MAP Estimation of Online Mapping Parameters in Ensemble Speaker and Speaking Environment Modeling

Yu Tsao [1], Shigeki Matsuda[1], Satoshi Nakamura[1], and Chin-Hui Lee[2]

[1] *Spoken Language Communication Group, National Institute of Information and Communications Technology, Kyoto, 619-0288, Japan*

[1] {yu.tsao, shigeki.matsuda, satoshi.nakamura}@nict.go.jp

[2] *School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332-0250, USA*

[2] chl@ece.gatech.edu

***Abstract*—Recently, an ensemble speaker and speaking environment modeling (ESSEM) framework was proposed to enhance automatic speech recognition performance under adverse conditions. In the online phase of ESSEM, the prepared environment structure in the offline stage is transformed to a set of acoustic models for the target testing environment by using a mapping function. In the original ESSEM framework, the mapping function parameters are estimated based on a maximum likelihood (ML) criterion. In this study, we propose to use a maximum a posteriori (MAP) criterion to calculate the mapping function to avoid a possible over-fitting problem that can degrade the accuracy of environment characterization. For the MAP estimation, we also study two types of prior densities, namely, clustered prior and hierarchical prior, in this paper. On the Aurora-2 task using either type of prior densities, MAP-based ESSEM can achieve better performance than ML-based ESSEM, especially under low SNR conditions. When comparing to our best baseline results, the MAP-based ESSEM achieves a 14.97% (5.41% to 4.60%) word error rate reduction in average at a signal to noise ratio of 0dB to 20dB over the three testing sets.**

I. INTRODUCTION

After decades of endeavor, the performance of automatic speech recognition (ASR) has been significantly improved [1]. However, an issue that has been around for a long time now still persists and limits the current applicability of ASR–its performance is considerably degraded when the training and testing conditions are mismatched. The difficulty with the handling of this issue is the fact that real-world distortion usually comes from an unknown combination of multiple distortion sources, including speaker variability and speaking environment distortions. Many approaches to reduce this acoustic mismatch have been proposed. Among them, a category of approaches calculates a new set of hidden Markov models (HMMs) that matches the testing condition. Effective examples include stochastic matching (SM) [2], maximum likelihood linear regression (MLLR) [3], and eigenvoice [4].

Recently, extensions of the abovementioned approaches have been proposed by utilizing the environmental knowledge. This environmental knowledge is obtained from a wide range of different acoustic conditions. These extensions can be summarized into two categories. For the first category, the environmental knowledge is used to prepare the prior density for the maximum a posteriori (MAP) estimation. The MAP criterion is known to show more stable performance with respect to estimating acoustic models or mapping functions compared to the maximum likelihood (ML) criterion, especially when only limited amount of adaptation statistics is available [5, 6]. Successful examples for this category include maximum a posteriori linear regression (MAPLR) [7] and MAP-based eigenvoice [8]. In the case of the second category, the environmental knowledge is used to build an environment structure. During the course of testing, a set or a group of acoustic models that are closest to the testing condition is selected for recognition or used as an initial set for another transformation to better match the testing condition. Examples include environmental sniffing [9] and piecewise-linear transformation (PLT) [10]. Another effective example belonging to this category is a recent ensemble speaker and speaking environment modeling (ESSEM) framework [11, 12].

In ESSEM, a complex environment structure is prepared in the offline phase. In the online phase, the most relevant sub-space in the environment structure is first selected. Thereafter, ESSEM estimates a mapping function to transform the selected sub-space to generate a set of acoustic models for the unknown testing condition [11, 12]. For the conventional ESSEM, the mapping function is estimated based on the ML criterion. In this study, we adopt the MAP criterion to further improve the conventional ESSEM framework. In order to perform the MAP estimation, we study two types of prior densities and compare their performance in this paper.

The paper is organized as follows. In Section II, we briefly review the ESSEM framework and its two extensions. In Section III, we introduce the MAP-based ESSEM algorithm and two types of prior densities. Then in Section IV, we present our experimental setup and results. Finally in Section V, we summarize our findings and discuss our future work.

II. REVIEW OF THE ESSEM FRAMEWORK

In this section, we briefly review the ESSEM framework and its two extensions, namely, environment clustering and environment partitioning algorithms.

A. Fundamentals of ESSEM

The ESSEM method is extended from the SM algorithm [2]. The goal of ESSEM is to estimate a mapping function to construct a set of HMMs that matches the testing condition.

