A LINEAR PROJECTION APPROACH TO ENVIRONMENT MODELING FOR ROBUST SPEECH RECOGNITION

Yu Tsao[1], Chien-Lin Huang[2], Shigeki Matsuda[2], Chiori Hori[2], Hideki Kashioka[2]
[1]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[2]SLC Group, National Institute of Information and Communications Technology, Kyoto, Japan

ABSTRACT

Use of a linear projection (LP) function to transform multiple sets of acoustic models into a single set of acoustic models is proposed for characterizing testing environments for robust automatic speech recognition. The LP function is an extension of the linear regression (LR) function used in maximum likelihood linear regression (MLLR) and maximum a posteriori linear regression (MAPLR) by incorporating local information in the ensemble acoustic space to enhance the environment modeling capacity. To estimate the nuisance parameters of the LP function, we developed maximum likelihood LP (MLLP) and maximum a posteriori LP (MAPLP) and derived a set of integrated prior (IP) densities for MAPLP. The IP densities integrate multiple knowledge sources from the training set, previously seen speech data, current utterance, and a prepared tree structure. We evaluated the proposed MLLP and MAPLP on the Aurora-2 database in an unsupervised model adaptation manner. Experimental results show that the LP function outperforms the LR function with both ML- and MAP-based estimates over different test conditions. Moreover, because the MAP-based estimate can handle over-fittings well, MAPLP has clear improvements over MLLP. Compared to the baseline result, MAPLP provides a significant 10.99% word error rate reduction.

Index Terms—Acoustic Model Adaptation, Robust Speech Recognition, Environment Modeling, Linear Projection

1. INTRODUCTION

An automatic speech recognition (ASR) system uses a set of acoustic models, which is estimated on a set of training data, to recognize testing utterances to word or sub-word sequences. If the training and testing conditions do not match, the ASR performance may be degraded. An effective way to handle the mismatch is to estimate a new set of acoustic models that characterizes the testing environment well. Then, the estimated acoustic models are used for performing recognition. Many environment modeling approaches have been proposed. Generally, these approaches prepare acoustic model sets, Ω, using training data. A mapping function, F_{φ}, is then calculated to transform Ω to a new set of acoustic models, Λ^{Y}, by

$$\Lambda^{\mathrm{Y}} = \mathrm{F}_{\varphi}(\Omega), \quad (1)$$

where Λ^{Y} is the acoustic model set for the testing condition. Ω may include a single set ($\Omega = \{\Lambda^{\mathrm{X}}\}$) or multiple sets ($\Omega = \{\Lambda^{1}, \Lambda^{2}, \ldots \Lambda^{P}\}$) of acoustic models, where Λ^{X} and Λ^{p} (p=1,2…P) are acoustic models estimated based on the training data. φ is the nuisance parameters of the mapping function, F_{φ} [1]. These approaches can be summarized into two categories based on the type of Ω.

The first category of approaches sets $\Omega = \{\Lambda^{\mathrm{X}}\}$, where Λ^{X} is generally calculated on the entire set of training data. For these approaches, the mapping function, F_{φ}, parameterizes the mismatch between training and testing conditions and transforms Λ^{X} to Λ^{Y}. Successful approaches include the stochastic matching algorithm [1], maximum likelihood linear regression (MLLR) [2], maximum a posteriori linear regression (MAPLR) [3], and joint compensation of additive and convolutive distortions (JAC) [4].

The second category of approaches sets $\Omega = \{\Lambda^{1}, \Lambda^{2}, \ldots \Lambda^{P}\}$, where each Λ^{p} (p=1,2…P) models a particular acoustic condition in the overall training set. Compared to the previous category, this one usually uses a simpler mapping function, such as best first (BF), linear combination (LC), or a linear combination with a correction bias (LCB) [5–8]. The collection of $\{\Lambda^{1}, \Lambda^{2}, \ldots \Lambda^{P}\}$ is usually pre-processed to enhance the efficiency and accuracy of the mapping procedure, such as through principal component analysis (PCA) and parameter structuring. Notable examples include cluster adaptive training (CAT) [7], eigenvoice [8], and ensemble speaker and speaking environment modeling (ESSEM) [5, 6].

