)|\mathbf{y}(t), q, m, j, \Theta)$, is given as $N(\mathbf{x}(t); \phi_{qmj}^x(t), \mathbf{\Psi}_{qmj}^x)$, where,

$$\begin{aligned} \phi_{qmj}^x(t) &= E_{\Theta}[\mathbf{x}(t)|\mathbf{y}(t), q, m, j] \\ &= \mathbf{\Psi}_{qmj}^x [(\mathbf{V}_{qj} + \mathbf{C}_q \mathbf{C}_q^T)^{-1} \xi_{qj} + \mathbf{\Lambda}_q^T \mathbf{\Sigma}_{qm}^{-1} (\mathbf{y}(t) - \mu_{qm})] \\ \mathbf{\Psi}_{qmj}^x &= E_{\Theta}[\delta \mathbf{x}(t) \delta \mathbf{x}(t)^T | \mathbf{y}(t), q, m, j] \\ &= [(\mathbf{V}_{qj} + \mathbf{C}_q \mathbf{C}_q^T)^{-1} + \mathbf{\Lambda}_q^T \mathbf{\Sigma}_{qm}^{-1} \mathbf{\Lambda}_q]^{-1} \end{aligned}$$

Denote $E_{\Theta}[\mathbf{x}(t)\mathbf{x}(t)^T | \mathbf{y}(t), q, m, j]$ as $\mathbf{\Phi}_{qmj}^x(t)$. It is given as $\mathbf{\Psi}_{qmj}^x + \phi_{qmj}^x(t)\phi_{qmj}^x(t)^T$. Then, the posterior mean vector $\phi_{qm}^x(t)$ and covariance matrix $\mathbf{\Psi}_{qm}^x(t)$ are respectively given as $\phi_{qm}^x(t) = \frac{\sum_j \gamma_{qmj}(t)\phi_{qmj}^x(t)}{\sum_j \gamma_{qmj}(t)}$ and $\mathbf{\Psi}_{qm}^x(t) = \frac{\sum_j \gamma_{qmj}(t)\mathbf{\Psi}_{qmj}^x(t)}{\sum_j \gamma_{qmj}(t)}$. These estimation formulae are similar for the posterior mean vector $\phi_q^x(t)$ and the posterior covariance matrix $\mathbf{\Psi}_q^x(t)$.

The posterior statistics of $\mathbf{z}(t)$ is $N(\mathbf{z}(t); \phi_{qmj}^z(t), \mathbf{\Psi}_{qj}^z(t))$, where,

$$\phi_{qmj}^z(t) = \mathbf{\Psi}_{qj}^z(t) \mathbf{C}_q^T \mathbf{V}_{qj}^{-1} (\phi_{qm}^x(t) - \xi_{qj}) \quad (16)$$

$$\mathbf{\Psi}_{qj}^z(t) = [\mathbf{I} + \mathbf{C}_q^T \mathbf{V}_{qj}^{-1} \mathbf{C}_q]^{-1} \quad (17)$$

Denote $E_{\Theta}[\mathbf{z}(t)\mathbf{z}(t)^T | \mathbf{y}(t), q, m, j]$ as $\mathbf{\Phi}_{qmj}^z(t)$, which is given as $\mathbf{\Psi}_{qj}^z(t) + \phi_{qmj}^z(t)\phi_{qmj}^z(t)^T$.

3.2. Parameter estimation

The parameters in Eq. (7) are updated by

$$\begin{aligned} &\sum_t \gamma_{qm}(t) \mathbf{\Sigma}_{qm}^{-1} \hat{\mathbf{\Lambda}}_q \mathbf{\Phi}_{qm}^x(t) \\ &= \sum_t \gamma_{qm}(t) \mathbf{\Sigma}_{qm}^{-1} (\mathbf{y}(t) - \mu_{qm}) \phi_{qm}^x(t)^T \quad (18) \end{aligned}$$

