IEICE TRANSACTIONS

on Information and Systems

VOL. E97-D NO. 6
JUNE 2014

A PUBLICATION OF THE INFORMATION AND SYSTEMS SOCIETY
The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER Special Section on Advances in Modeling for Real-world Speech Information Processing and its Application

Variable Selection Linear Regression for Robust Speech Recognition

Yu TSAO[†a)], **Ting-Yao HU**[††], ***Nonmembers***, **Sakriani SAKTI**[†††], **Satoshi NAKAMURA**[†††], ***Members***,
and* Lin-shan LEE**[††], ***Nonmember

SUMMARY This study proposes a variable selection linear regression (VSLR) adaptation framework to improve the accuracy of automatic speech recognition (ASR) with only limited and unlabeled adaptation data. The proposed framework can be divided into three phases. The first phase prepares multiple variable subsets by applying a ranking filter to the original regression variable set. The second phase determines the best variable subset based on a pre-determined performance evaluation criterion and computes a linear regression (LR) mapping function based on the determined subset. The third phase performs adaptation in either model or feature spaces. The three phases can select the optimal components and remove redundancies in the LR mapping function effectively and thus enable VSLR to provide satisfactory adaptation performance even with a very limited number of adaptation statistics. We formulate model space VSLR and feature space VSLR by integrating the VS techniques into the conventional LR adaptation systems. Experimental results on the Aurora-4 task show that model space VSLR and feature space VSLR, respectively, outperform standard maximum likelihood linear regression (MLLR) and feature space MLLR (fMLLR) and their extensions, with notable word error rate (WER) reductions in a per-utterance unsupervised adaptation manner.

key words: *variable selection, linear regression, MLLR, fMLLR, model space adaptation, feature space adaptation*

1. Introduction

A key point to the success of automatic speech recognition (ASR) is its ability to maintain satisfactory performance under various acoustic conditions. The difficulty in achieving this goal is that the acoustic mismatch between training and testing environments usually leads to considerable ASR performance degradation. Identifying a way to compensate for the acoustic mismatch effectively with constrained resources has become an important task. Many approaches have been developed to offer solutions to this task [1]–[12]. Among these approaches, adaptation methods have proven effective and are popularly used in ASR systems. Generally, adaptation methods compute a mapping function to characterize the mismatch effect between training and testing acoustic conditions. The computed mapping function is then used to adjust the parameters in the original acoustic model or to convert the testing acoustic features to reduce the acoustic mismatch. Among the mapping functions, the linear regression (LR) function is popularly adopted in many adaptation methods because it can effectively enhance the ASR accuracy with reasonable computational cost. Maximum likelihood linear regression (MLLR) [13]–[16] and feature space MLLR (fMLLR) [14], [17] are well-known methods using the LR function to perform adaptation in the model and feature spaces, respectively.

Although MLLR and fMLLR have been proven effective for most ASR tasks, they often encounter over-fitting issues when the amount of adaptation data is extremely limited and no correct transcription is available. To handle this issue, various solutions have been proposed. One direction is to predefine the structure of the LR mapping function. When a sufficient amount of training data is available, a complex (full) matrix is used; on the other hand, when the amount of adaptation data is limited, a simple (block or diagonal) matrix is used. Another successful direction to overcome over-fitting is to incorporate a regularization term into the objective function, which is used to estimate the LR mapping function. One popular regularization term is a prior distribution of the LR function [18]–[22]. Effective approaches include structural Bayesian linear regression (SBLR) [18] maximum a posteriori linear regression (MAPLR) [19]–[22] and feature-space MAPLR (fMAPLR) [23]. More recently, a class of approaches adopts the L2 and L1 norms of the rotation matrix of the LR mapping function as the regularization term; the corresponding approaches are Ridge-MLLR [24] and LASSO-MLLR [25], respectively. Because the regularization terms can stabilize the estimation of LR mapping functions, the regularized adaptation approaches effectively overcome the over-fitting issue and provide satisfactory performance with limited data. Different form the approaches described above (predefining the structure of rotation matrix or imposing a regularization term into the objective function), this study proposes to use the variable selection (VS) technique [26]–[28] to directly select the optimal subset of the LR mapping function parameters, according to the adaptation data, to perform adaptation.

In machine learning tasks, the main purpose of the VS techniques is to select a compact subset of variables to help improve performance. The selected variable subset excludes redundant and noisy variables, thereby improving the reliability of the learned models. Moreover, the time required

Manuscript received September 30, 2013.
Manuscript revised January 18, 2014.

[†] The author is with Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan.

[††] The authors are with the Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan.

[†††] The authors are with the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma-shi, 630–0192 Japan.

a) E-mail: yu.tsao@citi.sinica.edu.tw

DOI: 10.1587/transinf.E97.D.1477

for training and the storage requirements can be reduced, and data structures can be analyzed explicitly. The VS techniques can be divided into two categories—filter and wrapper methods. The filter methods first decide a ranking list for the entire variable set and then select those variables according to this list until a certain stop criterion is fulfilled. Since the selected variable subset has better representation capability for the data samples, the filter methods are also considered as pre-processing of the observed dataset. On the other hand, the wrapper methods evaluate all of the possible variable subsets using the performance score of the target task and then determine the optimal subset that achieves the best performance.

In this study, we propose to incorporate the VS techniques into the LR adaptation framework (named VSLR) and derive model space and feature space VSLR (named M-VSLR and F-VSLR, respectively) approaches. Both approaches consist of three phases. The first phase establishes and ranks regression variable subsets. The second phase determines the best variable subset according to a predefined performance evaluation criterion and calculates an LR mapping function based on the determined variable subset. The third phase performs adaptation in model or feature spaces. Comparing to the regularized LR methods that impose constraints on the LR function computation, the proposed VSLR framework directly selects suitable components in the LR functions according to the available adaptation statistics. We evaluated the proposed framework on the Aurora-4 database [29], [30]. Experimental results demonstrate that both M-VSLR and F-VSLR can provide better adaptation performance than conventional MLLR and fMLLR, as well as MAPLR and fMAPLR, and L2-norm regularized LR adaptation methods, in a per-utterance unsupervised adaptation mode. The results confirm that using the VS process to directly select informative variables in the LR mapping function can achieve better adaptation performance when only very limited amount of adaptation data is available.

The rest of this paper is organized as follows. We first review the conventional and regularized LR adaptation framework in Sect. 2. Then, we detail the proposed VSLR adaptation framework in Sect. 3. In Sect. 4, we report the experimental results and discussion. Finally, we conclude our findings in Sect. 5.

2. Conventional and Regularized Linear Regression Adaptation Framework

This section reviews the theories underlying conventional MLLR and fMLLR and their extensions that incorporate regularizations to improve adaptation performance.

2.1 Conventional MLLR and fMLLR

The conventional MLLR approach [13], [14] updates the mean vector, μ_i, for the i-th Gaussian mixture in the original acoustic model, Ω, to a new mean vector, $\tilde{\mu}_i$, using

$$\tilde{\mu}_i = \mathrm{W}\xi_i, \quad i = 1, 2, \ldots M, \tag{1}$$

where $\xi_i = [\mu_i' \ 1]'$ is the augmented vector, and M is the total number of Gaussian components in the acoustic model. Meanwhile, fMLLR [14], [17] updates the acoustic observation vector at the t-th frame, o_t, using

$$\tilde{o}_t = \mathrm{W}\zeta_t, \quad t = 1, 2, \ldots T, \tag{2}$$

where $\zeta_t = [o_t' \ 1]'$ is the augmented feature vector, $\tilde{o}_t$ is the updated feature vector, and T is the total number of frames.