In this paper, we propose a new mapping function, linear projection (LP), to transform $\{\Lambda^{1}, \Lambda^{2}, \ldots \Lambda^{P}\}$ to Λ^{Y}. The proposed LP function can be seen as an extension of the linear regression (LR) function of MLLR and MAPLR with incorporation of multiple sets of acoustic models. Local information in the entire training space is taken into account, so the proposed LP function provides better environment modeling capability than the LR function. To estimate the nuisance parameters in LP, we derived the maximum likelihood-based LP (MLLP) and maximum a posteriori-based LP (MAPLP). In addition, we developed integrated prior (IP) densities for MAPLP. Experimental results on the Aurora-2 task [9] indicate that with both ML- and MAP-based estimates, the LP function can achieve better performance than other mapping functions, including LR, BF, LC, and LCB.

2. USING THE LINEAR PROJECTION FUNCTION FOR ENVIRONMENT MODELING

This section first reviews the ML- and MAP-based environment modeling criteria; then, we introduce MLLP and MAPLP and show the prior density estimation for MAPLP.

2.1. ML- and MAP-based Environment Modeling Criteria

Calculation of the nuisance parameters, φ, in F_{φ}, in Eq. (1), requires speech, Y, from the testing condition and transcription, W_c, corresponding to Y. We can use an ML-based objective function:

$$L(\mathrm{Y}, \varphi, \Omega, W_c) = log\,[P(\mathrm{Y}|\varphi, \Omega, W_c)], \quad (2)$$

to estimate the nuisance parameters, φ_{ML}, in F_{φ} by

$$\varphi_{ML} = \underset{\varphi}{argmax}\, L(\mathrm{Y}, \varphi, \Omega, W_c). \quad (3)$$

When using an MAP-based objective function, we have

$$M(\mathrm{Y}, \varphi, \Omega, W_c) = log\,[P(\mathrm{Y}|\varphi, \Omega, W_c)p(\varphi, \Omega, W_c)], \quad (4)$$

and we can calculate the parameters, φ_{MAP}, in F_{φ} by

$$\varphi_{MAP} = \underset{\varphi}{argmax}\, M(\mathrm{Y}, \varphi, \Omega, W_c). \quad (5)$$

With the estimated F_{φ}, we can estimate a new set of acoustic models that matches the testing condition. For the m-th Gaussian, we estimate its mean vector, μ_m^{Y}, in the final acoustic models by

$$\mu_m^{\mathrm{Y}} = \mathrm{F}_{\varphi}(\xi_m), \quad (6)$$

where ξ_m is an extended vector. ξ_m can be a single mean vector, $\xi_m = \{\mu_m^{\mathrm{X}}\}$, where μ_m^{X} is from Λ^{X}, or a pool of mean vectors, $\xi_m = \{\mu_m^{1}, \mu_m^{2} \ldots \mu_m^{P}\}$, where μ_m^{p} is the m-th mean vector in Λ^{p}. We designed the prior density for the m-th mean vector by

$$p\{\mathrm{F}_{\varphi}(\xi_m)\} \sim \left\{\prod_{i=1}^{D} \exp\left[-\frac{1}{2\,\mathrm{V}_{m_{(i)}}}\left(\mathrm{F}_{\varphi}(\xi_m)_{(i)} - \eta_{m_{(i)}}\right)^2\right]\right\}, \quad (7)$$

where V_m and η_m are hyper-parameters of the prior density, and D is the number of feature dimension.

