

Noisy Speech Feature Estimation on the Aurora2 Database using a Switching Linear Dynamic Model

Jianping Deng, Martin Bouchard and Tet Hin Yeap

School of Information Technology and Engineering, University of Ottawa, Ottawa, Canada

Email: {jdeng, bouchard, tet}@site.uottawa.ca

Abstract— This paper presents an approach to enhance speech feature estimation in the log spectral domain under additive noise environments. A switching linear dynamic model (SLDM) is explored as a parametric model for the clean speech distribution, enforcing a state transition in the feature space and capturing the smooth time evolution of speech conditioned on the state sequence. Experimental results using the Aurora2 database show that the new SLDM approach can improve speech enhancement performance in terms of recognition accuracy.

Index Terms—speech feature enhancement, speech recognition, switching linear dynamic model, hidden Markov model

I. INTRODUCTION

The performance degradation of a speech recognizer in the presence of additive noise is one of the major problems that still remain unsolved in the real-field applications of speech recognition technology. Towards solving the noise robustness problem, in the past few years a variety of noise compensation techniques have been developed. One of the prevailing approaches is speech feature enhancement [1]-[7], where a noisy speech signal is processed in the feature enhancement. Instead of producing an enhanced waveform, an enhanced version of the recognition parameters is produced. It is believed that enhancement in the speech feature domain is most desirable if the purpose of speech enhancement is for robust speech recognition, since this is the domain as close as possible to the back end of the recognizer [5]. Since the most popular speech features for speech recognition are Mel-Frequency Cepstral Coefficients (MFCC), vectors of log-spectrum coefficients or cepstrum coefficients are enhanced, as they are an intermediate step in the computation of the MFCC coefficients.

One of the difficulties with speech feature enhancement is that the trajectory of speech features exhibits a complex and rich dynamic behavior that is both

nonlinear and time varying. There is no explicit dynamical model for a sequence of speech features. Therefore, most approaches use the idea of modeling the speech with a Gaussian mixture model (GMM) and consider noise as the variable to be estimated, whose observation is corrupted by speech [1]-[5]. After the environmental noise is estimated, the speech feature compensation procedure is completed by computing the clean feature estimate according to the minimum mean square error (MMSE) criterion. With this idea, the noise tracking and speech enhancement are separated.

One of the problems with these approaches is that, since the clean speech model is a Gaussian mixture model, each frame of data is enhanced independently. Without post-processing, this can result in artifacts, such as sharp single frame transitions, that are not part of the original clean speech signal. It is believed that joint noise and speech tracking should yield a better enhancement result, and that is the method followed by Droppo et al [6]. The use of a switching dynamical model to represent speech in the context of speech feature enhancement has probably first been proposed in that paper [6]. In their work, the time evolution of speech features is represented by a switching linear dynamic model (SLDM), which maintains the concept that as time progresses, the signal passes through several distinct states. This approach permits the feature enhancement not only at the level of individual frames but simultaneously at the level of local sequences of frames, and thus the performance will be improved. Recently, Kim et al [7] used a SLDM with a similar structure to the one in Droppo's work, to model the speech and then estimate the clean speech component simultaneously with the noise component. The graphical representation of the switching LDM used is shown in Fig. 1(a), where shaded nodes are observed, and clear nodes are hidden. Their SLDM assumes time dependence among the continuous speech features in one state, but not among the discrete states. Therefore, it could be treated as a combination of a Gaussian mixture model (GMM) with a linear dynamic model.

Our approach described in this paper uses a different SLDM structure which can be represented by Fig. 1(b). It is seen from Fig. 1(b) that unlike the SLDM presented in Fig. 1(a), the time-dependence among the discrete state variables and adjacent frames of speech are included. Our

Based on "Speech Feature Estimation under the Presence of Noise with a Switching Linear Dynamic Model", by J. Deng, M. Bouchard, and T. Yeap, which appeared in the Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing 2006, Toulouse, France, May 2006. © 2006 IEEE.

SLDM model could be treated as a combination of a hidden Markov model (HMM) with a linear dynamic model. Compared with the GMM, the state transition probabilities of an HMM tend to eliminate single frame errors in the output. This advantage also provides HMM-based switching models a potentially greater descriptive power over the GMM-based switching models. Kim's SLDM is closely related to the stochastic segment model [10], in which the segments are assumed to be independent of one another. While in our SLDM, the speech vector is propagated over the segment boundaries, which should provide a better model for co-articulation. At the same time, we use a different inference process to estimate the switching LDM parameters. In Kim's SLDM, the whole space of log spectra of clean speech is divided into a number of disjoint clusters. The statistics to estimate the SLDM parameters were computed based on a cluster index that was obtained by "hard" competition. We use a "soft" competition instead, hence it is possible to estimate at each time t the responsibility of each model and therefore detect a mode transition.

