

# A SWITCHING LINEAR GAUSSIAN HIDDEN MARKOV MODEL AND ITS APPLICATION TO NONSTATIONARY NOISE COMPENSATION FOR ROBUST SPEECH RECOGNITION

Jian Wu and Qiang Huo

Department of Computer Science and Information Systems  
The University of Hong Kong, Pokfulam Road, Hong Kong, China  
(Email: jwu@csis.hku.hk, qhuo@csis.hku.hk)

## ABSTRACT

The Switching Linear Gaussian (SLG) Models was proposed recently for time series data with nonlinear dynamics. In this paper, we present a new modelling approach, called SLGHMM, that uses a hybrid Dynamic Bayesian Network of SLG models and Continuous Density HMMs (CDHMMs) to compensate for the nonstationary distortion that may exist in speech utterance to be recognized. With this representation, the CDHMMs (each modelling mainly the linguistic information of a speech unit) and a set of linear Gaussian models (each modelling a kind of stationary distortion) can be jointly learnt from multi-condition training data. Such a SLGHMM is able to model approximately the distribution of speech corrupted by switching-condition distortions. The effectiveness of the proposed approach is confirmed in noisy speech recognition experiments on Aurora2 task.

## 1. INTRODUCTION

It is well known that the dynamic trend of a clean speech segment can be modelled reasonably well by a Continuous Density Hidden Markov Model (CDHMM) via the evolution process of its underlining state sequence. However, current automatic speech recognition systems are always compelled to be used in the unknown noisy environments where the performance degradation is observed. It is mainly due to the serious mismatches between the dynamic nature described by the provided CDHMM and that of the testing speech recorded in real environments, especially those with rapidly switching conditions, where the distortion sources are nonstationary and few samples are available for model adaptation. One possible solution is to adopt a model selection or fusion strategy, which first prepares offline a set of models, each hopefully “knowledgeable” to deal with a certain type of stationary distortion. During recognition, an appropriate model will be selected or composed from the set of pre-trained models based on the information embedded in the speech utterance to be recognized. In this way, the nonstationary distortion can be approximated by a series of models, each representing a stationary distortion. Such ideas have been extensively studied in speaker adaptation application (e.g. [4]), yet remains to be explored systematically for dealing with noise robustness in noisy speech recognition.

Another useful idea to address the robustness problem is to use the variability normalization based modelling techniques. One ex-

ample is the so-called *adaptive training* (e.g. [1]), that is designed to eliminate some irrelevant factors from the complex observations during training and thus create a set of generic CDHMMs that model mainly the linguistic information for phonetic discrimination. Such generic models, if used appropriately during recognition, can help improve the recognition performance. A recent development is to explain the principle of adaptive training from the viewpoint of the *Dynamic Bayesian Networks* (DBN) ([1]).

Inspired by these works, in this paper we present a *Switching Linear Gaussian HMM* (SLGHMM hereinafter) for nonstationary distortion compensation. Each SLGHMM consists of two coupled dynamic models: one stream of CDHMM to model the generic linguistic information of clean speech; and one set of parallel linear Gaussian dynamic streams, each representing a possible additive stationary distortion in feature vector space in conjunction with one discrete state Markov chain controlling the choice of the distortion source at each time step. Since the characteristic of each distortion source may slowly change with time, the real-valued state vector representing it in this DBN is allowed to evolve according to a linear Gaussian dynamic system.

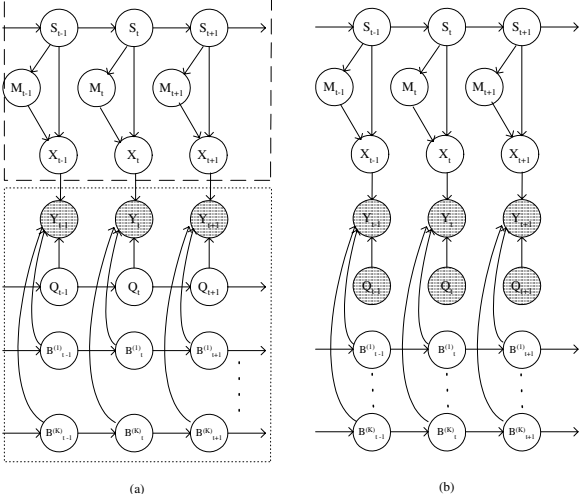
The rest of paper is organized as follows. In Section 2, the basic properties of the linear Gaussian model and the generative model for SLGHMM are described. In Section 3, a special case of SLGHMM tractable for learning and inference, namely Segmental SLGHMM, is presented as well as the detailed learning formulae. In Section 4, the illustrative experimental results on Aurora2 database are reported to demonstrate the effectiveness of the proposed approach. Finally, the paper is summarized in Section 5.