2.2. MLLP and MAPLP

For both MLLP and MAPLP, we set $\xi_m = \{\mu_m^1, \mu_m^2 \ldots \mu_m^P\}$, and μ_m^{Y} in Eq. (6) becomes

$$\mu_m^{\mathrm{Y}} = \mathrm{F}_{\varphi}(\xi_m) = \mathrm{A}^1\mu_m^1 + \mathrm{A}^2\mu_m^2 + \cdots \mathrm{A}^P\mu_m^P + b, \quad (8)$$

where b is a correction bias. Based on the MAP criterion in Eq. (5), we calculate the parameters in $\{\mathrm{A}^1, \mathrm{A}^2, \ldots \mathrm{A}^P\}$ and b by

$$[\mathrm{A}^1\ \mathrm{A}^2 \ldots \mathrm{A}^P\ b]'_{(i)} = (G_{(i)})^{-1}k_{(i)}, \quad (9)$$

with

$$G_{(i)} = \sum_{t=1}^{T}\sum_{s\in W_c} r_s(t)\frac{1}{\Sigma_{s_{(i)}}}\rho_s\rho_s' + \sum_{s\in S}\frac{\epsilon_s}{\mathrm{V}_{s_{(i)}}}\rho_s\rho_s', \quad (10)$$

$$k_{(i)} = \sum_{t=1}^{T}\sum_{s\in W_c} r_s(t)\frac{1}{\Sigma_{s_{(i)}}}o_{t_{(i)}}\rho_s + \sum_{s\in S}\frac{\epsilon_s}{\mathrm{V}_{s_{(i)}}}\eta_{s_{(i)}}\rho_s, \quad (11)$$

where o_t is the t-th observation; $r_s(t)$ is the occupation probability; $\rho_s = [\mu_s^{1\prime}\mu_s^{2\prime} \ldots \mu_s^{P\prime}1]'$; μ_s^p is the s-th Gaussian of the p-th set of acoustic models; $\Sigma_{s_{(i)}}$ is the i-th element of the covariance matrix; S is the entire set of Gaussians; $s \in W_c$ indicates that the s-th Gaussian is in the transcription reference, W_c; ϵ_s is a control factor determining the weight of prior information. In Eqs. (9)-(11), we present the MAPLP derivation. The MLLP derivation can be obtained by setting $\epsilon_s = 0, \forall s \in S$ in Eqs. (9)-(11). To reduce the complexity, we can use a simpler form for each $\mathrm{A}^p, p = 1,2 \ldots P$. When using diagonal matrix A^p, i.e, $\mathrm{A}^p = \mathrm{diag}\left[a_{(1)}^p, a_{(2)}^p, \ldots a_{(D)}^p\right]$, the parameters of $[\mathrm{A}^1\ \mathrm{A}^2 \ldots \mathrm{A}^P\ b]$ is estimated by

$$[a_{(i)}^1\ a_{(i)}^2 \ldots\ a_{(i)}^P\ b_{(i)}]' = (G_{(i)})^{-1}k_{(i)}, \quad (12)$$

with

$$G_{(i)} = \sum_{t=1}^{T}\sum_{s\in W_c} r_s(t)\frac{1}{\Sigma_{s_{(i)}}}\rho_s\rho_s' + \sum_{s\in S}\frac{\epsilon_s}{\mathrm{V}_{s_{(i)}}}\rho_s\rho_s', \quad (13)$$

$$k_{(i)} = \sum_{t=1}^{T}\sum_{s\in W_c} r_s(t)\frac{1}{\Sigma_{s_{(i)}}}o_{t_{(i)}}\rho_s + \sum_{s\in S}\frac{\epsilon_s}{\mathrm{V}_{s_{(i)}}}\eta_{s_{(i)}}\rho_s, \quad (14)$$

where $\rho_s = [\mu_{s_{(i)}}^1\mu_{s_{(i)}}^2 \ldots \mu_{s_{(i)}}^P\ 1]'$.

2.3. Prior Density Estimation for MAPLP

We derived four ways to calculate hyper-parameters in the prior densities in Eq. (7), namely clustered priors (CP), sequential priors (SP), hierarchical priors (HP), and integrated priors (IP).