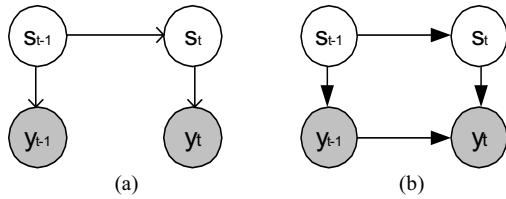


Figure. 1 Graphical representation of the switching LDM

In the paper we present experimental results demonstrating that, even with relatively small model sizes, substantial speech recognition rate improvements can be obtained compared with a baseline recognizer and previous methods. The rest of this paper is organized as follows: Section 2 describes how to model the sequence of clean speech features by a switching LDM. The method to compensate the noisy feature is presented in Section 3. Section 4 describes the experiments and presents the results. Conclusions are given in Section 5.

II. SLDM FOR CLEAN SPEECH FEATURES

To describe the clean speech features distribution, a switching linear dynamic model (LDM) obeys the system equation

$$\mathbf{y}_t = \boldsymbol{\mu}_{s_t=i} + \mathbf{B}_{1,s_t=i} \mathbf{y}_{t-1} + \dots + \mathbf{B}_{p,s_t=i} \mathbf{y}_{t-p} + \mathbf{e}_t \quad (1)$$

Equation (1) could be rewritten in the short-hand form

$$\mathbf{y}_t = \boldsymbol{\mu}_{s_t=i} + \mathbf{B}_{s_t=i} \mathbf{x}_t + \mathbf{e}_t \quad (2)$$

where \mathbf{x}_t is the column vector $[\mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-p}]^T$ and $\mathbf{B}_{s_t=i}$ is a matrix $(\mathbf{B}_{1,s_t=i} \dots \mathbf{B}_{p,s_t=i})$. \mathbf{B} and $\boldsymbol{\mu}$ are dependent on a hidden variable s_t at each time t . The state-dependent residual \mathbf{e}_t has a Gaussian distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}_{s_t}$. The graphic structure in Fig. 1(b) depicts the special case where $p=1$.

Assume that the discrete hidden states follow a Markov chain process with M states and that the state transition matrix is defined as

$$z(i, j) = p(s_{t+1} = j | s_t = i), \quad i, j \in (1, \dots, M).$$

Given the state sequence, the observation likelihood for \mathbf{y}_t given that the LDM is in state i at time t is

$$p(\mathbf{y}_t | \mathbf{x}_t, s_t = i) = (2\pi)^{-d/2} |\boldsymbol{\Sigma}_i|^{-1/2} \exp(-\frac{1}{2} [\mathbf{y}_t - \mathbf{B}_i \mathbf{x}_t - \boldsymbol{\mu}_i]^T \boldsymbol{\Sigma}_i^{-1} [\mathbf{y}_t - \mathbf{B}_i \mathbf{x}_t - \boldsymbol{\mu}_i]) \quad (3)$$

The parameters $\{\mathbf{B}_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ associated with the specified SLDM can be estimated from a set of clean speech training data using the standard EM algorithm [8]. The algorithm then iterates, using the current parameter estimate to compute the expected state occupancy

$$\gamma_t(i) = p(s_t = i | \mathbf{y}_{1:T}) \quad (4)$$

where $\gamma_t(i)$ is the probability that the HMM is in state i at time t , and it is calculated by the forward-backward algorithm [9]. The EM algorithm requires us to maximize the following expected log-likelihood, $Q(\lambda_0, \lambda)$, by choosing the parameters of the new model M .

$$Q(\lambda_0, \lambda) = \sum_i \sum_t P_{\lambda_0}(s_t = i | \mathbf{y}_{1:T}) \log P_{\lambda}(\mathbf{y}_t | \mathbf{x}_t, s_t) \quad (5)$$

λ_0 is the model corresponding to an initial estimate of the parameters, and $P_{\lambda_0}(s_t | \mathbf{y}_{1:T})$ stands for the probability of s_t conditioned on the observation sequence $\mathbf{y}_{1:T}$, calculated using the parameters of the model λ_0 . To present the result of this maximization, the following expected sufficient statistics are first introduced:

$$\begin{aligned} S_{YY',i} &= \sum_t \gamma_t(i) E_{ii}[\mathbf{y}_t \mathbf{y}_t'] \\ S_{Y,i} &= \sum_t \gamma_t(i) E_{ii}[\mathbf{y}_t] \\ S_{XX',i} &= \sum_t \gamma_t(i) E_{ii}[\mathbf{x}_t \mathbf{x}_t'] \\ S_{X,i} &= \sum_t \gamma_t(i) E_{ii}[\mathbf{x}_t] \\ S_{XY',i} &= \sum_t \gamma_t(i) E_{ii}[\mathbf{x}_t \mathbf{y}_t'] \\ S_{YX',i} &= \sum_t \gamma_t(i) E_{ii}[\mathbf{y}_t \mathbf{x}_t'] \end{aligned} \quad (6)$$

