Title	Unsupervised Online Adaptation of Segmental Switching Linear Gaussian Hidden Markov Models for Robust Speech Recognition
Author(s)	Huo, Q; Zhu, D; Wu, J
Citation	IEEE International Conference on Acoustics, Speech and Signal Processing Proceedings, Toulouse, France, 14-19 May 2006, v. 1, p. 1125-1128
Issued Date	2006
URL	http://hdl.handle.net/10722/45562
Rights	Creative Commons: Attribution 3.0 Hong Kong License

UNSUPERVISED ONLINE ADAPTATION OF SEGMENTAL SWITCHING LINEAR GAUSSIAN HIDDEN MARKOV MODELS FOR ROBUST SPEECH RECOGNITION

Qiang HUO, Donglai ZHU and Jian WU

Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong
(Email: qhuo@cs.hku.hk, dzhu@i2r.a-star.edu.sg, jianwu@microsoft.com)

ABSTRACT

In our previous works, a Segmental Switching Linear Gaussian Hidden Markov Model (SSLGHMM) was proposed to model "noisy" speech utterance for robust speech recognition. Both ML (maximum likelihood) and MCE (minimum classification error) training procedures were developed for training model parameters and their effectiveness was confirmed by evaluation experiments on Aurora2 and Aurora3 databases. In this paper, we present an ML approach to unsupervised online adaptation (OLA) of SSLGHMM parameters for achieving further performance improvement. An important implementation issue of how to initialize the switching linear Gaussian model parameters is also studied. Evaluation results on Finnish Aurora3 database show that in comparison with the performance of a baseline system based on ML-trained SSLGHMMs, unsupervised OLA yields a relative word error rate reduction of 4.3%, 9.1%, and 17.8% for well-matched, medium-mismatched, and high-mismatched conditions respectively.

1. INTRODUCTION

A Switching Linear Gaussian Hidden Markov Model (SLGHMM), as shown in Fig. 1(a), was proposed in [7] to accommodate the nonstationary distortion that may exist in a speech utterance to be recognized. It is a hybrid Dynamic Bayesian Network (DBN) with two coupled streams of dynamic models. One stream is a Continuous Density HMM (CDHMM) to model the generic linguistic information of "clean" speech $X = \{x_t\}$. Another stream is a Switching Linear Gaussian (SLG) model to model the nonstationary distortion mechanism with a set of parallel linear Gaussian dynamic streams, $B^{(k)} = \{b_t^{(k)}\}$ (each representing a possible additive stationary distortion in feature vector space), and a discrete-state Markov chain, $Q = \{q_t\}$ (controlling the choice of the distortion source at each time step). An SLGHMM with such a mechanism, is thus able to model approximately the distribution of speech, $Y = \{y_t\}$ corrupted by switching-condition distortions. In [6], a variational approach has been proposed to solve the approximate maximum likelihood (ML) parameter learning and probabilistic inference problems for SLGHMMs. Unfortunately, it is not computationally feasible for automatic speech recognition (ASR) applications that require prompt response. Therefore, a Segmental SLGHMM (SSLGHMM hereinafter), as illustrated in Fig. 1(b), was proposed in [7] and refined in [6].

This research was supported by grants from the RGC of the Hong Kong SAR (Project Numbers HKU7022/00E and HKU7039/02E). Donglai Zhu is now with Institute for Infocomm Research, Singapore. Jian Wu is now with Microsoft Corporation, Redmond, USA.

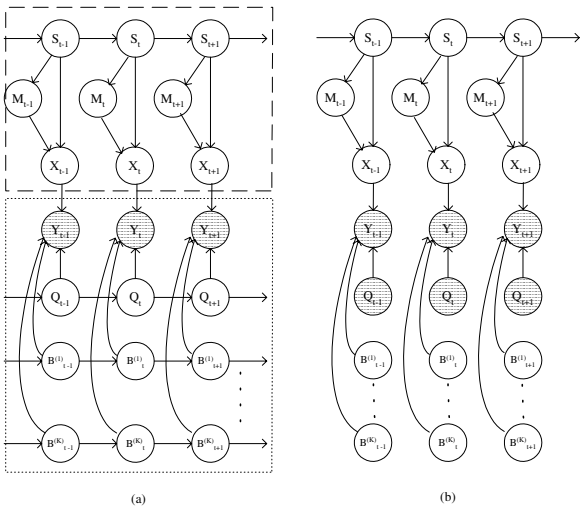

Fig. 1. Directed acyclic graph specifying conditional independence relations for (a) SLGHMM and (b) SSLGHMM.