The ESSEM process comprises two phases, namely offline and online phases. In the offline phase, we collect or artificially simulate speech data from a wide range of different speaker and speaking environments. With speech data from P different speaker and speaking environments, we can train P sets of HMMs. Each set of HMMs characterizes a particulate speaking and speaker environment. For ease of modeling, we concatenate the entire set of mean parameters within a set of HMMs into a super-vector, V_p, p=1,2,…,P. Thereafter, we collect these P super-vectors to construct an ensemble speaker and speaking structure, namely ESS space, $\Omega_{\mathbf{V}} = \{V_1 V_2 \ldots V_P\}$.

In the online phase, we estimate the target super-vector, V_Y, for the testing environment through a mapping function, $\mathbf{G}_\varphi$:

$$V_Y = \mathbf{G}_\varphi(\Omega_V) . \quad (1)$$

The form of $\mathbf{G}_\varphi$ depends on the amount of adaptation data and distortion types. In the previous study [11, 12], we estimate the mapping parameters $\hat{\varphi}$ in $\mathbf{G}_\varphi$ based on the ML criterion:

$$\hat{\varphi} = \underset{\varphi}{argmax}\ P(F_Y \mid \Omega_{\mathbf{V}}, \varphi, W) , \quad (2)$$

where W is the transcription corresponding to the testing utterances, F_Y. With the estimated target super-vector, V_Y, we can build the set of acoustic models for the testing condition.

B. Establishment of the ESSEM Environment Structure

To build a well-constructed environment structure for the ESSEM framework, we developed environment clustering (EC) and environment partitioning (EP) algorithms [12].

1) Environment Clustering (EC): The EC algorithm enables the establishment of several EC spaces from the ESS space in the offline phase. In our ESSEM framework, we utilize a hierarchical tree to facilitate the EC process. With the EC tree, we collect super-vectors belonging to the same node to form an EC space. For a hierarchical tree with a total of C nodes, we categorize the entire set of ESS space into C EC spaces: $\Omega_{\mathbf{V}} = \{\Omega_{\mathbf{V}^{(1)}} \cup \Omega_{\mathbf{V}^{(2)}} \ldots \cup \Omega_{\mathbf{V}^{(C)}}\}$. For each EC space, we specify a representative super-vector, $V_{\text{rep}}^{(c)}$, c=1,2,…C. During testing, these C representative super-vectors are used to select the best matched EC space corresponding to the testing condition. Thereafter, the selected EC space is taken to estimate the target super-vector by a mapping function as shown in Eq-(1).

2) Environment Partitioning (EP): The objective of EP is to use multiple transformations of Eq-(1) instead of a global one to better characterize testing conditions. In our previous study, we presented two types of partitioning; in this study, we discuss the mixture-based EP [12]. For the mixture-based EP, we use an EP tree to cluster mean vectors, similar to that used in SMAP [13]. Based on the clustering result, we partition each super-vector into S sub-vectors ($V_p = [V_{p,1}^T, V_{p,2}^T, \ldots, V_{p,S}^T]^T$, p=1, 2,…P). From the entire set of P environments, we collect these sub-vectors, and finally, build S sets of EP sub-spaces, $\Omega_{\mathbf{V}_s} = \{V_{1,s} V_{2,s} \ldots V_{P,s}\}$, s=1,2,…S. More details on the EC and EP algorithms have been provided in our previous study [12].

3) A Combination Structure: Figure 1 illustrates the overall system that combines the EC and EP algorithms. We first establish an EC tree to cluster the training environments into C clusters. Each EC node has another EP tree structure, where each EP tree is established based on the representative super-vector of the EC node. During testing, we first select an EC node from the EC tree; then, we identify the appropriate EP sub-space by the EP tree and perform ESSEM transformation on the selected EP sub-space [12]. By using a more complex environment structure of Fig. 1, ESSEM can better model testing conditions. However, the increased free parameters can cause an over-fitting issue when the adaptation data is limited. In this study, we handle this issue by using the MAP criterion.