2.3.1 Clustered Priors (CP)

For the m-th mean vector, we first calculate K sets of mean vectors $\{\mathrm{F}_{\varphi}(\xi_m)^{(1)}, \mathrm{F}_{\varphi}(\xi_m)^{(2)} \ldots \mathrm{F}_{\varphi}(\xi_m)^{(K)}\}$ using K subsets of the entire training set. We can segment the entire training set into K subsets based on speakers' genders or accents, the signal to noise ratio (SNR), or in a data-driven manner. Then, we estimate the hyper-parameters of the CP density, $\{\eta_m^{CP}, \mathrm{V}_m^{CP}\}$, by

$$\eta_{m_{(i)}}^{CP} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{F}_{\varphi}(\xi_m)_{(i)}^{(k)}, \quad (15)$$

$$\mathrm{V}_{m_{(i)}}^{CP} = \frac{1}{K}\sum_{k=1}^{K}\left[\mathrm{F}_{\varphi}(\xi_m)_{(i)}^{(k)} - \eta_{m_{(i)}}^{CP}\right]^2, \quad (16)$$

where $\eta_{m_{(i)}}^{CP}$ and $\mathrm{V}_{m_{(i)}}^{CP}$ are the i-th components of η_m^{CP} and V_m^{CP}.

2.3.2 Sequential Priors (SP)

By using the SP densities, we incorporate the information seen before for estimating the current mapping function [10]. In this study, we sequentially update η_m^{SP} and use a fixed V_m^{SP} to simplify online computation. For the first utterance, we set $\epsilon_s = 0, \forall s \in S$ in Eqs. (9)-(11), or Eqs. (12)-(14) for the diagonal form of A^p, to calculate $\mathrm{F}_{\varphi}(\xi_m)$; the estimated $\mathrm{F}_{\varphi}(\xi_m)$ is used as the hyper-parameters for the next utterance. Then, for the u-th utterance, we prepare the hyper-parameters $\eta_m^{SP(u)}$ by

$$\eta_m^{SP(u)} = \mathrm{F}_{\varphi}(\xi_m)^{(u-1)}, \quad (17)$$

where $\mathrm{F}_{\varphi}(\xi_m)^{(u-1)}$ is estimated from the (u-1)-th utterance.

2.3.3 Hierarchical Priors (HP)

We prepare a tree structure for calculating HP. The estimate of HP densities resembles that is performed in structural MAPLR (SMAPLR) [11]. When computing the HP densities, we first estimate a mapping function at the top node of the tree structure. The estimated mean parameters are propagated to the child nodes and used as the HP density in the next layer. The estimation and propagation processes iterate and finally stop at the desired layer of the tree structure. For the m-th mean vector in the q-th node at the n-th level, its hyper-parameter, $\eta_m^{HP(n)}$, is calculated by

$$\eta_m^{HP(n)} = \mathrm{F}_{\varphi}(\xi_m)^{(n-1)}, \quad (18)$$

where $\mathrm{F}_{\varphi}(\xi_m)^{(n-1)}$ is from the parent node of the q-th node.

2.3.4 Integrated Priors (IP)

We derive the IP density that combines the above three prior densities using a function, $\Gamma(.)$:

$$\eta_m^{IP} = \Gamma(\eta_m^{CP}, \eta_m^{SP}, \eta_m^{HP}). \quad (19)$$

Here, we use a linear combination function for $\Gamma(.)$ to estimate η_m^{IP}:

$$\eta_m^{IP} = w^{CP}\eta_m^{CP} + w^{SP}\eta_m^{SP} + w^{HP}\eta_m^{HP}, \quad (20)$$

where w^{CP}, w^{SP}, and w^{HP} are weighting coefficients. We optimize the coefficients using a set of development data. Note that CP, SP, and HP are estimated using the information from the training set, seen information, and the current testing utterance with a tree structure, respectively. Therefore, the IP densities incorporate multiple knowledge sources. In this study, we only online estimate η_m^{IP} and use a fixed V_m^{IP} for the IP densities.

3. CORRELATIONS OF LP WITH OTHER WELL-KNOWN MAPPING FUNCTIONS

This section discusses the correlations of LP with several other well-known mapping functions. These functions can be classified into: (1) single model input; (2) multiple models input. In the following discussion, we present the MAP-based derivations. The ML-based counterparts can be obtained by setting $\epsilon_s = 0, \forall s \in S$.

3.1 Single Model Input

This category of approaches sets $\Omega = \{\Lambda^{\mathrm{X}}\}$. Here, we discuss linear regression (LR) and compensation bias (BC).