## 2. SWITCHING LINEAR GAUSSIAN HMM

### 2.1. Linear Gaussian Models

Linear Gaussian model (e.g. [6]) is a special case of state-space model which represents the past information through a  $D$ -dimensional real valued hidden state vector  $b$ . Given a sequence of  $D$ -dimensional observation vectors,  $Y_1^T = \{y_1, \dots, y_T\}$ , at any time  $t$ , the past observations,  $Y_1^{t-1}$ , present observation,  $y_t$ , and future observations,  $Y_{t+1}^T$ , are rendered independent conditioned on current state vector  $b_t$  in the state-space model, which is known as the *Markov independence property*. The dependencies are specified through the dynamic equations of the system and the noise model as  $b_t = f(b_{t-1}, v_t)$  and  $y_t = g(b_t, u_t)$ . When these equations are linear as  $b_t = Ab_{t-1} + v_t$  and  $y_t = Cb_t + u_t$ , where  $A$  and  $C$  are both  $D \times D$  matrices, and the noise model  $u_t$  and  $v_t$  are zero-mean Gaussian noise with diagonal covariance matrices  $\Xi$  and  $\Omega$  respectively, the state-space model is called a linear Gaussian

This research was supported by grants from the RGC of the Hong Kong SAR (Project Numbers HKU7022/00E and HKU7039/02E). Authors would like to thank L. Deng, J. Droppo and A. Acero at Microsoft Research for their valuable discussions on some details of experiments.



**Fig. 1.** Directed acyclic graphs specifying conditional independence relations for (a)Switching Linear Gaussian Hidden Markov Models; (b)Segmental Switching Linear Gaussian Hidden Markov Model. (The nodes in shade represent observable variables.)

model. Such a model structure can be used to describe a stationary or slowly changing dynamic process.

## 2.2. Generative Model of SLGHMM

It is not surprised that most real-world processes cannot be characterized by either purely discrete dynamics (i.e., HMM), or purely linear Gaussian dynamics. Accordingly, Switching Linear Gaussian Model (e.g. [2]) was proposed for time series data with non-linear dynamics, which iteratively segments the data into regimes with approximately linear dynamics and learns the parameters of each regime. In this way, it is possible to approximate the nonstationary distortion by a set of stationary or slowly changing distortions with a switching mechanism, which inspires our work in this paper. The DBN illustrating the independence assumptions of the proposed SLGHMM is shown in Fig. 1(a), in which the switching linear Gaussian models (the graph within the dotted-line box) are used to determine the preference of current  $K$  possible additive distortion sources (each represented by  $\mathbf{B}^{(k)}$ ). Let's assume that the discrete state,  $q_t$ , is modelled as a multinomial random variable that can take on  $K$  values which stand for the index of the chosen distortion. The parameters of the discrete state chain  $\mathbf{Q}$  thus include the initial probability of value  $k$ ,  $\varpi_k$ , and the transition probability from value of  $k$  to  $k'$ ,  $h_{kk'}$ . The dynamic evolution equations of each linear Gaussian model stream are defined as  $b_t^{(k)} = A^{(k)}b_{t-1}^{(k)} + v_t^{(k)}$  and thus  $p(b_t^{(k)}|b_{t-1}^{(k)}) = \mathcal{N}(b_t^{(k)} - A^{(k)}b_{t-1}^{(k)}; 0, \Omega^{(k)})$ . The prior distribution of real-valued state vector  $b_1^{(k)}$  is assumed to be a normal distribution with mean vector  $r^{(k)}$  and diagonal covariance  $\mathfrak{R}^{(k)}$ .