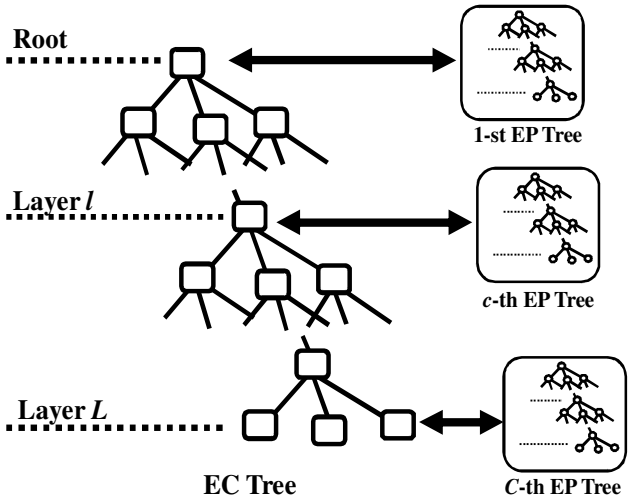

Fig. 1. Environment clustering and partitioning on ESSEM

III. MAP-BASED ESSEM ALGORITHM

In this section, we first present the MAP-based ESSEM framework. Next, we discuss two types of prior densities, namely clustered prior and hierarchical prior densities.

A. MAP Criterion on ESSEM

From Eq-(2), we derive the following equation to apply the MAP criterion into the conventional ESSEM framework:

$$\{\hat{\Omega}_{\mathbf{V}}, \hat{\varphi}\} = \underset{\Omega_{\mathbf{V}}, \varphi}{argmax}\ P(F_Y \mid \Omega_{\mathbf{V}}, \varphi, W)\, P(\Omega_{\mathbf{V}}, \varphi) , \quad (3)$$

where $P(\Omega_{\mathbf{V}}, \varphi)$ is the joint prior distribution of environment structure and mapping parameters. Similar to our previous study [14], we can use a two-stage optimization procedure to solve Eq-(3). First, we estimate the optimal set of mapping parameters by using the following equation:

$$\hat{\varphi} = \underset{\varphi}{argmax}\ P(F_Y \mid \Omega_{\mathbf{V}}, \varphi, W)\, P(\Omega_{\mathbf{V}}, \varphi) . \quad (4)$$

Thereafter, we put the calculated mapping parameters into Eq-(3) and adjust the ESSEM environment structure by:

$$\hat{\Omega}_{\mathbf{V}} = \underset{\Omega_{\mathbf{V}}}{argmax}\ P(F_Y \mid \Omega_{\mathbf{V}}, \varphi, W)\, P(\Omega_{\mathbf{V}}, \varphi) . \quad (5)$$

Following several iterations of Eq-(4) and Eq-(5), we can obtain the optimal solution for Eq-(3). In this paper, we limit our discussion to optimizing the mapping parameters (Eq-(4)). Therefore, the objective of applying MAP in ESSEM now is to increase the reliability of mapping function estimation, especially when adaptation data is limited and the ESSEM

transformation is too complex. For the joint prior distribution, $P(\Omega_{\mathbf{V}}, \varphi)$, we attempt to choose a conjugate prior for the likelihood function. Moreover, we intend to use a function that any mapping structure can be plugged in. Because the likelihood for individual mean vector (Gaussian component) is modeled by a multivariate Normal distribution, we set the prior density to be a multivariate Normal distribution.

To apply the abovementioned MAP estimation into the ESSEM framework (Fig. 1), we take the following steps. First, we located an EC node from the EC tree. Next, we decided the optimal layer of EP and estimated a mapping function on the EP sub-space to generate a sub-vector, $\mathrm{V}_{\mathrm{Y},s}$:

$$\mathrm{V}_{\mathrm{Y},s} = \mathbf{G}_{\varphi_s}(\Omega_{\mathrm{V},s}) \ , s = 1,2,..S. \tag{6}$$

Thereafter, we obtained the target super-vector for the testing condition by using a collection of these sub-vectors:

$$\mathrm{V}_{\mathrm{Y}} = [\ \mathrm{V}_{\mathrm{Y},1}^T,\ \mathrm{V}_{\mathrm{Y},2}^T, \ldots, \mathrm{V}_{\mathrm{Y},S}^T\ ]^T. \tag{7}$$

In the conventional ESSEM we solve Eq-(6) by:

$$\hat{\varphi}_s = \underset{\varphi_s}{argmax}\ P(F_{\mathrm{Y}} \mid \Omega_{\mathbf{V},s}, \varphi_s, W). \tag{8}$$

On the other hand, when using MAP to solve Eq-(6), we have:

$$\hat{\varphi}_s = \underset{\varphi_s}{argmax}\ P(F_{\mathrm{Y}} \mid \Omega_{\mathbf{V},s}, \varphi_s, W)\, P(\Omega_{\mathbf{V},s}, \varphi_s). \tag{9}$$