In an SSLGHMM, several assumptions are made to simplify the model. Each switching state q_t is assumed to be independent of all switching states at other time instances. Switching states are treated as observations rather than hidden variables as in SLGHMM. The values of q_t's are assigned by an appropriate *switching state segmentation* procedure (e.g., [9, 11]). For a particular stream k, $b_t^{(k)}$'s are assumed to follow an i.i.d. (identical independently distributed) Gaussian distribution $\mathcal{N}(b_t^{(k)}; r^{(k)}, \Xi^{(k)})$ with a D-dimensional mean vector $r^{(k)} = [r_1^{(k)}, \cdots, r_D^{(k)}]^{Tr}$ and a diagonal covariance matrix $\Xi^{(k)} = diag\{\xi_1^{(k)2}, \cdots, \xi_D^{(k)2}\}$. It is assumed that $y_t = x_t + b_t^{(k)} + u_t$ given $q_t = k$, where u_t is a Gaussian noise with a zero mean vector and a diagonal covariance matrix Ω.

Let's assume that in our speech recognizer, each basic speech unit is modeled by an SSLGHMM. Its SLG part, whose model parameters are denoted as $\Phi = \{r^{(k)}, \Xi^{(k)}, \Omega, k = 1 \ldots K\}$, is shared by all the speech units. Its CDHMM part is speech unit dependent, with model parameters denoted as $\lambda = \{\pi_s, a_{ss'}, c_{sm}, \mu_{sm}, \Sigma_{sm}; s, s' = 1, \cdots, N_s; m = 1, \cdots, N_m\}$, where N_s is the number of states, N_m is the number of Gaussian components for each state, $\{\pi_s\}$ is the initial state distribution, $a_{ss'}$'s are state transition probabilities, c_{sm}'s are Gaussian mixture weights, $\mu_{sm} = [\mu_{sm1}, \cdots, \mu_{smD}]^{Tr}$ is a D-dimensional mean vector, and $\Sigma_{sm} = diag\{\sigma_{sm1}^2, \cdots, \sigma_{smD}^2\}$ is a diagonal covariance matrix. Conse-

quently, we use $\Lambda = \{\lambda\}$ to denote the set of CDHMM parameters, and $\Gamma = \{\Lambda, \Phi\}$ to denote the set of SSLGHMM parameters respectively in our speech recognizer. Accordingly, the distribution of observation feature vector sequence Y given Q, Γ and a word sequence W can be easily derived as follows:

$$p(Y|Q,\Gamma,W) = \sum_S A_S^* \prod_{t=1}^{T} \sum_{m=1}^{N_m} c_{s_t m}$$
$$\mathcal{N}(y_t; \mu_{s_t m} + r^{(q_t)}, \Sigma_{s_t m} + \Xi^{(q_t)} + \Omega) \ , \qquad (1)$$

where A_S^* is the product of transition probabilities given the state sequence S of the CDHMMs for the word sequence W associated with the observation Y. The above term can be calculated by using a Forward-Backward algorithm [7, 6]. Viterbi algorithm can be used to calculate $\max p(Y, S|Q, \Gamma, W)$.

In our previous studies, an ML and an MCE training procedure were developed in [7, 6] and [6, 9] respectively for the estimation of SSLGHMM parameters. Their effectiveness was confirmed by evaluation experiments on both Aurora2 and Aurora3 databases [7, 6, 9, 11]. However, it was also observed that SSLGHMMs achieve less performance improvement when there exists a high mismatch between training and testing conditions. It is therefore natural to explore the idea of unsupervised online adaptation (OLA) of SSLGHMM parameters and verify whether a further performance improvement can be achieved. The main purpose of this paper is to report our study on this topic.

The rest of the paper is organized as follows. In Section 2, we present a new approach to setting the initial values of SLG model parameters for improving the convergence property of our ML training and adaptation procedures. In Section 3, we present an ML formulation for OLA of SLG model parameters. Evaluation results on Finnish Aurora3 database are reported in Section 4. Finally, we conclude the paper in Section 5.