Maximizing $Q(\lambda_0, \lambda)$ with respect to \mathbf{B}_i and setting the derivative to zero, the following equation is obtained:

$$\mathbf{B}_i = (S_{YX',i} - \boldsymbol{\mu}_i S_{XX',i}^{-1}) S_{XX',i}^{-1} \quad (7)$$

Likewise, setting the derivatives of the objective function with respect to $\boldsymbol{\mu}_i$ to zero, another vector equation is obtained:

$$\boldsymbol{\mu}_i = \frac{S_{Y,i} - \mathbf{B}_i S_{X,i}}{\gamma_i} \quad (8)$$

where $\gamma_i = \sum_t \gamma_t(i)$. The new estimates for $\boldsymbol{\mu}_i$ and \mathbf{B}_i can now be obtained by solving the pair of simultaneous equations. Likewise, the re-estimation formula for $\boldsymbol{\Sigma}_i$ is:

$$\boldsymbol{\Sigma}_i = \frac{1}{\gamma_i} (S_{YY,i} - S_{YX,i} \mathbf{B}_i' - \mathbf{B}_i S_{XY,i} + \mathbf{B}_i S_{XX,i} \mathbf{B}_i' - \gamma_i \boldsymbol{\mu}_i \boldsymbol{\mu}_i') \quad (9).$$

The re-estimation formula for the state transition matrix is the same as an HMM [9]:

$$Z(i, j) = \frac{\sum_t p(s_{t-1} = i, s_t = j | \mathbf{y}_{1:T})}{\gamma_i} \quad (10).$$

Throughout the above analysis, we have assumed that the model is trained using a single token. However, a left-to-right model cannot, in general, be reliably trained with a single token. The multiple token training case analysis is similarly derived. Suppose we have N tokens $\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(N)}$ of length $T^{(1)}, \dots, T^{(N)}$, we replace each “ t (time) summation” by a “ l (token) summation and t (time) summation”: $\sum_t \Rightarrow \sum_{l=1}^N \sum_{t=1}^{T^{(l)}}$. The subsequent analysis is exactly the same as previously described.

The switching linear dynamic model trained using the above EM algorithm is guaranteed to reach only a local maximum likelihood solution. Because there are many local maxima, such models are sensitive to how they are initialized. For this reason, we need to initialize the model rather carefully. Both the HMM and the linear dynamical model must be initialized. The key point is to start with a good partition of the data, each being modeled by a linear dynamic model. The initialization algorithm is summarized as follows:

A fuzzy C-means clustering technique [11] is applied to cluster all the feature vectors into M segments. The process calculates the cluster membership degree $\alpha_j(t)$, and updates the cluster centers V_j iteratively. The fuzzy membership degree $\alpha_j(t)$ to cluster j at a time point t can be interpreted as to what degree the data point belongs to a specific state j . The aforementioned EM algorithm is performed to obtain the initial estimate for the LDM parameters in cluster j using (7)-(9), with $\alpha_j(t)$ instead of $\gamma_t(i)$. The initial probability and transition probability are set to be $1/M$.

III CLEAN FEATURES ESTIMATION

Assume that speech and noise are mixed linearly in the time domain. This corresponds to a nonlinear mixing in the log spectrum feature space as follows [1]:

$$\mathbf{o}_t = f(\mathbf{y}_t, \mathbf{n}_t) = \mathbf{y}_t + \log(\mathbf{I} + \exp(\mathbf{n}_t - \mathbf{y}_t)) \quad (11)$$

in which \mathbf{o}_t , \mathbf{y}_t and \mathbf{n}_t , respectively represent the log spectrum of noisy speech, of the hypothetical clean speech and of noise at the t^{th} frame.

