Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 68 (2015) 55–68

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Ensemble environment modeling using affine transform group

Yu Tsao [a,*], Payton Lin [a], Ting-yao Hu [a], Xugang Lu [b]

[a] Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[b] Spoken Language Communication Laboratory, National Institute of Information and Communications Technology, Kyoto, Japan

Received 10 July 2014; received in revised form 22 November 2014; accepted 24 December 2014
Available online 8 January 2015

Abstract

The ensemble speaker and speaking environment modeling (ESSEM) framework was designed to provide online optimization for enhancing workable systems under real-world conditions. In the ESSEM framework, ensemble models are built in the offline phase to characterize specific environments based on local statistics prepared from those particular conditions. In the online phase, a mapping function is computed based on the incoming testing data to perform model adaptation. Previous studies utilized linear combination (LC) and linear combination with a correction bias (LCB) as simple mapping functions that only apply one weighting coefficient on each model. In order to better utilize the ensemble models, this study presents a generalized affine transform group (ATG) mapping function for the ESSEM framework. Although ATG characterizes unknown testing conditions more precisely using a larger amount of parameters, over-fitting issues occur when the available adaptation data is especially limited. This study handles over-fitting issues with three optimization processes: maximum a posteriori (MAP) criterion, model selection (MS), and cohort selection (CS). Experimental results showed that ATG along with the three optimization processes enabled the ESSEM framework to allow unsupervised model adaptation using only one utterance to provide consistent performance improvements.

Keywords: Ensemble modeling; Environment modeling; Prior knowledge; Maximum a posteriori; Model selection; Cohort selection

1. Introduction

Towards ubiquitous adoption of human–machine communication (Deng and Huang, 2004), robustness in automatic speech recognition (ASR) (Junqua et al., 1996) has been addressed by noise-robust techniques (Li et al., 2014), data reduction (O'Shaughnessy, 2008), and predictive classification (Huo and Lee, 2000). To address the technical challenges of performing according to the user's intention, selection, execution, and evaluation (Norman, 1984), environment modeling or model adaptation methods (Lee, 1998; Sankar and Lee, 1996,) extend workable systems to real-world situations by modeling specific speakers and acoustic environments with unlabeled and limited amounts of adaptation data. Fig. 1 presents the structure of environment modeling, where either one general model (Category-1) or multiple environment specific models (Category-2) are first prepared as a structure using the entire training data set. In the online phase, speech segments from incoming testing conditions are collected to derive a mapping function, $F_\varphi(\cdot)$, that performs model adaptation to obtain a target model, Λ^Y, minimizing the differences between training and testing conditions. Parameters in the mapping function can be estimated via criterion such as maximum likelihood (ML) and maximum a posteriori (MAP).

For Category-1, a single source model (Λ^X in Fig. 1) is built to reflect the average statistics of the whole training data set. The mapping function is then estimated to adapt the source model to the target model. Several estimation

* Corresponding author at: No 128, Academia Road, Section 2, Nankang, Taipei 11529, Taiwan. Tel.: +886 2 2787 2390; fax: +886 2 2787 2315.
E-mail address: yu.tsao@citi.sinica.edu.tw (Y. Tsao).

http://dx.doi.org/10.1016/j.specom.2014.12.007
0167-6393/

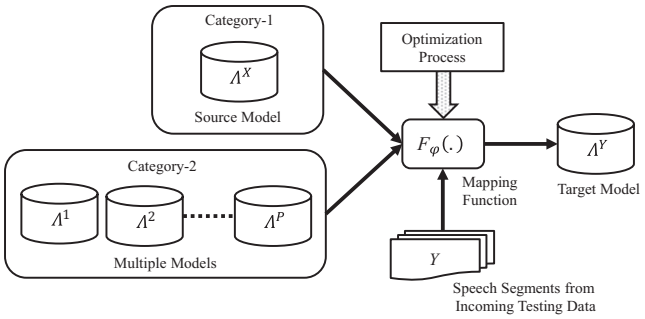


Fig. 1. Structure of environment modeling and model adaptation.

algorithms have been proposed such as linear and nonlinear stochastic matching approaches (Lee, 1998; Sankar and Lee, 1996; Suredran et al., 1999), signal bias removal (SBR) (Rahim and Juang, 1996), maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995; Gales, 1997), maximum a posteriori linear regression (MAPLR) (Chesta et al., 1999; Siohan et al., 2001; Siohan et al., 2002), structural Bayesian linear regression (SBLR) (Watanabe et al., 2014), VTS-based model adaptation (Kim et al., 1998), joint compensation of additive and convolutive distortions (JAC) (Gong, 2005; Hu and Huo, 2007; Li et al., 2009), and JAC with unscented transform (JAC-UT) (Hu and Huo, 2006; Li et al., 2010).

For Category-2, multiple models ($\{\Lambda^1, \Lambda^2, \ldots, \Lambda^P\}$ in Fig. 1) that are trained using subsets of the entire training data allow more effective local statistics of environment conditions. In these cases, the mapping function for adaptation needs to transform multiple models to the target model. For efficient estimation of mapping functions, several techniques have been proposed such as reference speaker weighting (RSW) (Hazen, 2000), eigenvoice (Kuhn et al., 2000), cluster adaptive training (CAT) (Gales, 2000; Yu and Gales, 2006), speaker clustering (Kosaka et al., 1996; Padmanabhan et al., 1998), probabilistic 2DPCA/GLRAM (Jeong, 2012), tensor voices (Jeong, 2014), and ensemble speaker and speaking environment modeling (ESSEM) (Tsao and Lee, 2009). Generally, a simple mapping function such as best first (BF) (Tsao et al., 2012), linear combination (LC) (Kuhn et al., 2000; Gales, 2000), or linear combination with correction bias (LCB) (Tsao et al., 2014) is used to perform adaptation. However, a mapping function that utilized more free parameters could enable more accurate model estimation when larger amounts of adaptation data become available. Therefore, the present study proposes an affine transform group (ATG) mapping function that applies an affine transform for each model in $\{\Lambda^1, \Lambda^2, \ldots, \Lambda^P\}$ to compute the target model. The ATG mapping function expands upon the previous ESSEM framework (Tsao et al., 2014; Tsao et al., 2012) and is denoted as ATG-ESSEM in the following discussion. While the usage of more free parameters can provide better environment modeling capabilities, over-fitting issues must be considered when the amount of adaptation data is insufficient. A previous study proposed to adopt the MAP criterion to handle over-fitting (Tsao et al., 2012). The present study proposes two additional approaches to enhance optimization processes: model selection (MS) and cohort selection (CS). This study also compares four different types of affine transform matrix: full, diagonal, scalar, and identity matrices, in order to evaluate the benefits of added complexity.

To verify effective model adaptation using ATG-ESSEM with the MAP criterion, MS, and CS, experiments were conducted on Aurora-4, a large vocabulary continuous speech recognition (LVCSR) task (Parihar and Picone, 2002; Parihar et al., 2004; Hirsch, 2001; Au Yeung and Siu, 2004). Unsupervised ESSEM adaptation could also enhance the parameter estimation of deep neural networks (DNNs) (Seltzer et al., 2013). Some adaptation methods have been proposed in DNN–HMM systems by using linear transformations (Neto et al., 1995; Li and Sim, 2010; Yao et al., 2012; Gemello et al., 2007; Ochiai et al., 2014). Due to the enormous amount of parameters, DNN has limited adaptation capability when only limited adaptation data is available. Since DNN parameter estimation is based on discriminative criterion, adaptation performance is sensitive to label errors. A combination of GMM and DNN has also effectively enhanced ASR performance (Liu and Sim, 2014) since the GMM–HMM framework is based on generative training paradigms for more robust unsupervised adaptation. This study evaluates the ATG-ESSEM framework using a difficult task designed to simulate "real-world" conditions: per-utterance unsupervised adaptation with lots of fluctuating SNRs. Experimental results confirmed the effective adaptation capability of ATG-ESSEM with only one adaptation utterance. Discussion related to adaptation under future DNN–HMM systems will be included following results of our GMM–HMM findings.

The rest of this paper is organized as follows: Section 2 reviews the ESSEM framework and the ATG mapping function, Section 3 derives three optimization processes to enhance ATG-ESSEM performance, Section 4 reports results and discusses extensions of ESSEM to DNN model parameter adaptation, and Section 5 offers concluding remarks.

2. Ensemble environment modeling and affine transform group (ATG)

In this section, the ESSEM framework is first described, followed by the presentation of the proposed ATG mapping function.