In Eqs. (1) and (2), W = [A b] is the LR mapping function, which comprises a rotation matrix, A, and a bias vector, b. With the adaptation data, O, MLLR and fMLLR use the maximum likelihood (ML) criterion to estimate the LR mapping function using

$$\mathrm{W}^* = \underset{\mathrm{W}}{\mathrm{argmax}}\, log\, P(\mathrm{O}|\Omega, \mathrm{W}, \mathrm{U}), \tag{3}$$

where U is the transcription corresponding to O.

Although MLLR and fMLLR have proven effective in many tasks, they suffer from over-fitting issues when the LR mapping function is poorly estimated, usually caused by limited adaptation data and incorrect transcriptions. In the next section, we introduce the regularized LR approaches that can effectively handle the over-fitting issue.

2.2 Regularized Linear Regression Adaptation

The goal of regularized LR adaptation approaches is to overcome the over-fitting issue of the conventional MLLR and fMLLR approaches by integrating a regularization term into the likelihood objective function. Generally, regularized LR adaptation approaches compute the LR mapping function using

$$\mathrm{W}^* = \underset{\mathrm{W}}{\mathrm{argmax}}(log\, P(\mathrm{O}|\Omega, \mathrm{W}, \mathrm{U}) + \lambda \mathrm{R}), \tag{4}$$

where R is the regularization term, and λ is an interpolation weight, which determines the scale of regularization.

A successful regularized LR adaptation approach incorporates the norm of rotation matrix, A, to form the regularization term as

$$\mathrm{R} = -||\mathrm{A}||^q. \tag{5}$$

Ridge-MLLR and Ridge-fMLLR [24], respectively, adopt the regularized objective function in Eq. (4) with setting $q = 2$ in Eq. (5), to perform model space and feature space adaptation. On the other hand, MAPLR [19]–[22] and fMAPLR [23] use prior densities as regularization terms, R, in Eq. (4), by

$$\mathrm{R} = log\, P(\mathrm{W}, \Omega), \tag{6}$$

where the hyper-parameters of $P(\mathrm{W}, \Omega)$ can be prepared by training data. It has been confirmed that the regularized LR adaptation approaches can overcome the over-fitting issue successfully and thus provide satisfactory adaptation performance when only a very limited amount of adaptation data is available [19]–[25].

3. Variable Selection Linear Regression Adaptation Framework

In this section, we introduce the proposed M-VSLR and F-VSLR adaptation approaches. Both of them comprise three phases—variable subset construction, subset selection, and adaptation.

3.1 Model Space Variable Selection Linear Regression (M-VSLR)

The goal of M-VSLR is to estimate an LR mapping function with the optimal form to adjust mean parameters in the original acoustic model to match the testing condition. In the following, we introduce the three phases of M-VSLR as shown in Fig. 1.

3.1.1 Variable Subset Construction

The first phase of M-VSLR prepares multiple variable subset candidates. Many algorithms can be used to perform this task, such as independent component analysis (ICA) [31] and k-means [32]. We adopt principal component analysis (PCA) [33]–[35] in this study.

We first construct a D-by-M matrix, Λ, whose columns are mean vectors with dimension D, and M is the total number of Gaussian components. By applying singular value decomposition (SVD) on Λ, we can calculate a matrix formed by eigenvectors, X, of the covariance matrix, $\Lambda\Lambda'$, where $\mathrm{X} = [e^{(1)}e^{(2)} \dots e^{(D)}]'$, and $e^{(d)}$ is the eigenvectors with the d-th largest eigenvalue. Accordingly, we prepare D variable subsets, $\mathrm{X}^{(d)} = [e^{(1)}e^{(2)} \dots e^{(d)}]'$, $d = 1, 2, \dots D$, and $\mathrm{X}^{(d)}$ is a d-by-D matrix. Then for the i-th Gaussian mixture in the original acoustic model, we estimate its D representation vectors, $\varepsilon_i^{(d)}$ $(d = 1 \dots D)$, corresponding to D variable subsets, by projecting the mean vector, μ_i, on the eigenspace

$$\varepsilon_i^{(d)} = \mathrm{X}^{(d)}\mu_i, \quad d = 1, 2, \dots D. \tag{7}$$

Because PCA preserves most of the variance of Gaussian mean vectors in the first few eigenvectors, we can remove redundancies and keep compact representations by selecting the first few rows of X. Therefore, the variable subset construction procedure in M-VSLR can be considered as a filter-based VS process [26].

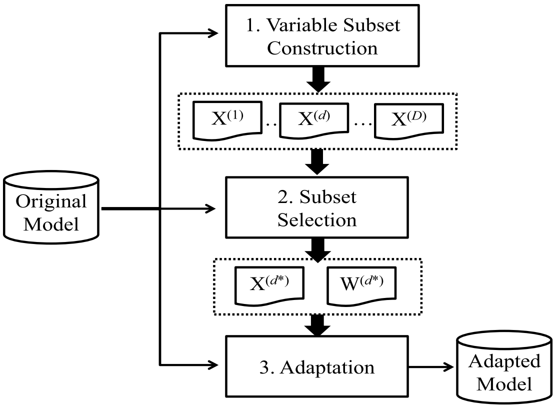


Fig. 1 Three phases of M-VSLR.

3.1.2 Subset Selection and Model Space Adaptation

The subset selection and adaptation phases of M-VSLR consist of the following four steps:
Step 1: computing D linear regressions
For each of the D subsets, $\mathrm{X}^{(d)} = [e^{(1)}e^{(2)} \dots e^{(d)}]'$, $d = 1, 2, \dots, D$, we compute an LR mapping function, $\mathrm{W}^{(d)} = \left[\mathrm{A}^{(d)} \ \mathrm{b}^{(d)}\right]$, where $\mathrm{A}^{(d)}$ is a D-by-d matrix, and $\mathrm{b}^{(d)}$ is a D-dimensional bias vector. In this study, we adopt the ML criterion to compute $\mathrm{W}^{(d)}$ using

$$\mathrm{W}^{(d)} = \underset{\mathrm{W}^{(d)}}{\operatorname{argmax}} \, log \, P(\mathrm{O}|\Omega, \mathrm{W}^{(d)}, \mathrm{U}), \\ d = 1, 2, \dots D. \tag{8}$$

Different from conventional MLLR, we define $\mathrm{W}^{(d)}$ as the LR mapping function to characterize the difference between $\tilde{\mu}_i$ and μ_i, and thus the adaptation formulation can be written as

$$\tilde{\mu}_i = \mu_i + \mathrm{W}^{(d)}\eta_i^{(d)}, \quad i = 1, 2, \dots M, \tag{9}$$

where $\eta_i^{(d)} = [\varepsilon_i^{(d)\prime} \ 1]'$. Equation (9) can be re-written as

$$\tilde{\mu}_i = \mu_i + \mathrm{A}^{(d)}\varepsilon_i^{(d)} + \mathrm{b}^{(d)}, \quad i = 1, 2, \dots M. \tag{10}$$

