A SAMPLING-BASED ENVIRONMENT POPULATION PROJECTION APPROACH FOR RAPID ACOUSTIC MODEL ADAPTATION

Yu Tsao, Shigeki Matsuda, Shinsuke Sakai, Ryosuke Isotani, Hisashi Kawai, and Satoshi Nakamura
Spoken Language Communication Group
National Institute of Information and Communications Technology, Kyoto, 619-0288, Japan

ABSTRACT

We propose an environment population projection (EPP) approach for rapid acoustic model adaptation to reduce environment mismatches with limited amounts of adaptation data. This approach consists of two stages: population construction and projection. In the population construction stage, we apply a sampling scheme on the adaptation data to construct an environment population based on acoustic models prepared in the training phase. With this sampling procedure, the environment samples in the population characterize diverse acoustic information embedded in the adaptation data. Next, the projection stage estimates a function to map the environment population into one set of acoustic models that matches the testing condition. With a well-constructed environment population, a simple projection function can enable the EPP approach to accurately characterize the testing environment even with a small amount of adaptation data. To examine the rapid adaptation ability of EPP, we used only one adaptation utterance and tested performance in both supervised and unsupervised adaptation modes on Aurora-2 and Aurora-2J tasks. It is found that EPP achieves satisfactory performance under both modes for both tasks. On the Aurora-2J task, for example, EPP gives a clear improvement of a 13.87% (8.58% to 7.39%) word error rate (WER) reduction over our baseline in the unsupervised adaptation mode.

Index Terms*—Stochastic matching, acoustic model adaptation, ensemble classification, environment population projection*

1. INTRODUCTION

For automatic speech recognition (ASR), enhancing performance robustness under training and testing mismatched conditions is a crucial task. Maximum likelihood (ML)-based model adaptation approaches have been proposed and shown strong ability to reduce such mismatch by adjusting acoustic model parameters to match the testing conditions. For these approaches, a mapping function is usually defined to characterize the mismatch, and the parameters in the function are estimated according to the available adaptation data based on the ML criterion. Successful methods include maximum likelihood-based stochastic matching algorithm [1] and maximum likelihood linear regression (MLLR) [2]. These ML-based methods can effectively characterize the environment mismatches when sufficient adaptation data with accurate transcription information are available. However, real-world ASR applications usually favor rapid model adaptation with small amounts of adaptation data with or even without correct transcription information. The insufficient adaptation samples along with possible imperfect transcriptions may deteriorate the ASR performance. Therefore, identifying ways to efficiently take advantage of the available adaptation data and to improve the transcription correctness is vital to the model adaptation capability.

More recently, a cross-validation (CV) based approach has been proposed to estimate multiple acoustic model sets with a CV scheme to improve the decoding hypothesis and accordingly enhance the performance for unsupervised adaptation [3]. Some other approaches incorporate *N*-best list to obtain better decoding hypothesis for the unsupervised adaptation mode [4] and to increase the discriminative power for supervised adaptation [5]. Another class of approaches develops a confidence measure and performs adaptation using only samples with high confidence scores [6]. In this study, we propose an environment population projection (EPP) approach that takes a sampling procedure to effectively utilize the available adaptation data. As will be presented later, the proposed EPP approach provides satisfactory performance in both supervised and unsupervised modes.

The EPP approach extends the ML-based stochastic matching algorithm by incorporating the ensemble classification concept [7]. In the implementation, EPP first performs a sampling procedure on the available adaptation data set to generate several adaptation data subsets. Each subset carries specific acoustic information embedded in the entire set of adaptation data. With these adaptation subsets and with acoustic models from the training phase, EPP calculates several environment-specific acoustic model samples. The ensemble samples then form an environment population. Finally, a projection function is estimated to map the environment population into one set of acoustic models that matches the testing condition.