2.1. Ensemble speaker and speaking environment modeling (ESSEM)

Fig. 2 illustrates the ESSEM framework, which consists of offline and online phases. In the offline phase, a single source model Λ^X is first estimated based on the entire set of training data. This source model is trained on speech data collected from a variety of environment conditions

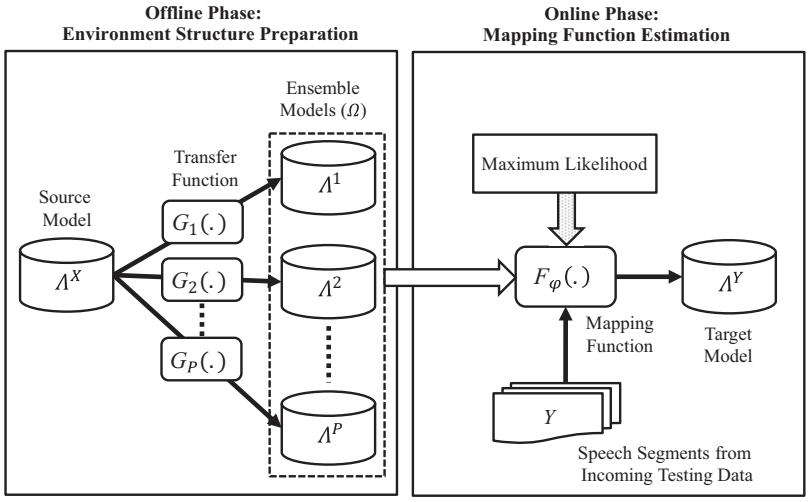


Fig. 2. Ensemble environment modeling framework.

and resamples the universal background model–Gaussian mixture models (UBM–GMM) that are used in speaker recognition tasks (Reynolds and Rose, 1995; Dehak et al., 2011). With the source model, Λ^X, multiple models (termed ensemble models in the ESSEM framework) $\{\Lambda^1, \Lambda^2, \ldots, \Lambda^P\}$ can be prepared using speech data from a wide range of conditions by

$$\Lambda^p = G_p(\Lambda^X), \quad p = 1, 2, \ldots P, \tag{1}$$

where $G_p(\cdot)$ denotes the transfer function that adapts Λ^X to form the p-th model, Λ^p. In this study, the MAP adaptation technique (Gauvian and Lee, 1994) is adopted for $G_p(\cdot)$, and only mean parameters are adapted. The source model can be considered as a speaker- and environment-independent model and is usually included as one of P models, where the transfer function in Eq. (1) becomes an identity transformation. Conventional ESSEM uses an environment clustering algorithm to group similar ensemble models to facilitate better environment modeling accuracy (Tsao and Lee, 2009). Since the goal of this paper is to introduce a novel mapping function, only one cluster is considered for ease of presentation.

During the online phase, a mapping function, $F_\varphi(\cdot)$, is calculated to transform Ω into a target model, Λ^Y, that better matches the incoming test condition data, Y, using

$$\Lambda^Y = F_\varphi(\Omega), \tag{2}$$

where φ is the parameter set in the mapping function.

The ML criterion is an effective learning method for computing parameters in the mapping function. When using the ML criterion to calculate φ, based on Eq. (2), we have

$$\hat{\varphi}_{ML} = \underset{\varphi}{\operatorname{argmax}}\, P(Y|\varphi, \Omega, W), \tag{3}$$

where Y denotes the adaptation data, and W represents the corresponding transcription reference. The ATG mapping function applies multiple affine transforms, namely $A^1, A^2, \ldots, A^P$, on the ensemble models $\Omega = \{\Lambda^1, \Lambda^2, \ldots, \Lambda^P\}$ in Eq. (2).

There are two main factors for ESSEM modeling capability: (1) form of the mapping function; (2) processes for estimating the mapping function. For (1), the BF, LC, and LCB mapping functions were previously utilized to perform model adaptation (Tsao et al., 2012). The present study proposes the ATG mapping function (discussed in Section 2.2) to perform model adaptation. For (2), the MAP, MS, and CS processes (discussed in Section 3) are derived for estimating the parameters of the online mapping function.

2.2. The affine transform group (ATG) mapping function

The ATG formulates the mapping function as

$$\mu_m^Y = A^1\mu_m^1 + A^2\mu_m^2 + \cdots A^P\mu_m^P + b, \quad m = 1, 2, \ldots, M, \tag{4}$$

where μ_m^Y is the mean vector of the m-th Gaussian in the target model (Λ^Y), μ_m^p is the mean vector for the m-th Gaussian in the p-th ensemble model (Λ^p), M is the total number of Gaussians in one model, and the parameter set in the mapping function, φ in Eq. (2), becomes $\{A^1, A^2, \ldots, A^P, b\}$. To compute φ, we rewrite Eq. (4) as:

$$\mu_m^Y = \Gamma\rho_m, \quad m = 1, 2, \ldots, M, \tag{5}$$

where $\Gamma = [A^1, A^2, \ldots, A^P b]$ denotes the ATG mapping function, and $\rho_m = \left[\mu_m^{1\prime}\mu_m^{2\prime}\ldots\mu_m^{P\prime}1\right]'$ is formed by the mean vector of the m-th Gaussian components in $\Lambda^1, \Lambda^2, \ldots, \Lambda^P$.

Based on Eq. (3), we derive the objective function for the ATG mapping function as:

$$\begin{aligned} Q(\Gamma) &= \sum_{t=1}^{T}\sum_{m\in W} r_m(t) log\mathcal{N}(y_t; \Gamma\rho_m, \Sigma_m^Y) \\ &= \frac{-1}{2}\sum_{t=1}^{T}\sum_{m\in W} r_m(t)\left[(y_t - \Gamma\rho_m)'\Sigma_m^{Y^{-1}}(y_t - \Gamma\rho_m)\right] \\ &\quad + \Psi, \end{aligned} \tag{6}$$

where y_t is the t-th observation; $r_m(t)$ is the occupation probability; Σ_m^Y denotes the covariance matrix of the m-th

Gaussian in the target model (Λ^Y); $m \in W$ indicates that the m-th Gaussian is in the transcription reference, W, and Ψ denotes the terms that are independent of Γ. This study does not adapt the covariance matrix, and thus we denote Σ_m^Y as Σ_m in the following discussion. In implementation, Λ^X is used to prepare the statistics $r_m(t)$ and Σ_m. The same objective function of Eq. (6) has been used in previous model adaptation approaches (Leggetter and Woodland, 1995; Gales, 1997; Tsao et al., 2014).

By taking the derivative of $Q(\Gamma)$ in Eq. (6) with respect to Γ, we have

$$\frac{\partial Q(\Gamma)}{\partial \Gamma} = \sum_{t=1}^{T}\sum_{m\in W} r_m(t)[\Sigma_m^{-1}(y_t - \Gamma\rho_m)\rho_m'], \tag{7}$$

and set $\frac{\partial Q(\Gamma)}{\partial \Gamma} = 0$ in Eq. (6). Then, we obtain the solution of Γ by:

$$\Gamma_{(d)} = \left[A^1 A^2 \ldots A^P b\right]_{(d)} = k_{(d)}(G_{(d)})^{-1}, \quad d = 1, \ldots, D \tag{8}$$

where

$$G_{(d)} = \sum_{t=1}^{T}\sum_{m\in W} r_m(t)\frac{1}{\Sigma_{m_{(dd)}}}\rho_m\rho_m', \tag{9}$$

and

$$k_{(d)} = \sum_{t=1}^{T}\sum_{m\in W} r_m(t)\frac{1}{\Sigma_{m_{(dd)}}}y_{t_{(d)}}\rho_m', \tag{10}$$

where D is the number of feature dimensions, $\Gamma_{(d)}$ is the d-th row of Γ, accordingly $\left[A^1 A^2 \ldots A^P b\right]_{(d)}$ is the d-th row of $[A^1 A^2 \ldots A^P b]$, and $\Sigma_{m_{(dd)}}$ is the dd-th diagonal element of the covariance matrix, Σ_m. With the computed Γ, we calculate μ_m^Y using Eq. (5).

To reduce the computational complexity, we can simplify the form for each $A^p, p = 1, 2 \ldots P$. By transforming subsets of coefficients in the mean vector independently, block-diagonal matrices can be used instead of a full matrix for A^p (Gales, 1997). For the simplest case, all the coefficients in the mean vector are assumed independent, a diagonal matrix is used for A^p (Leggetter and Woodland, 1995), and thus $A^p = \mathrm{diag}[a_{(1)}^p, a_{(2)}^p, \ldots, a_{(D)}^p]$. Similar to the derivation of full matrix ATG as presented above, the parameters of $\{A^1, A^2, \ldots, A^P, b\}$ using diagonal matrices are estimated by

$$[a_{(d)}^1 a_{(d)}^2 \ldots a_{(d)}^P b_{(d)}] = k_{(d)}\left(G_{(d)}\right)^{-1}, \quad d = 1, \ldots, D \tag{11}$$

with

$$G_{(d)} = \sum_{t=1}^{T}\sum_{m\in W} r_m(t)\frac{1}{\Sigma_{m_{(dd)}}}\xi_m\xi_m', \tag{12}$$

$$k_{(d)} = \sum_{t=1}^{T}\sum_{m\in W} r_m(t)\frac{1}{\Sigma_{m_{(dd)}}}y_{t_{(d)}}\xi_m', \tag{13}$$

where $\xi_m = \left[\mu_{m_{(d)}}^1 \mu_{m_{(d)}}^2 \ldots \mu_{m_{(d)}}^P 1\right]'$. With the estimated parameter set, $\{a_{(d)}^1, a_{(d)}^2, \ldots, a_{(d)}^P, b_{(d)}, d = 1, \ldots, D\}$, the multiple affine transforms $\{A^1, A^2, \ldots, A^P, b\}$ are obtained accordingly. Finally, μ_m^Y is calculated using Eq. (4).

In Eqs. (8)–(10) and Eqs. (11)–(13), we present the case that only a global ATG mapping function is used for model adaptation. Previous studies have shown the effectiveness of using a regression tree to extend the global to multiple mapping functions (Gales, 1997; Tsao et al., 2014). The proposed global ATG mapping function can be extended to multiple ATG mapping functions in the same manner. When using C sets of ATG mapping function to perform model adaptation, we prepare $\{\Gamma^1, \Gamma^2, \ldots, \Gamma^c, \ldots, \Gamma^C\}$, where Γ^c denotes the mapping function for the c-th class. To compute Γ^c, Eqs. (9) and (10) and Eqs. (12) and (13) are used while replacing $m \in W$ by $m \in \{W, Z^c\}$ in the summation terms, where Z^c denotes the c-th node in the regression tree. The estimated mapping function Γ^c is then used to adapt the mean parameters belonging to the c-th node. Details about the use of a regression tree to prepare multiple ATG mapping functions are described in Section 4.1.4.

2.3. Relation of ATG with approaches in Category-1 and Category-2

This section presents the relation of ATG-ESSEM with several well-known adaptation methods belonging to Category-1 and Category-2.