With the observation, O, $\mathrm{W}^{(d)}$ can be computed in a similar manner to conventional MLLR using

$$\mathrm{W}_j^{(d)} = k_j^{(d)}(\mathrm{G}_j^{(d)})^{-1}, \tag{11}$$

with

$$\mathrm{G}_j^{(d)} = \sum_i \frac{1}{\sigma_{ij}} \sum_t \gamma_i(t)\eta_i^{(d)}(\eta_i^{(d)})',$$
$$k_j^{(d)} = \sum_i \frac{1}{\sigma_{ij}} \sum_t \gamma_i(t)(o_{tj} - \mu_{ij})(\eta_i^{(d)})',$$

where $\mathrm{W}_j^{(d)}$ is the j-th row of $\mathrm{W}^{(d)}$; $\gamma_i(t)$ is the occupation probability of the i-th Gaussian at time t; μ_{ij} and σ_{ij} are the j-th element of mean vector and the (j, j) element of covariance matrix for the i-th Gaussian, respectively; o_{tj} is the j-th element of the observation vector at time t. Notably, elements in $k_j^{(D)}$ and $\mathrm{G}_j^{(D)}$ include those in $k_j^{(d)}$ and $\mathrm{G}_j^{(d)}$, $d = 1, 2, \dots D$. Accordingly only the statistics, $\mathrm{G}_j^{(D)}$ and $k_j^{(D)}$, need to be calculated to compute the entire D sets of LR functions. Therefore, the online computation is not increased much comparing to conventional MLLR [13], [14]. Figure 2 shows D sets of LR mapping function by using $\varepsilon_i^{(d)}$ $(d = 1 \dots D)$, which is obtained by Eq. (7).
Step 2: calculating performance scores
We define the performance score for the d-th variable subset, $S^{(d)}$, as:

$$S^{(d)} = \mathrm{F}(\mathrm{O}, \mathrm{W}^{(d)}), \tag{12}$$

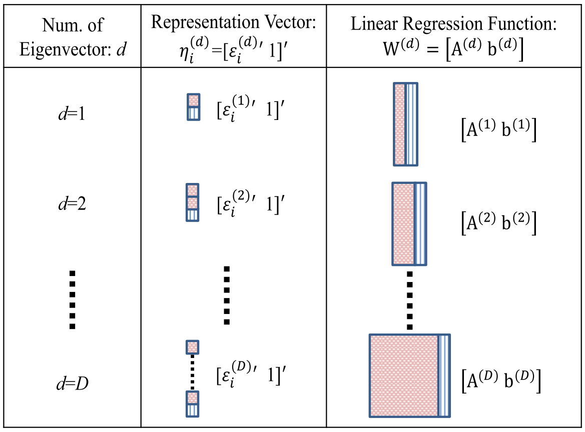

Num. of Eigenvector: d
Representation Vector: $\eta_i^{(d)}=[\varepsilon_i^{(d)\prime}\ 1]'$
Linear Regression Function: $W^{(d)}=[A^{(d)}\ b^{(d)}]$
d=1
$[\varepsilon_i^{(1)\prime}\ 1]'$
$[A^{(1)}\ b^{(1)}]$
d=2
$[\varepsilon_i^{(2)\prime}\ 1]'$
$[A^{(2)}\ b^{(2)}]$
d=D
$[\varepsilon_i^{(D)\prime}\ 1]'$
$[A^{(D)}\ b^{(D)}]$

Fig. 2 M-VSLR with D sets of LR mapping function.

where $F(O, W^{(d)})$ is related to the observations of O and $W^{(d)}$. Many criteria can be used for $F(O, W^{(d)})$, and we define it as a combination of a fitness measure (log-likelihood) and a regularization term in this study. Accordingly, Eq. (12) becomes

$$S^{(d)} = log\,P(O|\Omega, W^{(d)}, U) - \lambda \left\| A^{(d)} \right\|^2, \qquad (13)$$

Step 3: selecting the best variable subset
With the computed $W^{(d)} = \left[A^{(d)}b^{(d)}\right]$, $d = 1, 2, \dots D$, we can compute the performance scores, $S^{(d)}$, $d = 1, 2, \dots D$, based on Eq. (13). Then, the optimal variable subset can be determined using

$$d^* = \underset{d}{\operatorname{argmax}}\, S^{(d)}, \quad d = 1, 2, \dots D. \qquad (14)$$

Finally, we obtain $X^{(d^*)} = [e^{(1)}e^{(2)}\dots e^{(d^*)}]'$, $\varepsilon_i^{(d^*)}$ and, $W^{(d^*)}$. Notably, the subset selection process can be considered as a wrapper-based VS process [26].
Step 4: performing model space adaptation
The adaptation phase transforms the original acoustic model to an updated one. With the selected $\varepsilon_i^{(d^*)}$ and $W^{(d^*)}$ from the previous step, we can update the mean vector, $\tilde{\mu}_i$, using Eq. (9)

3.2 Feature Space Variable Selection Linear Regression (F-VSLR)

The goal of F-VSLR is to estimate an LR function in the optimal form to convert the testing feature sequences to match the acoustic model. In this section, we introduce the three phases of F-VSLR as shown in Fig. 3.

3.2.1 Variable Subset Construction

Similarly to M-VSLR, F-VSLR applies PCA to construct multiple variable subsets. Let $O = [o_1, o_2, \dots, o_t, \dots o_T]$ be the testing feature sequence. We can calculate the matrix of eigenvectors, $X = [e^{(1)}e^{(2)}\dots e^{(D)}]'$ of OO' and then obtain D variable subsets using the following representation:

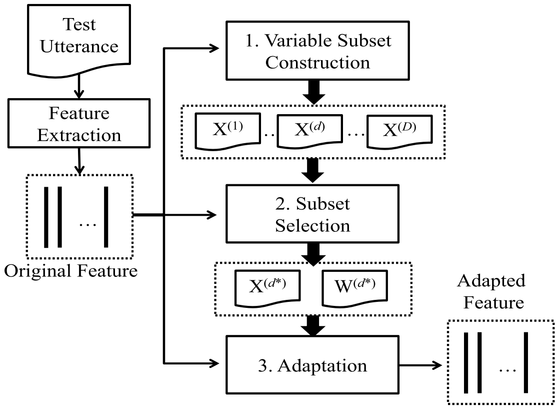


Fig. 3 Three phases of F-VSLR.

$$\varepsilon_t^{(d)} = X^{(d)} o_t, \quad d = 1, 2, \dots D, \qquad (15)$$

where $X^{(d)} = [e^{(1)}e^{(2)}\dots e^{(d)}]'$. Similar to M-VSLR, the variable subset construction procedure in F-VSLR can be considered as a filter-based VS process [26].

3.2.2 Subset Selection and Feature Space Adaptation

The subset selection and adaptation phases of F-VSLR consist of the following four steps:
Step 1: computing D linear regressions
For each of the D subsets, we compute an LR mapping function, $W^{(d)}$, $d = 1, 2, \dots D$, based on the ML criterion

$$W^{(d)*} = \underset{W^{(d)}}{\operatorname{argmax}}\, log\,P(O|\Omega, W^{(d)}, U),$$
$$d = 1, 2, \dots D. \qquad (16)$$

To calculate $W^{(d)}$, we derive the adaptation formulation as

$$\tilde{o}_t = o_t + W^{(d)}\eta_t^{(d)}, \quad t = 1, 2, \dots T, \qquad (17)$$

where $\eta_t^{(d)} = [\varepsilon_t^{(d)\prime}\ 1]'$. Equation (17) can be re-written as

$$\tilde{o}_t = o_t + A^{(d)}\varepsilon_t^{(d)} + b^{(d)}, \quad t = 1, 2, \dots T. \qquad (18)$$

With the observation, O, $W^{(d)}$ in Eq. (16) can be computed in a similar manner to conventional fMLLR [23] using