The graph within the dash-line box in the figure can be treated as a generative model of usual CDHMM representing mainly linguistic information, where node  $\mathbf{S}$  stands for the hidden state of CDHMM,  $\mathbf{M}$  the hidden indicator of Gaussian component within each state and  $\mathbf{X}$  the generated vectors that are hidden too. Let's

further assume that the parameter set of each CDHMM is  $\Lambda = \{\pi_i, a_{ij}, c_{sm}, \mu_{sm}, \Sigma_{sm}, i, j, s = 1 \dots S, m = 1 \dots M\}$ . It consists of  $S$  states with transition probability  $a_{ij}$  from state  $i$  to state  $j$  and initial probability  $\pi_i$  of state  $i$ . Each state has  $M$  Gaussian components with mean vectors  $\mu_{sm}$  and diagonal covariance matrices  $\Sigma_{sm}$ .  $c_{sm}$  denotes the weight of  $m$ -th Gaussian component in the  $s$ -th state.

Assume that, given  $q_t = k$ , the observation  $y_t$  is produced with  $y_t = x_t + C^{(k)}b_t^{(k)} + u^{(k)}$ . Then we have  $p(y_t|x_t, b_t^{(1)}, \dots, b_t^{(K)}, q_t = k) = \mathcal{N}(y_t - C^{(k)}b_t^{(k)} - x_t; 0, \Xi^{(k)})$ , where  $\Xi^{(k)}$  is the diagonal covariance matrix of zero mean noise  $u^{(k)}$ . According to the independence relations illustrated in Fig. 1(a), the joint distribution of observations and hidden variables can be factored as (Let's use  $\mathbf{B}$  to generically denote  $\{\mathbf{B}^{(k)}\}_{k=1}^K$ )

$$\begin{aligned}
& p(\mathbf{Y}, \mathbf{X}, \mathbf{B}, \mathbf{S}, \mathbf{M}, \mathbf{Q}) \\
&= p(\mathbf{Y}|\mathbf{X}, \mathbf{Q}, \mathbf{B})p(\mathbf{X}|\mathbf{S}, \mathbf{M})P(\mathbf{M}|\mathbf{S})P(\mathbf{Q})P(\mathbf{S})p(\mathbf{B}) \\
&= P(s_1) \prod_{t=2}^T P(s_t|s_{t-1}) \cdot P(q_1) \prod_{t=2}^T P(q_t|q_{t-1}) \\
&\cdot \prod_{k=1}^K p(b_1^{(k)}) \prod_{t=2}^T p(b_t^{(k)}|b_{t-1}^{(k)}) \cdot \prod_{t=1}^T P(m_t|s_t) \cdot \prod_{t=1}^T p(x_t|s_t, m_t) \\
&\cdot \prod_{t=1}^T p(y_t|x_t, b_t^{(1)}, \dots, b_t^{(K)}, q_t). \tag{1}
\end{aligned}$$

Therefore, the parameters to be learnt are  $\Gamma = \{\Lambda, \Phi\}$  where  $\Phi = \{\varpi_k, h_{kk'}, A^{(k)}, \Omega^{(k)}, r^{(k)}, \mathfrak{R}^{(k)}, C^{(k)}, \Xi^{(k)}, k, k' = 1, \dots, K\}$ . Although the exact probability propagation algorithms exist for learning the parameters of general graphical models (e.g., [5]), these algorithms are intractable for densely-connected models such as SLGHMM. An alternative approximate resolvent is to use variational approach (e.g. [2]), but we will not explore them in this paper. In the next section, we demonstrate a special case of SLGHMM, which eliminates some dependence between the variables so that it becomes tractable for the exact EM algorithm.

## 3. SEGMENTAL SLGHMM

### 3.1. Parameter Learning

The Segmental SLGHMM (SSLGHMM hereinafter) is illustrated in Fig. 1(b), where the value of switch state  $q_t$  is independent of all switch states at other time and treated as observations to this DBN. It means that the values of  $q_t$  are assigned by an appropriate pre-segmentation. One example of such segmentation is given later in this paper. Another structure retrogression of SSLGHMM is that  $b_t^{(k)}$  is assumed to be equal to  $b_{t-1}^{(k)}$  and in practice,  $\mathfrak{R}^{(k)}$  is set as zero, which results in  $p(b_t^{(k)}) = p(b_1^{(k)}) = \delta(b_t^{(k)} - r^{(k)})$ , where  $\delta(\cdot)$  denotes the Kronecker delta function. Besides,  $C^{(k)}$  is assumed to be an identity matrix. Accordingly,