3.1.1 Linear Regression (LR)

When using LR, we set $\xi_m = \{\mu_m^{\mathrm{X}}\}$, and μ_m^{Y} in Eq. (6) becomes

$$\mu_m^{\mathrm{Y}} = \mathrm{F}_{\varphi}(\xi_m) = \mathrm{A}\mu_m^{\mathrm{X}} + b. \quad (21)$$

The parameters in A and b in Eq. (21) can be calculated by

$$[\mathrm{A}\ b]'_{(i)} = (G_{(i)})^{-1}k_{(i)}, \quad (22)$$

with

$$G_{(i)} = \sum_{t=1}^{T}\sum_{s\in W_c} r_s(t)\frac{1}{\Sigma_{s_{(i)}}}\rho_s\rho_s' + \sum_{s\in S}\frac{\epsilon_s}{\mathrm{V}_{s_{(i)}}}\rho_s\rho_s', \quad (23)$$

$$k_{(i)} = \sum_{t=1}^{T}\sum_{s\in W_c} r_s(t)\frac{1}{\Sigma_{s_{(i)}}}o_{t_{(i)}}\rho_s + \sum_{s\in S}\frac{\epsilon_s}{\mathrm{V}_{s_{(i)}}}\eta_{s_{(i)}}\rho_s, \quad (24)$$

where $\rho_s = [\mu_s^{\mathrm{X}\prime}\ 1]'$.

When we use a diagonal matrix A, $\mathrm{A} = \mathrm{diag}\left[a_{(1)}, a_{(2)}, \ldots a_{(D)}\right]$, in Eq. (21), we can solve A and b by

$$[a_{(i)}\ b_{(i)}]' = (G_{(i)})^{-1}k_{(i)}, \quad (25)$$

with

$$G_{(i)} = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) \frac{1}{\Sigma_{s_{(i)}}} \rho_s \rho_s' + \sum_{s \in S} \frac{\epsilon_s}{V_{s_{(i)}}} \rho_s \rho_s', \quad (26)$$

$$k_{(i)} = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) \frac{1}{\Sigma_{s_{(i)}}} o_{t_{(i)}} \rho_s + \sum_{s \in S} \frac{\epsilon_s}{V_{s_{(i)}}} \eta_{s_{(i)}} \rho_s, \quad (27)$$

where $\rho_s = [\mu^{X}_{s_{(i)}}\ 1]'$.

3.1.2 Bias Compensation (BC)

When using BC, we set $\xi_m = \{\mu^{X}_m\}$, and μ^{Y}_m in Eq. (6) becomes

$$\mu^{Y}_m = F_{\varphi}(\xi_m) = \mu^{X}_m + b. \quad (28)$$

The compensation bias, b, can be solved by

$$b_{(i)} = (G_{(i)})^{-1} k_{(i)}, \quad (29)$$

with

$$G_{(i)} = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) \frac{1}{\Sigma_{s_{(i)}}} + \sum_{s \in S} \frac{\epsilon_s}{V_{s_{(i)}}}, \quad (30)$$

$$k_{(i)} = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) \frac{(o_{t_{(i)}} - \rho_{s_{(i)}})}{\Sigma_{s_{(i)}}} + \sum_{s \in S} \frac{\epsilon_s (\eta_{s_{(i)}} - \rho_{s_{(i)}})}{V_{s_{(i)}}}, \quad (31)$$

where $\rho_{s_{(i)}} = \mu^{X}_{s_{(i)}}$.

3.2 Multiple Models Input

For this category of approaches, we use $\Omega = \{\Lambda^1, \Lambda^2, \ldots \Lambda^P\}$. Here, we discuss the correlations of LP with linear combination with a correction bias (LCB), linear combination (LC), and best first (BF) mapping functions. Compared to LP, these mapping functions use simpler forms of matrices, $\{A^1, A^2, \ldots A^P\}$, in Eq. (8).