2.3.1. Model adaptation methods in Category-1

First, we discuss the relation of ATG with two well-known mapping functions: linear regression (LR) and compensation bias (BC), both of which use a single source model (Λ^X in Fig. 1) which is usually a speaker- and environment-independent acoustic model. For LR and BC, the model adaptation is formulated as

$$\mu_m^Y = A\mu_m^X + b, \quad m = 1, 2, \ldots, M, \tag{14}$$

and the parameters of $\{A, b\}$ stand for φ in Eq. (2). As introduced in Section 2.2, by using a regression tree, we can use multiple LR and BC mapping functions to perform model adaptation. To facilitate presentation, this section only presents the case of using a global mapping function.

2.3.1.1. Linear regression (LR). Linear regression (LR) is often used for acoustic model adaptation (Leggetter and Woodland, 1995; Gales, 1997; Chesta et al., 1999; Siohan et al., 2001). When applying the ML criterion, the parameters in $\{A, b\}$ in Eq. (14) can be calculated by

$$[A\,b]_{(d)} = k_{(d)}(G_{(d)})^{-1}, \quad d = 1, \ldots, D, \tag{15}$$

where $[A\,b]_{(d)}$ is the d-th raw of $[A\,b]$, $G_{(d)}$ and $k_{(d)}$ are computed based on Eqs. (9) and (10) with $\rho_m = \left[\mu_m^{X\prime} 1\right]'$. Notably, the solution of LR resembles that of ATG except LR uses a single rotation matrix, A (Eq. (14)), and ATG uses multiple rotation matrices, $\{A^1, A^2, \ldots, A^P\}$ (Eq. (4)).

When using diagonal matrix A, $A = \mathrm{diag}[a_{(1)}, a_{(2)}, \ldots, a_{(D)}]$, in Eq. (14), we can also solve A and b using

$$[a_{(d)}\ b_{(d)}] = k_{(d)}(G_{(d)})^{-1}, \quad d = 1, \ldots, D, \tag{16}$$

where $G_{(d)}$ and $k_{(d)}$ are computed based on Eqs. (12) and (13) with $\xi_m = \left[\mu^X_{m_{(d)}} 1\right]'$. With the estimated parameter set, $\{A, b\}$, we can calculate μ^Y_m using Eq. (14).

2.3.1.2. Bias compensation (BC). For BC (Rahim and Juang, 1996), we set $A = I_{D\times D}$ in Eq. (14), where $I_{D\times D}$ is a $D \times D$ identity matrix. Accordingly, the parameters to be estimated for BC is $\{b\}$. The parameters of $\{b\}$ can be computed by

$$b_{(d)} = \sum_{t=1}^{T}\sum_{m\in W} \frac{r_m(t)\left(y_{t_{(d)}} - \mu^X_{m_{(d)}}\right)}{\Sigma_{m_{(dd)}}} \Bigg/ \sum_{t=1}^{T}\sum_{m\in W} \frac{r_m(t)}{\Sigma_{m_{(dd)}}}, \tag{17}$$

where $b_{(d)}$ is the d-th element of b.

In contrast to LR and BC, ATG uses multiple rotation matrices. Therefore, prior information obtained from ensemble models are incorporated for modeling particular environments. Better environment modeling capability is expected from ATG when sufficient testing condition data is available.

2.3.2. Model adaptation methods in Category-2

In ensemble modeling, three well-known transform methods were initially proposed: linear combination with correction bias (LCB), linear combination (LC), and best first (BF). These three approaches use the same adaptation function as in Eq. (4), namely $\mu^Y_m = A^1\mu^1_m + A^2\mu^2_m + \cdots A^P\mu^P_m + b$, $m = 1, 2, \ldots, M$. The proposed ATG differs from LCB, LC, and BF in the form of the transform matrix for $A^1, A^2, \ldots, A^P$ and the use of the compensation bias, b. Similar to that presented in Sections 2.2 and 2.3.1, multiple sets of LCB, LC, and BF mapping functions can be used along with a regression tree for model adaptation. This section presents the case of using a global mapping function to facilitate presentation.

2.3.2.1. Linear combination with correction bias (LCB). In contrast to the ATG mapping function, the LCB mapping function simplifies each A^p to a scalar matrix: $A^p = \omega^p \cdot I_{D\times D}$, $p = 1, 2, \ldots, P$ in Eq. (4). The LCB performs model adaptation by

$$\mu^Y_m = H_m\theta, \quad m = 1, 2, \ldots, M, \tag{18}$$

where $H_m = [\mu^1_m \mu^2_m \ldots \mu^P_m I_{D\times D}]$ is constructed by the mean vector of the m-th Gaussian components in $A^1, A^2, \ldots, A^P$, and $\theta = [\omega^1\omega^2 \ldots \omega^P b']'$ denotes the mapping function. The mapping function, θ, is estimated by (Tsao et al., 2014):

$$\theta = G^{-1}k, \tag{19}$$

with

$$G = \sum_{t=1}^{T}\sum_{m\in W} r_m(t) H'_m \Sigma^{-1}_m H_m, \tag{20}$$

$$k = \sum_{t=1}^{T}\sum_{m\in W} r_m(t) H'_m \Sigma^{-1}_m y_t. \tag{21}$$

With the estimated θ, we obtain the adapted mean parameter μ^Y_m by Eq. (4).

2.3.2.2. Linear combination (LC). Similar to LCB, the LC mapping function uses $A^p = \omega^p \cdot I_{D\times D}$, $p = 1, 2, \ldots, P$, in Eq. (4), but the combination bias (b) is not used. Likewise, Eqs. (19)–(21) are used to compute the LC mapping function, where $\theta = [\omega^1\omega^2 \ldots \omega^P]'$ and $H_m = [\mu^1_m \mu^2_m \ldots \mu^P_m]$. With the estimated θ, we can obtain $\{A^1, A^2, \ldots, A^P\}$ and then compute μ^Y_m using Eq. (4).

2.3.2.3. Best first (BF). BF is used as a hard-decision based model selection from ensemble models. For BF, A^l is set as an identity matrix, $I_{D\times D}$ ($A^l = I_{D\times D}$), and all the other matrices are zero matrices, $\varnothing$, ($A^p = \varnothing$, $\forall p \neq l$) in Eq. (4). In this condition, we search for the P sets of models to find l using

$$l = \underset{p}{\operatorname{argmin}} \sum_{t=1}^{T}\sum_{m\in W} r_m(t)\left[\left(y_t - \mu^p_m\right)'\Sigma^{-1}_m\left(y_t - \mu^p_m\right)\right], \quad p = 1, 2, \ldots, P. \tag{22}$$

With the calculated l from Eq. (22), we can determine $A^l = I_{D\times D}$ with $A^p = \varnothing$, $\forall p \neq l$, and then obtain μ^Y_m using Eq. (4).

For ATG, LCB, LC, and BF methods, prior knowledge of the target environment is incorporated since $\Omega = \{A^1, A^2, \ldots, A^P\}$. However, ATG uses a more complex form for each A^p, $p = 1, 2, \ldots, P$. Therefore, ATG has potential to characterize testing conditions more accurately compared to LCB, LC, and BF, when sufficient testing data is available.

3. Three optimization processes for ATG-ESSEM

With sufficient availability of incoming testing data, ATG computed by the ML criterion should provide optimal adaptation performance since ATG-ESSEM is an extension of the model adaptation methods in Category-1 and Category-2. However, over-fitting issues could degrade adaptation performance when the adaptation data is limited. Therefore, this ATG study presents three optimization processes for handling over-fitting: MAP criterion, MS, and CS. The integration of the three processes within the ATG-ESSEM framework is shown in Fig. 3.

3.1. ATG-ESSEM with MAP criterion

The MAP criterion is often applied in model adaptation techniques to overcome over-fitting issues (Chesta et al., 1999; Siohan et al., 2001, 2002; Shinoda and Lee 2001; Tsao et al., 2014). When using the MAP criterion to calculate φ in Eq. (2), we have

$$\hat{\varphi}_{\text{MAP}} = \underset{\varphi}{\operatorname{argmax}}\, P(Y|\varphi, \Omega, W)p(\varphi, \Omega). \tag{23}$$

To calculate the transform matrix using the MAP criterion, we first define the prior density as

$$p(\mu_m^Y) = \mathcal{N}(\mu_m^Y; \eta_m, V_m), \tag{24}$$

where η_m and V_m are hyper-parameters of the prior density. In this study, we assume that V_m is a diagonal matrix.

With the prior density in Eq. (24), the MAP-based ATG mapping function is estimated by

$$\left[A^1 A^2 \dots A^P b\right]_{(d)} = k_{(d)} (G_{(d)})^{-1}, \quad d = 1, \dots D \tag{25}$$

where

$$G_{(d)} = \sum_{t=1}^{T} \sum_{m \in W} r_m(t) \frac{1}{\Sigma_{m_{(dd)}}} \rho_m \rho_m' + \sum_{m=1}^{M} \frac{\epsilon_m}{V_{m_{(dd)}}} \rho_m \rho_m', \tag{26}$$

and

$$k_{(d)} = \sum_{t=1}^{T} \sum_{m \in W} r_m(t) \frac{1}{\Sigma_{m_{(dd)}}} y_{t_{(d)}} \rho_m' + \sum_{m=1}^{M} \frac{\epsilon_m}{V_{m_{(dd)}}} \eta_{m_{(d)}} \rho_m', \tag{27}$$

where ϵ_m is a controlling factor. The MAP estimation [Eqs. (25)–(27)] becomes the ML estimation [Eqs. (8)–(10)] when setting $\epsilon_1 = \epsilon_2 = \cdots = \epsilon_M = 0$ in Eqs. (26) and (27).