We further develop two schemes to improve the accuracy and diversity of the environment population. First, a specially designed sampling and resampling scheme is used as the sampling procedure to enhance the confidence level of each adaptation subset. Therefore, each environment sample can be estimated more accurately toward a specific acoustic condition. Second, to further increase the coverage of environment population, we prepare multiple acoustic model sets from the training phase instead of a single set. From our experimental results in both supervised and unsupervised modes, EPP gives better recognition performance not only than the baseline but also than the conventional bias compensation and MLLR with a further confirmation by significance testing. The improvements suggest that EPP possesses better model adaptation ability by taking advantage of important diverse information from the adaptation data that is usually averaged out in the conventional direct estimation approaches. The rest of this paper is organized as follows. Section 2 introduces the EPP framework. Section 3 presents the implementation of EPP, and then section 4 shows the experimental setup and results. Finally, we provide our conclusions in section 5.

2. ENVIRONMENT POPULATION PROJECTION (EPP)

In this section, we first review the ML-based stochastic matching algorithm and then introduce the proposed EPP approach.

2.1. ML-based Stochastic Matching

The ML-based stochastic matching algorithm [1] characterizes an unknown combination of speaker variability and environment distortions by a mapping function, G_{φ}. By this mapping function, the original acoustic model, Λ_{X}, is transformed into a new acoustic model, $\overline{\Lambda}_{\mathrm{Y}}$, that matches the testing condition:

$$\overline{\Lambda}_{\mathrm{Y}} = \mathrm{G}_{\varphi}(\Lambda_{\mathrm{X}}). \tag{1}$$

The parameters of the mapping function,φ, are estimated based on the speech utterances,0_{Y}, in an ML manner:

$$\hat{\varphi} = \underset{\varphi}{argmax}\ P(0_{\mathrm{Y}}|\varphi, \Lambda_{\mathrm{X}}). \tag{2}$$

2.2. EPP Framework

Figure 1 shows the EPP framework, which consists of two stages: Stage-1 prepares an environment population, and Stage-2 estimates a set of acoustic models that matches the testing condition.

2.2.1 Environment Population Construction

For Stage-1, we first take a sampling and resampling scheme to generate S adaptation subsets. The scheme performs three steps on the available adaptation data of T speech frames. Step1 decodes the adaptation utterances to acquire state alignment and posterior probability for each frame. Step2 pools these T frame samples and randomly draws T' ($T'<T$) frames without replacement from the T frames. Step3 resamples and draws ($T-T'$) frames from the T' frames that are obtained in Step2. Finally, we collect T frames (same amount as the original adaptation set) for one adaptation subset. We perform the three steps S times and obtain S adaptation subsets. Moreover in Step3, we devise an additional procedure that draws more frames from the sample pool for those frames having higher posterior probabilities; frames with posterior probabilities lower than a certain threshold are removed from the subset. This procedure enhances confidence level of each adaptation subset. With the collection of S adaptation subsets, we calculate S sets of acoustic models by:

$$\overline{\Lambda}^{s} = \mathrm{G}_{\varphi^s}(\Lambda_{\mathrm{X}}),\ s = 1,2 \dots, S, \tag{3}$$

where Λ_{X} and $\overline{\Lambda}^{s}$ are, respectively, the acoustic models from the training phase and transformed acoustic models using the s-th adaptation subset. We call G_{φ^s} population construction function, and the parameters, φ^s, in G_{φ^s} are calculated by the ML criterion:

$$\hat{\varphi}^s = \underset{\varphi^s}{argmax}\ P(\mathrm{O}^s_{\mathrm{Y}}|\varphi^s, \Lambda_{\mathrm{X}}), \tag{4}$$

where $\mathrm{O}^s_{\mathrm{Y}}$ is the s-th adaptation subset. The ensemble S environment samples form an environment population, $\Theta = \{\overline{\Lambda}^{1}, \overline{\Lambda}^{2} \dots, \overline{\Lambda}^{S}\}$. We call this population constructed using information from adaptation data adaptation-phase environment (AE) population.