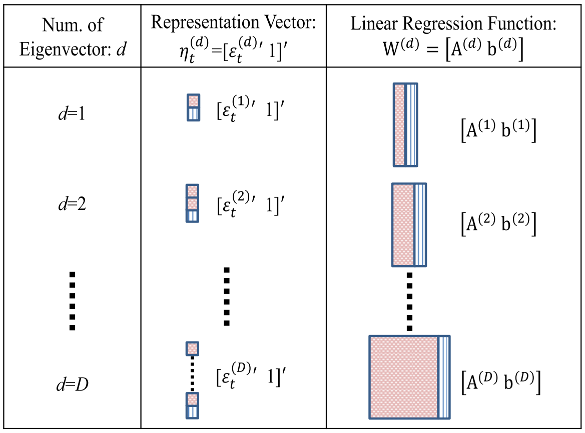
$$W_j^{(d)} = (k_j^{(d)} + \alpha p_j Y^{(d)\prime})(G_j^{(d)})^{-1}, \qquad (19)$$

with

$$G_j^{(d)} = \sum_i \frac{1}{\sigma_{ij}} \sum_t \gamma_i(t)\eta_t^{(d)}(\eta_t^{(d)})',$$
$$k_j^{(d)} = \sum_i \frac{1}{\sigma_{ij}} \sum_t \gamma_i(t)(o_{tj} - \mu_{ij})(\eta_t^{(d)})',$$

where $p_j = [c_{j1}, c_{j2}, \dots, c_{jD}]$, $c_{jk} = cofactor(A_{jk})$, and $Y^{(d)} = [(X^{(d)})'\vec{0}]'$. $W_j^{(d)}$ and p_j can be updated iteratively. The parameter α is obtained by solving the following quadratic function:

$$\alpha^2(p_j Y^{(d)\prime}(G_j^{(d)})^{-1}Y^{(d)}p_j')$$



Num. of Eigenvector: d	Representation Vector: $\eta_t^{(d)}=[\varepsilon_t^{(d)\prime}\ 1]'$	Linear Regression Function: $\mathrm{W}^{(d)}=[\mathrm{A}^{(d)}\ \mathrm{b}^{(d)}]$
d=1	$[\varepsilon_t^{(1)\prime}\ 1]'$	$[\mathrm{A}^{(1)}\ \mathrm{b}^{(1)}]$
d=2	$[\varepsilon_t^{(2)\prime}\ 1]'$	$[\mathrm{A}^{(2)}\ \mathrm{b}^{(2)}]$
d=D	$[\varepsilon_t^{(D)\prime}\ 1]'$	$[\mathrm{A}^{(D)}\ \mathrm{b}^{(D)}]$

Fig. 4 F-VSLR with D sets of LR mapping function.

$$+\alpha \mathrm{p}_j(\mathrm{Y}^{(d)\prime}(\mathrm{G}_j^{(d)})^{-1}k_j^{(d)\prime} + I_j) - \beta = 0, \qquad (20)$$

where $\beta = \sum_{i,t}\gamma_i(t)$, and I_j is the j-th column of a D-by-D identity matrix. Similarly to M-VSLR, only the statistics, $\mathrm{G}_j^{(D)}$ and $k_j^{(D)}$, need to be calculated to compute the entire D sets of LR functions. Figure 4 shows D sets of LR mapping function by using $\varepsilon_t^{(d)}$ ($d = 1, 2, \ldots D$), which is obtained by Eq. (16).

Step 2: calculating performance scores

For F-VSLR, we also define and compute a performance score for each of the D sets of LR functions. The performance score for the d-th variable subset, $S^{(d)}$, is again specified in Eq. (13).

Step 3: selecting the best variable subset

Next, with the computed $\mathrm{W}^{(d)} = \left[\mathrm{A}^{(d)}\mathrm{b}^{(d)}\right]$, $d = 1, 2, \ldots D$, we compute D performance scores, $S^{(d)}$, $d = 1, 2, \ldots D$, by following Eq. (13). Then, we determine the optimal variable subset using

$$d^* = \operatorname*{argmax}_d S^{(d)}, \quad d = 1, 2, \ldots D. \qquad (21)$$

Finally, we obtain $\mathrm{X}^{(d^*)} = [e^{(1)}e^{(2)} \ldots e^{(d^*)}]'$, as well as $\varepsilon_t^{(d^*)}$, and $\mathrm{W}^{(d^*)}$. It is noted that the subset selection process can be considered as a wrapper-based VS process [26].

Step 4: performing feature space adaptation

The adaptation phase of F-VSLR transforms the original acoustic feature sequences into updated ones. With the selected $\varepsilon_t^{(d^*)}$ and $\mathrm{W}^{(d^*)}$ from the subset selection phase, we compute the new feature, $\tilde{o}_t$, using Eq. (17).

3.3 Related Approaches

In previous studies, PCA has been used to extract prior knowledge of speakers for model adaptation. Successful examples include Eigenvoice [36]–[38] and Eigen-MLLR [39], [40]. These approaches prepare multiple sets of acoustic models or regression matrices to incorporate prior knowledge of training condition, such as speaker and speaking environments, in the offline. Then, a linear combination (LC) function is estimated online to perform model adaptation. Because the LC function is simpler than the LR function, fewer adaptation data is required to calculate the mapping function accurately; nevertheless, the performance converges quickly when more adaptation data becomes available.

This study aims to apply VS on the LR function to improve the adaptation performance. In contrast to Eigenvoice and Eigen-MLLR, the proposed VSLR framework applies PCA on the original acoustic model (for M-VSLR) and feature sequence (for F-VSLR) to construct regression variable subsets. Each variable subset increased by one element from its preceding subset, based on the ranking list generated using PCA. Hence, each prepared variable subset can be expected to contain more information about the structure of the original acoustic models and features than any alternatives of the same variable size.

After the construction step, the proposed VSLR adaptation framework defines a performance score and determines the best variable subset according to the adaptation statistics. Unlike the exhaustive search used in general wrapper methods, VSLR only evaluates the prepared variable subset candidates. Since PCA enables the prepared subsets to be used to characterize the data distribution of the original model, the selection capability of VSLR is comparable to an exhaustive search process, but the online search space is considerably reduced. It is also noted that for an unsupervised self-adaptation task, no development set is available to compute the performance scores in Eqs. (12) and (13). Accordingly VSLR cannot perform the subset selection by Eqs. (14) and (21). In this study, we use the first-pass decoding result as a reference and to compute the performance scores.

4. Experiment

In this section, we present the experimental setup and report the recognition results of the proposed M-VSLR and F-VSLR approaches, along with several well-known adaptation methods.

4.1 Experimental Setup

The proposed VSLR adaptation framework was evaluated on the Aurora-4 task [29], [30], a standardized database for evaluating ASR performance in various noise types and channel conditions. The original clean speech utterances in Aurora-4 were acquired from the Wall Street Journal (WSJ0) corpus [41]. Then, different noises were artificially added to the clean speech to generate noisy data. Aurora-4 included data that were obtained at two sampling rates, 8 kHz and 16 kHz. Data sampled at 8 kHz were chosen herein for both training and testing. Aurora-4 provided two training sets, namely clean-condition and multi-condition training sets. In this study, the baseline acoustic model was trained on the multi-condition training set [29], [30], which consisted of 7,138 training utterances. The training utterances were divided into two groups, one recorded with the

Sennheiser microphone and the other recorded with a different microphone. Each group of utterances was then artificially contaminated by six different noises (car, babble, restaurant, street, airport, and train) at SNR levels between 10 dB and 20 dB. The testing dataset comprised 14 sets under different noise and channel conditions, and 166 utterances for each test set were used to test recognition as suggested in [29]. The test sets also included six types of noise, including car, babble, restaurant, street, airport, and train. The 14 sets were further categorized into four larger sets: set A (clean speech with the Sennheiser microphone; set 1), set B (noisy speech of six noises at 5 dB to 15 dB SNRs with the Sennheiser microphone; sets 2-7), set C (clean speech with a different microphone; set 8), and set D (noisy speech of six noises at 5 dB to 15 dB SNRs with a different microphone; sets 9-14). In addition to the four test sets, we report the average results of these 14 sets, denoted as set Avg, in the following discussion.