$$\hat{\mu}_{qm} = \frac{1}{\sum_t \gamma_{qm}(t)} \sum_t \gamma_{qm}(t) [\mathbf{y}(t) - \hat{\mathbf{\Lambda}}_q \phi_{qm}^x(t)] \quad (19)$$

$$\begin{aligned} \hat{\mathbf{\Sigma}}_{qm} &= \text{diag} \frac{1}{\sum_t \gamma_{qm}(t)} \sum_t \gamma_{qm}(t) [(\mathbf{y}(t) - \hat{\mu}_{qm})(\mathbf{y}(t) - \hat{\mu}_{qm})^T \\ &\quad - \hat{\mathbf{\Lambda}}_q \phi_{qm}^x(t)(\mathbf{y}(t) - \hat{\mu}_{qm})^T \\ &\quad - (\mathbf{y}(t) - \hat{\mu}_{qm}) \phi_{qm}^x(t)^T \hat{\mathbf{\Lambda}}_q^T + \hat{\mathbf{\Lambda}}_q \mathbf{\Phi}_{qm}^x(t) \hat{\mathbf{\Lambda}}_q^T] \quad (20) \end{aligned}$$

The parameters in Eq. (3) can be updated by

$$\hat{\mathbf{C}}_q \left[\sum_t \sum_j \gamma_{qj}(t) \mathbf{\Phi}_{qj}^z(t) \right] = \quad (21)$$

$$\hat{\xi}_{qj} = \frac{\sum_t \sum_j [\phi_{qj}^x(t) - \xi_{qj}] \phi_{qj}^z(t)^T}{\sum_t \gamma_{qj}(t)} \sum_t \gamma_{qj}(t) (\phi_{qj}^x(t) - \hat{\mathbf{C}}_q \phi_{qj}^z(t)) \quad (22)$$

$$\hat{\mathbf{V}}_{qj} = \text{diag} \frac{1}{\sum_t \gamma_{qj}(t)} \sum_t \gamma_{qj}(t) \{ \Psi_{qj}^x(t) + \hat{\mathbf{C}}_q \Psi_{qj}^z(t) \hat{\mathbf{C}}_q^T + [\phi_{qj}^x(t) - \hat{\xi}_{qj} - \hat{\mathbf{C}}_q \phi_{qj}^z(t)] [\phi_{qj}^x(t) - \hat{\xi}_{qj} - \hat{\mathbf{C}}_q \phi_{qj}^z(t)]^T \}$$

In addition to the above parameters, the weights of the components m and j can be found by $\hat{\pi}_{qm} = \frac{\sum_{t=1}^T \sum_{j=1}^{M_q^x} \gamma_{qmj}(t)}{\sum_{t=1}^T \sum_{m=1}^{M_q^y} \sum_{j=1}^{M_q^x} \gamma_{qmj}(t)}$, and $\hat{c}_{qj} = \frac{\sum_{t=1}^T \sum_{m=1}^{M_q^y} \gamma_{qmj}(t)}{\sum_{t=1}^T \sum_{m=1}^{M_q^y} \sum_{j=1}^{M_q^x} \gamma_{qmj}(t)}$.

4. Experimental results

We compared the proposed GFA-HMM with the traditional HMM by performing experiments on a subset of Aurora 2 database. Features for recognition were 39-dimensional MFCC plus C0 and its first- and second-order coefficients. That is $N = 39$. One thousand utterances¹ from the clean training set of the database were used for training acoustic models. Testing was conducted with 1,000 clean utterances from the testing set of the database.

Acoustic models were trained by EM algorithm with six iterations. Number of states, S , was eight for digit models and one for silence model.

Given the fixed number of state, traditional HMM could only adjust the number of mixture components M_q^y . Accordingly, the number of free parameters (NoFP) for a model was $S \times (2N) \times M_q^y$. The structure of GFA-HMM is flexible. In this work, we set $M_q^y = 1$. This reduced GFA-HMM to a model with only one observation mixture. Therefore, if any performance improvement over traditional HMM can be observed, much of the gain should be attributed to the introduction of the continuous-valued latent representation. For the configuration of the latent representation, the dimension of $\mathbf{z}(t)$, K , was set to one. The model parameters $\{\mathbf{A}_q, c_{qj}, \xi_{qj}, \mathbf{V}_{qj}\}$ were shared among states in the word level. We varied the number of mixture components for $j(t)$, M_q^x , and the dimension of vector $\mathbf{x}(t)$, L . For this configuration, the GFA-HMM has NoFP as $S \times (2N) + (N + 1) \times L + (2 \times L) \times M_q^x$.