Recently, a class of studies indicates that the diverse information from the training data provide crucial prior knowledge for model adaptation [8-10]. Here, we further include the training information by using multiple anchor model sets to increase the coverage of environment population. For the p-th anchor model, Λ_p, with the s-th adaptation subset, we obtain a new model, $\overline{\Lambda}^s_p$, by a function, $\mathrm{G}_{\varphi^s_p}$:

$$\overline{\Lambda}^s_p = \mathrm{G}_{\varphi^s_p}(\Lambda_p). \tag{5}$$

Similar to Eq-(2), we estimate the parameters, φ^s_p, in $\mathrm{G}_{\varphi^s_p}$ by:

$$\hat{\varphi}^s_p = \underset{\varphi^s_p}{argmax}\ P(O^s_{\mathrm{Y}}|\varphi^s_p, \Lambda_p). \tag{6}$$

With P anchor models and S adaptation subsets, we get P×S samples and form an environment population, $\Theta = \{\overline{\Lambda}^1_1, \overline{\Lambda}^2_1 \dots, \overline{\Lambda}^1_p, \dots, \overline{\Lambda}^S_P\}$. We name it training-adaptation-phase environment (TAE) population because both training and adaptation information are incorporated.

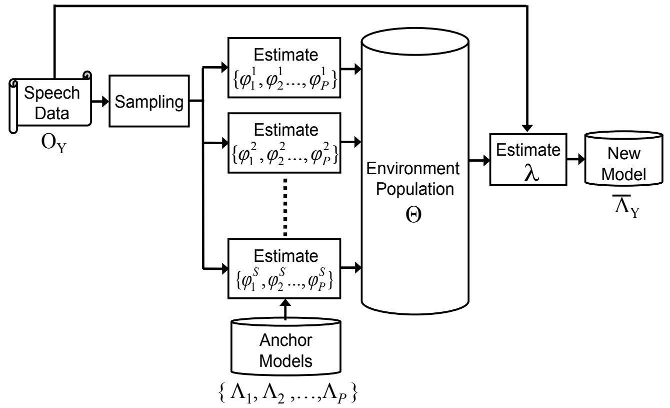
Speech Data
O_{Y}
Sampling
Estimate $\{\varphi^1_1, \varphi^1_2 \dots, \varphi^1_P\}$
Estimate $\{\varphi^2_1, \varphi^2_2 \dots, \varphi^2_P\}$
Estimate $\{\varphi^S_1, \varphi^S_2 \dots, \varphi^S_P\}$
Anchor Models
$\{\Lambda_1, \Lambda_2, \dots, \Lambda_P\}$
Environment Population Θ
Estimate λ
New Model
$\overline{\Lambda}_{\mathrm{Y}}$

Figure 1: Environment population projection framework

2.2.2. Environment Population Projection

For Stage-2, we calculate a projection function, G_λ, to map the environment population, Θ, to one set of acoustic models, $\overline{\Lambda}_{\mathrm{Y}}$, by:

$$\overline{\Lambda}_{\mathrm{Y}} = \mathrm{G}_\lambda(\Theta). \tag{7}$$

We estimate the parameter set, λ, of the projection function, G_λ, based on the ML criterion:

$$\hat{\lambda} = \underset{\lambda}{argmax}\ P(\mathrm{O}_{\mathrm{Y}}|\lambda, \Theta). \tag{8}$$

3. IMPLEMENTATION OF THE EPP APPROACH

In this section, we introduce our implementation steps of the EPP approach. An affine transformation is used as the construction function, and a linear combination is adopted to be the projection function. We only consider adaptation on mean vectors in this study. To enhance adaptation accuracy, we design a hierarchical tree structure to cluster mean vectors in our EPP implementation. The leaf nodes are individual means in Λ_{X}, and each intermediate node contains a group of means. The top node includes the entire set of means. Before adaptation, we perform a searching procedure through the tree to locate a node (the c-th node in the following) that contains sufficient adaptation data for each mean vector.