Similar to the derivations of Eqs. (11)–(13), when using a diagonal matrix for each matrix A^p, $p = 1, 2, \dots, P$, we can obtain the MAP solution by

$$[a_{(d)}^1 a_{(d)}^2 \dots a_{(d)}^P b_{(d)}] = k_{(d)} (G_{(d)})^{-1}, \quad d = 1, \dots, D \tag{28}$$

with

$$G_{(d)} = \sum_{t=1}^{T} \sum_{m \in W} r_m(t) \frac{1}{\Sigma_{m_{(dd)}}} \xi_m \xi_m' + \sum_{m=1}^{M} \frac{\epsilon_m}{V_{m_{(dd)}}} \xi_m \xi_m', \tag{29}$$

$$k_{(d)} = \sum_{t=1}^{T} \sum_{m \in W} r_m(t) \frac{1}{\Sigma_{m_{(dd)}}} y_{t_{(d)}} \xi_m' + \sum_{m=1}^{M} \frac{\epsilon_m}{V_{m_{(dd)}}} \eta_{m_{(d)}} \xi_m', \tag{30}$$

where $\xi = \left[\mu_{m_{(d)}}^1 \mu_{m_{(d)}}^2 \dots \mu_{m_{(d)}}^P 1\right]'$. With the estimated parameter set, $\{a_{(d)}^1, a_{(d)}^2, \dots, a_{(d)}^P, b_{(d)}, d = 1, \dots, D\}$, the multiple affine transform set $\{A^1, A^2, \dots, A^P, b\}$ is obtained accordingly. Finally, μ_m^Y is calculated using Eq. (4).

3.2. ATG-ESSEM with model selection (MS)

In addition to adopting the MAP criterion, over-fitting issues can also be effectively handled by directly reducing the free parameters in the mapping function. This study proposes the MS process (Burnham and Anderson, 2002; Cotter et al., 2001; Kadane and Lazar, 2004; Sutter and Kalivas, 1993; Yuan and Lin, 2006) to select a compact subset of models that excludes redundant and noisy variables, thereby improving the accuracy of the adapted models. Various selection approaches and scores can be used to perform the selection. This study uses the backward elimination criterion (Cotter et al., 2001; Sutter and Kalivas, 1993) along with the Akaike information criterion (AIC) (Posada and Buckley, 2004; Yamaoka et al., 1978; Sakamoto et al., 1986). Algorithm 1 outlines the proposed MS process

Algorithm 1: Model Selection (MS)

Step-1: Initialization: the original score, L^0, is computed by

$$L^0 = 2K[\Omega] - 2ln[P(Y|\Lambda^Y, W)], \tag{31}$$

where $\Lambda^Y = F_\varphi(\Omega)$, $\Omega = \{\Lambda^1, \Lambda^2, \dots, \Lambda^{p-1}, \Lambda^p, \Lambda^{p+1}, \dots, \Lambda^P\}$. In Eq. (31), $ln[P(Y|\Lambda^Y, W)]$ is the log-likelihood of test data Y given the target model Λ^Y, and $K[\Omega]$ denotes the number of models in Ω. At the beginning stage, $K[\Omega] = P$.

Step-2: For the i-th iteration, one model is removed from Ω, and then an AIC value is computed. Accordingly, P AIC scores, L^p, $p = 1 \dots P$, are computed by

$$L^p = 2K[\Omega^{\notin p}] - 2ln[P(Y|\Lambda^Y, W)], \quad p = 1, 2, \dots P \tag{32}$$

where $\Lambda^Y = F_\varphi(\Omega^{\notin p})$, $\Omega^{\notin p} = \{\Lambda^1, \Lambda^2, \dots, \Lambda^{p-1}, \Lambda^{p+1}, \dots, \Lambda^P\}$. Among the P scores, the lowest score is selected by

$$L^i = \underset{p}{\text{argmin}}(L^1, L^2, \dots, L^{p-1}, L^p, L^{p+1}, \dots, L^P). \tag{33}$$

Next, either **Step-3a** or **Step-3b** will be taken.

Step-3a: If $L^i < L^0$, Λ^i is deleted, and we set the new score $L^0 = L^i$, $\Omega = \Omega^{\notin i}$, and $P = P - 1$. Then we go back to Step-2.

Step-3b: If $L^i > L^0$, the original Ω is selected, and the iteration stops. The computed model, Λ^Y is used to test recognition.

3.3. ATG-ESSEM with cohort selection (CS)

The CS process has been successfully integrated into previous ESSEM framework (Tsao et al., 2009). Experimental results confirmed that CS can improve ESSEM performance with ML-based LC and LCB (Tsao et al., 2009). This study expands by integrating CS with both ML- and MAP-based ATG-ESSEM.

The goal of CS is to construct a space that provides good coverage over the test condition (Rosenberg et al., 1992; Mak et al., 2006). The main concept of CS resembles that of subset selection methods, where a subset of components from the entire set of components is determined to model a signal of interest (Chen et al., 1998). The CS process can also be considered as an extension of the BF function (Section 2.3.2), except instead of locating the most matched environment, CS locates N training environments (cohort sets) that are closest to the testing environment. In this study, the likelihood score is used as the distance to determine the cohort set by

$$\Omega^{(\text{cohort})} = \text{RS}[\ln P(Y|\Lambda^p, W), \quad p = 1, 2, \dots P], \tag{34}$$

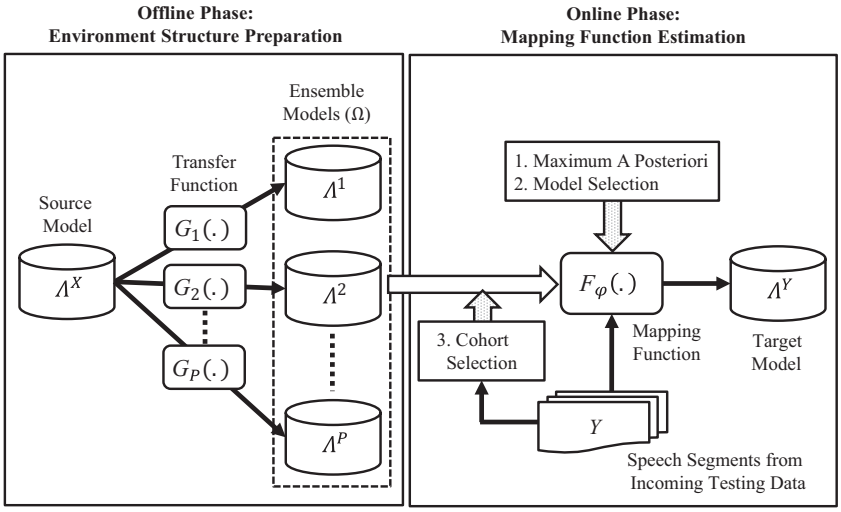


Fig. 3. ATG-ESSEM with three optimization processes: MAP criterion, MS, and CS.

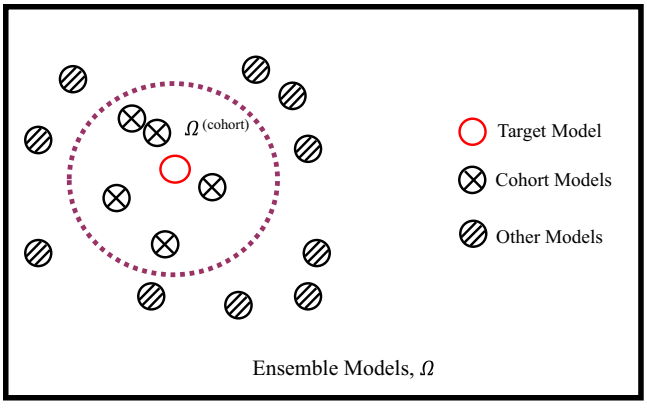


Fig. 4. Structure of ensemble models and the cohort models located closest to target.

where RS[·] denotes the ranking and selection function, which ranks the likelihood scores and selects the top N models out of the prepared P models. The entire structure is illustrated in Fig. 4. The cohort model set, $\Omega^{(\text{cohort})}$, represents a subset of the original ensemble model set, Ω, and has better coverage over the test condition.

After performing the CS process, the target model Λ^Y, is computed by:

$$\Lambda^Y = F_\varphi(\Omega^{(\text{cohort})}). \tag{35}$$

Since the number of models in $\Omega^{(\text{cohort})}$ is smaller than that in Ω, the free parameters in the mapping function in Eq. (35) are less than that in Eq. (2). Therefore, over-fitting issues can be handled when the amount of adaptation data is limited.

4. Experimental setup and results

The proposed ATG-ESSEM, three optimization processes, and related approaches were evaluated on the Aurora-4 database (Parihar and Picone, 2002; Parihar et al., 2004; Hirsch, 2001; Au Yeung and Siu, 2004).

4.1. Experimental setup

In this section, we introduce the experimental setup, including the Aurora-4 database, model topologies, features, and the optimization processes for the evaluation.

4.1.1. The Aurora-4 databases

Aurora-4 is a standardized database for evaluating LVCSR performance under different noise types and channel conditions. The original clean speech utterances in Aurora-4 were acquired from the Wall Street Journal (WSJ0) corpus (Paul and Baker, 1992). Different noises were later artificially added to the clean speech to generate noisy data. Two sampling rates, 8 kHz and 16 kHz, are provided in Aurora-4. The 8 kHz data was selected for both training and testing. Two training sets were provided: multi-condition and clean-condition training sets. In this study, we adopted the multi-condition training set. This set includes 7138 speech utterances with various microphone and noise distortions. The training set was divided into two parts: the first part of training data was recorded with the Sennheiser microphone, and the second part was recorded with different microphones. Each of the two parts included 893 clean speech utterances and 2676 noisy speech utterances that were artificially contaminated by six different noise types (car, babble, restaurant, street, airport, and train) at random SNR levels between 10 and 20 dB (Parihar and Picone, 2002; Parihar et al., 2004).