To take into account the time-varying characteristics of the background noise, we model the sequence of noise features as the output of a first-order auto-regressive (AR) system excited by a zero mean Gaussian process \mathbf{v} with a covariance matrix \mathbf{Q}_n as follows [2]:

$$\mathbf{n}_t = \mathbf{A} \mathbf{n}_{t-1} + \boldsymbol{\mu}_n + \mathbf{v}_t \quad (12)$$

Combining equation (2) with (11) and (12) leads to a nonlinear state space model conditioned on $s_t = j$ as:

$$\mathbf{z}_t = \boldsymbol{\Phi}_j \mathbf{z}_{t-1} + \bar{\boldsymbol{\mu}}_j + \bar{\mathbf{e}}_{t,j} \quad (13)$$

$$\mathbf{o}_t = f(\mathbf{z}_t) = \mathbf{y}_t + \log(\mathbf{I} + \exp(\mathbf{n}_t - \mathbf{y}_t)) \quad (14)$$

where

$$\mathbf{z}_t = [\mathbf{y}_t^T \quad \mathbf{n}_t^T]^T \quad \bar{\mathbf{e}}_{t,j} = [\mathbf{e}_{t,j}^T \quad \mathbf{v}_t^T]^T \quad (15)$$

$$\boldsymbol{\Phi}_j = \begin{bmatrix} \mathbf{B}_j & \mathbf{0} \\ \mathbf{0} & \mathbf{A} \end{bmatrix} \quad \bar{\boldsymbol{\mu}}_j = \begin{bmatrix} \boldsymbol{\mu}_j \\ \boldsymbol{\mu}_n \end{bmatrix} \quad \mathbf{Q}_j = \begin{bmatrix} \boldsymbol{\Sigma}_j & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_n \end{bmatrix} \quad (16)$$

Both noise and clean speech components are incorporated in the state space, and are estimated simultaneously.

The major obstacle to use the switching LDM for enhancement is the computational burden that it brings. Optimal minimum mean squared error estimators involve a bank of filters tuned to all the possible parameter histories, which makes the cost in computations grow exponentially with data length [12]. To solve this problem, we show in this paper how the Interacting Multiple Model (IMM) algorithm [12] can be adapted to the nonlinear state-space model of feature dynamics presented above, to provide a sub-optimal approximation solution. A block diagram of the IMM algorithm is shown in Fig. 2.

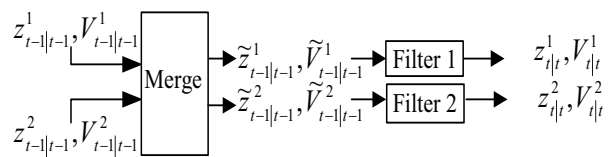


Figure. 2 Diagram of the IMM algorithm

There are M filters, each of which is supplied with a different input. Let us define

$$\hat{\mathbf{z}}_{t-1|t-1} = E(\mathbf{z}_{t-1} | \mathbf{o}_{1:t-1}) \quad (17)$$

$$\hat{\mathbf{V}}_{t-1|t-1} = \text{cov}(\mathbf{z}_{t-1} | \mathbf{o}_{1:t-1}) \quad (18)$$

$$\tilde{\mathbf{z}}_{t-1|t-1}^j = E(\mathbf{z}_{t-1} | s_t = j, \mathbf{o}_{1:t-1}) \quad (19)$$

$$\tilde{\mathbf{V}}_{t-1|t-1}^j = \text{cov}(\mathbf{z}_{t-1} | s_t = j, \mathbf{o}_{1:t-1}) \quad (20)$$

$$W_{t-1|t-1}^{ij} = p(s_{t-1} = i | s_t = j, \mathbf{o}_{1:t-1}) \quad (21)$$

$$L_t^j = p(\mathbf{o}_t | s_{t-1} = i, s_t = j, \mathbf{o}_{1:t-1}) \quad (22)$$

in which $\mathbf{o}_{1:t} = \{\mathbf{o}_1, \mathbf{o}_2, \dots, \mathbf{o}_t\}$ denotes the noisy speech observation sequence up to time t . The algorithm mixes the estimates according to the Markov transition probability, in order to allow the system to react to changes in the model in force. In that way, the input to the j^{th} filter becomes the best estimate of the state at time instant $t-1$, conditioned on the event that model j is in force at time instant t (the new sample time). $\tilde{\mathbf{z}}_{t-1|t-1}^j$ and $\tilde{\mathbf{V}}_{t-1|t-1}^j$ are then obtained according to

$$\tilde{\mathbf{z}}_{t-1|t-1}^j = \sum_i W_{t-1|t-1}^{ij} \hat{\mathbf{z}}_{t-1|t-1}^i \quad (23)$$

$$\begin{aligned} \tilde{\mathbf{V}}_{t-1|t-1}^j &= \sum_i W_{t-1|t-1}^{ij} \hat{\mathbf{V}}_{t-1|t-1}^i \\ &+ \sum_i W_{t-1|t-1}^{ij} (\hat{\mathbf{z}}_{t-1|t-1}^i - \tilde{\mathbf{z}}_{t-1|t-1}^j)(\hat{\mathbf{z}}_{t-1|t-1}^i - \tilde{\mathbf{z}}_{t-1|t-1}^j)^T \end{aligned} \quad (24)$$