$$\begin{aligned}
& p(\mathbf{Y}, \mathbf{X}, \mathbf{B}, \mathbf{S}, \mathbf{M}, \mathbf{Q}) \\
&= P(s_1) \prod_{t=2}^T P(s_t|s_{t-1}) \cdot \prod_{t=1}^T P(m_t|s_t)p(x_t|s_t, m_t)p(b_t^{(q_t)}) \\
& p(y_t|b_t^{(q_t)}, x_t)
\end{aligned}$$

$$\begin{aligned}
&= \pi_{s_1} \prod_{t=2}^T a_{s_t, s_{t-1}} \cdot \prod_{t=1}^T c_{s_t m_t} \mathcal{N}(x_t; \mu_{s_t m_t}, \Sigma_{s_t m_t}) \\
&\quad \cdot \prod_{t=1}^T \mathcal{N}(y_t - x_t - b_t^{(q_t)}; 0, \Xi^{(q_t)}) \cdot \delta(b_t^{(q_t)} - r^{(q_t)}) .(2)
\end{aligned}$$

Now the parameters  $\Gamma$  include  $\Lambda$  and  $\Phi = \{r^{(k)}, \Xi^{(k)}, k = 1 \dots K\}$ . It is possible to iteratively improve the likelihood on an initial model  $\Gamma$  and find a new model  $\bar{\Gamma}$  such that  $p(\mathbf{Y}|\bar{\Gamma}) \geq p(\mathbf{Y}|\Gamma)$ . The auxiliary  $Q$ -function is accordingly defined as  $Q(\Gamma, \bar{\Gamma}) = E\{\log p(\mathbf{Y}, \mathbf{X}, \mathbf{B}, \mathbf{S}, \mathbf{M}, \mathbf{Q}|\bar{\Gamma})|\mathbf{Y}, \mathbf{Q}, \Gamma\}$  where  $E\{\cdot|\mathbf{Y}, \mathbf{Q}, \Gamma\}$  is the conditional expectation taken over all of the hidden variables.

In the E-step of EM algorithm, the sufficient statistics about the unknown parameters should be estimated. Similar to the forward backward algorithm used for the conventional mixture of Gaussian HMM, given a training utterance  $Y = \{y_1, \dots, y_T\}$  and its corresponding segmentation of  $Q = \{q_1, \dots, q_T\}$ , the following auxiliary variables can be worked out recursively:

$$\tilde{b}_{sm}(y_t) = c_{sm} \mathcal{N}(y_t; r^{(q_t)} + \mu_{sm}, \Xi^{(q_t)} + \Sigma_{sm}) \quad (3)$$

$$\tilde{b}_s(y_t) = \sum_{m=1}^M \tilde{b}_{sm}(y_t) \quad (4)$$

$$\tilde{\alpha}_1(j) = \pi_j \tilde{b}_j(y_1) \quad (5)$$

$$\tilde{\alpha}_t(j) = \tilde{b}_j(y_t) \sum_{i=1}^S \tilde{\alpha}_{t-1}(i) \quad (6)$$

$$\tilde{\beta}_T(i) = 1 \quad (7)$$

$$\tilde{\beta}_{t-1}(i) = \sum_{j=1}^S a_{ij} \tilde{b}_j(y_t) \tilde{\beta}_t(j) \quad (8)$$

$$\tilde{p}(Y) = \sum_{j=1}^S \tilde{\alpha}_T(j) . \quad (9)$$

Accordingly some posterior probabilities related to the sufficient statistics can be obtained as follows:

$$\tilde{\gamma}_t(i) = P(s_t = i|Y, Q, \Gamma) = \frac{\tilde{\alpha}_t(i) \tilde{\beta}_t(i)}{\tilde{p}(Y)} \quad (10)$$

$$\begin{aligned}
\tilde{\xi}_t(i, j) &= P(s_{t-1} = i, s_t = j|Y, Q, \Gamma) \\
&= \frac{\tilde{\alpha}_{t-1}(i) a_{ij} \tilde{b}_j(y_t) \tilde{\beta}_t(j)}{\tilde{p}(Y)} \quad (11)
\end{aligned}$$

$$\begin{aligned}
\tilde{\zeta}_t(s, m) &= P(s_t = s, m_t = m|Y, Q, \Gamma) \\
&= \frac{c_{sm} b_{sm}(y_t) \sum_{i=1}^S a_{is} \tilde{\alpha}_{t-1}(i) \tilde{\beta}_t(s)}{\tilde{p}(Y)} . \quad (12)
\end{aligned}$$