3.1. Construction Function Estimation

When using an affine transformation as the construction function in Eq-(5), the estimation of parameters is equivalent to the MLLR transformation solution [2]. With the m-th mean $\mu_{p,m}$ in the p-th anchor model, Λ_p, we intend to calculate $\bar{\mu}^s_{p,m}$ based on the s-th adaptation subset. If we know the m-th mean vector belonging to the *c-th* node, EPP estimates an affine transformation, $\Gamma^s_{p,c}$, to perform:

$$\bar{\mu}^s_{p,m} = \Gamma^s_{p,c}\xi_{p,m}\ , \tag{9}$$

where $\xi_{p,m}$ is the augmented vector: $[1, \mu^{(1)}_{p,m}, \dots, \mu^{(D)}_{p,m}]'$ (each vector has D coefficients). We can calculate $\Gamma^s_{p,c}$ by the ML criterion:

$$\hat{\Gamma}^s_{p,c} = \underset{\Gamma^s_{p,c}}{argmin} \sum_{t=1}^{T} \sum_{k\in c} L^s(t) r_k(t)[(y_t - \Gamma^s_{p,c}\xi_{p,k})\Sigma_k^{-1}(y_t - \Gamma^s_{p,c}\xi_{p,k})], \tag{10}$$

where y_t is the t-th observation vector. Σ_k is the covariance matrix, and $r_k(t)$ is the posterior probability of the k-th Gaussian that belongs to the c-th node. $L^s(t)$ is the sampling unit that generates the s-th adaptation subset. We set T'=0.7×T, and by the resampling step each adaptation subset has the same amount T frames as the original adaptation set. By collecting P×S means, we get a matrix, $\overline{\mathrm{M}}_m = \{\bar{\mu}^1_{1,m}, \bar{\mu}^2_{1,m} \dots, \bar{\mu}^1_{p,m}, \dots, \bar{\mu}^S_{P,m}\}$, as the population for the m-th Gaussian.

3.2. Projection Function Estimation

With the prepared environment population, we estimate a projection function to find the final acoustic models. From a previous study, we can have several projection function candidates [4]. In this paper, we choose a linear combination function as the projection function. For the m-th Gaussian, we first follow the searching process to determine a node (the c-th node) and then calculate the final mean vector, $\bar{\mu}_{\mathrm{Y},m}$, that matches the testing condition by:

$$\bar{\mu}_{\mathrm{Y},m} = \overline{\mathrm{M}}_m w_c\ , \tag{11}$$

where w_c is the coefficient vector of the linear combination function for the c-th node. Similarly, we calculate w_c using the ML criterion:

$$\hat{w}_c = \underset{w_c}{argmin} \sum_{t=1}^{T} \sum_{k\in c} r_k(t)[(y_t - \overline{\mathrm{M}}_k w_c)'\Sigma_k^{-1}(y_t - \overline{\mathrm{M}}_k w_c)]\,. \tag{12}$$

The presented EPP implementation using affine transformation as the construction function and linear combination as the projection function is only an example. Other types of construction and projection functions can be implemented in a similar manner.

4. EXPERIMENTS

In this section, we first briefly introduce the experimental setup. Then, we present and discuss our experimental results.

4.1. Experimental Setup

We evaluated the EPP approach on two speech databases, Aurora-2 [11] and Aurora-2J [12]. Aurora-2 is a well-known English connected digit recognition task that is often used to evaluate ASR's noise robustness. Aurora-2J is a Japanese version digit recognition task and is designed to have the same structure as Aurora-2. Both tasks have two training sets, multi-condition and clean condition, and 70 different testing conditions (ten noise types at seven SNR levels).

In this paper, we use the multi-condition training sets for both the Aurora-2 and Aurora-2J tasks. For both tasks, the multi-condition training set includes 17 different speaking environments that are from the same four types of noise as in test SetA, at different SNR levels: 5dB, 10dB, 15dB, 20dB, and clean condition. In the following discussions, we report performance using 50 testing conditions (ten noise types at five SNR levels, 0dB, 5dB, 10dB, 15dB, 20dB). Each condition has 1001 utterances collected from 104 testing speakers (52 male and 52 female). Each speaker pronounced nine or ten utterances. The testing speakers did not participate in the training phase. Here, we report our results in average word error rate (WER).