3.2.1 Linear Combination with a Correction Bias (LCB)

For LCB, we set $\xi_m = \{\mu^1_m, \mu^2_m \ldots \mu^P_m\}$, and μ^{Y}_m in Eq. (6) becomes

$$\mu^{Y}_m = F_{\varphi}(\xi_m) = A^1 \mu^1_m + A^2 \mu^2_m + \cdots A^P \mu^P_m + b, \quad (32)$$

where μ^p_m is the m-th mean in Λ^p. $A^p = \mathrm{diag}[\omega^p, \omega^p, \ldots \omega^p]$, $p = 1,2 \ldots P$. We estimate the nuisance parameters in Eq. (32) by

$$[\omega^1\ \omega^2 \ldots \omega^P\ b']' = G^{-1} k\,, \quad (33)$$

with

$$G = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t)\ H_s' \Sigma_s^{-1} H_s + \sum_{s \in S} \epsilon_s\ H_s' V_s^{-1} H_s, \quad (34)$$

$$k = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) H_s' \Sigma_s^{-1} o_t + \sum_{s \in S} \epsilon_s\ H_s' V_s^{-1} \eta_s, \quad (35)$$

where $H_s = [\mu^1_s\ \mu^2_s \ldots \mu^P_s\ I]$. I is a $D \times D$ identity matrix.

3.2.2 Linear Combination (LC)

For LC, we set $\xi_m = \{\mu^1_m, \mu^2_m \ldots \mu^P_m\}$, and μ^{Y}_m in Eq. (6) becomes

$$\mu^{Y}_m = F_{\varphi}(\xi_m) = A^1 \mu^1_m + A^2 \mu^2_m + \cdots A^P \mu^P_m, \quad (36)$$

where $A^p = \mathrm{diag}[\omega^p, \omega^p, \ldots \omega^p]$, $p = 1,2 \ldots P$. We estimate the nuisance parameters in Eq. (36) by

$$[\omega^1\ \omega^2\ \ldots \omega^P]' = G^{-1} k, \quad (37)$$

with

$$G = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) H_s' \Sigma_s^{-1} H_s + \sum_{s \in S} \epsilon_s\ H_s' V_s^{-1} H_s, \quad (38)$$

$$k = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) H_s' \Sigma_s^{-1} o_t + \sum_{s \in S} \epsilon_s\ H_s' V_s^{-1} \eta_s, \quad (39)$$

where $H_s = [\mu^1_s\ \mu^2_s \ldots \mu^P_s]$.

3.2.3 Best First (BF)

The BF function can be considered as a hard-decision version of LC. For BF, we set $\xi_m = \{\mu^1_m, \mu^2_m \ldots \mu^P_m\}$, and μ^{Y}_m becomes

$$\mu^{Y}_m = F_{\varphi}(\xi_m) = A^1 \mu^1_m + A^2 \mu^2_m + \cdots A^P \mu^P_m, \quad (40)$$

where the l-th matrix, A^l, is an identity matrix, I ($A^l = I$); all the other matrices are zero matrices, $\emptyset$, ($A^p = \emptyset, \forall p \neq l$). We search for the P sets of acoustic models to find l by

$$l = \underset{p}{argmin} \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) \left[(o_t - \mu^p_s)' \Sigma_s^{-1} (o_t - \mu^p_s) \right], \quad p = 1, 2, \ldots P. \quad (41)$$

4. EXPERIMENTS

This section presents the experimental setup and discusses the experimental results. The proposed ML- and MAP-based LP and other mapping functions were evaluated on the Aurora-2 database. For the MAP estimates, we adopted the IP densities in this study.

4.1. Experimental Setup

We used the multi-condition training set to prepare acoustic models. This training set includes 8440 utterances from four types of noise, at 5- to 20-dB SNRs, and a clean condition. We clustered the training set by speakers' genders and accordingly obtained female and male training subsets, with each subset containing 4220 utterances. With the entire training set and two gender dependent (GD) subsets, we prepared a gender independent (GI) and two GD acoustic models. We followed the complex back-end hidden Markov model (HMM) topology presented in [12] to train the three sets of acoustic models. Each digit was modeled with 20 mixtures per state, and the silence and short pauses were modeled with 36 mixtures per state. We used a modified European Telecommunications Standards Institute (ETSI) advanced front-end (AFE) for feature extraction [12]. Every feature vector comprised 13 static plus their first- and second-order time derivatives.