2. AN IMPROVED INITIALIZATION APPROACH FOR SLG MODEL PARAMETERS

In our previous studies, the SLG model parameters, $\Phi = \{r^{(k)}, \Xi^{(k)}, \Omega, k = 1 \dots K\}$, are initialized as follows:

- Ω is set as a zero matrix, and is fixed during the training process;
- The bias means, $r^{(k)}$'s, are initialized by running the first two steps of an ML training procedure for environment compensated training as described in [8];
- The diagonal elements of $\Xi^{(k)}$'s (referred to as "bias variances" hereinafter) are initialized as a small positive number κ. During the ML training process, κ is also used as a floor value for "bias variances".

The CDHMM parameters of the SSLGHMMs, Λ, are initialized as described in Section 4. The above initialization method is hereinafter referred to as "Old-Init" method. By examining the reestimation formulas of $r^{(k)}$'s in [7, 6], it can be seen that the convergence of training $r^{(k)}$'s is slow if the value of κ is small. This is very similar to situations in a model-space stochastic matching approach for robust ASR as described in [5, 1]. How to set κ appropriately remains an open problem. Inspired by a technique described in [5], in the following, we propose a new way of initializing $\Xi^{(k)}$'s.

First, we define a likelihood function over the training data set $\{Y\}$ as follows:

$$\mathcal{L}(\{\alpha_i^{(k)}\}|\{Y\}; \Gamma^{(0)}) = \prod_{\{Y\}} \sum_S A_S^* \prod_{t=1}^{T} \sum_{m=1}^{N_m} c_{s_t m}^{(0)}$$
$$\mathcal{N}(y_t; \mu_{s_t m}^{(0)} + r^{(q_t)(0)}, \Sigma_{s_t m}^{(0)} + \Theta_{s_t m}^{(q_t)}) \ , \qquad (2)$$

where $\Theta_{sm}^{(k)} = diag\{\theta_{sm1}^{(k)2}, \cdots, \theta_{smD}^{(k)2}\}$ with $\theta_{smi}^{(k)2} = \alpha_i^{(k)} \sigma_{smi}^{(0)2}$, and $\Gamma^{(0)}$ denotes the set of SSLGHMM parameters with initial values specified according to "Old-Init" method.

Then, we set the initial values for $\alpha_i^{(k)}$'s as

$$\alpha_i^{(k)(0)} = \frac{\xi_i^{(k)(0)2}}{\sum_s (\sum_m c_{sm}^{(0)} \sigma_{smi}^{(0)2}) p_s^{(k)}}$$

with

$$p_s^{(k)} = \frac{\sum_{\{Y\}} \sum_{t,m} \delta(q_t - k) \tilde{\zeta}_t(s,m)}{\sum_{\{Y\}} \sum_{t,s,m} \delta(q_t - k) \tilde{\zeta}_t(s,m)} \ ,$$

where $\delta(\cdot)$ denotes the Kronecker delta function, and $\tilde{\zeta}_t(s,m) = P(s_t = s, m_t = m|Y, Q, \Gamma^{(0)})$ that can be calculated by using the formulas listed in [7, 6]. Starting from the above initial values and running one EM iteration to increase $\mathcal{L}(\{\alpha_i^{(k)}\}|\{Y\}; \Gamma^{(0)})$ with respect to $\alpha_i^{(k)}$'s, it can be derived

$$\alpha_i^{(k)} = \frac{\sum_{\{Y\}} \sum_{t,s,m} \zeta_t(s,m) \delta(q_t - k) \frac{\left(y_{ti} - r_i^{(k)(0)} - \mu_{smi}^{(0)}\right)^2}{\sigma_{smi}^{(0)2}}}{\sum_{\{Y\}} \sum_{t,s,m} \zeta_t(s,m) \delta(q_t - k)} - 1 \ , \qquad (3)$$

where y_{ti} is the i-th element of y_t, and $\zeta_t(s,m) = P(s_t = s, m_t = m|Y, Q, \alpha_i^{(k)(0)})$ that can be calculated by using the Forward-Backward algorithm. To avoid $\alpha_i^{(k)}$ from taking a negative value, we use a small positive number ε as a floor as follows:

$$\tilde{\alpha}_i^{(k)} = \max(\alpha_i^{(k)}, \varepsilon) \ . \qquad (4)$$

Finally, the initial values of "bias variances" are computed empirically as follows:

$$\xi_i^{(k)2} = (\sum_s (\sum_m c_{sm}^{(0)} \sigma_{smi}^{(0)2}) p_s^{(k)}) \tilde{\alpha}_i^{(k)} \ . \qquad (5)$$

This method is hereinafter referred to as "New-Init" method.