Let  $o_t = y_t - x_t$  and we can have the rest of the sufficient statistics:

$$\begin{aligned}
\tilde{x}_{smk}(t) &= E\{x_t|y_t, s_t = s, m_t = m, q_t = k\} \\
&= \mu_{sm} + \Delta_{smk}^{(1)} \epsilon_{smk}(t) \quad (13)
\end{aligned}$$

$$\begin{aligned}
\tilde{o}_{smk}(t) &= E\{o_t|y_t, s_t = s, m_t = m, q_t = k\} \\
&= r^{(k)} + \Delta_{smk}^{(2)} \epsilon_{smk}(t) \quad (14)
\end{aligned}$$

$$\begin{aligned}
\tilde{U}_{smk}(t) &= E\{x_t x_t' | y_t, s_t = s, m_t = m, q_t = k\} \\
&= \tilde{x}_{smk}(t) \tilde{x}_{smk}'(t) + \Delta_{smk}^{(3)} \quad (15)
\end{aligned}$$

$$\begin{aligned}
\tilde{V}_{smk}(t) &= E\{o_t o_t' | y_t, s_t = s, m_t = m, q_t = k\} \\
&= \tilde{o}_{smk}(t) \tilde{o}_{smk}'(t) + \Delta_{smk}^{(3)} , \quad (16)
\end{aligned}$$

where

$$\Delta_{smk}^{(1)} = \Sigma_{sm} (\Sigma_{sm} + \Xi^{(k)})^{-1} \quad (17)$$

$$\Delta_{smk}^{(2)} = I - \Delta_{smk}^{(1)} \quad (18)$$

$$\Delta_{smk}^{(3)} = \Delta_{smk}^{(1)} \Xi^{(k)} \quad (19)$$

$$\epsilon_{smk}(t) = y_t - r^{(k)} - \mu_{sm} , \quad (20)$$

with  $I$  being a  $D \times D$  identity matrix. Again, the re-estimation formulae for the HMM parameters  $\Lambda$  are similar to the standard one except that the observations are replaced by the expected value of  $x_t$ . If  $L$  utterances, in total, are used for training, then

$$\bar{\pi}_i = \sum_{l=1}^L \tilde{\gamma}_1(i) / L \quad (21)$$

$$\bar{a}_{ij} = \frac{\sum_{l=1}^L \sum_{t=2}^T \tilde{\xi}_t(i, j)}{\sum_{l=1}^L \sum_{t=2}^T \tilde{\gamma}_{t-1}(i)} \quad (22)$$

$$\bar{c}_{sm} = \frac{\sum_{l=1}^L \sum_{t=1}^T \tilde{\zeta}_t(s, m)}{\sum_{l=1}^L \sum_{t=1}^T \tilde{\gamma}_t(s)} \quad (23)$$

$$\begin{aligned}
\bar{\mu}_{sm} &= \frac{\sum_{l=1}^L \sum_{t=1}^T \sum_{k=1}^K \tilde{\zeta}_t(s, m) \delta(q_t - k) \tilde{x}_{smk}(t)}{\sum_{l=1}^L \sum_{t=1}^T \sum_{k=1}^K \tilde{\zeta}_t(s, m) \delta(q_t - k)} \\
&= \mu_{sm} + \frac{\sum_{l,t,k} \tilde{\zeta}_t(s, m) \delta(q_t - k) \Delta_{smk}^{(1)} \epsilon_{smk}(t)}{\sum_{l,t,k} \tilde{\zeta}_t(s, m) \delta(q_t - k)} \quad (24)
\end{aligned}$$

$$\bar{\Sigma}_{sm} = \text{diag} \left[ \frac{\sum_{l,t,k} \tilde{\zeta}_t(s, m) \delta(q_t - k) \tilde{U}_{smk}(t)}{\sum_{l,t,k} \tilde{\zeta}_t(s, m) \delta(q_t - k)} - \bar{\mu}_{sm} \bar{\mu}_{sm}' \right] . \quad (25)$$