For both the Aurora-2 and Aurora-2J tasks, we used a modified ETSI advanced front-end (AFE) for feature extraction [13]. Every feature vector comprised 13 static plus their first and second order time derivatives. Meanwhile, we followed a complex back-end topology presented in [13] to train baseline hidden Markov models (HMMs). Each digit was modeled with 20 mixtures per state, and the silence and short pause were modeled with 36 mixtures per state.

Because we focused on rapid model adaptation, only one utterance was used for adaptation. We tested EPP performance in both supervised and unsupervised adaptation modes on both Aurora-2 and Aurora-2J tasks. Similar results were obtained for all the evaluations. Due to the limited space, we only report supervised experiments on Aurora-2 and unsupervised on Aurora-2J. For the supervised mode, we used the first single utterance to adapt model parameters for each speaker. The adapted model was then used to test recognition on the remaining eight or nine utterances from that same speaker. For each of the 50 testing conditions, we performed the adaptation and testing procedures 104 times and finally reported the average WER over 897 (1001-104) testing utterances. For the unsupervised mode, we conducted a per-utterance self-adaptation scheme: each testing utterance was first decoded into *N*-best lists; the *N*-best hypotheses were then used for adaptation. Finally, the adapted models were used to recognize the same testing utterance. Accordingly, each condition contained 1001 testing utterances.

4.2. Experimental Results

As mentioned in Section 3, we built a tree structure to facilitate model adaptation. For the Aurora-2 task, we prepared a three-layered tree structure (including root, intermediate, and leaf nodes) based on the distances between mean vectors in the Aurora-2 baseline HMM set. For all the experiments on Aurora-2 in the following discussions, we used this tree structure to perform model adaptation. In the same way, we built another three-layered tree for Aurora-2J and used the tree to conduct all the model adaptation experiments on Aurora-2J.

4.2.1. Supervised Experiments

First, we compare the EPP performance using different environment populations. In addition to AE and TAE, we designed another environment population by taking P anchor model sets in the training phase, $\{\Lambda_1, \Lambda_2 \ldots, \Lambda_P\}$, and the original adaptation data set (without doing the sampling procedure) to calculate the new model set, $\overline{\Lambda}_p$, by:

$$\overline{\Lambda}_p = \mathrm{G}_{\varphi_p}(\Lambda_p),\ p = 1,2 \ldots., P, \tag{13}$$

where

$$\hat{\varphi}_p = \underset{\varphi_p}{argmax}\, P(\mathrm{O}_\mathrm{Y}|\varphi_p, \Lambda_p). \tag{14}$$

Then, the ensemble P sets of transformed acoustic models form a population:$\Theta = \{\overline{\Lambda}_1, \overline{\Lambda}_2 \ldots, \overline{\Lambda}_P\}$. We call this population the training-phase environment (TE) population. Here, we set P=5 by using five anchor HMM sets. In addition to the baseline HMM set, we prepared other four anchor HMM sets that represented four different acoustic characteristics in the training set, including anchor HMM sets for high SNR, low SNR, male speakers, and female speakers.

Table I presents the EPP testing results using TE, AE, and TAE populations in a supervised adaptation mode on Aurora-2. To get the environment populations, EPP-AE used one anchor HMM set (the baseline HMMs) and 30 adaptation subsets (P=1,S=30); EPP-TE used five anchor HMM sets and the original adaptation utterance without sampling (P=5,S=1); EPP-TAE used five anchor HMM sets with six adaptation sampling subsets (P=5,S=6). In this set of experiments, we used the affine transformation as the construction function in Eq-(5) and the linear combination function as the projection function in Eq-(7). SetA, SetB, and SetC in Table I show the average WERs over 20, 20, and 10 conditions. We also list Baseline and MLLR results in Table I for comparison. For Baseline, we directly used the baseline HMM set to test recognition without performing adaptation. For MLLR, we performed MLLR adaptation to calculate new HMMs. Then, the testing utterances were tested recognition with the adapted HMM set. It is noted that we can consider MLLR as EPP with (P=1,S=1) in this condition.