3. AN ML APPROACH TO ONLINE ADAPTATION OF SLG MODEL PARAMETERS

In the SSLGHMM, the CDHMM stream models mainly the information useful for phonetic discrimination, while the SLG streams are used to model the possible switching "distortions" caused by other factors irrelevant for phonetic classification. For those "unseen" distortions that are not covered in training conditions but exist in testing conditions, the pre-trained SLG model may not work as effectively as expected. To mitigate the problem, one solution is to perform an unsupervised online adaptation (OLA) using the utterance to be recognized to adapt the SLG model parameters to characterize the new environment better. Apparently, there are many ways of doing OLA. As a first step, we tried a simple ML approach described as follows:

Step 1. Given an unknown utterance, first do switching state segmentation [9, 11], and then recognize the utterance via Viterbi decoding with existing ML-trained SSLGHMMs.

Step 2. Given the recognized transcription, update $r^{(k)}$ to maximize the likelihood function defined in Eq. (1) as follows [7, 6]:

$$\bar{r}^{(k)} = r^{(k)} + \frac{\sum_{t,s,m} \tilde{\zeta}_t(s,m)\delta(q_t - k)\Delta_{smk}\epsilon_{smk}(t)}{\sum_{t,s,m} \tilde{\zeta}_t(s,m)\delta(q_t - k)} , \tag{6}$$

where

$$\Delta_{smk} = \Xi^{(k)}(\Sigma_{sm} + \Xi^{(k)} + \Omega)^{-1} , \tag{7}$$

$$\epsilon_{smk}(t) = y_t - r^{(k)} - \mu_{sm} . \tag{8}$$

Step 3. Recognize the utterance with the updated SSLGHMMs again.

Step 4. Repeat Steps 2 and 3 until a pre-specified criterion is satisfied (e.g., a fixed number of cycles).

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

We use Finnish Aurora3 database [2] to verify our algorithm. Aurora3 contains utterances of connected digits that were recorded by using both close-talking (CT) and hands-free (HF) microphones in cars under several driving conditions to reflect some realistic scenarios for typical in-vehicle ASR applications. There are roughly three conditions: *quiet*, *low noise*, and *high noise*. The database is divided into following three subsets according to matching degree between training data and test data:

- **Well-Matched (WM) condition**: Both training and testing data include utterances recorded by both CT and HF microphones from all conditions;
- **Medium-Mismatched (MM) condition**: Training data includes utterances recorded by HF microphone in the *quiet* and *low noise* conditions. Testing data includes utterances recorded by HF microphone in the *high noise* condition;
- **High-Mismatched (HM) condition**: Training data includes utterances recorded by CT microphone from all conditions. Testing data includes utterances recorded by HF microphone in the *low noise* and *high noise* conditions.

Therefore, the MM condition simulates mainly the mismatch caused by a noisy environment due to different driving speeds and possible background music. The HM condition simulates mainly the mismatch caused by different transducers.

In our experiments, the ETSI Advanced Front-End (AFE) as described in [3] is used for feature extraction from a speech utterance. A feature vector sequence is extracted from the input speech utterance via a sequence of processing modules that include noise reduction, waveform processing, cepstrum calculation, blind equalization, and "server feature processing". Each frame of feature vector has 39 features that consists of 12 MFCCs (C_1 to C_{12}), a combined log energy and C_0 term, and their first and second order derivatives. Although all the feature vectors are computed from a given speech utterance, the feature vectors that are sent to the speech recognizer and the training module are those corresponding to speech frames, as detected by a VAD module described in Annex A of [3].

Table 1. A comparison of word error rates (in %) of different SSLGHMM-based systems trained from initial models specified by "Old-Init" method with different κ.

