SWITCHING LINEAR DYNAMIC TRANSDUCER FOR
STEREO DATA BASED SPEECH FEATURE MAPPING

Chang Woo Han #, Tae Gyoon Kang #, Doo Hwa Hong #, Nam Soo Kim #, Kiwan Eom *, and Jaewon Lee *

# School of Electrical Engineering and INMC
Seoul National University, Seoul 151-742, Korea
* Samsung Electronics, Suwon 443-742, Korea
E-mail: {cwhan, tgkang, dhhong}@hi.snu.ac.kr, nkim@snu.ac.kr {kiwan.eom, jwonlee}@samsung.com

ABSTRACT

The performance of a speech recognition system may be degraded even without any background noise because of the linear or non-linear distortions incurred by recording devices or reverberations. One of the well-known approaches to reduce this channel distortion is feature mapping which maps the distorted speech feature to its clean counterpart. The feature mapping rule is usually trained based on a set of stereo data which consists of the simultaneous recordings obtained in both the reference and target conditions. In this paper, we propose a novel approach to speech feature sequence mapping based on the switching linear dynamic transducer (SLDT). The proposed algorithm enables us a sequence-to-sequence mapping in a systematic way, instead of the traditional vector-to-vector mapping. The proposed approach is applied to compensate channel distortion in speech recognition and shows improvement in recognition performance.

Index Terms—Switching linear dynamic transducer, feature mapping, channel compensation, stereo data

1. INTRODUCTION

There exist numerous factors that cause mismatches between the input speech signals and those used for training the acoustic model for speech recognition. This mismatch of acoustic features usually causes a degradation of the speech recognition performance. The factors that affect acoustic mismatch are largely classified into two categories: system and environmental factors [1]. The system factors include speech capturing devices such as microphones, analog circuits, A/D converters and data compression modules. On the other hand, the environmental factors such as additive background noise, acoustic reverberations and various interfering signals affect the speech quality.

There are two major approaches to alleviate this type of performance degradation: feature mapping and model adaptation techniques. In the feature mapping techniques, the input signal waveforms or feature vectors are enhanced during front-end processing while the model adaptation techniques modify the parameters of acoustic recognition models to fit the input speech signal more closely. In this work, we focus on the feature mapping technique in which the input speech features such as the cepstrum vectors are converted to their enhanced version before being decoded through the acoustic recognition models that were trained on a different system and in a different environment.

From a system theoretic viewpoint, feature mapping is considered a transducer as shown in Fig. 1 in which the input feature vector sequence \((x_1, x_2, \cdots, x_T)\) is converted to the target sequence \((y_1, y_2, \cdots, y_T)\). Based on this viewpoint, the design of the feature mapping rule can be handled as the system identification problem with a set of input and corresponding output feature vector streams. There are two approaches for estimating the parameters for feature mapping: stereo data based and blind techniques. In the stereo data based technique, a database of simultaneous recordings obtained in both the reference and target conditions is given and feature mapping rules are derived from the difference between the associated feature vectors [2]-[4]. On the other hand, in the blind techniques, only the input feature vectors are given and the information related to the target feature vectors is provided in the form of 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 according to either the minimum mean square error (MMSE) or the maximum likelihood (ML) criterion.

In this paper, we propose a novel approach to speech feature sequence mapping based on the switching linear dynamic transducer (SLDT). We also propose a method to train the SLDT parameters based on a given stereo database. SLDT is considered an extension of SLDM [6]. In SLDT, since there is an exogenous input feature vector sequence, it can be assumed to be a transducer. One of the prominent advantages of the proposed method is that it enables a systematic implementation of sequence-to-sequence mapping instead of the traditional vector-to-vector mapping. In the stereo data, one is captured with the same conditions as used in the speech recognition system training and the other is collected with a different device. The performance of the proposed method is evaluated with speech recognition experiments. The proposed speech feature mapping algorithm based on SLDT shows better performance than other approaches when evaluated with a Korean continuous speech recognition task.

