Advanced switching linear dynamic system using enhanced clustering method for speech feature mapping

Chang Woo Han¹, Tae Gyoong Kang², Doo Hwa Hong³, and Nam Soo Kim⁴
School of Electrical Engineering and INMC
Seoul National University, Seoul 151-744, KOREA

ABSTRACT
It is generally known that the performance of a speech recognition system deteriorates in the presence of acoustic mismatch caused by system and environmental factors. In order to ameliorate the degradation in recognition performance, we can either suppress the mismatch in the feature domain or adapt the model to match the input. One of the successful previous stereo-data-based feature mapping approaches is switching linear dynamic system (SLDS), which is considered as an extension of switching linear dynamic model (SLDM). The main advantage of the SLDS is that it enables a systematic implementation of the sequence-to-sequence mapping instead of the traditional vector-to-vector mapping. To achieve a better performance of the SLDS, assigning proper cluster to each frame should be taken into consideration. In this paper, we propose an enhanced clustering method to improve the performance of the SLDS. In the proposed clustering method, we utilize the local trajectory of the input feature vector stream. The proposed approach is applied to compensate channel distortion in speech recognition and shows significant improvement in recognition performance.
Keywords: Switching Linear Dynamic System, Feature Mapping, Channel Compensation

1. INTRODUCTION
In general, the performance of a speech recognition system degrades when there is a mismatch between test and training conditions. There are several causes that lead to acoustic mismatch such as differences of channel characteristics, audio devices, reverberations, etc., which are categorized into two factors: system and environmental factors [1]. The system factors are caused by mismatch of speech capturing devices such as microphones, analog circuits, A/D converters and data compression modules while the environmental factors include additive background noise and reverberation that affect the speech quality.

In order to ameliorate the degradation in recognition performance, we can either suppress the mismatch in the feature domain or adapt the model to match the input. In the feature mapping technique, a feature extracted in the test environment is mapped closer to the training feature. In the model adaptation technique, on the other hand, acoustic models are modified to match the input speech feature more closely.

Depending on the type of training or adaptation data, parameter estimation approaches for feature mapping can be divided into stereo data based and blind techniques. Stereo data based technique is applied when there exists a database of simultaneous recordings obtained in both the reference and target conditions, and feature mapping rules are derived from the difference between the pair of feature vectors [2]-[4]. In the blind technique, on the other hand, only the input feature vectors are given and the information related to the target feature vectors is usually provided by statistical models such as the Gaussian mixture model (GMM), hidden Markov model (HMM) and switching linear dynamic model (SLDM) [5]-[7]. In general, feature mapping for the blind technique is done

¹ cwhan@hi.snu.ac.kr
² tgkang@hi.snu.ac.kr
³ dhhong@hi.snu.ac.kr
⁴ nkim@snu.ac.kr
based on either the minimum mean square error (MMSE) or the maximum likelihood (ML) criterion.

In this paper, we focus on the stereo data based feature mapping technique when the channel
distortion is mainly caused by an environmental factor. The stereo data set consists of data captured
with the same conditions as used in the speech recognition system training and data collected in
different environments. Among traditional stereo data based feature mapping techniques, SPLICE
and switching linear dynamic system (SLDS) are successfully implemented. SPLICE is a
frame-based bias removal algorithm for feature enhancement under additive noise distortion, channel
distortion or a combination of the two [3]. SLDS is considered as an extension of SLDM [6]. One of
the prominent advantages of the SLDS is that it enables a systematic implementation of
sequence-to-sequence mapping instead of the traditional vector-to-vector mapping, which makes it
unique in that SPLICE and SLDM are vector-to-vector mapping approaches.

In this paper, we propose an enhanced clustering method to improve the performance of the SLDS.
In the SLDS, assigning the appropriate cluster to each frame should be taken into consideration to
achieve a better performance. To determine a proper cluster, past inputs as well as current input
should be considered especially when there is channel distortion caused by environmental factors
such as reverberation. In the proposed clustering method, we utilize not only a current frame but the
local trajectory of the input feature vector stream. The performance of the proposed method is
evaluated with speech recognition experiments. The proposed algorithm shows better performance
than other approaches when evaluated with the AURORA-5 task where various kinds of mismatches
between the training and test data caused by background noises, different microphones and acoustic
reverberation exist.

2. SWITCHING LINEAR DYNAMIC SYSTEM

Let \( x_t \) and \( y_t \) respectively denote a \( d_x \)-dimensional input feature vector and
\( d_y \)-dimensional output feature vector at time \( t \). Then our goal is to predict the output feature vector
sequence, \( Y = (y_1, y_2, \cdots, y_T) \), through some process when only the input sequence,
\( X = (x_1, x_2, \cdots, x_T) \), is given.