Mixture components were incrementally obtained by mixture splitting, which repeatedly splits the mixture with the largest weight until the desired number of components is obtained. In the training stage, variance of elements in noise vectors $\zeta_q(t)$ and $\mathbf{v}_q(t)$ were floored to 1.0 and 0.001, respectively.

4.1. Results

Performances by traditional HMM and the GFA-HMM are shown in Table 1. Varying number of mixture components M_q^y can change recognition word accuracy (W.A.). In particular, the highest W.A. for traditional HMM was attained to 88.96% with $M_q^y = 4$.

The GFA-HMM can achieve higher recognition accuracy over traditional HMM with the same amount of training data. For example, word accuracy increased consistently by increasing mixture component M_q^x while keeping $L = 1$. The highest W.A.

¹We limited number of training utterances to 1000, which was less than the standard number, 8440, of training utterances in Aurora 2 database, in order to show the efficacy of the method.

Table 1: For the given test set, we compare the performance between traditional HMM and the GFA-HMM in terms of the number of free parameters (NoFP) for one digit model and word accuracy (W.A. in %).

	L	M_q^y	1	2	3	4
Traditional HMM	0	NoFP	624	1248	1872	2496
		W.A.	88.48	88.64	88.96	88.96
GFA-HMM ($M_q^y = 1$, $K = 1$)		M_q^x	1	2	3	4
	1	NoFP	666	668	670	672
		W.A.	88.80	89.73	90.30	90.93
	2	NoFP	708	712	716	720
	W.A.	86.44	89.09	89.73	89.66	

was 90.93% by setting $L = 1$ and $M_q^x = 4$. This is compared to the highest W.A. achieved by traditional HMM, leading to a relative error rate reduction of 18%. Moreover, the NoFPs could be much lower than those by traditional HMMs. For example, in the situation where the highest word accuracies were achieved by traditional HMM and GFA-HMM, the NoFP for GFA-HMM was 672, whereas the NoFP for traditional HMM was 2496.

As shown in Table 1, increasing the dimension of latent vector $\mathbf{x}(t)$, L , did not result in improvement of recognition accuracy. This comes from the flooring scheme used in the experiments. The variances of elements in noise vector $\zeta_q(t)$ in the latent vector $\mathbf{x}(t)$ were floored to 1.0. Increasing dimension L of the latent vector may result in a more noisy model if the underlying generative process of the observations are different from the structure of GFA-HMM specified for experiments. However, the GFA-HMM still outperformed traditional HMM in the situation of $L = 2$. For example, the GFA-HMM achieved 89.66% W.A. when $M_q^x = 4$, a relative error rate reduction of 6% over the highest W.A. achieved by traditional HMM.

5. Discussions and Summary

The proposed generative factor analyzed HMM (GFA-HMM) incorporates a discrete latent representation, as the traditional HMM, and a hierarchical continuous-valued latent representation. This representation may achieve compact representation of statistics of acoustic vectors. We plan to apply this model to other speech recognition tasks.

6. Acknowledgement

The first author would like to thank Dr. Satoshi Nakamura and Dr. Jianwu Dang at ATR for helpful discussions.

7. References

- [1] L. K. Saul and M. G. Rahim, "Maximum likelihood and minimum classification error factor analysis for automatic speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 2, pp. 115 – 125, March 2000.
- [2] A.-V.I. Rosti and M.J.F. Gales, "Factor analyzed hidden Markov models," in *ICASSP*, 2002.
- [3] H. Attias, "Independent factor analysis," *Neural Computation*, vol. 11, pp. 803–851, 1998.