The value of  $r^{(k)}$  is updated with

$$\begin{aligned}
\bar{r}^{(k)} &= \frac{\sum_{l=1}^L \sum_{t=1}^T \sum_{s=1}^S \sum_{m=1}^M \tilde{\zeta}_t(s, m) \delta(q_t - k) \tilde{o}_{smk}(t)}{\sum_{l=1}^L \sum_{t=1}^T \sum_{s=1}^S \sum_{m=1}^M \tilde{\zeta}_t(s, m) \delta(q_t - k)} \\
&= r^{(k)} + \frac{\sum_{l,t,s,m} \tilde{\zeta}_t(s, m) \delta(q_t - k) \Delta_{smk}^{(2)} \epsilon_{smk}(t)}{\sum_{l,t,s,m} \tilde{\zeta}_t(s, m) \delta(q_t - k)} \quad (26)
\end{aligned}$$

and the diagonal covariance matrix of noise model is updated using the following formulae

$$\bar{\Xi}^{(k)} = \text{diag} \left[ \frac{\sum_{l,t,s,m} \tilde{\zeta}_t(s, m) \delta(q_t - k) \tilde{V}_{smk}(t)}{\sum_{l,t,s,m} \tilde{\zeta}_t(s, m) \delta(q_t - k)} - (\bar{r}^{(k)}) (\bar{r}^{(k)})' \right] . \quad (27)$$

### 3.2. Probabilistic Inference

The statistical inference of SLGHMM consists of computing the marginalized likelihood  $P(\mathbf{Y}|\Gamma)$  given the word sequence  $W$ . In Segmental SLGHMM, since each switch state  $q_t$  is an observed input and  $p(b_t^{(k)}) = \delta(b_t^{(k)} - r^{(k)})$ ,

$$\begin{aligned}
P(\mathbf{Y}|\Gamma) &= \sum_{\mathbf{S}} \sum_{\mathbf{M}} \int P(\mathbf{Y}|\mathbf{X}, \mathbf{Q}, \mathbf{B}) P(\mathbf{X}|\mathbf{S}, \mathbf{M}) \\
&\quad P(\mathbf{M}|\mathbf{S}) P(\mathbf{S}) p(\mathbf{B}) d\mathbf{B} d\mathbf{X} \\
&= \sum_J A_J^* \prod_{t=1}^T \sum_{m=1}^M c_{s_t m} \\
&\quad \mathcal{N}(y_t; \mu_{s_t m} + r^{(q_t)}, \Sigma_{s_t m} + \Xi^{(q_t)}) , \quad (28)
\end{aligned}$$

where  $A_j^*$  is the product of transition probabilities given the state sequence  $J$  of the CDHMMs for the word sequence  $W$ .

#### 4. EXPERIMENTS AND RESULTS

The task used to verify our idea is the speaker independent recognition of connected digit strings. The recognition results presented in this section are produced on the Aurora2 database using the modified reference of WI007 for Aurora front-end evaluation [3]. In this modified front-end, for each frame, a 39-dimensional feature vector is generated, which consists of 12 MFCCs and MFCC of order 0, plus their first and second order derivatives. Another modification on WI007 is that the cepstra are computed based on the power spectral density instead of the magnitude spectrum. In all of our tests the complex back-end with 20 Gaussian components in each state of CDHMM is used. In the BASELINE-CT system shown in Table 1, all of the CDHMMs are trained from the clean speech while in the BASELINE-MT system, they are trained from the collection of 8440 utterances that come from 20 subsets representing 4 different noise scenarios (i.e., *suburban train*, *babble*, *car* and *exhibition hall*) at 4 different SNRs (i.e., 20dB, 15dB, 10dB and 5dB) and the *clean condition*. Obviously the BASELINE-CT is much worse than that of BASELINE-MT because the mismatch between its training and testing is more serious.

In order to assign a meaningful value to each switch state  $q_t$  of SSLGHMM, 16 sets of GMMs are trained for each condition with different noise types and levels and one for clean speech, each consists of 256 Gaussian components. We assume that each Gaussian component corresponds to a subspace distorted by a kind of noise source. Therefore there are  $K = 17 \times 256 = 4352$  linear Gaussian dynamic streams in the SSLGHMM. Accordingly, Given an unknown utterance  $Y$ , the most similar training environment is first identified as that having the maximum likelihood of the GMM. Then the closest subspace within that environment is chosen for each frame of feature,  $y_t$ , and the index of the linear Gaussian stream associated with the subspace is assigned to  $q_t$ . The detail of above process can also be found in [7].