B. MAP Estimation on ESSEM Mapping Functions

In this study, we discuss the MAP estimation on two types of ESSEM mapping functions—linear combination (LC) and linear combination with a correction bias (LCB) [11]. First, when using LC as the ESSEM mapping function, we arrange the set of mapping parameters (φ in Eq-(1)) to be a vector $\boldsymbol{\theta}=\{\mathbf{w}\}$, where $\mathbf{w}$ is a P dimensional vector of weighting coefficients. For the m-th mean vector (of the m-th Gaussian), we build a structure of $\boldsymbol{\Gamma}_m = \{\boldsymbol{\mu}_m^1, \boldsymbol{\mu}_m^2, ..\boldsymbol{\mu}_m^P\}$, where $\boldsymbol{\mu}_m^p$ is the m-th mean vector for the p-th speaker and speaking environment.

When using LCB as the mapping function, we again arrange the parameters to a vector $\boldsymbol{\theta}=\{\mathbf{w}^T, \mathbf{b}^T\}^T$. Similarly, $\mathbf{w}$ is a weighting coefficient vector, and $\mathbf{b}$ is a D dimensional correction bias vector (each feature vector has D components). Now for the m-th mean vector, we build an environment structure $\boldsymbol{\Gamma}_m = \{\boldsymbol{\mu}_m^1, \boldsymbol{\mu}_m^2, ..\boldsymbol{\mu}_m^P, \mathbf{I}\}$, where $\mathbf{I}$ is a $D \times D$ identity matrix.

As presented in the previous section, the mapping parameter set $\boldsymbol{\theta}$ can be shared by a selected group of parameters, (for example: ESS space and EP sub-space). When the selected group includes M mean vectors, we build a structure of $\boldsymbol{\Gamma}=\{\boldsymbol{\Gamma}_1\ \boldsymbol{\Gamma}_2 \ldots \boldsymbol{\Gamma}_M\}$. With this environment structure $\boldsymbol{\Gamma}$ and mapping parameters $\boldsymbol{\theta}$, we can calculate the m-th mean vector for the testing environment $\boldsymbol{\mu}_m^{\mathrm{Y}}$ by:

$$\boldsymbol{\mu}_m^{\mathrm{Y}} = \boldsymbol{\Gamma}_m \boldsymbol{\theta}. \tag{10}$$

When solving $\boldsymbol{\theta}$ by ML, we use the following equation:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{argmax}\ P(F_{\mathrm{Y}} \mid \boldsymbol{\Gamma}, \boldsymbol{\theta}). \tag{11}$$

Based on the EM algorithm [15], we can estimate $\boldsymbol{\theta}$ by:

$$\boldsymbol{\theta} = \boldsymbol{G}^{-1}\boldsymbol{k}, \tag{12}$$

with

$$\boldsymbol{G} = \sum_{t=1}^{N}\sum_{m=1}^{M} r_m(t)\, \boldsymbol{\Gamma}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{\Gamma}_m, \tag{13}$$

$$\boldsymbol{k} = \sum_{t=1}^{N}\sum_{m=1}^{M} r_m(t)\, \boldsymbol{\Gamma}_m^T \boldsymbol{\Sigma}_m^{-1} \mathbf{y}_t, \tag{14}$$

where $\mathbf{y}_t$ is the t-th feature vector, N is the total number of feature frames, and $r_m(t)$ is the posterior probability of m-th Gaussian on the t-th frame. $\boldsymbol{\Sigma}_m$ is the covariance for the m-th Gaussian; we can obtain it from the representative HMM sets.

When solving $\boldsymbol{\theta}$ by MAP, we use the following equation:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{argmax} P(F_{\mathrm{Y}} \mid \boldsymbol{\Gamma}, \boldsymbol{\theta})\, P(\boldsymbol{\Gamma}, \boldsymbol{\theta}). \tag{15}$$