The mixing probability $W_{t-1|t-1}^{ij} = p(s_{t-1} = i | s_t = j, \mathbf{o}_{1:t-1})$ is computed recursively with Bayes' rule

$$\begin{aligned} W_{t-1|t-1}^{ij} &= p(s_{t-1} = i | s_t = j, \mathbf{o}_{1:t-1}) \\ &= \frac{p(s_t = j | s_{t-1} = i, \mathbf{o}_{1:t-1}) p(s_{t-1} = i | \mathbf{o}_{1:t-1})}{\sum_i p(s_t = j | s_{t-1} = i, \mathbf{o}_{1:t-1}) p(s_{t-1} = i | \mathbf{o}_{1:t-1})} \end{aligned} \quad (25)$$

Given a noisy speech feature vector \mathbf{o}_t ($t=1, \dots, T$), $\hat{\mathbf{z}}_{t|t}$ is found from the interacting multiple model algorithm using a bank of Extended Kalman Filters (EKF) with the following steps performed in sequence:

for the j^{th} ($j=1, \dots, M$) extended Kalman filter:

$$(\hat{\mathbf{z}}_{t|t}^j, \mathbf{V}_{t|t}^j, L_t^j) = \text{EKF}(\tilde{\mathbf{z}}_{t-1|t-1}^j, \tilde{\mathbf{V}}_{t-1|t-1}^j, \mathbf{o}_t; \Phi_j, \mathbf{u}_j, \mathbf{Q}_j) \quad (26)$$

The model probabilities $p(s_t = j | \mathbf{o}_{1:t})$ are updated according to

$$\begin{aligned} p(s_t = j | \mathbf{o}_{1:t}) &= \frac{1}{\Omega} \sum_i p(s_{t-1} = i, s_t = j, \mathbf{o}_{1:t}) \\ &= \frac{1}{\Omega} \sum_i p(\mathbf{o}_t | s_{t-1} = i, s_t = j, \mathbf{o}_{1:t-1}) p(s_{t-1} = i, s_t = j | \mathbf{o}_{1:t-1}) \\ &= \frac{1}{\Omega} \sum_i p(\mathbf{o}_t | s_{t-1} = i, s_t = j, \mathbf{o}_{1:t-1}) p(s_t = j | s_{t-1} = i, \mathbf{o}_{1:t-1}) p(s_{t-1} = i | \mathbf{o}_{1:t-1}) \end{aligned} \quad (27)$$

where Ω is a scale factor.

$$\Omega = \sum_j \sum_i p(\mathbf{o}_t | s_{t-1} = i, s_t = j, \mathbf{o}_{1:t-1}) p(s_t = j | s_{t-1} = i, \mathbf{o}_{1:t-1}) p(s_{t-1} = i | \mathbf{o}_{1:t-1})$$

The estimated output at time t is calculated according to

$$\hat{\mathbf{z}}_{t|t} = \sum_j p(s_t = j | \mathbf{o}_{1:t}) \hat{\mathbf{z}}_{t|t}^j \quad (28)$$

The extended Kalman filter algorithm in (26) is defined as:

$$\hat{\mathbf{z}}_{t|t-1}^j = \Phi_j \hat{\mathbf{z}}_{t-1|t-1}^j + \mathbf{u}_j \quad (29)$$

$$\hat{\mathbf{V}}_{t|t-1}^j = \Phi_j \hat{\mathbf{V}}_{t-1|t-1}^j \Phi_j^T + \mathbf{Q}_j \quad (30)$$

$$\mathbf{K}_t^j = \hat{\mathbf{V}}_{t|t-1}^j \mathbf{H}^T (\mathbf{H} \hat{\mathbf{V}}_{t|t-1}^j \mathbf{H})^{-1} \quad (31)$$

$$\hat{\mathbf{z}}_{t|t}^j = \hat{\mathbf{z}}_{t|t-1}^j + \mathbf{K}_t^j (\mathbf{o}_t - f(\hat{\mathbf{z}}_{t|t-1}^j)) \quad (32)$$

$$\hat{\mathbf{V}}_{t|t}^j = (\mathbf{I} - \mathbf{K}_t^j \mathbf{H}) \hat{\mathbf{V}}_{t|t-1}^j \quad (33)$$

$$L_t^j = N(f(\hat{\mathbf{z}}_{t|t-1}^j), \mathbf{H} \hat{\mathbf{V}}_{t|t-1}^j \mathbf{H}) \quad (34)$$

where $\mathbf{H} = \left[\frac{\partial f}{\partial \mathbf{y}} \Big|_{(z_{t|t-1}^j)} \quad \frac{\partial f}{\partial \mathbf{n}} \Big|_{(z_{t|t-1}^j)} \right]$

$$\frac{\partial f}{\partial \mathbf{y}} = \text{diag} \left(\frac{1}{1 + \exp(\mathbf{n} - \mathbf{y})} \right) \quad (35)$$

$$\frac{\partial f}{\partial \mathbf{n}} = \text{diag} \left(1 - \frac{1}{1 + \exp(\mathbf{n} - \mathbf{y})} \right) \quad (36)$$