Another important implementation issue is the choice of initial value for each parameter. In all of our experiments, the initial values for  $r^{(k)}$  are zero and  $\Xi^{(k)}$  diagonal matrices with small values. As for the initial values of  $\Lambda$ , two kind of configurations are investigated in our experiments. One starts from the values of the CDHMM trained on clean speech, which is labeled as "SSLGHMM-CT", and another from those of the BASELINE-MT system, which is labeled as "SSLGHMM-MT". It is observed that the performance can be improved greatly no matter which configuration is used although "SSLGHMM-CT" still can not achieve comparable performance with that of "SSLGHMM-MT".

The test set consists of three different parts. For the Test Set A, the same four types of noises as those in training set are added to its subsets, but with 7 different SNRs. For the Test Set B, another 4 types of noises (i.e., *restaurant*, *street*, *airport* and *train station*) are added to its subset with also 7 SNRs. For the Test Set C, *suburban train* and *street* noises are used as the additive noise sources but the speech and noise are filtered with a MIRS characteristic while the G.712 characteristic is used in training set as well as the first two test sets. Therefore, Test Sets B and C are seriously mismatched from that of training speech. However, from the results in Table 1, the information in the linear Gaussian dynamic streams learnt from the training data are helpful to reduce the word error rate of speech distorted by unknown sources in Test Sets B and C,

Table 1. Aurora2 Word Error Rate

	Set A	Set B	Set C	Overall
BASELINE-CT	31.42%	26.47%	30.50%	29.26%
BASELINE-MT	8.07%	9.40%	10.30%	9.04%
SSLGHMM-CT	8.69%	10.01%	9.71%	9.42%
SSLGHMM-MT	6.51%	7.65%	6.83%	7.03%

which proves that some nonstationary distortions can be approximated by a set of stationary distortions with a systematic approach like SLGHMM.

#### 5. DISCUSSION AND CONCLUSION

In this paper, we present a novel structure of dynamic Bayesian network, SLGHMM, to compensate for the nonstationary distortion caused by environmental noise, which integrates a switching state-space model and CDHMMs. With this framework, the nonstationary distortion can be approximated by a set of stationary or slowly changing distortions and hopefully a normalized CDHMM is generated with the strategy of adaptive training. The experimental results of a special case of our proposal on Aurora2 have shown the potential of this new approach. Considering the fact of that our systems based on minimum classification error training in [7, 8] have provided a further error reduction compared with that based on maximum likelihood criterion as in this paper, it would be interesting to verify whether an even better performance can be achieved by a joint MCE estimation of all the parameters in SLGHMM.

#### 6. REFERENCES

- [1] M.J.F. Gales, "Adaptive training for robust ASR," *Proc. IEEE ASRU-01*, Italy, 2001.
- [2] Z. Ghahramani and G.E. Hinton, "Variational learning for switching state-space models," *Neural Computation*, vol. 12, no. 4, pp. 831-864, 2000.
- [3] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions", *ISCA ITRW ASR2000*, Paris, France, September 2000.
- [4] Q. Huo and B. Ma, "Robust speech recognition based on off-line elicitation of multiple priors and on-line adaptive prior fusion," *Proc. ICSLP-2000*, Beijing, China, 2000.
- [5] F.V. Jensen, S.L. Lauritzen and K.G. Olesen, "Bayesian updating in recursive graphical models by local computations," *Computational Statistical Quarterly*, vol. 4, pp. 269-282, 1990.
- [6] S. Roweis and Z. Ghahramani, "A unifying review of Linear Gaussian Models," *Neural Computation*, vol. 11, pp. 305-345, 1999
- [7] J. Wu and Q. Huo, "An environment compensated minimum classification error training approach and its evaluation on Aurora2 database," *Proc. ICSLP-2002*, 2002, pp.1-453-456.
- [8] J. Wu and Q. Huo, "Modelling uncertainty in stochastic vector mapping with minimum classification error training for robust speech recognition," *Proc. ICASSP-2003*, HK, 2003.