Fourteen test sets were provided for evaluating performances, and 166 utterances for each test set were used as suggested in (Parihar and Picone, 2002). The testing set included six types of noise: N1: car, N2: babble, N3: restaurant, N4: street, N5: airport, N6: train. The 14 test sets were classified into four groups (Au Yeung and Siu,

2004): set A (clean data with the Sennheiser microphone), set B (N1–N6 noisy data at 5–15 dB SNRs with the Sennheiser microphone), set C (clean data with different microphones), and set D (N1–N6 noisy data at 5–15 dB SNRs with different microphones).

4.1.2. Generating environment data for estimating ensemble models

With the 893 clean speech utterances recorded by the Sennheiser microphone, the Monte Carlo method (Metropolis and Ulam, 1949) was used to artificially generate environment data with four different environment types: subway, exhibition, pink, and white Gaussian noise (WGN) at 5, 10, 15, and 20 dB SNR levels. These four different environment types were intentionally selected because they were not included in the test set of Aurora-4. For this reason, fair performance comparisons can be made between approaches throughout this study without experimenter's bias about environment types. The data belonging to the same environment types were also split into high SNRs (15 and 20 dB) and low SNRs (5 and 10 dB). We thus generated eight artificial environment datasets, each containing 1786 utterances.

4.1.3. Acoustic and language models

For the multi-condition training set of Aurora-4, we applied ML training to estimate a set of context-dependent triphone models. Each triphone was characterized by an HMM, which consisted of three states, with eight Gaussian mixtures per state. This "environment-independent" model is denoted as the "EI" model. With the eight artificially generated environment datasets (Section 4.1.2), we applied MAP adaptation (Gauvian and Lee, 1994) on the EI model to transform eight sets of "environment-dependent" ED models. In addition, a clean condition trained model was estimated by applying MAP adaptation on the EI model using the 893 clean utterances. This environment-dependent model is denoted as the Clean-ED model. We thus have one Clean-ED and eight artificially generated ED models. Since the Clean-ED and eight artificially generated ED models were adapted from the EI model, the Gaussian components in all of the ten models are organized in the same order. Combining the EI model, Clean-ED, and eight artificially generated ED models, ten models were prepared for the ensemble models ($P = 10$ for the ESSEM framework in Fig. 2). Additionally, a tri-gram language model was prepared based on the reference transcription of training utterances (Hilger and Ney, 2006; Tuske et al., 2011).

4.1.4. Features, regression tree, and evaluation metrics

A modified European Telecommunications Standards Institute (ETSI) advanced front-end (AFE) (ETSI, 2007; Wu and Huo, 2006) was used for feature extraction. Each feature vector was characterized by standard 39-dimensional feature components, consisting of 13 static coefficients, and their first and second derivatives. In the experiments, we built a regression tree to cluster mean parameters in the model. The regression tree was constructed based on the EI models and consisted of one root, three intermediate, and six leaf nodes. We used the tree to determine the number of mapping functions. For the c-th node in the regression tree, we estimate its accumulated statistics, $R_c = \sum_{t=1}^{T} \sum_{m \in \{W, Z^c\}} r_m(t)$, where $m \in \{W, Z^c\}$ represents that the m-th Gaussian is in the transcription reference, W, and belongs to the c-th node, Z^c. If R_c is larger than a predefined threshold, we use the mapping function for the c-th node. If not, we check the accumulated statistics at the parent node, and the process repeats until we find a node that has sufficient statistics. For fair comparison, the same set of transcription references generated by the EI models were used to calculate mapping functions throughout the evaluations. For each utterance, the number of mapping functions used for environment modeling was thereby the same for the different types of mapping functions. The performance differences were determined only by the modeling capabilities among the mapping functions. When performing the MAP criterion in Section 3.1, the prior density was prepared based on the tree by following the steps used in a previous study (Tsao et al., 2012). On the other hand, for the MS process, likelihood scores are computed under the condition that the optimal number of mapping functions is used and determined by the adaptation data and the regression tree.

Besides the baseline, all of the following experimental results were obtained by performing environment modeling in an unsupervised per-utterance self-learning manner: Every incoming test utterance was first recognized by the EI model; the generated transaction and the speech data were then used to perform model adaptation to get a target model; the same utterance was then recognized by the target model, and the recognition result was used to evaluate performance. Word error rates (WERs) are reported as the performance measure.

4.2. Experimental results

For each experiment, we present the Aurora-4 results for set A, set B, set C, and set D, including 166, 996, 166, and 996 testing utterances, respectively. In addition, the average WER over the 14 testing conditions is reported and denoted as "Avg". The result Avg thus represents the average WER of 14 test sets, 2324 (166×14) utterances, and 38,010 (2715×14) words.

4.2.1. Confirmation of artificially generated environment data

We confirmed fair recognition performance using the models estimated by the artificially generated environment datasets (which were intentionally different from those in Aurora-4 to prevent experimenter's bias, as introduced in Section 4.1.2). Table 1 lists results for the Baseline using the EI model without model adaptation, Clean-ED model, and the eight artificially generated ED models at high SNRs (15 and 20 dB) and low SNRs (5 and 10 dB).

Table 1
Average WERs (%) of ten models on Aurora-4 test sets. The best result for each test set is shown with bold digits.

Test condition	Set A	Set B	Set C	Set D	Avg
EI (Baseline)	9.80	**16.99**	14.00	**23.11**	**18.88**
Clean-ED	**9.54**	20.42	**13.78**	27.49	22.20
Subway_high	10.24	18.08	15.54	24.73	20.19
Exhibition_high	10.28	18.64	14.88	24.73	20.38
Pink_high	10.31	19.69	14.88	26.15	21.44
WGN_high	10.72	18.96	15.10	26.46	21.31
Subway_low	11.42	17.63	15.76	24.30	19.91
Exhibition_low	11.05	17.98	15.73	24.27	20.02
Pink_low	11.31	18.36	15.65	25.03	20.52
WGN_low	11.49	18.68	15.99	25.62	20.95

As expected, the EI model which was trained on the multi-condition training data achieved the best performance for the noisy conditions (sets B and D), while Clean-ED achieved the best performance in sets A and C containing some data recorded from clean conditions. Table 1 also confirms that the eight artificially generated ED models cannot initially provide satisfactory performance since they provide poor coverage over the testing sets of Aurora-4. This confirms the need for model adaptation for preparing environment-specific models. Real-world environments (subway and exhibition) provided better recognition compared to pink and WGN.

4.2.2. ML-based mapping function estimation

This section presents recognition results of using the ATG mapping function estimated by the ML criterion. Table 2 shows the ML-based ATG results, where ML-ATG (*Full*) denotes results obtained using a full matrix and ML-ATG (*Diag*) denotes results obtained using a diagonal matrix in each A^p in Eq. (4). All $P = 10$ ensemble models prepared by the 10 conditions in Table 1 were used in this set of model adaptation results.

Table 2
Average WERs (%) of ATG estimated by ML criterion, $P = 10$ ensemble models. The best result for each test set is shown with bold digits.

Test condition		Set A	Set B	Set C	Set D	Avg
$P = 10$	ML-ATG (*Full*)	17.94	27.93	22.95	33.06	29.06
	ML-ATG (*Diag*)	**9.83**	**16.78**	**13.11**	**22.43**	**18.44**

Table 2 shows ML-ATG (*Diag*) outperforms EI (Baseline in Table 1) for almost all of the five test sets, confirming the effectiveness of the ATG mapping function for model adaptation. Since Table 1 demonstrated that the artificially generated environment conditions had poor coverage over the testing condition, it can be concluded that ML-ATG (*Diag*) effectively estimated a target model that better matched the test condition. However, Table 2 also showed that ML-ATG (*Full*) underperforms EI (Baseline). These results demonstrate the detrimental effects of over-fitting issues when full matrices were used for ML-ATG (*Full*) adaptation but only one adaptation utterance was available in the task.

Next, we compare the ATG mapping function with related mapping functions. Table 3 lists the results of ML-LCB, ML-LC, and ML-BF (for $P = 10$ ensemble models), and those of ML-LR (*Diag*), ML-LR (*Full*), and ML-BC (for $P = 1$, single source model). The results in Table 3 show that all the environment modeling approaches outperform EI (Baseline in Table 1), with the exception of ML-LR (*Full*) due to the over-fitting issue. Results also demonstrate that using ensemble models ($P = 10$, for ML-LCB and ML-LC) facilitated better environment modeling accuracy compared to using only a single source model ($P = 1$, for ML-LR and ML-BC), especially under noisy conditions (sets B and D) in this task. The correction bias (ML-LCB vs. ML-LC) was effective for reducing residual interferences. In fact, ML-LCB also outperforms ML-ATG (*Diag*) from Table 2. Since the numbers of mapping functions (determined based on the regression tree and the amount of adaptation data) are equivalent in Tables 2 and 3, performance differences only correspond to the different modeling capabilities among the mapping functions. Therefore, the only difference between ML-ATG (*Diag*) and ML-LCB was the additional regression terms used in each A^p, indicating a simpler mapping function could be more favorable for avoiding over-fitting issues when the adaptation data is especially limited.

4.2.3. ATG-ESSEM with the MAP criterion, model selection, and cohort selection

This section presents the results of ATG estimated with the MAP criterion, MS, and CS. As mentioned in Section 3, these processes are used to overcome over-fitting issues

Table 3
Average WERs (%) of EI (Baseline) and ML-based related methods. The best result for each test set is shown with bold digits.

Test condition		Set A	Set B	Set C	Set D	Avg
EI (Baseline)		9.80	16.99	14.00	23.11	18.88
$P = 10$	ML-LCB	9.77	**16.46**	**13.44**	**22.10**	**18.18**
	ML-LC	9.79	16.49	14.07	22.50	18.41
	BF	**9.45**	17.10	13.78	23.05	18.87
$P = 1$	ML-LR (*Full*)	10.50	19.76	15.14	25.55	21.25
	ML-LR (*Diag*)	9.98	16.89	13.78	22.84	18.72
	ML-BC	9.78	16.89	13.50	22.94	18.73

to enhance adaptation performance. Since Table 2 indicated that ATG with diagonal matrices provides better performance than ATG with full matrices, the following discussion only presents ATG with diagonal matrices.