From the results, we first observe that MLLR and EPP with any of the three populations can provide clear improvements over Baseline for all sets, especially SetC. Next, we observe that EPP-AE and EPP-TE give improvements over MLLR. The results indicate that either by using multiple adaptation subsets (S=1 to S=30) or by incorporating multiple training anchor models (P=1 to P=5), we can achieve better adaptation performance. When comparing EPP-AE with EPP-TAE, we see that by using a same number of environment samples (both are 30), EPP-TAE achieves better performance. Finally by comparing EPP-TE with EPP-TAE, we find that EPP-TAE outperforms EPP-TE by using more adaptations subsets (S=1 to S=6). The results suggest that by including both training and adaptation information in the environment population, EPP can model the testing conditions more accurately. Compared to Baseline, EPP-TAE gives a 6.05% (6.45% to 6.06%) average WER reduction.

TABLE I. SUPERVISED MODE AVERAGE WER (%) ON AURORA-2

Test Condition	SetA	SetB	SetC	Overall
Baseline	5.88	6.70	7.08	6.45
MLLR(P=1,S=1)	5.74	6.57	6.53	6.23
EPP-AE(P=1,S=30)	5.66	6.46	6.35	6.12
EPP-TE(P=5,S=1)	5.57	6.56	6.26	6.10
EPP-TAE(P=5,S=6)	5.50	6.49	6.33	6.06

From Table I, the overall performance of EPP-TAE (S=6,P=5) seems only marginally better than that of MLLR(S=1,P=1) with an average WER reduction of 2.73% (6.23% to 6.06%). In this study, we further took a matched pair t-Test significance test [14] to verify the performance improvement. Because each SNR condition has ten results in the Aurora-2 test set, we conducted t-Test by ten pair-wise results. Instead of using a fixed threshold to determine the t-Test results, we directly present P-values of the matched pair t-Test. A small P-value indicates consistent performance improvements over ten results. We list the MLLR and EPP-TAE WER results along with P-values at different SNR condition in Table II. Each block in the first three columns lists an average WER over ten different noise types. Each block in the fourth column lists a P-value of EPP-TAE versus MLLR. From Table II, small P-values are observed for almost every condition. This observation indicates that EPP-TAE is consistently better than MLLR for each particular SNR condition.

TABLE II. SUPERVISED MODE WER (%) AND P-VALUES ON AURORA-2

dB	Baseline	MLLR	EPP-TAE	P-value
20	0.69	0.69	0.62	0.015
15	1.19	1.17	1.10	0.070
10	2.71	2.59	2.49	0.061
5	6.79	6.55	6.41	0.047
0	20.87	20.15	19.69	0.001
All	6.45	6.23	6.06	

4.2.2. Unsupervised Experiments

In this set of experiments, we intentionally used another function, bias compensation, as the construction function to test the EPP performance. The linear combination is still used as the projection function. When using a compensation bias, $b^s_{p,c}$, Eq-(5) becomes:

$$\bar{\mu}^s_{p,m} = \mu_{p,m} + b^s_{p,c}\,. \tag{15}$$

Similar to Eq-(10), we estimate the bias compensation for the p-th environment with given the s-th adaptation subset by :

$$\hat{b}^s_{p,c} = \underset{b^s_{p,c}}{argmin} \sum_{t=1}^{T} \sum_{k\in c} L^s(t) r_k(t) [(y_t - \mu_{p,k} - b^s_{p,c})' \Sigma_k^{-1} (y_t - \mu_{p,k} - b^s_{p,c})]\,. \tag{16}$$

Table III reports the EPP results with three different environment populations in the unsupervised mode on Aurora-2J. We also list Baseline and Bias results for comparison. Similar to the supervised mode, for Baseline, we directly used the baseline HMMs to test recognition; for Bias, we used a bias compensation function to adapt HMMs and then decoded the same testing utterance using the adapted HMMs. From Table III, we observe similar phenomena to that from Table I. First, both Bias (P=1,S=1) and the three EPP setups give clear improvements over Baseline. Next, EPP-TAE achieves the best performance comparing to Bias and the other two EPP setups. For the overall testing conditions, EPP-TAE provides a 13.87% (8.58% to 7.39%) average WER reduction over Baseline.