2. SWITCHING LINEAR DYNAMIC TRANSDUCER

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 transducers (LDTs). In our proposed SLDT, when the \( k \)-th LDT is applied, the feature mapping process is approximated as follows:

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

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), \( u^{(k)}_t \) and \( u^{(k)}_t \) are random vectors with a Gaussian distribution as follows:

\[
\begin{align*}
u^{(k)}_t & \sim \mathcal{N}(\mu^{(k)}, Q^{(k)}) \\
u^{(k)}_t & \sim \mathcal{N}(\mu^{(k)}, R^{(k)})
\end{align*}
\]

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

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

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

Determing an appropriate LDT among the \( K \) candidate models at each time is very important in SLDT-based feature mapping. The LDT selection rule should be solely dependent on the input feature vector sequence because the output feature vector sequence is not available at runtime. Simply, we divide the input vector \( x_t \) into \( K \) disjoint clusters. In our implementation, a GMM-based clustering technique is applied. Since we can compute the a posteriori probability \( p(k|y_t) \) in the GMM-based technique, by taking advantage of these posterior probabilities, a soft decision is adopted. Then the output feature vector stream is generated by following

\[
\begin{align*}
z_{t+1} & = \sum_{k=1}^{K} p(k|x_t) \left[ A^{(k)} z_t + B^{(k)} x_t + \mu^{(k)}_u \right] \\
y_t & = \sum_{k=1}^{K} p(k|x_t) \left[ C^{(k)} z_t + D^{(k)} x_t + \mu^{(k)}_u \right].
\end{align*}
\]

3. SLDT PARAMETER ESTIMATION

The SLDT parameters \( \lambda = \{ \lambda_1, \lambda_2, \cdots, \lambda_K \} \) are estimated from a set of stereo speech data. 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 LDT 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 SLDT. 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 in (5) 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 \( Z = (z_1, z_2, \cdots, z_T) \) is obtained conditioned on the current SLDT parameters and then the parameters are updated so as to maximize the complete data likelihood. Given the input and output feature vector 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 [8].

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

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

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

where \( \hat{\lambda} \) and \( \hat{\lambda} \) represent the updated and current SLDT parameters, respectively, and \( p(Z|X, Y, \lambda) \) is the posterior PDF of the hidden state sequence derived from the Kalman filtering.
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

$$\frac{\partial}{\partial \lambda} \Phi(\lambda, \hat{\lambda})|_{\lambda=\hat{\lambda}} = 0. \quad (7)$$

For convenience of the formulation, we assume that $Y$ is generated from $X$ through a single LDT. Once the update equations of single LDT parameters are derived, it is not difficult to extend these to the case of SLDT 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 LDT. Then, the solution to (7) is given as (8) and (9) with

$$\hat{z}_t = E[z_t|X, Y, \hat{\lambda}]$$

$$\hat{z}_t \hat{z}_t' = E[z_t z_t'|X, Y, \hat{\lambda}]$$

$$\hat{z}_t \hat{z}_{t+1} = E[z_t z_{t+1}'|X, Y, \hat{\lambda}] \quad (10)$$

where $\hat{z}_t$, $\hat{z}_t \hat{z}_t'$ and $\hat{z}_t \hat{z}_{t+1}$ are obtained during the Kalman filtering step, and $E[\cdot]$ denotes the expectation operation. Finally, the covariance matrices, $Q$ and $R$, are updated as (11) and (12).

4. EXPERIMENTAL RESULTS

The performance of the proposed approach was evaluated with a large vocabulary Korean continuous speech recognition DB. For the training of the baseline system, about 150 hours of speech was used. We extracted 13-dimensional cepstrum and the corresponding $\Delta$- and $\Delta\Delta$-cepstra which were used as the feature vector for speech recognition. The total number of words in the vocabulary was 24,017 and trigram language model was applied. All the experiments were carried out using the CMU PocketSphinx speech recognition system, which was built based on semi-hidden HMM.

In our feature mapping system, SLDT was applied in the cepstral domain. The total number of LDTs was 128 and we employed a GMM-based soft-decision scheme given by (4). The dimensions of the input feature vector $d_u$, and the hidden state of each LDT $d_z$ were set to 13, 13 and 65, respectively.

As reference systems, with which we compared the performance, we implemented the maximum likelihood linear regression (MLLR) and SPLICE [2] algorithm which is a well-known model adaptation and stereo data based feature mapping technique, respectively. We applied the MLLR algorithm using the Sphinx toolkit which simply provides the adaptation of the Gaussian means. In SPLICE, as the distribution of the input, the same GMM at the SLDT was applied.