We assume that the feature mapping process is modeled by \( K \) different linear dynamic systems
(LDSs). In our proposed SLDS, when the \( k \)-th LDS is applied, the feature mapping process is
approximated as follows [2]:

\[
\begin{align*}
z_{t+1} &= A^{(k)} z_t + B^{(k)} x_t + u_t^{(k)} \quad (1) \\
y_t &= C^{(k)} z_t + D^{(k)} x_t + w_t^{(k)} \quad (2)
\end{align*}
\]

where \( A^{(k)} \), \( B^{(k)} \), \( C^{(k)} \) and \( D^{(k)} \) are matrices with the dimension \( d_z \times d_z \), \( d_z \times d_x \), \( d_y \times d_z \)
and \( d_y \times d_x \), respectively, and \( z_t \) is the \( d_z \)-dimensional vector which is called the hidden state. In
(1) and (2), \( u_t^{(k)} \) and \( w_t^{(k)} \) are random vectors with a Gaussian distribution as follows:

\[
\begin{align*}
u_t^{(k)} &\sim \mathcal{N}(\mu_u^{(k)}, Q^{(k)}) \quad (3) \\
w_t^{(k)} &\sim \mathcal{N}(\mu_w^{(k)}, R^{(k)}) \quad (4)
\end{align*}
\]

where \( \mathcal{N}(\mu, \Sigma) \) means a Gaussian PDF with the mean vector \( \mu \) and covariance matrix \( \Sigma \).

Once the parameters of \( k \)-th LDS, \( \lambda^{(k)} = \{ A^{(k)}, B^{(k)}, C^{(k)}, D^{(k)}, \mu_u^{(k)}, \mu_w^{(k)}, Q^{(k)}, R^{(k)} \} \), are given,
the output feature vector sequence can be generated from the input sequence, \( X \), as follows:

\[
z_{t+1} = A^{(k)} z_t + B^{(k)} x_t + \mu_u^{(k)} \quad (5)
\]
3. ENHANCED CLUSTERING METHOD

Determining an appropriate LDS among the $K$ candidate models at each time is very important in SLDS-based feature mapping. The LDS selection rule should be solely dependent on the input feature vector sequence because the output feature vector sequence is not available at runtime.

In the previous study, the input vector $x_t$ was simply divided into $K$ disjoint clusters [2]. However, especially when a frame is influenced by the surrounding frame such as reverberation environment, it is advantageous to consider the local trajectory of the input feature vector stream. For this reason, here we propose an enhanced clustering method and apply the principal component analysis (PCA) method for data reduction.

Let $\bar{X}_{t, \tau} = [x_{t-\tau}, x_{t-\tau+1}, \ldots, x_{t}]^T$ and $\mathbf{\Sigma}$ denote the $M$-dimensional concatenation of $\tau + 1$ feature vectors around time $t$ and its covariance matrix, respectively, with the prime denoting the transpose of a vector or a matrix. Then the eigenvalue and eigenvector matrices, $\mathbf{\Lambda}$ and $\mathbf{V}$, can be obtained from a singular value decomposition of the $\mathbf{\Sigma}$ as follows [8]:

$$ \mathbf{V}^T \mathbf{\Sigma} \mathbf{V} = \mathbf{\Lambda} \tag{7} $$

with

$$ \Lambda(p, q) = \begin{cases} e_m, & p = q = m \\ 0, & p \neq q \\ \end{cases}, \quad e_i \geq e_j \text{ for } i < j \tag{8} $$

$$ \mathbf{V} = [v_1, v_2, \ldots, v_{M-1}, v_M] \tag{9} $$

where $e_m$ and $v_m$ are ordered eigenvalue and corresponding eigenvector, respectively, and $m = 1, 2, \ldots, d_M$.

Let $\mathbf{W}$ be a $M \times L$ PCA transformation matrix. Then, an $L$-dimensional projected feature vector $\bar{x}_{t, \tau}^L$ can be calculated as follows [8]:

$$ \bar{x}_{t, \tau}^L = \mathbf{W}(\bar{x}_{t, \tau} - E(\bar{x}_{t, \tau})) \tag{10} $$

with

$$ \mathbf{W} = [v_1, v_2, \ldots, v_{L-1}, v_L], \quad 1 \leq L \leq M. \tag{11} $$

Since $L$ is usually set much smaller than the dimension of $\bar{x}_{t, \tau}$, $\bar{x}_{t, \tau}^L$ is a lower-dimensional vector. We then train a codebook for $\{\bar{x}_{t, \tau}^L\}$ with the use of a conventional GMM training algorithm.