TABLE III. UNSUPERVISED MODE AVERAGE WER (%) ON AURORA-2J

Test Set	SetA	SetB	SetC	Overall
Baseline	6.80	10.67	7.95	8.58
Bias(P=1,S=1)	6.43	9.23	7.32	7.73
EPP-AE (P=1,S=30)	6.27	8.87	7.32	7.52
EPP-TE (P=5,S=1)	6.24	8.85	7.34	7.50
EPP-TAE(P=5,S=6)	6.20	8.67	7.22	7.39

Similarly, we verify the performance improvements of EPP-TAE over Bias by using the matched pair t-Test. Table IV shows WERs and P-values at different SNR levels. We find similar observations to that from the supervised experiments: though the overall average improvement of EPP-ATE over Bias is 4.40% (7.73% to 7.39%), the small P-values listed in Table IV indicate that the improvements are consistent among ten results for each of the five SNR levels.

TABLE IV. UNSUPERVISED MODE WER (%) AND P-VALUES ON AURORA-2J

dB	Baseline	Bias	EPP-TAE	P-value
20	0.49	0.41	0.36	0.035
15	1.01	0.88	0.72	0.001
10	3.12	2.56	2.25	0.019
5	9.87	8.51	7.90	0.001
0	28.39	26.28	25.73	0.017
All	8.58	7.73	7.39	

5. CONCLUSION

We propose an EPP approach to perform rapid model adaptation for reducing environment mismatches. EPP uses a sampling scheme to prepare an environment population and then estimates a projection function to map the population to one set of acoustic models that matches the testing condition. We evaluated the EPP approach in both supervised and unsupervised modes on Aurora-2 and Aurora-2J tasks. To investigate the rapid adaptation capability of EPP, we used only one adaptation utterance for all the evaluations. We observed similar results from all the evaluations and reached three major conclusions:(1) EPP provides significant improvements over the baseline under both supervised and unsupervised modes; (2) EPP using affine transformation or bias compensation as the construction function can give better performance than their direct estimation counterparts; (3) Incorporating additional training information by using multiple anchor model sets in the training phase can enable EPP to achieve even better performance. In the future, we will explore other sampling schemes and the correlation between number of samples and achievable performance. Moreover, other types of construction and projection functions will be further studied.

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 Language*, vol. 9, pp.171-185, 1995.

[3] T. Shinozaki, Y. Kubota, and S. Furui, "Unsupervised cross-validation algorithm for improved adaptation performance," in *Proc. ICASSP*, pp. 4377-4380, 2009.

[4] 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.

[5] P. Nguyen, P. Gelin, J.-C. Junqua, and J.-T. Chien, "N-best based supervised and unsupervised adaptation for native and non-native speakers in cars," in *Proc. ICASSP*, pp. 173-176, 1999.

[6] W. K. Lo and F. K. Soong, "Generalized posterior probability for minimum error verification of recognized sentences," in *Proc. ICASSP*, pp.85-88, 2005.

[7] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, New York, Wiley, 2001.

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

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

[10] 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.

[11] 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.

[12] S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa, N. Kitaoka, T. Nishiura, A. Sasou, M. Mizumachi, C. Miyajima, M. Fujimoto, and T. Endo, "AURORA-2J: An evaluation framework for Japanese noisy speech recognition," *IEICE Trans. Inf. Syst.*, vol. E88-D, pp. 535-544, 2005.

[13] 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.

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