We used hidden Markov toolkit (HTK) [42] with ML training to establish a set of context-dependent triphone acoustic models. Each triphone was characterized by a hidden Markov model (HMM), which comprised three states, with eight Gaussian mixtures per state. A tri-gram language model was prepared based on the reference transcription of the training utterances. Each utterance was characterized by 39 dimensional MFCCs, consisting of 13 static coefficients, and their first and second derivatives. In the following, all of the experimental results, except the baseline, were obtained by performing per-utterance unsupervised self-adaptation. Word error rates (WERs) are reported as the performance metric.

4.2 Experimental Results

This section presents our experimental results. We intended to investigate the effects of the first and second phases of VSLR and thus conducted two sets of VSLR experiments. First, for both M-VSLR and F-VSLR, we tested recognitions using a fixed subset dimension, *d*, for each testing utterance, to perform adaptation using Eqs. (9) and (17); the setups are denoted as M-VSLR (C) and F-VSLR (C), respectively, where (C) represents variable subset construction. Thus, the results for M-VSLR (C) and F-VSLR (C) show the effects achieved by the first phase of VSLR alone. Second, we tested recognitions of M-VSLR and F-VSLR using variable subset construction followed by subset selection by Eqs. (14) and (21), and performing adaptation by Eqs. (9) and (17); the setups are denoted as M-VSLR (C+S) and F-VSLR (C+S), respectively, where (C+S) indicates the variable subset construction and subset selection phases. Accordingly, M-VSLR (C+S) and F-VSLR (C+S) use the combination of two phases to perform adaptation.

4.2.1 Comparison of Model Space Adaptation Methods

Table 1 shows the results of the baseline (with no adaptation, denoted as Baseline) and several LR adaptation approaches, including MLLR using full, diagonal, and identity rotation matrixes (denoted as Full-MLLR, Diag-MLLR and Bias-MLLR, respectively), Ridge-MLLR [24], MAPLR [19]–[22], and principle component MLLR (PCMLLR) [35]. Please note that to obtain the results in Table 1, we tested the performances using different parameters (interpolation weights for Ridge-MLLR, hyper-parameters for MAPLR, and PC numbers for PCMLLR). We only reported the best results for each method in Table 1.

Table 1 WERs (%) of Baseline, Full-MLLR, Bias-MLLR, Ridge-MLLR, MAPLR, and PCMLLR for five test sets on Aurora-4. The best result for each test set is shown with bold digits.

Test set	A	B	C	D	Avg
Baseline	10.98	19.56	17.05	28.33	22.53
Full-MLLR	12.15	20.14	17.09	28.09	22.76
Diag-MLLR	10.87	20.02	14.73	**26.16**	21.62
Bias-MLLR	11.20	19.64	**14.18**	26.66	21.66
Ridge-MLLR	11.31	19.66	15.43	26.67	21.76
MAPLR	**10.76**	**19.04**	15.10	26.75	**21.47**
PCMLLR	11.82	19.78	16.24	27.23	22.15

From Table 1, we notice that Full-MLLR underperforms Baseline for most sets, indicating that over-fitting occurs when directly applying Full-MLLR to perform unsupervised self-adaptation. Next, Bias-MLLR, Diag-MLLR, Ridge-MLLR, MAPLR, and PCMLLR all outperform Baseline and Full-MLLR, confirming that the complexity of rotation matrices indeed affects the adaptation performance.

Next, we compare the performance of M-VSLR with Baseline and LR approaches presented in Table 1. Before the quantitative comparison, we first analyze the theoretical differences of M-VSLR with Bias-MLLR, Ridge-MLLR, and PCMLLR.

Bias-MLLR performs adaptation by:

$$\tilde{\mu}_i = \mu_i + \mathrm{b}, \quad i = 1, 2, \ldots M, \tag{22}$$

where b is the compensation bias. Different from Bias-MLLR, M-VSLR characterizes the difference but using an LR mapping function as indicated in Eqs. (9) and (10). Notably when not using $\varepsilon_i^{(d)}$ in Eq. (10), M-VSLR becomes Bias-MLLR.

Ridge-MLLR uses a regularized objective function (log-likelihood and a regularization term) to compute an LR function as indicated in Eqs. (4) and (5). On the other hand, M-VSLR uses a regularized function (log-likelihood and a regularization term) to select the optimal LR function as shown in Eqs. (13) and (14).

PCMLLR performs adaptation by:

$$\tilde{\mu}_i = \mathrm{A}^{(d)}\varepsilon_i^{(d)} + \mathrm{b}^{(d)}, \quad i = 1, 2, \ldots M, \tag{23}$$

where $\varepsilon_i^{(d)}$ is the representation vector of the *i*-th mean vector, which is the same as that in Eq. (7). From Eqs. (9), (10), and (23), it is observed that M-VSLR uses a similar adaptation function to that used in PCMLLR, but two differences are noted: (1) the regression targets of M-VSLR and PCMLLR are $(\tilde{\mu}_i - \mu_i)$ and $(\tilde{\mu}_i)$, respectively; (2) PCMLLR

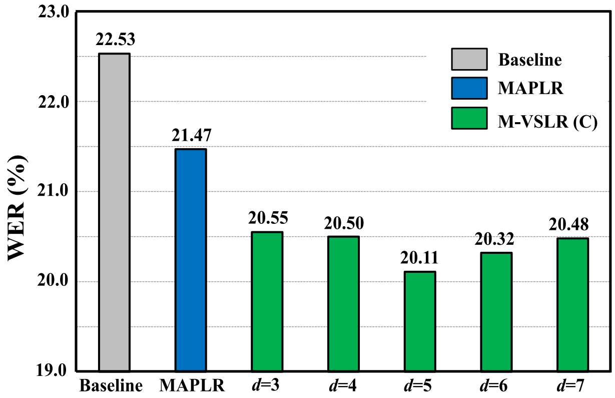


Fig. 5 WERs (%) of Baseline, MAPLR, and M-VSLR (C) with different subset sizes (d) for set Avg on Aurora-4.

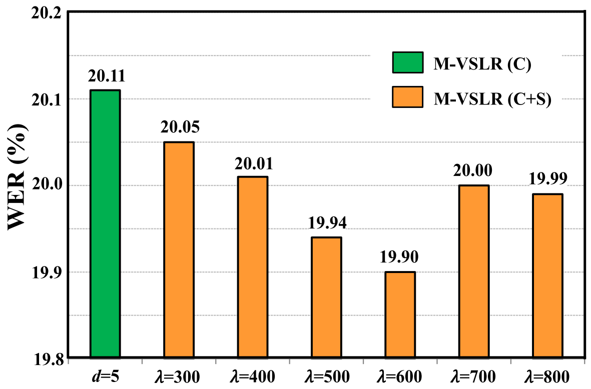


Fig. 6 WERs (%) of M-VSLR (C+S) with different weights (λ) and the best M-VSLR (C) (d = 5 in Fig. 5) for the set Avg on Aurora-4.