For each time $t$, we first get the local trajectory vector $\bar{x}_{t, \tau}$ and project it onto the subspace spanned by the $L$ PCA basis vectors, which results in the low-dimensional vector $\bar{x}_{t, \tau}^L$. If the hard decision technique is employed, the LDS identity, $k_t$, which corresponds to the nearest codeword at time $t$ is found and the output feature vector stream is generated by the following

$$ z_{t+1} = A^{(k)}z_t + B^{(k)}x_t + \mu^{(k)} \tag{12} $$

$$ y_t = C^{(k)}z_t + D^{(k)}x_t + \mu^{(k)}. \tag{13} $$

In contrast, we can compute the a posteriori probability, $p(k | \bar{x}_{t, \tau}^L)$ of each cluster $k$ when soft decision is adopted. By taking advantage of these posterior probabilities we can further modify (12)
and (13) as

$$z_{t+1} = \sum_{k=1}^{K} p(k | x_{t}^{k}) \left[ A(k) z_{t} + B(k) x_{t} + \mu_{u}^{(k)} \right]$$

(14)

$$y_{t} = \sum_{k=1}^{K} p(k | x_{t}^{k}) \left[ C(k) z_{t} + D(k) x_{t} + \mu_{u}^{(k)} \right].$$

(15)

4. SLDS PARAMETER ESTIMATION

The SLDS parameters $\lambda = \{ \lambda^{(1)}, \lambda^{(2)}, \ldots, \lambda^{(K)} \}$ are estimated from a set of stereo speech data [2]. In the stereo data set, a reference feature vector stream and a target feature vector sequence that we want to predict respectively correspond to $X$ and $Y$ in the previous section. For simplicity, we assume that a hard decision clustering scheme is employed to estimate the parameters. Then the LDS identity, $k_{t}$, varies with time and is determined to the nearest codeword at time $t$.

We apply the ML criterion for parameter estimation in SLDS. Since the state variable $z_{t}$ is hidden, it is impractical to maximize the likelihood function directly. Instead, we apply the expectation maximization (EM) algorithm which iteratively increases the likelihood. The complete data log-likelihood is given by

$$L(X, Y, Z | \lambda) = -\sum_{t=1}^{T} \left( z_{t+1} - A(k) z_{t} + B(k) x_{t} + \mu_{u}^{(k)} \right)^{T} \left[ Q(k) \right]^{-1} \left( z_{t+1} - A(k) z_{t} + B(k) x_{t} + \mu_{u}^{(k)} \right)$$

$$- \sum_{t=1}^{T} \left( y_{t} - C(k) z_{t} + D(k) x_{t} + \mu_{u}^{(k)} \right)^{T} \left[ R(k) \right]^{-1} \left( y_{t} - C(k) z_{t} + D(k) x_{t} + \mu_{u}^{(k)} \right)$$

$$- \sum_{t=1}^{T} \log|Q(k)| - \sum_{t=1}^{T} \log|R(k)| + \text{Constant.}$$

(16)

where $| \cdot |$ means the determinant of a square matrix.

The general approach to estimate parameters is similar to the technique proposed in [6]. At first, the smoothed estimate for the hidden state sequence $Z = (z_{1}, z_{2}, \ldots, z_{T})$ is obtained conditioned on the current SLDS parameters and then the parameters are updated so as to maximize the complete data likelihood. Given the input and output feature vectors sequences, $X$ and $Y$, smoothed estimates for the hidden state sequence $Z$ and some of its statistics are obtained by means of the traditional Kalman filtering algorithm [9].

After the Kalman filtering step is completed, the parameters are updated according to the following criterion:

$$\hat{\lambda} = \arg \max_{\lambda} \Phi(\lambda, \hat{\lambda})$$

$$= \arg \max_{\lambda} \int L(X, Y, Z | \lambda) p(Z | X, Y, \lambda) dZ$$

(17)

where $\hat{\lambda}$ and $\lambda$ represent the updated and current SLDS parameters, respectively, and $p(Z | X, Y, \lambda)$ is the posterior PDF of the hidden state sequence derived from the Kalman filtering step. The two procedures of Kalman filtering and parameter updating are iterated until convergence.