4.2.3.1. ATG estimated by MAP criterion. Table 4 shows the results of MAP-based ATG using a diagonal matrix, denoted as MAP-ATG. For the MAP criterion, the hyper-parameters, η_m and V_m, of the prior density are obtained from the training data. Since MAP-ATG outperforms ML-ATG from Table 2, this confirms the effectiveness of the MAP criterion for ATG mapping function estimation.

Fig. 5 shows MAP-ATG along with the other mapping functions estimated based on the MAP criterion and compares them with their ML-based counterparts on test set Avg. Although LCB previously outperformed ATG with the ML criterion, Fig. 5 clearly shows MAP-ATG outperforms MAP-LCB, demonstrating how MAP is preferable for incorporating prior knowledge in the more complicated mapping function of ATG. Results also demonstrate that using ensemble models ($P = 10$, for MAP-LCB and MAP-LC) facilitated better environment modeling accuracy compared to using only a single source model ($P = 1$, for MAP–LR and MAP-BC). Moreover, the MAP-based methods of Fig. 5 all outperform their ML-based counterparts, confirming the effectiveness of prior information for providing better adaptation performance. From all the presented results, MAP-ATG achieved the best performance among all the different mapping functions and criteria.

Table 4
Average WERs (%) of ATG estimated by MAP criterion, $P = 10$ ensemble models.

Test condition		set A	set B	set C	set D	Avg
$P = 10$	MAP-ATG	**9.72**	**16.36**	**13.26**	**21.71**	**17.96**

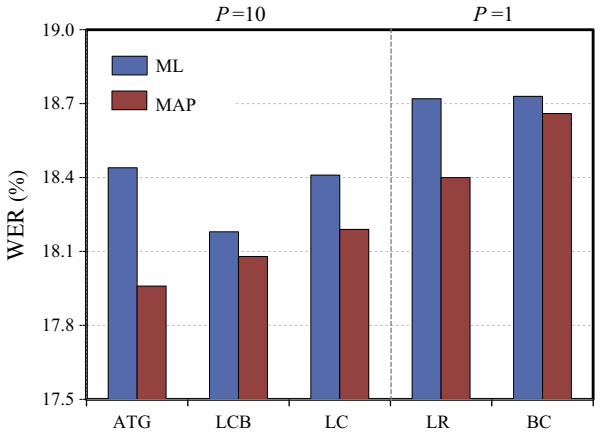


Fig. 5. WERs (%) of test set Avg of five different mapping functions estimated with the ML and MAP criteria.

4.2.3.2. ATG with model selection (MS). This section presents the results of ATG with the MS process, denoted as ATG-MS in the following discussion. In our preliminary experiments, we found that when using the AIC criterion as defined in Eq. (31), the likelihood part, namely $2ln[P(Y|\Lambda^Y, W)]$, dominates the AIC score. Therefore, we derived a modified AIC score for our task:

$$L^{\text{mod}} = 2\lambda K[\Omega] - 2ln[P(Y|\Lambda^Y, W)], \tag{36}$$

where λ is a scalar which determines the complexity term and log-likelihood score. When $\lambda = 0$, the L^{mod} becomes the log-likelihood value. The modified score in Eq. (36) is used in the three steps of backward elimination criterion for ATG-MS (Algorithm 1). It is noted that by using a larger λ, fewer models are used for model adaptation, and by using a smaller λ, more models are used for ATG. The baseline of ML-ATG (*Diag*) (average WER = 18.44%, Table 2) should serve for comparison since the goal of the MS process is to optimize the number of ensemble models by eliminating redundant models based on the adaptation data.

Fig. 6 shows results for ATG-MS with various λ in Eq. (36). From Fig. 6, optimal performance of ATG-MS occurred when $\lambda = 4.0$, confirming that ESSEM can achieve better performance by using the MS process to optimally determine the complexity of the ATG mapping function (instead of using all $P = 10$ sets of ensemble models).

Table 5 lists the detailed results of five test sets of ATG-MS with $\lambda = 4.0$. Table 5 confirms the effectiveness of the MS process for overcoming over-fitting issues since ATG-MS outperforms ML-ATG (*Diag*) of Table 2 for most test

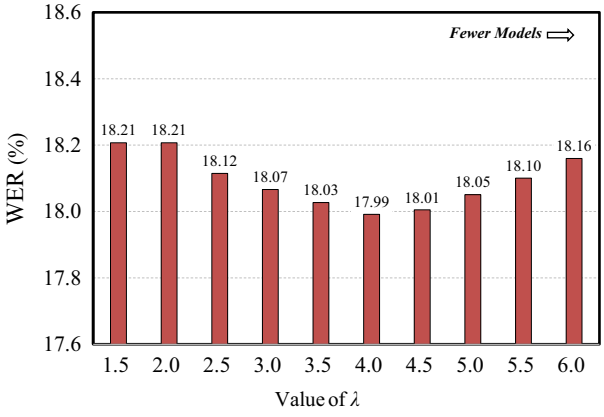


Fig. 6. WERs (%) of test set Avg of ATG-MS using different λ values.

Table 5
Average WERs (%) of ATG-MS with $\lambda = 4.0$, LCB-MS, and LC–MS. The best result for each test set is shown with bold digits.

Test condition	Set A	Set B	Set C	Set D	Avg
ATG-MS	**9.61**	**16.38**	**13.19**	**21.80**	**17.99**
LCB-MS	9.76	16.48	13.52	21.85	18.09
LC–MS	9.80	16.49	13.78	22.11	18.22

conditions and the result Avg. In addition, when the MS process was integrated with the LCB and LC mapping functions, both LCB-MS and LC–MS outperform their ML-based counterparts of Table 3. This confirms that the MS process can also enhance the adaptation performance for Category-2 mapping functions.

4.2.3.3. ATG with cohort selection (CS). This section presents the results of ATG with the CS process, denoted as ATG-CS in the following discussion. As mentioned in Section 3.3, the log-likelihood score is used as the measurement to find the cohort set. The N models in Ω that give higher log-likelihoods are selected. Fig. 7 shows the ATG-CS with various numbers ($N = 1$–10) of cohort models. When $N = 1$, only the one model which gives the highest likelihood is selected. When $N = 10$, all of the ensemble models are used for ATG. Therefore, $N = 10$ should be considered the baseline of ATG-CS since all of the 10 models are used to generate the cohort set without any selection process, which is equivalent to the ML-ATG (*Diag*) condition in Table 2. Fig. 7 confirms that an optimal number of ensemble models is required ($N = 5$ or $N = 6$) to facilitate better adaptation performance, consistent with the reported set of results from Fig. 6.

To determine whether the combination of CS and the MAP criterion could further enhance performance, we used CS to select a cohort set and then applied the MAP-ATG in Eq. (23) for the ESSEM adaptation. Table 6 shows that the five test sets of the best performing ATG-CS condition ($N = 6$) could be further enhanced via integration with the MAP criterion.

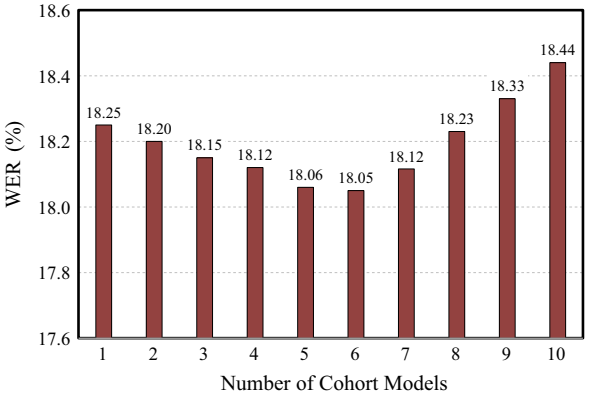


Fig. 7. WERs (%) of test set Avg of ATG-CS with different numbers of cohort models.

Table 6
Average WERs (%) of ATG with CS or CS + MAP. The best result for each test set is shown with bold digits.

Test condition	Set A	Set B	Set C	Set D	Avg
ATG-CS	9.67	16.46	13.18	21.86	18.05
MAP-ATG-CS	**9.61**	**16.30**	**13.00**	**21.75**	**17.93**

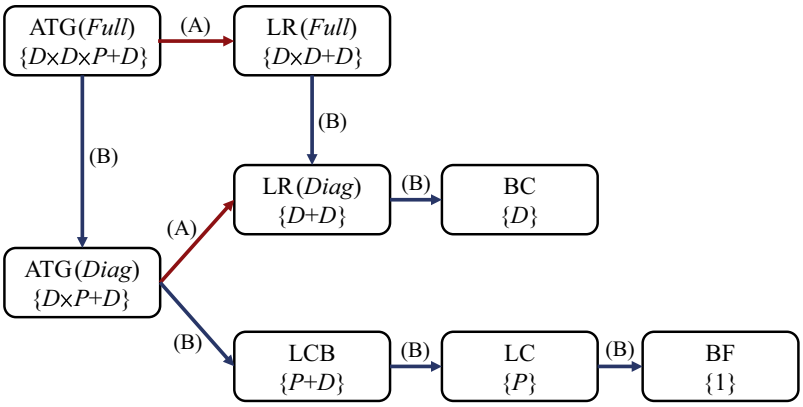
4.3. Mapping function complexity and performance summary

Fig. 8 presents the complexity of the eight different types of mapping functions used in this paper, where D is the number of feature dimensions, P is the number of ensemble models, and the total number of free parameters is shown in the brace. For example in our experimental setup, $D = 39$ and $P = 10$. In Fig. 8, both ATG (*Full*) and ATG (*Diag*) used $\Omega = \{\Lambda^1, \Lambda^2, \ldots, \Lambda^P\}$. When using a single source model in Ω, namely $\Omega = \{\Lambda^X\}$, ATG (*Full*) and ATG (*Diag*) becomes LR (*Full*) and LR (*Diag*), respectively. By further reducing the affine transform, LR (*Diag*) simplifies to BC. On the other hand, ATG becomes LCB by simplifying each A^p to a scalar matrix. When the bias b in LCB is not used, ATG simplifies to LC. Since BF uses only one parameter in the mapping function, it can be considered as the hard-decision version of LC. With sufficient availability of adaptation data, ATG should characterize the testing condition most precisely because it has the most complex form.