Table 2 WERs (%) of M-VSLR(C) and M-VSLR(C+S) for five test sets on Aurora-4. The best result for each test set is shown with bold digits.

Test set	A	B	C	D	Avg
M-VSLR (C)	11.27	18.43	13.74	24.32	20.11
M-VSLR (C+S)	**11.23**	**18.35**	**13.63**	**23.94**	**19.90**

directly uses a fixed subset for adaptation, while M-VSLR performs an additional subset selection before adaptation.

Figure 5 illustrates the WERs (on set Avg) of M-VSLR (C) with different feature dimensions (d = 3, 4, 5, 6, 7). Figure 5 also shows the results of Baseline and MAPLR, which gives the best performance in Table 1, for comparison. From Fig. 5, we observe that M-VSLR (C) with any variable dimension (d = 3, 4, 5, 6, 7) outperforms MAPLR. Of various dimensions, d = 5 achieves the best result, representing the optimal performance of M-VSLR (C) with a fixed variable dimension on this task. This set of results indicates that M-VSLR, which uses a PCA-based reduced LR mapping function, can already yield better performance than Baseline, Full-MLLR, Bias-MLLR, Ridge-MLLR, MAPLR, and PCMLLR.

Three conclusions can be drawn based on the experimental results in Table 1 and Fig. 5. First, using a complexity-reduced rotation matrix facilitates better adaptation capability with limited adaptation data. Second, using the LR function to model the difference of the original and updated means (as shown in Eq. (9)) can preserve better power to distinguish mean vectors when using a low-dimensional variable representation. Third, when using a small d, the regression result of $\mathrm{W}^{(d)}\eta_i^{(d)}$ (PCMLLR in Eq. (23)) may be constrained in a low-dimension subspace; on the other hand, the regression result of $\mu_i + \mathrm{W}^{(d)}\eta_i^{(d)}$ (M-VSLR in Eq. (9)) is represented in the original space. For example when d = 1, M-VSLR performs adaptation by $\tilde{\mu}_i = \mu_i + \mathrm{A}^{(1)}\varepsilon_i^{(1)} + \mathrm{b}^{(1)}$, while PCMLLR performs adaptation by $\tilde{\mu}_i = \mathrm{A}^{(1)}\varepsilon_i^{(1)} + \mathrm{b}^{(1)}$. Since $\mathrm{A}^{(1)}$ and $\mathrm{b}^{(1)}$ (as shown in Fig. 2) are shared by all the mean vectors, it is clear that the PCMLLR adapted mean vectors, $\tilde{\mu}_i$, $i = 1, 2, \ldots M$, may have limited discriminative power between each other. Therefore, M-VSLR can achieve better performance than PCMLLR when using a small d. On the other hand, when using a large d, the performances of M-VSLR and PCMLLR become similar. Because this study focuses on adaptation with very limited adaptation data, a small d is favored. Thus it is reasonable that the best M-VSLR performance (d = 5) in Fig. 5 outperforms the best PCMLLR performance (d = 25) in Table 1.

Next, the performance of M-VSLR (C+S), which conducts both offline construction and online selection, is evaluated. Figure 6 presents the results (on set Avg) of M-VSLR (C+S) with different λ in Eq. (13). The result of M-VSLR (C) with the best setting, d = 5 is also listed for comparison. From Fig. 6, M-VSLR (C+S) outperforms M-VSLR (C) with different λ, and the best performance is obtained when using λ = 600.

Table 2 further lists the results of M-VSLR (C+S) using λ = 600 of the five tests on Aurora-4. As mentioned earlier, for M-VSLR (C+S), the subset dimension, d, was determined based on adaptation statistics according to Eq. (14). The results of M-VSLR (C) with d = 5 is also listed for comparison.

The results in Fig. 6 and Table 2 show that M-VSLR (C+S) outperforms M-VSLR (C), confirming that M-VSLR using the selection phase achieves better performance than that using a fixed variable subset for each utterance. As mentioned in Sect. 3.1.1, the subset construction can be considered as the filter-based VS process, and subset selection can be considered as the wrapper-based VS process. Therefore, M-VSLR (C) incorporates the filter-based VS process, and M-VSLR (C+S) incorporates a combination of filter- and wrapper-based VS processes. The better results achieved by M-VSLR (C+S) suggest that applying the combination of filter- and wrapper-based VS processes provides better performance than applying a single filter-based VS process on the LR mapping function to perform model adaptation.

Moreover from Tables 1 and 2, the proposed M-VSLR (C+S) outperforms Baseline, Full-MLLR, Bias-MLLR, Ridge-MLLR, MAPLR, and PCMLLR, verifying the outstanding adaptation capability of M-VSLR for this

Table 3 WERs (%) of Baseline, Full-fMLLR, Bias-fMLLR, Ridge-fMLLR, and fMAPLR for five test sets on Aurora-4. The best result for each test set is shown with bold digits.

Test set	A	B	C	D	Avg
Baseline	10.98	19.56	17.05	28.33	22.53
Full-fMLLR	11.12	19.93	15.29	28.43	22.61
Bias-fMLLR	11.20	19.64	**14.18**	26.66	21.66
Ridge-fMLLR	**10.13**	**18.70**	14.22	**26.11**	**20.93**
fMAPLR	10.50	19.68	15.29	27.98	22.27

unsupervised self-adaptation task. In the meanwhile, we notice that the absolute improvements (WER reductions) of M-VSLR (C+S) over MAPLR are somehow limited. To verify the significance of the improvements, we conducted a t-test analysis [43], [44]. Because the entire Aurora-4 test set has 14 different sets, we conducted the t-test on 14 pair-wise results. For the t-test, we assumed that for H_0, "method-II is not better than method-I," and for H_1, "method-II is better than method-I." We used the P-value as the t-test results [43], [44]. Small P-values imply consistent improvements of method-II over method-I across the 14 sets of results. By testing the t-test on M-VSLR (C+S) versus MAPLR, we obtained the P-value = 0.01, which is smaller than a significance level of 0.05. The result verifies that M-VSLR (C+S) consistently outperform MAPLR over the 14 test sets.

4.2.2 Comparison of Feature Space Adaptation Methods

Table 3 lists the performance of fMLLR with full-rank and identity rotation matrices, denoted as Full-fMLLR, and Bias-fMLLR, respectively. We also tested and listed the performances of Ridge-fMLLR [24] and fMAPLR [23] in Table 3. From Table 3, we note that Full-fMLLR underperforms Baseline for most test sets, again showing the existence of over-fitting problem when applying Full-fMLLR on the unsupervised self-adaptation task. Next, we notice that Ridge-fMLLR performs the best among the five results in Table 3. The results suggest that incorporating a regularization term can effectively enhance the performance when the amount of adaptation data is very limited (only one utterance in this task).

Next, Fig. 7 shows the WERs of F-VSLR (C) with different variable subset numbers on set Avg. In addition to the baseline, the results of Ridge-fMLLR, which performs the best in Table 3, are also listed in Fig. 7 for comparison. From Fig. 7, we notice that variable subset $d = 6$ yields the best performance. By further comparing Table 3 and Fig. 7, we observe that F-VSLR (C) can achieve notable improvements over Baseline, Full-fMLLR, Bias-fMLLR, Ridge-fMLLR, and fMAPLR.