As mentioned in the previous section, we choose the multivariate Normal distribution and define the prior density for the m-th Gaussian by the following function:

$$p(\boldsymbol{\Gamma}_m, \boldsymbol{\theta}) \propto \exp\{-\frac{1}{2}(\boldsymbol{\Gamma}_m\boldsymbol{\theta} - \boldsymbol{\eta}_m)^T \boldsymbol{\Psi}_m^{-1}(\boldsymbol{\Gamma}_m\boldsymbol{\theta} - \boldsymbol{\eta}_m)\ \}, \tag{16}$$

where $\boldsymbol{\eta}_m$ and $\boldsymbol{\Psi}_m^{-1}$ are the hyperparameters of the prior density. In the next section, we will present two procedures to prepare these hyperparameters. From Eq-(15) and Eq-(16), we can estimate the mapping parameter set $\boldsymbol{\theta}$ by using the same equation of Eq-(12), i.e., $\boldsymbol{\theta} = \boldsymbol{G}^{-1}\boldsymbol{k}$, but now we have:

$$\boldsymbol{G} = \sum_{t=1}^{N}\sum_{m=1}^{M} r_m(t)\, \boldsymbol{\Gamma}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{\Gamma}_m + \sum_{t=1}^{N}\sum_{m=1}^{M} r_m(t)\, \boldsymbol{\Gamma}_m^T \boldsymbol{\Psi}_m^{-1} \boldsymbol{\Gamma}_m, \tag{17}$$

$$\boldsymbol{k} = \sum_{t=1}^{N}\sum_{m=1}^{M} r_m(t)\, \boldsymbol{\Gamma}_m^T \boldsymbol{\Sigma}_m^{-1} \mathbf{y}_t + \sum_{t=1}^{N}\sum_{m=1}^{M} r_m(t)\, \boldsymbol{\Gamma}_m^T \boldsymbol{\Psi}_m^{-1} \boldsymbol{\eta}_m. \tag{18}$$

C. Two Types of Prior Densities

In this section, we discuss two types of prior densities—clustered prior (CP) and hierarchical prior (HP) densities—for MAP estimation of mapping parameters in ESSEM.

1) Clustered Prior (CP) Densities: We specify a CP density for each EP node in the ESSEM environment structure (Eq-(9)). The preparation process of each CP density is very similar to that is used in MAPLR [7, 12]. In the offline, the entire set of training data is used to calculate hyperparameters of the CP densities. By selecting the best matched EP sub-space in the online phase, we also obtain the most relevant CP density for MAP. Since each CP density corresponds to a certain cluster of parameters (in the EP tree) for a particular group of environments (in the EC tree), it provides important local information for environment modeling. Therefore by using the most relevant CP density, MAP-based ESSEM can estimate the online mapping parameters more accurately.

2) Hierarchical Prior (HP) Densities: The process of calculating HP densities for MAP-based ESSEM resembles that is performed in structural MAP (SMAP) [13] and structural maximum a posteriori linear regression (SMAPLR) [16]. The EP trees in the ESSEM environment structure (Fig. 1) are used to compute the HP densities. In the online, the mapping parameters for the top node of the EP tree are first estimated (no prior density for that top level). Thereafter, the estimated mapping parameters are propagated to the next layer to form the prior density. The estimation and propagation processes iterate and finally stop at the desired level of the EP tree; then the super-vector for the testing condition is obtained. Different to the CP densities, we only need to prepare the EP tree structures for the ESSEM framework, and the preparation of the HP densities is not required in the offline phase.

IV. EXPERIMENTS

This section introduces our experimental setup and presents the MAP-based ESSEM performance using two different forms of mapping functions along with CP and HP densities.

A. Experimental Setup

We evaluated the proposed method on the Aurora-2 database [17]. The multi-condition training set was used to obtain environment-specific HMMs and to build the environment structure. The training set contains 8440 speech utterances from individuals of both genders under 17 different speaking environments. Accordingly, the training data was classified into 34 different acoustic conditions. We adopted a modified ETSI advanced front-end (AFE) to perform feature extraction [18] and used a complex back-end topology [17] to train 34 sets of speaker and speaking environment-specific HMMs. Thus, we obtained 34 super-vectors from the 34 sets of HMMs. We further increased the discriminations within each and between each pair of these super-vectors by soft margin estimation (SME) [19] and minimum classification error (MCE) methods [20]. We have included additional details on the experimental setup in our previous study [11].

We tested performance in a per-utterance unsupervised compensation mode on a gender dependent system. For this gender dependent system, we prepared an automatic gender identification (AGI) unit to determine each testing speaker's gender identity [11]. In this paper, we report the results of 50 testing conditions (including three testing sets, ten different noise types, 0dB to 20dB SNR) from the Aurora-2 test set.