Testing Conditions	κ in "Old-Init" Method for SSLGHMMs				
	0.001	0.01	0.1	1.0	10.0
WM($\times 40\%$)	3.36	3.36	3.40	3.47	3.72
MM($\times 35\%$)	17.78	17.78	17.58	17.37	13.68
HM($\times 25\%$)	16.18	16.18	16.29	15.23	9.51
Average	11.61	11.61	11.59	11.28	9.61

Each digit is modeled as a whole word left-to-right SSLGHMM with 16 emitting states, 3 Gaussian mixture components with diagonal covariance matrices per state. Besides, two pause models, "sil" and "sp", are created to model the silence before/after the digit string and the short pause between any two digits, respectively. The "sil" model is a 3-emitting state SSLGHMM with a flexible transition structure as that of HMM described in [4]. Each state is modeled by a mixture of 6 Gaussian components with diagonal covariance matrices. The "sp" model consists of 2 dummy states and a single emitting state which is tied with the middle state of "sil".

During recognition, an utterance can be modeled by any sequence of digits with the possibility of a "sil" model at the beginning and at the end and a "sp" model between any two digits. All of the recognition experiments are performed with the search engine of HTK3.0 toolkit [10] and a modified version for supporting SSLGHMMs.

Before training SSLGHMM-based system, traditional CDHMMs that have the same model structure as their CDHMM counterparts in SSLGHMMs are trained first by running the training scripts published in Aurora3 CDs, i.e., the standard ML training implemented in HTK. The Word Error Rates (WERs) of this CDHMM-based system are 3.95%, 19.70%, 14.28% under WM, MM, HM conditions respectively.

For SSLGHMM-based system, in the process of switching state segmentation [9, 11], the number of conditions is set to 8, i.e., each input utterance is labeled to one of 8 conditions. In frame labeling, the number of classes in each condition is set to 32, i.e., each speech frame in an utterance is classified to one of 32 classes in the corresponding condition. Therefore, there are $8 \times 32 = 256$ linear Gaussian dynamic streams in the SSLGHMM.

In ML training of SSLGHMMs, the initial values of CDHMM parameters are set to those of the above traditional CDHMMs. The initial values of SLG model parameters are specified by using the the approaches described in Section 2. Starting from the above initial values, five EM iterations are performed. After each iteration, the parameters of CDHMMs and SLG models are both updated. In OLA of the SLG model parameters, only one EM iteration is performed. In the following subsection, we compare the effect of using different initialization approaches for "bias variances".

4.2. Effects of Different Initialization Approaches

Table 1 summarizes WERs of different SSLGHMM-based systems trained from initial models specified by the "Old-Init" method with different values of control parameter κ. It is observed that a larger κ gives a better performance in MM and HM conditions, while a smaller κ is more desirable for WM condition. However, if the κ is too small, bias means and variances converge very slowly in the ML training process. After five EM iterations, values of

Table 2. A comparison of word error rates (in %) of different SSLGHMM-based systems trained from initial models specified by "New-Init" method with different ε.

Testing Conditions	ε in "New-Init" Method for SSLGHMMs				
	0.001	0.01	0.1	1.0	10.0
WM($\times 40\%$)	3.35	3.36	3.42	3.51	4.01
MM($\times 35\%$)	17.17	16.96	17.24	15.05	14.02
HM($\times 25\%$)	16.29	16.47	16.68	13.29	12.40
Average	11.42	11.40	11.57	9.99	9.61

most bias means and variances remain unchanged or have changed only slightly. If OLA starts from such a pre-trained SSLGHMM-based system, little change will be made on "bias means", therefore no performance improvement can be achieved. We have done a detailed analysis of the trained "bias variances" in all the cases shown in Table 1. It is observed that most of "bias variances" don't change after ML training and take the floor value κ in all cases. This fact implies that "Old-Init" method fails to give an appropriate setting for "bias variances" that could affect the learning behavior to learn more informative "bias variances" from training data. This motivates us to develop the "New-Init" method as described in Section 2 as well as several others. Among them, the "New-Init" method achieves the best performance, thus we only report its results here.