Finally, Table 4 compares the performance of F-VSLR (C) and F-VSLR (C+S). F-VSLR (C+S) sets $\lambda = 100$ in Eq. (13), which gives the best performance in this task. Similar to the results in Table 2, F-VSLR (C+S) can yield lower WERs than F-VSLR (C) in most test sets of Aurora-4. The same as M-VSLR (C) and M-VSLR (C+S), F-VSLR (C) uses the filter-based VS process alone, and F-VSLR (C+S) incorporates the combination of filter- and wrapper-based VS processes. The results from Table 4 confirm that the combination of filter- and wrapper-based VS processes in F-VSLR (C+S) can successfully determine the optimal variable subset and further improve the performance achieved by F-VSLR (C) that uses a single filter-based VS process.

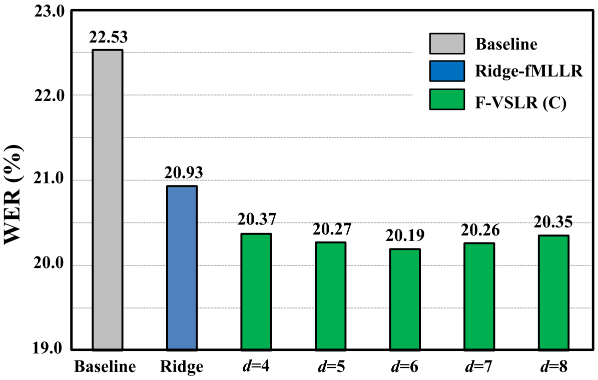


Fig. 7 WERs (%) of Baseline, Ridge-fMLLR, and F-VSLR (C) with different subset sizes (d) for set Avg on Aurora-4.

Table 4 WERs (%) of F-VSLR (C) and F-VSLR (C+S) for five test sets on Aurora-4. The best result for each test set is shown with bold digits.

Test set	A	B	C	D	Avg
F-VSLR (C)	9.83	18.62	13.15	24.66	20.19
F-VSLR (C+S)	**9.66**	**18.59**	**12.83**	**24.53**	**20.09**

From Tables 3 and 4, the proposed F-VSLR (C+S) outperforms Baseline, Full-fMLLR, Bias-fMLLR, Ridge-fMLLR, and fMAPLR, confirming the outstanding adaptation capability of F-VSLR for this unsupervised self-adaptation task. Similar to M-VSLR in the previous section, we conducted a t-test analysis to verify the significance of the improvements. By testing the t-test on F-VSLR (C+S) in Table 4 versus Ridge-fMLLR in Table 3, we obtain P-value = 0.02, which is smaller than a significance level of 0.05. The result verifies that F-VSLR (C+S) outperform Ridge-fMLLR consistently over the 14 test sets in Aurora-4.

5. Conclusion

In this paper, we proposed a VSLR adaptation framework to improve the speech recognition performance under mismatched conditions with very limited adaptation resources (small number of adaptation samples and no correct transcription). The M-VSLR and F-VSLR approaches are developed to perform adaptation in the model- and feature spaces, respectively. Both of these two approaches consist of three phases—variable subset construction, subset selection, and adaptation. We evaluated the proposed approaches on the Aurora-4 database in a per-utterance unsupervised self-adaptation mode. Experimental results demonstrated that M-VSLR and F-VSLR, respectively, outperformed MLLR and fMLLR, Ridge-MLLR and Ridge-fMLLR, and

MAPLR and fMAPLR, with notable WER reductions. The results confirm that the variable subset preparation and subset selection phases enable the VSLR adaptation framework to determine the optimal form of mapping functions according to the adaptation statistics, and thus to achieve satisfactory performance when only limited adaptation resources are available.

This study applies PCA to prepare multiple variable subset candidates. In the future, we will investigate other filtering criteria to preparing better variable subset candidates. Next, this study focused on applying the VS techniques to directly optimize the LR mapping function. We believe that the VS techniques can also be used to optimize other forms of mapping function, such as the LC mapping function used in Eigenvoice and Eigen-MLLR. Moreover, we will further compare the performances of using the L2 norm (used in this study) and other regularization terms, such as Bayesian information criterion (BIC) or minimum description length (MDL), for the proposed VSLR adaptation framework.

References

[1] L. Deng and X. Huang, "Challenges in adopting speech recognition," Commun. ACM, vol.47, pp.69–75, 2004.

[2] D. O'Shaughnessy, "Automatic speech recognition: History, methods and challenges," Pattern Recognit., pp.2965–2979, 2008.

[3] L.-C. Sun and L.-S. Lee, "Modulation spectrum equalization for improved robust speech recognition," IEEE Trans. Audio Speech Language Process., vol.20, pp.828–843, 2012.

[4] J.L. Gauvian and C.-H. Lee, "Maximum a posterier estimation for multivariate Gaussian mixture observations of Markov chains," IEEE Trans. Speech Audio Process., vol.2, pp.291–298, 1994.

[5] Y. Gong, "Speech recognition in noisy environments: A survey," Speech Commun., vol.16, pp.261–291, 1995.

[6] J.C. Junqua, J.P. Haton, and H. Wakita, Robustness in Automatic Speech Recognition, Kluwer, 1996.

[7] Y. Tsao and C.-H. Lee, "An ensemble speaker and speaking environment modeling approach to robust speech recognition," IEEE Trans. Audio Speech Language Process., vol.17, pp.1025–1037, 2009.

[8] A. Sankar and C.-H. Lee, "A maximum-likelihood approach to stochastic matchingfor robust speech recognition," IEEE Trans. Speech Audio Process., vol.4, pp.190–202, 1996.

[9] L. Lee and R. Rose, "Speaker normalization using efficient frequency warping procedures," Proc. ICASSP, pp.353–356, 1996.

[10] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," Proc. ICASSP, pp.346–348, 1996.

[11] D.Y. Kim, C.K. Un, and N.S. Kim, "Speech recognition in noisy environments using first order vector Taylor series," Speech Commun., vol.24, pp.39–49, 1998.

[12] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions," Computer Speech and Language, vol.23, pp.389–405, 2009.

[13] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, vol.9, pp.171–185, 1995.

[14] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Computer Speech and Language, vol.12, pp.75–98, 1998.

[15] K. Visweswariah, V. Goel, and R. Gopinath, "Structuring linear transforms for adaptation using training time information," Proc. ICASSP, pp.585–588, 2002.

[16] Z. Zhang and S. Furui, "Piecewise-linear transformation-based HMM adaptation for noisy speech," Speech Commun., vol.42, pp.43–58, 2004.

[17] Y. Li, H. Erdogan, Y. Gao, and E. Marcheret, "Incremental on-line feature space MLLR adaptation for telephony speech recognition," Proc. Interspeech, 2002.

[18] S. Watanabe, A. Nakamura, and B.-H. Juang, "Structural Bayesian linear regression for hidden Markov models," J. Signal Processing Systems, pp.1–18, 2013.

[19] O. Siohan, C. Chesta, and C.-H. Lee, "Joint maximum a posteriori adaptation of transformation and HMM parameters," IEEE Trans. Speech Audio Process., vol.9, pp.417–428, 2001.

[20] C.-H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," Proc. IEEE, vol.88, pp.1241–1269, 2000.

[21] O. Siohan, C. Chesta, and C.-H. Lee, "Joint maximum a posteriori adaptation of transformation and HMM parameters," IEEE Trans. Speech Audio Process., vol.9, pp.417–428, 2001.

[22] W. Chou, "Maximum a posterior linear regression with elliptically symmetric matrix variate priors," Proc. Eurospeech, pp.1–4, 1999.