We obtained results for 50 testing conditions (10 noise types, 0- to 20-dB SNR) from the Aurora-2 test set; each condition had 1,001 utterances. The 50 conditions were divided into SetA, SetB, and SetC. SetA included the same four types of noise as those in the multi-condition training set, SetB contained four unseen types of noise, and SetC had an additional channel distortion. Word error rate (WER) was used to evaluate the ASR performance. All the results reported in this paper, except for the baseline, were tested in a per-utterance unsupervised model compensation mode.

To enhance the accuracy of the environment modeling, we built a tree structure to cluster mean parameters in the acoustic models. The tree was constructed based on the GI acoustic models and consisted of one root, three intermediate, and six leaf nodes. For both ML and MAP estimates, we used the tree to determine the number of mapping functions. For the q-th node in the tree, we estimate its accumulated statistics, $R_q = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t), \forall s \in q$. If R_q is larger than a predefined threshold, we use the mapping function for the q-th node. If not, we check the accumulated statistics at the parent node. The process repeats until we find a node with sufficient statistics. For a fair comparison, a same set of transcription reference, generated by the GI acoustic models, was used to calculate nuisance parameters. Therefore, for each utterance, the number of mapping functions used for model compensation was the same across different types of mapping functions. The differences in performance simply represent the environment modeling capabilities among them.

Table I lists the amount of nuisance parameters in the mapping functions that we tested in the experiments. D is the number of feature dimension, and P is the number of acoustic models. In our experiments, we used a diagonal matrix for A in LR and each A^p in $\{A^1, A^2 \ldots A^P\}$ in LP and set D=39 and P=3. Note that Table I only presents the complexity of a single mapping function. The number of mapping functions used for a particular testing utterance is determined by the tree structure and the accumulated statistics, R_q.

TABLE I. COMPLEXITY OF EACH MAPPING FUNCTION

Function	LR	BC	LCB	LC	LP
Complexity	$D+D$	D	$P+D$	P	$D\times P+D$

4.2. Experimental Results

This section presents our experimental results. For each experiment set, we present the results for SetA, SetB, SetC, and All test conditions, where All is the average WER over 50 testing results.

4.2.1 Baseline and BF

Table II lists the baseline results. We used the GI HMM set to decode testing utterances without performing model compensation to obtain the results. The BF results are also listed in Table II as another baseline. Note that the BF results were not from a parallel decoding procedure through the three sets of HMMs (one GI and two GDs) but through the BF function as presented in Eq. (41).

TABLE II. AVERAGE WER (%) OF BASELINE AND BF

Test Condition	SetA	SetB	SetC	All
Baseline	5.92	6.69	7.11	6.46
BF	5.68	6.48	6.90	6.24

4.2.2 ML-based LP versus BC and LR

Table III presents the results of the LR, BC, and LP mapping functions, where the nuisance parameters were estimated based on the ML criterion. Note that both LR and BC take a single set of HMMs. Therefore, we can obtain three sets of testing results for both MLLR and MLBC by using the three HMM sets (one GI and two GD HMMs). In Table III, we only present the best MLLR and MLBC performances from their individual three sets of results.

From Table III, we can observe that MLLP achieves better performance than MLLR and MLBC for almost all the test conditions. This set of results verifies that incorporating local information in the training space can enhance the environment modeling capacity. With a further investigation, we found that MLLP sometimes generates poor results due to over-fittings. We handled this over-fitting issue by using the MAP-based estimates. The experimental results are presented in the next sub-section.

TABLE III. AVERAGE WER (%) OF ML-BASED ESTIMATE

Test Condition	SetA	SetB	SetC	All
MLLR	5.65	6.20	6.33	6.01
MLBC	5.89	6.10	6.62	6.12
MLLP	5.58	6.14	6.32	5.95

4.2.3 MAP-based LP versus LCB, LC, and LR

Table IV lists the MAP-based estimates for LCB, LC, and LP. First, we observe that MAPLCB clearly performs better than MAPLC, and MAPLP outperforms MAPLCB. Second, by comparing Tables III and IV, we can see that by using the MAP-based estimate, LP can be clearly enhanced. This confirms that a MAP-based estimate provides better result when the amount of statistics is limited.