B. EC and EP Structures

For the EC algorithm, we constructed a two-layer binary tree structure to cluster the 34 environments into seven groups (one root, two intermediate, and four leaf nodes). In the first layer, the 34 environments were divided into two groups of two genders. Each node of the second layer was then classified roughly according to high/low SNR levels [12].

Next, we built a two-layer EP tree for each EC node. Each EP tree consisted of one root, three intermediate and six leaf nodes. Accordingly, we prepared ten EP sub-spaces for each EC node. In the online, we performed a search procedure on the EP tree to find a layer with sufficient adaptation statistics. Therefore, the total number of EP sub-spaces used for ESSEM transformation was not predetermined but rather was decided with respect to the amount of adaptation data [12].

C. Experimental Results

In the following experiments, the average word error rate (WER) is used as the evaluation measure. First, we list the baseline results as "Baseline" in Table I. To obtain this set of results, we located one cluster from the EC tree and used the corresponding representative HMM set to decode the testing utterance [11]. In addition to the baseline, Table I also lists ML-based ESSEM (as "ML") and MAP-based ESSEM (as "MAP(HP)") results using LC as the mapping function. In this set of experiments, the HP densities are used for the MAP estimation. The results of using CP densities show similar properties, so we did not include them here in this paper.

From Table I, we first observe that ML-based ESSEM already provides significant improvements over the baseline system. Next, by comparing "ML" and "MAP(HP)" results in Table I, it can be observed that MAP-based ESSEM gives further improvements over the ML-based ESSEM in average over 50 testing conditions (WER is reduced from 4.81% to 4.77%). To further investigate this improvement, we used a matched pair t-Test for hypothesis testing [21]. In this test, under hypothesis H_0, it is considered that "method-II is not better than method-I" while under hypothesis H_1, it is considered that "method-II is better than method-I". Here, we consider MAP-based ESSEM as method-II and ML-based ESSEM as method-I. Since each SNR condition has ten results, we used ten pair-wised samples for t-Test. Instead of presumptuously setting a threshold to determine the results, we list P-values of the matched t-Test in Table I.

For the matched pair t-Test, a smaller P-value indicates a more significant improvement. From Table I, we can see that the P-values are smaller for 0dB and 5dB SNR comparing to other SNR conditions. Especially, it is noted that for 10dB SNR, although "MAP(HP)" achieves lower WER than "ML", the P-value is relatively large (0.372). The reason for this result is that the improvements are not consistent across the ten testings. For 0dB and 5 dB SNR conditions, on the other hand, "MAP(HP)" gives consistent improvements over "ML", and therefore the corresponding P-values are relatively small.

In summary, the results from Table I indicate that more prominent improvements can be provided by MAP-based ESSEM under low SNR conditions (0dB and 5dB SNR); it is intuitively reasonable since the accuracy of the ESSEM mapping function is more critical for lower SNR conditions.

TABLE I
AVERAGE WORD ERROR RATE (%) AND P-VALUE FOR EACH SNR CONDITION

dB	Baseline	ML	MAP(HP)	P-value
20	0.55	0.45	0.45	0.399
15	0.88	0.69	0.69	0.296
10	2.05	1.70	1.69	0.372
5	5.55	4.80	4.75	0.062
0	18.01	16.43	16.28	0.001
All	5.41	4.81	4.77	

Table II lists the performance of ML-based ESSEM (as "ML") and MAP-based ESSEM (as "MAP(CP)"), both using LCB as the mapping function. "Baseline" is also listed as a reference. Here, the CP densities were used for the MAP estimation. Again, we present the P-values of matched pair t-Test for MAP-ESSEM versus ML-ESSEM. We already know that by using LCB mapping function, we can obtain better performance than using LC [11]. Therefore, it is not surprising that the "ML" results in Table II are better than those in Table I. Next, from the Table II, we obtain similar observations to those from Table I. First, when compared with ML-based ESSEM, MAP-based ESSEM gives lower average WER over 50 testing conditions (WER is reduced from 4.66% to 4.62%). Second, more significant improvements (smaller P-values) are achieved under lower SNR conditions (0dB and 5dB SNR).