Tables 1–3 showed over-fitting degraded ML-ATG performance at first, since our model adaptation experiments were conducted in a per-utterance unsupervised manner. Therefore, this ATG-ESSEM study showed how over-fitting issues could be handled by using either the MAP criterion, MS, or CS. In fact, MAP effectively enabled ATG to achieve better performance than all related approaches including LCB, LC, BF, LR, and BC. The MS process was also verified to improve ML-ATG as well as the LCB and LC mapping functions. The CS process improved both ML-ATG and MAP-ATG when used as a preceding stage. To verify the significance of all these improvements, we conducted a t-test analysis (Hayter, 2006) using the 14 pair-wise results from the 14 distinct test sets to compare EI (Baseline) and the performances of ATG-ESSEM with MAP, MS, and CS processes. The t-test results verified that ATG-ESSEM with MAP, MS, and CS processes all provided P-values ($p < 0.01$) indicating consistent performance improvements compared to EI (Baseline) over the 14 test sets. The consistency of our preliminary results were especially encouraging and demonstrated that the ESSEM framework effectively utilized prior knowledge and optimized to provide adaptation. To simulate "real world" conditions, the evaluation task was designed to be especially difficult: using only one available adaptation utterance with no available SNR labels and potential errors in the decoded transcription. If environment and SNR information were given or potentially detected, better ensemble models could be prepared or selected by processes such as CS, thus enabling better adaptation performance for ATG-ESSEM.

4.4. Discussion

The MAP, MS, and CS processes have their advantages and potential extensions for optimizing the mapping func-



(A): Reducing the number of ensemble models.
(B): Simplifying the mapping function.

Fig. 8. ATG and related mapping functions with corresponding complexities {shown in the brace}, where D denotes feature dimensions and P denotes the number of ensemble models used for adaptation.

tion estimation and determining the appropriate number of ensemble models for adaptation. For MAP, since a direct regularization is applied in the mapping function estimation, the online computation is almost the same as the ML counterpart. This study adopts the Gaussian distribution for the prior density, but other forms of prior density could be explored to provide the prior information better. For MS, the main advantage is that it allows direct control of the complexity of the mapping function according to the amount of adaptation data. Notably, both MAP and AIC are derived based on the Bayesian framework (Burnham and Anderson, 2002) but use different approaches to handle over-fitting issues. In addition to AIC, other penalization terms could be designed such as discrimination and regularization (L1 or L2 norms) scores in order to insert into Eq. (36) for different purpose. For CS, a compact model set is generated from the original model set by filtering out irrelevant models. Therefore, CS can be regarded as a pre-processing step and could be combined with other processes (such as how this study combined the MAP criterion). While the cohort set was selected based on the likelihood score, future work could explore alternative criteria to determine the cohort set.

The present study focused more on optimizing the number of ensemble models for adaptation and only compared two fundamental affine transform structures for ATG: full and diagonal matrices. Several options could be used to optimize the structure of affine transforms. By considering the correlations of coefficients in a mean vector, a full matrix can be divided into several block-diagonal matrices for transformation (Gales, 1997). Moreover, regularization terms (Li et al., 2010; Li et al., 2011) and variable selection methods (Tsao et al., 2014) can be used to reduce redundant components in a full matrix according to the available adaptation data.

Current state-of-the-art techniques for ASR involve the hybrid DNN–HMM framework, where several adaptation algorithms have been proposed. For example, linear input network (LIN) and linear output network (LON) adopt affine transformations (akin to the present study's LR mapping function) to the inputs or outputs of a neural networks for adaptation (Neto et al., 1995; Li and Sim, 2010; Yao et al., 2012), and linear hidden network (LHN) applies affine transformations to the activations of the internal hidden layers (Gemello et al., 2007). Affine transforms can also be adopted to conduct speaker adaptive training (SAT) for DNN (Ochiai et al., 2014). These DNN–HMM parameter adaptation approaches all use the same linear transformation as those discussed in the present study based on the original model parameter and adaptation data. Therefore, many of the findings from the ESSEM framework can also be generalized to the DNN–HMM adaptation framework. For instance, a group of adaptation transform functions could be utilized to capture more structures of adaptation data and model parameters in DNN model parameter adaptation. Since the GMM–HMM framework is based on the generative training paradigm, robustness for unsupervised adaptation could also be enabled by combining GMM and DNN to enhance the ASR performance (Liu and Sim, 2014). In addition, DNN–HMM systems involve enormous quantities of neural network parameters and special optimization processes are often necessary to avoid overfitting of the limited adaptation data. For instance, the Kullback–Leibler divergence regularization (Yu et al., 2013) and L2 prior regularization (Liao 2013) can be integrated to perform parameter adaptation for DNN. The optimization processes described in the ESSEM framework can also be applied to optimize the adaptation process, especially when adaptation data is limited. Moreover, similar concepts of ATG-ESSEM can be applied to entirely different tasks with limited availability of data, such as in ensemble modeling of denoising deep autoencoder based spectral restoration (Lu et al., 2014) and in vector-space based acoustic echo cancelation (Tsao et al., 2015).

5. Conclusion

The ESSEM framework was designed to handle real-world conditions by preparing ensemble models to cover a wide range of environments in the offline phase. The ATG-ESSEM framework effectively characterized unknown combinations of multiple distortion sources from environment conditions by applying an affine transform for each ensemble model to compute the final target model. Experimental results on the Aurora-4 task in a per-utterance unsupervised model adaptation mode showed ATG achieved consistent performance improvements compared to baseline and most mapping functions in Category-1 and Category-2, even when only unlabeled and limited adaptation data (i.e., only one utterance) from the test condition was available. Performance of ATG was also shown to optimize using either the MAP criterion, MS, or CS. Furthermore, directly imposing constraints on the ATG mapping function was shown to allow flexible extension to simpler approaches such as LR, BC, LCB, LC, and BF to deliver a variety of workable systems depending on the user's environments.

Acknowledgement

This work was supported by the National Science Council of Taiwan under contracts NSC101-2221-E-001-020-MY3.

References

AuYeung, S.-K., Siu, M.-H., 2004. Improved performance of Aurora-4 using HTK and unsupervised MLLR adaptation. In: Proceedings of ICSLP'04, pp. 161–164.

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer.

Chen, S.S., Donoho, D.L., Saunders, M.A., 1998. Atomic decomposition by basis pursuit. SIAM J. Scient. Comput. 20 (1), 33–61.

Chesta, C., Siohan, O., Lee, C.-H., 1999. Maximum a posteriori linear regression for hidden Markov model adaptation. In: Proceedings of Eurospeech'99, pp. 211–214.

Cotter, S.F., Kreutz-Delgado, K., Rao, B.D., 2001. Backward sequential elimination for sparse vector subset selection. Sig. Process. 81 (9), 1849–1864.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front-end factor analysis for speaker verification. IEEE Trans. Audio, Speech Lang. Process. 19 (4), 788–798.

Deng, L., Huang, X., 2004. Challenges in adopting speech recognition. Commun. ACM 47 (1), 69–75.

ETSI ES 202 050 V1.1.5, 2007. Speech Processing, Transmission and Quality-Aspects (STQ); Distributed Speech Recognition; Advanced Frontend Feature Extraction Algorithms. ETSI standard.

Gales, M.J.F., 1997. Maximum likelihood linear transformations for HMM-based speech recognition. Comp. Speech Lang. 12, 75–98.

Gales, M.J.F., 2000. Cluster adaptive training of hidden Markov models. IEEE Trans. Speech Audio Process. 8 (4), 417–428.

Gauvian, J., Lee, C.-H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech Audio Process. 2 (2), 291–298.

Gemello, R., Mana, F., Scanzio, S., Laface, P., De Mori, R., 2007. Linear hidden transformations for adaptation of hybrid ANN/HMM models. Speech Commun. 49 (10), 827–835.

Gong, Y., 2005. A method of joint compensation of additive and convolutive distortions for speaker-independent speech recognition. IEEE Trans. Speech Audio Process. 13 (5), 975–983.

Hayter, A.J., 2006. Probability and Density for Engineers and Scientists, third ed. Duxbury Press.

Hazen, T.J., 2000. A comparison of novel techniques for rapid speaker adaptation. Speech Commun. 31 (1), 15–33.

Hilger, F., Ney, H., 2006. Quantile based histogram equalization for noise robust large vocabulary speech recognition. IEEE Trans. Audio, Speech, Lang. Process. 14 (3), 845–854.

Hirsch, G., 2001. Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task. ETSI STQ Aurora DSR Working Group.

Hu, Y., Huo, Q., 2006. An HMM compensation approach using unscented transformation for noisy speech recognition. In: Chin. Spoken Lang. Process. Springer, Berlin Heidelberg, pp. 346–357.

Hu, Y., and Huo, Q., 2007. Irrelevant variability normalization based HMM training using VTS approximation of an explicit model of environmental distortions. In: Proceedings of Interspeech' 07, pp. 1042–1045.

Huo, Q., Lee, C.-H., 2000. A Bayesian predictive classification approach to robust speech recognition. IEEE Trans. Speech Audio Process. 8 (2), 200–204.