The extended Kalman filter (EKF) algorithm in (26) can also be replaced by the Unscented Kalman filter (UKF) algorithm, which normally provides an improvement in performance [13],[14]. The Unscented Kalman filter (UKF) is a straightforward extension of the unscented transformation (UT). The UT is a method for calculating the statistics of a random variable by propagating a small set of deterministically chosen (Sigma) points through a nonlinear system [13]. No explicit calculations of Jacobians or Hessians are necessary to implement this algorithm. Compared with the EKF, the UKF can handle nonlinearities without using numerical derivatives and provides higher order approximation for both Gaussian and non-Gaussian distributions. We will show next how to replace the extended Kalman filter (EKF) algorithm in (26) to the Unscented Kalman filter (UKF).

The UKF algorithm is given as follows. Firstly, form $\hat{\mathbf{z}}_{t-1|t-1}^a, \mathbf{V}_{t-1|t-1}^a, \chi^a$ as in [14]:

$$\hat{\mathbf{z}}_{t-1|t-1}^a = [\tilde{\mathbf{z}}_{t-1|t-1}^j; \mathbf{u}_j] \quad (37)$$

$$\mathbf{V}_{t-1|t-1}^a = \begin{bmatrix} \tilde{\mathbf{V}}_{t-1|t-1}^j & \mathbf{0}; & \mathbf{0} & \mathbf{Q}_j \end{bmatrix} \quad (38)$$

$$\chi^a = [(\chi^z)^T \quad (\chi^v)^T]^T \quad (39)$$

Sigma matrices: $i=1, \dots, L$, L is the dimension of $\hat{\mathbf{z}}_{t-1|t-1}^a$

$$\begin{aligned} \chi_0^a(t-1|t-1) &= \hat{\mathbf{z}}_{t-1|t-1}^a \\ \chi_i^a(t-1|t-1) &= \hat{\mathbf{z}}_{t-1|t-1}^a + \left(\sqrt{(L+\lambda) \mathbf{V}_{t-1|t-1}^a} \right)_i \\ \chi_{i+L}^a(t-1|t-1) &= \hat{\mathbf{z}}_{t-1|t-1}^a - \left(\sqrt{(L+\lambda) \mathbf{V}_{t-1|t-1}^a} \right)_i \end{aligned} \quad (40)$$

Time update equations: $i = 0, \dots, 2L$

$$\chi_i^z(t|t-1) = \mathbf{A} \chi_i^z(t-1|t-1) + \chi_i^v(t-1|t-1) \quad (41)$$

$$\hat{\mathbf{z}}_{t|t-1}^j = \sum_{i=0}^{2L} W_i^{(m)} \chi_i^z(t|t-1) \quad (42)$$

$$\mathbf{V}_{t|t-1}^j = \sum_{i=0}^{2L} W_i^{(c)} [\chi_i^z(t|t-1) - \hat{\mathbf{z}}_{t|t-1}^j][\chi_i^z(t|t-1) - \hat{\mathbf{z}}_{t|t-1}^j]^T \quad (43)$$

$$Y_i(t|t-1) = f(\chi_i^z(t|t-1)) \quad i = 0, \dots, 2L \quad (44)$$

$$\hat{\mathbf{o}}_{t|t-1}^j = \sum_{i=0}^{2L} W_i^{(m)} Y_i(t|t-1) \quad (45)$$

Measurement update equations:

$$\mathbf{P}_{\hat{\mathbf{o}}\hat{\mathbf{o}}}(t) = \sum_{i=0}^{2L} W_i^{(c)} [Y_i(t|t-1) - \hat{\mathbf{o}}_{t|t-1}^j][Y_i(t|t-1) - \hat{\mathbf{o}}_{t|t-1}^j]^T \quad (46)$$

$$\mathbf{P}_{\hat{\mathbf{z}}\hat{\mathbf{z}}}(t) = \sum_{i=0}^{2L} W_i^{(c)} [\chi_i^z(t|t-1) - \hat{\mathbf{z}}_{t|t-1}^j][Y_i(t|t-1) - \hat{\mathbf{o}}_{t|t-1}^j]^T \quad (47)$$

$$\mathbf{K}_t = \mathbf{P}_{\hat{\mathbf{z}}\hat{\mathbf{o}}}(t) \mathbf{P}_{\hat{\mathbf{o}}\hat{\mathbf{o}}}(t)^{-1} \quad (48)$$