Table 2 summarizes WERs of different SSLGHMM-based systems trained from initial models specified by the "New-Init" method with different values of control parameter ε. It is also observed that with the increasing value of ε, the recognition performance gets worse for the WM condition, but becomes better for the MM and HM conditions. A detailed analysis of the trained "bias variances" reveals that they are indeed very informative. When the floor value ε is 1.0, a good compromise is achieved. Therefore, in the following OLA experiments, we start from SSLGHMM-based systems trained from initial models specified by "New-Init" method with $\varepsilon = 1.0$.

4.3. Results of Unsupervised Online Adaptation

For each test utterance, unsupervised online adaptation of "bias means" is performed according to the procedure described in Section 3. The adaptation process is carried out for two cycles. Table 3 summarizes WERs of the adapted systems after each cycle. For comparison, we also list the results of the CDHMM-based baseline system and the SSLGHMM-based baseline system. It is shown that unsupervised OLA can indeed improve the performance further. Actually, we have conducted a comparative study with several existing robust ASR approaches in literature that include MLLR, Stochastic Matching, and Model Fusion. The SSLGHMM-based system with unsupervised OLA achieves the best performance. Due to the space limitation, we can only report the detailed results elsewhere.

5. SUMMARY

In this paper, we have studied an important implementation issue of how to initialize the SLG model parameters of our previously proposed SSLGHMM for robust ASR. An ML approach to unsupervised online adaptation (OLA) of SLG model parameters is also studied. Evaluation results on Finnish Aurora3 database show that the SSLGHMM-based system with unsupervised OLA achieves a

Table 3. A comparison of word error rates (in %) of a CDHMM-based baseline system, an SSLGHMM-based baseline system, and the adapted SSLGHMM-based systems with different OLA cycles.

Testing Conditions	CDHMM Baseline	SSLGHMM Baseline	OLA Cycles	
			1	2
WM($\times 40\%$)	3.95	3.51	3.42	3.36
MM($\times 35\%$)	19.70	15.05	14.16	13.68
HM($\times 25\%$)	14.28	13.29	11.98	10.92
Average	12.05	9.99	9.32	8.86

WER of 3.36%, 13.68%, and 10.92% for WM, MM, and HM conditions respectively. In comparison with the performance of the baseline system based on ML-trained SSLGHMMs, unsupervised OLA yields a relative word error rate reduction of 4.3%, 9.1%, and 17.8% respectively. The relative word error rate reduction will become 15%, 30.6%, and 23.5% respectively if the comparison is made with the CDHMM-based baseline system.

6. REFERENCES

[1] M. Afify, Y. Gong and J.-P. Haton, "A general joint additive and convolutive bias compensation approach applied to noisy Lombard speech recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 6, No. 6, pp.524-538, 1998.

[2] Aurora document AU/217/99, "Availability of Finnish SpeechDat-Car database for ETSI STQ WI008 front-end standardisation," Nokia, Nov 1999.

[3] ETSI standard document, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," ETSI ES 202 050 v1.1.1 (2002-10), 2002.

[4] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," *ISCA ITRW ASR-2000*, Paris, France, September 2000.

[5] A. Sankar and C.-H. Lee, "A maximum likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 4, No. 3, pp.190-202, 1996

[6] J. Wu, "Discriminative speaker adaptation and environmental robustness in automatic speech recognition," Ph.D Thesis, Department of Computer Science, The University of Hong Kong, July 2004.

[7] J. Wu and Q. Huo, "A switching linear Gaussian hidden Markov model and its application to nonstationary noise compensation for robust speech recognition," *Proc. Eurospeech-2003*, 2003, pp.977-980.

[8] J. Wu, Q. Huo and D. Zhu, "An environment compensated maximum likelihood training approach based on stochastic vector mapping," *Proc. ICASSP-2005*, 2005.

[9] J. Wu, D. Zhu and Q. Huo, "A study of minimum classification error training for segmental switching linear Gaussian hidden Markov models," *Proc. ICSLP-2004*, 2004.

[10] S. Young, et al., *The HTK Book (for HTK V3.0)*, July 2000.

[11] D. Zhu, Q. Huo and J. Wu, "A study of switching state segmentation in segmental switching linear Gaussian hidden Markov models for robust speech recognition," *Proc. ISCSLP-2004*, Hong Kong, 2004, pp.97-100.