Jeong, Y., 2012. Adaptation of hidden markov models using model-as-matrix representation. IEEE Trans. Audio, Speech, Lang. Process. 20 (8), 2352–2364.

Jeong, Y., 2014. Joint speaker and environment adaptation using TensorVoice for robust speech recognition. Speech Commun. 58, 1–10.

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

Kadane, J.B., Lazar, N.A., 2004. Methods and criteria for model selection. J. Am. Statist. Assoc. 99 (465), 279–290.

Kim, D.Y., Un, C.K., Kim, N.S., 1998. Speech recognition in noisy environments using first order vector Taylor series. Speech Commun. 24, 39–49.

Kosaka, T., Matsunaga, S., Sagayama, S., 1996. Speaker-independent speech recognition based on tree structured speaker clustering. Comp. Speech Lang. 10, 55–74.

Kuhn, R., Junqua, J.-C., Nguyen, P., Niedzielski, N., 2000. Rapid speaker adaptation in Eigenvoice space. IEEE Trans. Speech Audio Process. 8 (6), 695–707.

Lee, C.-H., 1998. On stochastic feature and model compensation approaches to robust speech recognition. Speech Commun. 25 (1–3), 29–47.

Leggetter, C., Woodland, P., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Comp. Speech Lang. 9 (2), 171–185.

Li, B., Sim, K.C., 2010. Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems. In: Proceedings of Interspeech'10, pp. 526–529.

Li, J., Deng, L., Yu, D., Gong, Y., Acero, A., 2009. A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions. Comp. Speech Lang. 23 (3), 389–405.

Li, J., Yu, D., Gong, Y., Deng, L. 2010a. Unscented transform with online distortion estimation for HMM adaptation. In: Proceedings of Interspeech'10, pp. 1660–1663.

Li, J., Tsao, Y., Lee, C.-H., 2010b. Shrinkage model adaptation in automatic speech recognition. In: Proceedings of Interspeech'10, pp. 1656–1659.

Li, J., Yuan, M., Lee, C.-H., 2011. LASSO model adaptation for automatic speech recognition. In: Proceedings of ICML'11, Workshop on Learning Architectures, Representations, and Optimization for Speech and Visual Information Processing.

Li, J., Deng, L., Gong, Y., Haeb-Umbach, R., 2014. An overview of noise-robust automatic speech recognition. IEEE/ACM Trans. Audio, Speech, Lang. Process. 22 (4), 745–777.

Liao, H., 2013. Speaker adaptation of context dependent deep neural networks. In: Proceedings of IC ASSP'13, pp. 7947–7951.

Liu, S., Sim, K.C., 2014. On combining DNN and GMM with unsupervised speaker adaptation for robust automatic speech recognition. In: Proceedings of IC ASSP'14, pp. 195–199.

Lu, X., Tsao, Y., Matsuda, S., Hori, C., 2014. Ensemble modeling of denoising autoencoder for speech spectrum restoration. In: Proceedings of Interspeech'14, pp. 885–889.

Mak, B., Lai, T.C., Hsiao, R., 2006. Improving reference speaker weighting adaptation by the use of maximum-likelihood reference speakers. In: Proceedings of ICASSP' 06, pp. 229–232.

Metropolis, N., Ulam, S., 1949. The Monte Carlo method. J. Am. Statist. Assoc. 44 (247), 335–341.

Neto, J., Almeida, L., Hochberg, M., Martins, C., Nunes, L., Renals, S., Robinson, T., 1995. Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system. In: Proceedings of Eurospeech'95, pp. 2171–2174.

Norman, D.A., 1984. Stages and levels in human-machine interaction. Int. J. Man–Mach. Stud. 21 (4), 365–375.

Ochiai, T., Matsuda, S., Lu, X., Hori, C., Katagiri, S., 2014. Speaker adaptive training using deep neural networks. In: Proceedings of IC ASSP'14, pp. 6349–6353.

O'Shaughnessy, D., 2008. Invited paper: automatic speech recognition: history, methods and challenges. Pattern Recog. 41 (10), 2965–2979.

Padmanabhan, M., Bahl, L.R., Nahamoo, D., Picheny, M.A., 1998. Speaker clustering and transformation for speaker adaptation in speech recognition systems. IEEE Trans. Speech Audio Process. 6 (1), 71–77.

Parihar, N., Picone, J., 2002. Aurora working group: DSR front end LVCSR evaluation au/384/02. In: Institute for Signal and Information Processing Report.

Parihar, N., Picone, J., Pearce, D., Hirsch, H.G., 2004. Performance analysis of the Aurora large vocabulary baseline system. In: Proceedings of EUSIPCO'04, pp. 553–556.

Paul, D.B., Baker, J.M., 1992. The design for the wall street journal-based CSR corpus. In: Proceedings of ICSLP'92, pp. 357–362.

Posada, D., Buckley, T.R., 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53 (5), 793–808.

Rahim, M., Juang, B.-H., 1996. Signal bias removal by maximum likelihood estimation for robust telephone speech recognition. IEEE Trans. Speech Audio Process. 4 (1), 19–30.

Reynolds, D.A., Rose, R., 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Process. 3 (1), 72–83.

Rosenberg, A.E., DeLong, J., Lee, C.H., Juang, B.H., Soong, F.K., 1992, The use of cohort normalized scores for speaker verification. In: Proceedings of ICSLP'92, pp. 599–602.

Sakamoto, Y., Ishiguro, M., Kitagawa, G., 1986. Akaike Information Criterion Statistics. D. Reidel, Dordrecht, The Netherlands.

Sankar, A., Lee, C.-H., 1996. A maximum-likelihood approach to stochastic matching for robust speech recognition. IEEE Trans. Speech Audio Process. 4 (3), 190–202.

Seltzer, M.L., Yu, D., Wang, Y., 2013. An investigation of deep neural networks for noise robust speech recognition. In: Proceedings of I CASSP'13, 7398–7402.

Shinoda, K., Lee, C.-H., 2001. A structural Bayes approach to speaker adaptation. IEEE Trans. Speech Audio Process. 9 (3), 276–287.

Siohan, O., Chesta, C., Lee, C.-H., 2001. Joint maximum a posteriori adaptation of transformation and HMM parameters. IEEE Trans. Speech Audio Process. 9 (4), 417–428.

Siohan, O., Myrvoll, T.A., Lee, C.-H., 2002. Structural maximum a posteriori linear regression for fast HMM adaptation. Comp. Speech Lang. 16 (1), 5–24.

Suredran, A.C., Lee, C.-H., Rahim, M., 1999. Nonlinear compensation for stochastic matching. IEEE Trans. Speech Audio Process. 7 (6), 643–655.

Sutter, J.M., Kalivas, J.H., 1993. Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection. Microchem. J. 47 (1), 60–66.

Tsao, Y., Lee, C.-H., 2009. An ensemble speaker and speaking environment modeling approach to robust speech recognition. IEEE Trans. Audio, Speech, Lang. Process. 17 (5), 1025–1037.

Tsao, Y., Li, J., Lee, C.H., 2009. Ensemble speaker and speaking environment modeling approach with advanced online estimation process. In: Proceedings of ICASSP'09, pp. 3833–3836.

Tsao, Y., Huang, C.-L., Matsuda, S., Hori, C., Kashioka, H., 2012. A linear projection approach to environment modeling for robust speech recognition. In: Proceedings of ICASSP'12, pp. 705–708.

Tsao, Y., Matsuda, S., Hori, C., Kashioka, H., Lee, C.-L., 2014a. A MAP-based online estimation approach to ensemble speaker and speaking environment modeling. IEEE/ACM Trans. Audio, Speech, Lang. Process. 22 (2), 403–416.

Tsao, Y., Hu, T.-Y., Sakti, S., Nakamura, S., Lee, L.-S., 2014b. Variable selection linear regression for robust speech recognition. IEICE Trans. Inform. Syst. 97 (6), 1477–1487.

Tsao, Y., Lu, X., Dixon, P., Hu, T.-Y., Matsuda, S., Hori, C., 2014c. Incorporating local information of the acoustic environments to MAP-based feature compensation and acoustic model adaptation. Comp. Speech Lang. 28 (3), 709–726.

Tsao, Y., Fang, S.H., Shiao, Y., 2015. Acoustic echo cancellation using a vector-space-based adaptive filtering algorithm. Signal Process. Lett., IEEE 22 (3), 351–355.

Tuske, Z., Golik, P., Schluter, R., Drepper, F.R., 2011. Non-stationary feature extraction for automatic speech recognition. In: Proceedings of ICASSP'11, pp. 5204–5207.

Watanabe, S., Nakamura, A., Juang, B.-H., 2014. Structural Bayesian linear regression for hidden Markov models. J. Signal Process. Syst. 74 (3), 341–358.

Wu, J., Huo, Q., 2006. An environment-compensated minimum classification error training approach based on stochastic vector mapping. IEEE Trans. Audio, Speech, Lang. Process. 14 (6), 2147–2155.

Yamaoka, K., Nakagawa, T., Uno, T., 1978. Application of Akaike's information criterion (AIC) in the evaluation of linear pharmacokinetic equations. J. Pharmacokinet. Biopharmaceut. 6 (2), 165–175.

Yao, K., Yu, D., Seide, F., Su, H., Deng, L., Gong, Y., 2012. Adaptation of context-dependent deep neural networks for automatic speech recognition. In: Proceedings of SLT'12, pp. 366–369.

Yu, K., Gales, M.J.F., 2006. Discriminative cluster adaptive training. IEEE Trans. Audio, Speech, Lang. Process. 4 (5), 1694–1703.

Yu, D., Yao, K., Su, H., Li, G., Seide, F., 2013. KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In: Proceedings of ICASSP'13, pp. 7893–7897.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. Roy. Statist. Soc.: Ser. B (Statist. Method.) 68 (1), 49–67.