TABLE II
AVERAGE WORD ERROR RATE (%) AND P-VALUE FOR EACH SNR CONDITION

dB	Baseline	ML	MAP(CP)	P-value
20	0.55	0.45	0.44	0.135
15	0.88	0.67	0.67	0.363
10	2.05	1.68	1.67	0.316
5	5.55	4.59	4.56	0.034
0	18.01	15.92	15.77	0.009
All	5.41	4.66	4.62	

Finally, Table III presents the MAP results (using LCB as the mapping function) with CP and HP densities for three test sets over 0dB to 20dB SNR. From Table III, we can see that by using "ML" results as a basis, "MAP(CP)" provides relatively less improvements than "MAP(HP)" for SetA. The reason underlying such a difference in improvements should be that the environment structure already provides rich information for observed testing conditions. Using additional CP densities only gives marginal improvements. However in the case of SetB that has different types of noise, "MAP(CP)" can provide better improvements than "MAP(HP)". Finally, both two MAP approaches achieve improvements for SetC while "MAP(HP)" gives relatively more improvements than "MAP(CP)". The results show that the HP densities might be a better choice for testing conditions containing distortion types not prepared in the training set (here: channel distortion).

TABLE III
AVERAGE WORD ERROR RATES (%) FOR THREE TESTING SETS

Test Condition	SetA	SetB	SetC	All
Baseline	5.05	5.31	6.31	5.41
ML	4.44	4.69	5.07	4.66
MAP(CP)	4.43	4.64	4.97	4.62
MAP(HP)	4.36	4.69	4.91	4.60

V. CONCLUSION

In this paper, we propose the MAP estimation of online mapping parameters in the ESSEM framework. Two types of prior densities, clustered prior (CP) and hierarchical prior (HP), are introduced. To have a better comparison, we conducted MAP-based ESSEM experiments using two forms of mapping functions, namely linear combination (LC) and linear combination with a correction bias (LCB). Based on the experimental results, it is observed that for both LC and LCB mapping functions, MAP-based ESSEM achieves better performance than the conventional ML-based ESSEM by using either type of prior densities. Moreover, because the accuracy of ESSEM mapping function is more critical to the performance in noisier environments, the MAP-based ESSEM provides better improvements under lower SNR conditions. Among all the testing results, MAP-based ESSEM using LCB as the mapping function along with the HP densities achieves the best performance of 4.60% WER in average of 50 testing conditions, which corresponds to a 14.97% (5.41% to 4.60%) WER reduction over our best baseline result (5.41% WER).

Our first future work would be testing MAP-based ESSEM using other forms of online mapping functions. Moreover in this study, we only use MAP to estimate the ESSEM mapping parameters and keep the environment structure unadjusted. In the future, we will research on a joint MAP adaptation on both mapping parameters and environment structure.

ACKNOWLEDGEMENT

The authors would like to thank Sakai Shinsuke of NICT for helpful discussions and comments.

REFERENCES

[1] B.-H. Juang, W. Chou, and C.-H. Lee, "Statistical and discriminative methods for speech recognition," In: C.-H. Lee, K.K. Soong, F.K. Paliwal, *Automatic Speech and Speaker Recognition: Advanced Topics, Ch. 5.* Kluwer Academic Publishers, Dordrecht, 1996.

[2] 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.

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

[4] 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.

[5] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", *IEEE Trans. Speech Audio Proc.*, vol. 2, pp. 291-99, 1994.

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

[7] 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.

[8] C.-H. Huang, J.-T. Chien, and H.-M. Wang "A new eigenvoice approach to speaker adaptation," in *Proc. ISCSLP*, pp. 109 - 112, 2004.

[9] M. Akbacak and J. H. L Hansen, "Environmental sniffing: noise knowledge estimation for robust speech systems," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 15, pp.465-477, 2007.

[10] Z. Zhang and S. Furui, "Piecewise-linear transformation-based HMM adaptation for noisy speech," *Speech Comm.*, vol. 24, pp. 43-58, 2004.

[11] 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.

[12] 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.

[13] K. Shinoda and C.-H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Trans. Speech Audio Proc.*, vol. 9, pp. 276-287, 2001.

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

[15] A. P. Dempster, N. M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B*, vol. 39, pp. 1-38, 1977.

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

[17] 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.

[18] 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.

[19] J. Li, M. Yuan, and C.-H. Lee, "Approximate test risk bound minimization through soft margin estimation," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 15, pp. 2393-2404, 2007.

[20] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. on Speech and Audio Proc.*, vol. 5, pp. 257-265, 1997.

[21] A. J. Hayter, *Probability and Statistics for Engineers and Scientists*, Duxbury Press; 3rd edition, 2006.