[23] X. Lei, J. Hamaker, and X. He, "Robust feature space adaptation for telephony speech recognition," Proc. Interspeech, pp.773–776, 2006.

[24] J. Li, Y. Tsao, and C.-H. Lee, "Shrinkage model adaptation in automatic speech recognition," Proc. Interspeech, pp.1656–1659, 2010.

[25] J. Li, M. Yuan, and C.-H. Lee, "LASSO model adaptation for automatic speech recognition," Proc. ICML Workshop on Learning Architectures, Representations, and Optimization for Speech and Visual Information Processing, 2011.

[26] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," J. Machine Learning Research, vol.3, pp.1157–1182, 2003.

[27] M. Dash and H. Liu, "Feature selection for classification," Intelligent Data Analysis, vol.1, pp.131–156, 1997.

[28] H. Liu and H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, Springer, 1998.

[29] N. Parihar and J. Picone, "Aurora Working Group: DSR Front End LVCSR Evaluation AU/384/02," Institute for Signal and Information Processing report, 2002.

[30] N. Parihar, J. Picone, D. Pearce, and H.G. Hirsch, "Performance analysis of the Aurora large vocabulary baseline system," Proc. EUSIPCO, pp.553–556, 2004.

[31] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and application," Neural Netw., vol.13, pp.411–430, 2000.

[32] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[33] I.T. Jolliffe, Pricipal Component Analysis, Springer-Verlag, 1986.

[34] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, Wiley, 2001.

[35] S. Doh and R.M. Stern, "Weighted principal component MLLR for speaker adaptation," Proc. ASRU, 1999.

[36] R. Kuhn, P. Nguyen, J.-C. Junqua, R. Boman, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "Fast speaker adaptation using a priori knowledge," Proc. ICASSP, pp.749–752, 1999.

[37] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," IEEE Trans. Speech Audio Process., vol.8, pp.695–707, 2000.

[38] Y. Tsao, S.-M. Lee, and L.-S. Lee, "Segmental eigenvoice with delicate eigenspace for improved speaker adaptation," IEEE Trans. Speech Audio Process., vol.13, pp.399–411, 2005.

[39] K.-T. Chen, W.-W. Liau, H.-M. Wang, and L.-S. Lee, "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression," Proc. ICSLP, pp.742–745, 2000.

[40] N. Wang, S. Lee, F. Seide, and L.S. Lee, "Rapid speaker adaptation using a priori knowledge by eigenspace analysis of MLLR parameters," Proc. ICASSP, pp.345–348, 2001.

[41] D. Paul and J. Baker, "The design of Wall Street Journal-based CSR

corpus," Proc. ICSLP, pp.899–902, 1992.
[42] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book, Cambridge University Engineering Department, 2005.
[43] A.J. Hayter, Probability and Statistics for Engineers and Scientists, Duxbury Press; 3rd ed., 2006.
[44] A. Agresti and C.A. Franklin, Statistics: The Art and Science of Learning from Data (MyStatLab Series), Prentice Hall, 2008.

Yu Tsao received the B.S. and M.S. degrees in Electrical Engineering from National Taiwan University in 1999 and 2001, respectively, and the Ph.D. degree in Electrical and Computer Engineering from Georgia Institute of Technology in 2008. From 2009 to 2011, Dr. Tsao was a researcher at National Institute of Information and Communications Technology (NICT), Japan, where he engaged in research and product development in automatic speech recognition for multi-lingual speech-to-speech translation. Currently, he is an assistant research fellow of the Research Center for Information Technology Innovation (CITI) at Academia Sinica, Taiwan. Dr. Tsao's research interests include speech and speaker recognition, acoustic and language modeling, multimedia signal and information processing, pattern recognition and machine learning.

Ting-Yao Hu received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, R.O.C., in 2011 and 2013, respectively. His research interests include multimedia signal processing, machine learning, and acoustic modeling.

Sakriani Sakti received her B.E. degree in Informatics (cum laude) from Bandung Institute of Technology, Indonesia, in 1999. In 2000, she received an MSc scholarship award from the "DAAD-Siemans Program Asia 21st Century", to study in Communication Technology, University of Ulm, Germany, and received her MSc degree in 2002. During her thesis work, she worked with Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. Between 2003–2009, she was a researcher at ATR SLC Labs, Japan, and during 2006–2011, she worked as an expert researcher at NICT SLC Groups, Japan. While working with ATR-NICT, Japan, she continued her study (2005–2008) with Dialog Systems Group University of Ulm, Germany, and received her PhD degree in 2008. Her pioneering work in Indonesian language processing has established the first Indonesian speech recognition system, the first HMM-based Indonesian speech synthesis, and the first Indonesian speech-to-speech translation system. She played as a central role in research and development collaboration activities such as Asian Pacific Telecommunity Project in 2003–2007, as well as the Asian Speech Translation Advanced Research (A-STAR) and Universal Speech Translation Advanced Research (U-STAR) Consortium in 2006–2011. She also served as a visiting professor of Computer Science Department, University of Indonesia (UI), Indonesia, in 2009–2011. Currently, she is an assistant professor of the Augmented Human Communication Lab, Graduate School of Information Science, Nara Institute of Science and Technology, Japan. She is an active member of Society of Neuroscience (SFN), Japan Neuroscience Society (JNS), Institute of Electrical and Electronics Engineers (IEEE) Computer Society, Acoustical Society of Japan (ASJ), International Speech Communication Association (ISCA). Her reseach interests include statistical pattern recognition, speech recognition, spoken language translation, cognitive communication, and graphical modeling framework.

Satoshi Nakamura received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was Associate Professor of Graduate School of Information Science at Nara Institute of Science and Technology in 1994–2000. He was Director of ATR Spoken Language Communication Research Laboratories in 2000–2008, and Director General of Keihanna Research Laboratories and the Executive Director of Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan in 2009–2010. He is ATR Fellow. He is currently Professor of Graduate School of Information Science at Nara Institute of Science and Technology. He received Yamashita Research Award, Kiyasu Award from the Information Processing Society of Japan, Telecom System Award, AAMT Nagao Award, Docomo Mobile Science Award in 2007, ASJ Award for Distinguished Achievements in Acoustics, the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology, and The Commendation for Informatization Promotion by the Minister of Internal Affairs and Communications. He also received LREC Antonio Zampoli Award 2012. He is a convenor of Oriental COCOSDA from 2011. He has been Elected Board Member of International Speech Communication Association, ISCA, since June 2011 and IEEE Signal Processing Magazine Editorial Board Member since April 2012.

Lin-shan Lee received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA. He has been a Professor of electrical engineering and computer science at the National Taiwan University, Taipei, Taiwan, since 1982 and holds a joint appointment as a Research Fellow of Academia Sinica, Taipei. His research interests inlude digital communications and spoken language processing. He developed several of the earliest versions of Chinese spoken language processing systems in the world including text-to-speech systems, natural language analyzers, dictation systems, and voice information retrieval systems. Dr. Lee was Vice President for International Affairs (1996–1997) and the Awards Committee chair (1998–1999) of the IEEE Communications Society. He was a member of the Board of International Speech Communication Association (ISCA 2002–2009), a Distinguished Lecture (2007–2008) and a member of the Overview Paper Editorial Board (since 2009) of the IEEE Signal Processing Society, and the general chair of ICASSP 2009 in Taipei. He is a fellow of ISCA since 2010, and received the Meritorious Service Award from IEEE Signal Processing Society in 2011.