The maximization of the auxiliary function $\Phi(\lambda, \hat{\lambda})$ is possible by taking the gradient such that

$$\frac{\partial}{\partial \lambda} \Phi(\lambda, \hat{\lambda}) \bigg|_{\lambda=\hat{\lambda}} = 0.$$  

(18)

For convenience of the formulation, we assume that $Y$ is generated from $X$ through a single LDS. Once the update equations of single LDS parameters are derived, it is not difficult to extend these to the case of SLDS parameters. Let $\hat{\lambda} = \{ \hat{A}, \hat{B}, \hat{\mu}_{u}, \hat{C}, \hat{D}, \hat{\mu}_{u}, \hat{Q}, \hat{R} \}$ be the updated
parameters of this LDS. Then, the solution to (18) is gives us the following equations:

\[
\begin{bmatrix}
\sum_{t=1}^{T-1} \hat{z}_t z_i \\
\sum_{t=1}^{T-1} x_t z_i \\
\sum_{t=1}^{T-1} z_i \\
\end{bmatrix}
\begin{bmatrix}
\sum_{t=1}^{T-1} \hat{z}_t x_i' \\
\sum_{t=1}^{T-1} x_t x_i' \\
\sum_{t=1}^{T-1} x_i' \\
\end{bmatrix}
\begin{bmatrix}
\sum_{t=1}^{T-1} \hat{z}_t z_i z_{r+1} \\
\sum_{t=1}^{T-1} x_t z_i z_{r+1} \\
\sum_{t=1}^{T-1} z_i z_{r+1} \\
\end{bmatrix}
\begin{bmatrix}
\hat{A}' \\
\hat{B}' \\
\hat{C}' \\
\end{bmatrix}
= 
\begin{bmatrix}
\sum_{t=1}^{T-1} \hat{z}_t z_i z_{r+1} \\
\sum_{t=1}^{T-1} x_t z_i z_{r+1} \\
\sum_{t=1}^{T-1} z_i z_{r+1} \\
\end{bmatrix}
\begin{bmatrix}
\mu_w \\
\mu_w \\
\mu_w \\
\end{bmatrix}.
\]

(19)

\[
\begin{bmatrix}
\sum_{t=1}^{T} \hat{z}_t z_i' \\
\sum_{t=1}^{T} x_t z_i' \\
\sum_{t=1}^{T} z_i' \\
\end{bmatrix}
\begin{bmatrix}
\sum_{t=1}^{T} \hat{z}_t x_i' \\
\sum_{t=1}^{T} x_t x_i' \\
\sum_{t=1}^{T} x_i' \\
\end{bmatrix}
\begin{bmatrix}
\sum_{t=1}^{T} \hat{z}_t z_i z_{r+1} \\
\sum_{t=1}^{T} x_t z_i z_{r+1} \\
\sum_{t=1}^{T} z_i z_{r+1} \\
\end{bmatrix}
\begin{bmatrix}
\hat{C}' \\
\hat{D}' \\
\end{bmatrix}
= 
\begin{bmatrix}
\sum_{t=1}^{T} \hat{z}_t y_i' \\
\sum_{t=1}^{T} x_t y_i' \\
\sum_{t=1}^{T} y_i' \\
\end{bmatrix}
\begin{bmatrix}
\mu_w \\
\mu_w \\
\mu_w \\
\end{bmatrix}.
\]

(20)

In (19) and (20),

\[
\hat{z}_i = E[z_i | X, Y, \hat{\lambda}] 
\]

(21)

\[
\hat{z}_i z_i' = E[z_i z_i' | X, Y, \hat{\lambda}] 
\]

(22)

\[
\hat{z}_i z_{r+1} = E[z_i z_{r+1} | X, Y, \hat{\lambda}] 
\]

(23)

where \( \hat{z}_i \), \( z_i z_i' \) and \( z_i z_{r+1} \) are obtained during the Kalman filtering step, and \( E[\cdot] \) denotes the expectation operation. Finally, the covariance matrices, \( \hat{Q} \) and \( \hat{R} \), are updated as follows:

\[
\hat{Q} = \frac{1}{T-1} \sum_{t=1}^{T-1} E \left[ \left( z_{r+1} - \hat{A} z_i - \hat{B} x_i - \hat{\mu}_w \right)' \left( z_{r+1} - \hat{A} z_i - \hat{B} x_i - \hat{\mu}_w \right) | X, Y, \hat{\lambda} \right] 
\]