We performed two experiments to evaluate the robustness of the proposed approach to channel distortion. In the first, we evaluated the performance when the channel mismatch was caused by a system factor. For this, we captured 4,582 utterances of stereo data in which each utterance consisted of the simultaneous recordings obtained in both the reference and target conditions. The reference data was obtained from a hand-held mobile device and the target data was recorded at the same microphone used for the training material. Among the collected stereo data, 3,000 utterances were used to train the MLLR, SPLICE and SLDT parameters, and the remaining 1,582 utterances were applied for the performance evaluation. In the second experiment, we evaluated the performance when the channel distortion was caused by an environmental factor. For this, channel mismatches were artificially created from the target data by simulating a room reverberant environment. The reference speech was obtained by convolution between the target data and room impulse response. To generate the impulse response, image method [9] was applied for a rectangular room which was 15 m in length, 30 m in width and 4 m in height. The reflection coefficient of the walls was set to 0.6. The microphone was assumed to be located at (8, 5, 1, 5) and the speech source was placed at (7, 6, 1, 6).

The results are shown in Fig. 2. Word recognition accuracy obtained from the target speech was 91.03% which was considered a performance upper bound of any feature mapping algorithm because it is a matched conditions. In Fig. 2, we can see that the proposed SLDT approach provided better performance than the other algorithms. From the results, it can be concluded that the proposed algorithm is more robust to channel distortions compared with other approaches.
\[
\begin{bmatrix}
\sum_{t=1}^{T-1} \hat{z}_t z'_t \\
\sum_{t=1}^{T-1} x_t z'_t \\
\sum_{t=1}^{T-1} z'_t \\
\sum_{t=1}^{T-1} x'_t \\
\sum_{t=1}^{T} z'_t \\
\sum_{t=1}^{T} x'_t \\
\sum_{t=1}^{T} x_t \\
\sum_{t=1}^{T} x'_t \\
\sum_{t=1}^{T} x_t \\
\end{bmatrix}
\begin{bmatrix}
\sum_{t=1}^{T-1} \hat{z}_t z_t \\
\sum_{t=1}^{T-1} x_t z_t \\
\sum_{t=1}^{T-1} z_t \\
\sum_{t=1}^{T-1} x_t \\
\sum_{t=1}^{T} z_t \\
\sum_{t=1}^{T} x_t \\
\sum_{t=1}^{T} x_t \\
\sum_{t=1}^{T} x'_t \\
\sum_{t=1}^{T} x'_t \\
\end{bmatrix}
\begin{bmatrix}
\bar{y} \\
\bar{B} \\
\bar{\mu}_y \\
\end{bmatrix}
= 
\begin{bmatrix}
\sum_{t=1}^{T-1} \hat{z}_t z_t \\
\sum_{t=1}^{T-1} x_t z_t \\
\sum_{t=1}^{T} z_t \\
\sum_{t=1}^{T} x_t \\
\sum_{t=1}^{T} x_t \\
\sum_{t=1}^{T} x_t \\
\sum_{t=1}^{T} x_t \\
\sum_{t=1}^{T} x'_t \\
\sum_{t=1}^{T} x'_t \\
\end{bmatrix}
. 
\]

(8)

\[
\hat{Q} = \frac{1}{T-1} \sum_{t=1}^{T-1} E \left[ (z_{t+1} - \hat{A} z_t - \hat{B} x_t - \hat{\mu}_u) \left( z_{t+1} - \hat{A} z_t - \hat{B} x_t - \hat{\mu}_u \right)' | X, Y, \lambda \right]. 
\]

(11)

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

(12)

Fig. 2. Word recognition accuracies with channel distortion.

5. CONCLUSIONS

In this paper, we have proposed a speech feature mapping algorithm based on SLDT. In contrast to the conventional vector-to-vector mapping approach, SLDT can describe the sequence-to-sequence mapping in a systematic way. The proposed algorithm has been applied to stereo data based speech feature mapping for channel distorted speech recognition. From a number of experiments, it has been shown that the proposed method outperforms the conventional feature mapping approach.

6. REFERENCES