Filtered estimate of the state vector:

$$\hat{\mathbf{z}}_{t|t}^j = \hat{\mathbf{z}}_{t|t-1}^j + \mathbf{K}_t(\mathbf{o}_t - \hat{\mathbf{o}}_{t|t-1}^j) \quad (49)$$

Filtered state-error covariance matrix:

$$\mathbf{V}_{t|t}^j = \mathbf{V}_{t|t-1}^j - \mathbf{K}_t \mathbf{P}_{\hat{\mathbf{o}}\hat{\mathbf{o}}}(t) \mathbf{K}_t^T \quad (50)$$

$$L_t^j = N(\mathbf{o}_t; \hat{\mathbf{o}}_{t|t}^j, \mathbf{P}_{\hat{\mathbf{o}}\hat{\mathbf{o}}}(t)) \quad (51)$$

For each frame of incoming noisy speech, the algorithm described above gives the estimate of the speech in the frame.

IV EXPERIMENTS

The proposed speech enhancement scheme has been evaluated on the Aurora2 database, using the standard recognition tasks designed for this database [15]. The recognition system used in our evaluation experiments is obtained using the HTK scripts provided by the Aurora2 database. To conform to the observation model presented in (11), the feature generation was modified slightly from the reference implementation. In particular, we changed from using the spectral magnitude to using the power spectral density as the input to the Mel-frequency filterbank. For each noise type, a specific noise model was built. The parameters for the noise model were trained on the first ten frames of the noisy utterances. All the experimental conditions were maintained the same as those in [6][7].

A global SLDM model with 16 hidden linear dynamic models (LDMs) has been trained to describe the time movement of the speech features. It should be noted that the training is off-line and the speech model obtained will

be kept fixed during the whole database testing. Detailed recognition rates (%) using our proposed SLDM approach for each of the four noise conditions and for each of the SNRs in the Aurora2 set-A are provided in Table 1. If the EKF in this approach is replaced by the UKF, the results obtained are shown in Table 2. The simulation results are summarized in Table 3, averaged across 0dB to 20dB. For the purpose of comparison, the results reported in [6],[7] are also listed in Table 3. The baseline results without noise processing are also given. The number of LDMs used in the SLDM to obtain the results presented in Table 2 is given in Table 4.

TABLE 1 RECOGNITION RATES (%) WITH SPEECH ENHANCED BY EXTENDED KALMAN FILTER

	Subway	Babble	Car	Exhibition	Average
20dB	98.25	98.67	98.78	97.78	98.37
15dB	96.68	97.10	98.48	96.73	97.24
10dB	91.10	93.53	94.63	91.61	92.72
5dB	81.67	81.68	84.19	79.57	81.78
0dB	56.80	47.37	55.98	51.16	52.83
Average	84.90	83.67	86.41	83.37	84.59

TABLE 2 RECOGNITION RATES (%) WITH SPEECH ENHANCED BY UNSCENTED KALMAN FILTER

	Subway	Babble	Car	Exhibition	Average
20dB	98.25	98.67	98.78	97.78	98.37
15dB	96.76	97.07	98.39	96.42	97.16
10dB	93.52	94.13	96.77	93.40	94.46
5dB	86.37	85.12	89.66	86.70	86.96
0dB	70.73	56.41	74.25	68.03	67.36
Average	89.13	86.28	91.57	88.47	88.86

TABLE 3 SUMMARY OF RECOGNITION RATES (%)

	Subway	Babble	Car	Exhibition	Average
Baseline	70.22	51.35	60.28	65.39	61.81
SLDM [6]	80.55	67.52	86.19	77.82	78.02
SLDM [7]	80.48	80.98	86.82	84.53	83.20
Proposed EKF	84.90	83.67	86.41	83.37	84.59
Proposed UKF	89.13	86.28	91.57	88.47	88.86

TABLE 4 NUMBER OF LDMS IN THE SLDM

Number of LDMS in the SLDM	
SLDM [7]	128
SLDM [6]	16
Proposed	16

Our SLDM approach based on the linear predictive HMM outperforms the SLDM approach from [6],[7]. More specifically, with the same number of LDMS in the SLDM, the proposed SLDM approach outperforms the SLDM approach in [6], especially when the environmental noise is babble noise. As we compare our results with those obtained in [7], even with only 1/8 of the computational load used for the experiments in [7], a better performance is obtained, especially when the

environmental noise is babble noise and subway noise. If we replace the Extended Kalman filter with the Unscented Kalman filter, a clear performance improvement can also be observed. It has been shown in [6] that with the increase of the number of hidden LDMs, better results could be achieved. With the increase of the number of LDMs, the SLDM can describe the acoustic space more accurately. Therefore, it is promising that if the number of linear dynamic models in the SLDM is increased, even better results than our present results should be obtained.