(24)

\[
\hat{R} = \frac{1}{T} \sum_{t=1}^{T} E \left[ \left( y_i - \hat{C} z_i - \hat{D} x_i - \hat{\mu}_w \right)' \left( y_i - \hat{C} z_i - \hat{D} x_i - \hat{\mu}_w \right) | X, Y, \hat{\lambda} \right] 
\]

(25)

5. Experimental Results

We performed experiments to evaluate the robustness of the proposed approach to channel distortion caused by system and environmental factors with the AURORA-5 DB [10]. In the AURORA-5, the test data consisted of two sets: G. 712 filtered and non-filtered sets. Both of the sets comprised clean and noisy speech utterances where noisy speech utterances are summation of clean speech and randomly selected interior, car or public space noise samples at signal-to-noise ratio (SNR) levels 0 to 15 dB. Furthermore, to simulate the hands-free speech in a room, the clean speech signals are convoluted with the impulse responses of different acoustic scenarios. There are three different hands-free input conditions: hands-free in office (HFO), hands-free in living room (HFL) and hands-free in car (HFC). In the G. 712 filtered set, the GSM radio channel is also applied to simulate an influence for transmitting the noisy speech over a cellular telephone network.
In the experiments, we focused on the performance of the speech recognition system in a clean training condition. Baseline recognition systems were built based on the clean speech data provided by the G.712 filtered and non-filtered data sets. The number of utterances used for HMM training was 8623 per data set. In our implementation, we employed the conventional frontend (FE) feature specified in the ETSI standard [11] as the basic feature vectors. A 13-dimensional cepstrum and the corresponding $\Delta$- and $\Delta\Delta$-cepstra were extracted from each frame and used as the feature vector for speech recognition.

The performances of the proposed and the reference feature mapping algorithms were compared in terms of average word recognition accuracies when the SNR is 0, 5, 10 and 15 dB. For convenience, we denote the SLDS with the proposed enhanced clustering method by SLDS-PCA and with conventional simple clustering method [2] by SLDS-BASE. The total number of LDSs $K$ was 128 and we employed a GMM-based soft-decision scheme given by (14) and (15). The dimensions of the input feature vector $d_x$, output feature vector $d_y$, and the hidden state of each LDS $d_z$ were set to 13, 13 and 39, respectively. In the SLDS-PCA, $\tau$ is assigned to 2 and 4, and the number of PCA basis vectors, $L$ was held fixed at 13 which equals the dimension of a single cepstrum. As reference systems, we also implemented SPLICE [3] algorithm which is a well-known stereo data based feature mapping technique. In SPLICE, as the distribution of the input, the same GMM at the SLDS-BASE was applied.

The results are shown in Table 1. From the results, we can see that the SLDS algorithm provided better performance than the SPLICE algorithm. We can also observe that the proposed SLDS-PCA methods are more robust to channel distortions compared with the SLDS-BASE. In exceptional cases, when there is no channel distortion caused by reverberation, the performance of the proposed SLDS-PCA is slightly worse than that of SLDS-BASE. The observation reflects the fact that considering neighboring feature vectors jointly is useful in reverberant environments especially when the reverberation time is long.

<table>
<thead>
<tr>
<th></th>
<th>Non-Filtered</th>
<th>G. 712 Filtered</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Interior Noise</td>
<td>Car Noise</td>
</tr>
<tr>
<td></td>
<td>HFO</td>
<td>HFL</td>
</tr>
<tr>
<td>Baseline</td>
<td>44.23</td>
<td>35.00</td>
</tr>
<tr>
<td>SPLICE</td>
<td>77.81</td>
<td>61.72</td>
</tr>
<tr>
<td>SLDS-BASE</td>
<td>78.30</td>
<td>64.17</td>
</tr>
<tr>
<td>SLDS-PCA ($\tau=2$)</td>
<td>78.25</td>
<td>64.28</td>
</tr>
<tr>
<td>SLDS-PCA ($\tau=4$)</td>
<td>77.74</td>
<td>65.77</td>
</tr>
</tbody>
</table>

6. CONCLUSIONS

In this paper, we have proposed an enhanced clustering method for SLDS-based feature mapping algorithm. In contrast to the conventional simple clustering method, the proposed clustering method utilizes the local trajectory of the input feature vector stream. The proposed algorithm has been applied to stereo data based speech feature mapping for channel distorted speech recognition. From speech recognition experiments, it has been shown that the proposed method outperforms the conventional approaches.

ACKNOWLEDGEMENTS

This work was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 20100018858) and by the Advanced Industrial Technology Development Program funded by the Ministry of Knowledge Economy (No. 10031489).

REFERENCES