TABLE IV. AVERAGE WER (%) OF MAP-BASED ESTIMATE

Test Condition	SetA	SetB	SetC	All
MAPLCB	5.46	6.22	5.93	5.86
MAPLC	5.51	6.44	6.55	6.09
MAPLP	5.34	6.07	5.94	5.75

Table V lists the MAP-based estimates for LR, LCB, and LP. For MAP-based LR, we followed the derivations of Eqs. (21)-(27); this implementation is slightly different to the conventional MAPLR [3]. Therefore, we name MAP-based LR here in our experiment MAPLR with mean prior (MAPLR-MP). From Table V, we can observe that MAPLP consistently outperforms MAPLR-MP and MAPLCB. Based on an additional set of hypothesis testing results (matched pair t-Test [13]), we confirm that the improvements of MAPLP over both MAPLR-MP and MAPLCB are significant with P-values smaller than 0.05 under SNR=0 and 10 dB conditions.

TABLE V. AVERAGE WER (%) OF MAP-BASED LR, LCB, AND LP

SNR (dB)	0	5	10	15	20	All
MAPLR-MP	19.33	6.26	2.48	1.06	0.65	5.95
MAPLCB	19.20	6.11	2.36	1.00	0.62	5.86
MAPLP	18.76	6.03	2.35	1.00	0.61	5.75

5. SUMMARY

We proposed MLLP and MAPLP for environment modeling for robust ASR. For MAPLP, we derived the IP densities that include multiple knowledge sources. We also discussed the correlation between LP with other famous mapping functions and compared their performances. For a fair comparison, a same transcription reference and a same number of mapping functions were used for model compensation for each testing utterance. Experimental results on Aurora-2 indicated that for both ML- and MAP-based estimates, LP outperforms other mapping functions. Of these approaches, MAPLP gives the best performance with a significant 10.99% (6.46% to 5.75%) WER reduction over the baseline result.

We believe that the investigated approaches can be combined to further enhance model compensation capability. For example, LC or BF could be used as the first stage to generate a good initial HMM set, and then LR or BC as the second stage can further refine the generated HMMs. In addition, we could use LR or BC to adapt the multiple HMM sets and then apply a second stage LP or LCB to get better performance. Testing and comparing different possible combinations of these approaches will be our future work.

6. REFERENCES

[1] A. Sankar and C.-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. on Speech Audio Proc.*, vol. 4, pp. 190–202, 1996.

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

[3] C. Chesta, O. Siohan, and C.-H. Lee, "Maximum a posteriori linear regression for hidden Markov model adaptation," in *Proc. Eurospeech*, pp. 211–214, 1999.

[4] 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 Lang.*, vol. 23, pp. 389–405, 2009.

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

[6] Y. Tsao, J. Li, and C.-H. Lee, "Ensemble speaker and speaking environment modeling approach with advanced online estimation process," in *Proc. ICASSP*, pp. 3833–3836, 2009.

[7] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Trans. Speech Audio Proc.*, vol. 8, pp. 417–428, 2000.

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

[9] D. Macho, L. Mauuary, B. Noe, Y. M. Cheng, D. Ealey, D. Jouver, H. Kelleher, D. Pearce, and F. Saadoun, "Evaluation of a noise-robust DSR front-end on Aurora databases," in *Proc. ICSLP*, pp. 17-20, 2002.

[10] Q. Huo and C.-H. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate," *IEEE Trans. Speech Audio Proc.*, vol. 5, pp. 161–172, 1997.

[11] O. Siohan, T. A. Myrvoll, and C.-H. Lee, "Structural maximum a posteriori linear regression for fast HMM adaptation," *Computer Speech and Lang.*, vol. 16, pp. 5–24, 2002.

[12] J. Wu and Q. Huo, "Several HKU approaches for robust speech recognition and their evaluation on Aurora connected digit recognition tasks," in *Proc. Eurospeech*, pp. 21–24, 2003.

[13] A. J. Hayter, Probability and Density for Engineers and Scientists, Duxbury Press; 3rd edition, 2006.