V CONCLUSIONS

This paper explores the use of a switching linear dynamic model (SLDM) for speech features enhancement in the log-spectral domain under the presence of additive noise. The SLDM can capture the temporal correlations among adjacent frames of speech in a more accurate way compared with previous work. The simulation results have confirmed the improved performance of the proposed approach over baseline recognition and over previously published SLDM methods.

REFERENCES

- [1] P.J. Moreno, Speech recognition in noisy environments, *Ph.D Thesis*, Carnegie Mellon University, 1996
- [2] B. Raj, R. Singh, and R. Stern "On tracking noise with linear dynamical system models", *IEEE ICASSP'04*, Vol. 1, pp. I-965- I-968, May 2004
- [3] S. Kim, "IMM-based estimation for slowly evolving environments", *IEEE Signal Processing Letters*, Vol. 5, pp. 146-149, June 1998
- [4] J. Deng, M. Bouchard, and T.H. Yeap, "Noise compensation using interacting multiple Kalman filters", *InterSpeech 2005*, pp. 949-952, Sept. 2005.
- [5] L. Deng, J. Droppo and A. Acero, "Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features", *IEEE Trans. Speech and Audio Processing*, Vol. 12, pp. 218-233, May 2004.
- [6] J. Droppo and A. Acero, "Noise robust speech recognition with a switching linear dynamic model", *IEEE ICASSP'04*, Vol. 1, pp. I-953- I-956, May 2004
- [7] N.S. Kim, W. Lim, and R. Stern, "Feature compensation based on switching linear dynamic model", *IEEE Signal Processing Letters*, Vol. 12, pp. 473-476, June 2005.
- [8] P. Kenny, M. Lenning, and P. Mermelstein, "A linear predictive HMM for vector-valued observations with applications to speech recognition", *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 38, pp. 220-225, Feb. 1990
- [9] L.R. Rabiner, "A tutorial on Hidden Markov Models and selected application in speech recognition", *Proc. IEEE*, Vol. 77, pp. 257-286, Feb. 1989
- [10] V. Digalakis, J.R. Rohlicek, and M. Ostendorf, "ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition", *IEEE Transactions on Speech and Audio Processing*, Vol. 1, No. 4, pp. 431-442, 1993.
- [11] J. Bezdek, *Pattern recognition with fuzzy objective function algorithms*, Plenum press, New York, 1981
- [12] Y. B. Shalom and X. R. Li, *Estimation and Tracking*, Boston Artech house, 1993
- [13] S. J. Julier, and J. K. Uhlmann, "A new extension of the Kalman Filter to Nonlinear Systems." *Proceedings of AeroSense: The 11th Int. Symp. On Aerospace/Defence Sensing, Simulation and Controls*, 1997 (Available at <http://citeseer.ist.psu.edu/julier97new.html>)
- [14] E. A. Wan, R. V. D. Merwe, "The unscented Kalman filter for nonlinear estimation", *IEEE Adaptive Systems for Signal Processing, Communications, and Control Symposium AS-SPCC.*, pp.153 – 158, Oct. 2000
- [15] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *Proceedings of ISCA ITRW ASR2000 on Automatic Speech recognition: Challenges for the next Millennium*, Paris, France, Sept. 2000.

Jianping Deng received the bachelor degree from North China University of Electric Power, China and the Master degree from Nanyang Technological University, Singapore. She is currently a PhD student in the School of Information Technology and Engineering, University of Ottawa, Canada. Her research interests include adaptive filtering, speech enhancement, speech and speaker recognition.

Martin Bouchard received the B.Eng., M.Sc.A. and Ph.D. degrees in electrical engineering from Sherbrooke University in 1993, 1995 and 1997 respectively. He previously worked for Bechtel-Lavalin in 1991, for CAE Electronics in 1992-1993, and from 1995 to 1997 for SoftdB, which he co-founded. In January 1998, he joined the School of Information Technology and Engineering (SITE) at the University of Ottawa, where is he currently an associate professor. Some of his current research interests are the applications of signal processing to speech / audio / acoustics / hearing aids, noise suppression, adaptive filtering and blind source separation.

Tet Hin Yeap received the B.A.Sc. degree in electrical engineering from Queen's University, Kingston, ON, Canada, in 1982 and the M.A.Sc. and Ph.D. degrees in electrical engineering from the University of Toronto, Toronto, ON, Canada, in 1984 and 1991, respectively. He is an Associate Professor in the School of Information Technology and Engineering, University of Ottawa, Ottawa, ON, Canada. He is also a Director of the Bell Canada Advanced Research Laboratory, Ottawa. His research interests include broadband access architecture, neural networks, multimedia, parallel architectures, and dynamics and control.