INCREASING DISCRIMINATIVE CAPABILITY ON MAP-BASED MAPPING FUNCTION ESTIMATION FOR ACOUSTIC MODEL ADAPTATION

Yu Tsao, Ryosuke Isotani, Hisashi Kawai, and Satoshi Nakamura
SLC Group, National Institute of Information and Communications Technology, Kyoto, 619-0288, Japan

ABSTRACT

In this study, we propose increasing discriminative power on the maximum a posteriori (MAP)-based mapping function estimation for acoustic model adaptation. Based on the effective and stable learning advantages of MAP-based estimation, we incorporate a discriminative term and derive a new objective function. By applying the new function for online mapping function estimation, we developed discriminative maximum a posteriori (DMAP) linear regression (DMAPLR) and DMAP-based ensemble speaker and speaking environment modeling (DMAP-based ESSEM). We evaluate the DMAPLR and DMAP-based ESSEM on the Aurora-2 task in a supervised adaptation mode. The experimental results show that both DMAPLR and DMAP-based ESSEM consistently provide improvements over their ML-based and MAP-based counterparts irrespective of using one, two, or three adaptation utterances. From the improvements, we confirm the strong effect of increasing discriminative capability on the MAP-based mapping function estimation. Moreover, we verify that including multiple knowledge sources in the objective function can efficiently enhance model adaptation performance. When compared with the baseline result, DMAP-ESSEM achieves a 15.96% (9.21% to 7.74%) average word error rate (WER) reduction using only one adaptation utterance.

Index Terms-*Automatic speech recognition, MLLR, ESSEM, MAPLR, MAP-based ESSEM, discriminative training*

1. INTRODUCTION

Increasing discriminative power on acoustic models is known as an effective way to improve speech recognition performance under imperfect testing environments [1-3]. Many discriminative training (DT) methods have recently been proposed. These methods typically first define a particular objective function that measures the separation between parameters in the acoustic models. An optimization procedure is then performed on the objective function with the available training set to increase the separation between parameters. Popular examples include minimum classification error (MCE) [1], minimum phone error (MPE) [2], and soft margin estimation (SME) [3] methods. Because these DT methods target increasing parameter separations, we usually apply them to refine acoustic models after a maximum likelihood-based training.

In addition to DT, model adaptation is another successful way to reduce acoustic mismatches between training and testing conditions. Model adaptation approaches usually adopt some transformation functions to characterize the environment mismatches. A particular optimization criterion is used to find the parameters of the transformation function with a set of adaptation data that contain acoustic information of the testing conditions. Maximum likelihood (ML) is a popular criterion for performing the optimization. Maximum likelihood linear regression (MLLR) [4] and ML-based ensemble speaker and speaking environment modeling (ESSEM) [5] are two successful examples for the ML-based model adaptation approaches. Though these ML-based approaches can provide satisfactory results when sufficient adaptation data are available, their performance may become unstable due to an over-fitting issue when the amount of adaptation data is too limited. To avoid over-fittings, a class of approaches adopts prior knowledge and uses a maximum a posteriori (MAP) criterion to estimate transformation function. Corresponding to the previous two examples, their MAP-based counterparts, maximum a posteriori linear regression (MAPLR) [6, 7] and MAP-based ESSEM [8], have been proposed.

More recently, approaches combining the advantages of DT and acoustic model adaptation have been developed. Minimum classification error linear regression (MCELR) [9] and soft margin estimation-based linear regression (SMELR) [10] are proposed to refine transformation functions by increasing their discriminative power. In this paper, we propose increasing the discriminative capability on the MAP-based model adaptation methods by deriving a new objective function. We apply the new function on the two MAP-based approaches, MAPLR and MAP-based ESSEM, and develop discriminative MAPLR (DMAPLR) and discriminative MAP-based ESSEM (DMAP-based ESSEM), respectively. We evaluated DMAPLR and DMAP-based ESSEM on the Aurora-2 task [11] in a supervised adaptation mode. Experimental results indicate that both DMAPLR and DMAP-based ESSEM can provide clearly better performance than their MAP-based counterparts.

In a previous study, a confidence score of a combination of likelihood and likelihood ratio (LR) has been used to extend the conventional speech recognizer to a hybrid decoder [12]. Because the score integrates multiple knowledge sources, the hybrid decoder provides better recognition results than the conventional decoder that uses either the likelihood or LR score alone. The proposed DMAPLR and DMAP-based ESSEM share a similar concept that adopts a combined score in the objective function to estimate transformation functions. To investigate the effect of using multiple knowledge sources, we designed another set of experiments. We compared the objective functions using a combined score, namely, DMAP-based approaches, versus using likelihood and LR scores alone (all these objective functions integrate priors to improve performance stability). Our experimental results indicated that the DMAP-based approaches do give better performance than the counterpart methods using likelihood or LR alone. This set of results confirmed that by incorporating multiple knowledge sources in the objective function, we can further enhance the model adaptation ability.

2. DISCRIMINATIVE MAP-BASED OBJECTIVE FUNCTION

Typically, model adaptation techniques use a mapping function, G_φ, to transform parameters in the original acoustic model sets, Ω_X, (Ω_X may include one or multiple sets of acoustic models) to a new set of acoustic models, $\overline{\Lambda}_Y$, that matches the testing condition by:

$$\overline{\Lambda}_Y = G_\varphi(\Omega_X). \quad (1)$$

If the transcription, W_c, corresponding to the adaptation data, O_Y, is available, ML-based adaptation defines a objective function by:

$$L(O_Y, \varphi, \Omega_X, W_c) = log\,[P(O_Y|\varphi, \Omega_X, W_c)]. \quad (2)$$

We estimate the parameters, $\hat{\varphi}_{ML}$, in the function, G_φ, by:

$$\hat{\varphi}_{ML} = \underset{\varphi}{argmax}\, L(O_Y, \varphi, \Omega_X, W_c). \quad (3)$$

The MAP-based model adaptation methods, on the other hand, use the following objective function:

$$M(\mathrm{O_Y}, \varphi, \Omega_\mathrm{X}, W_c) = log\,[P(\mathrm{O_Y}|\varphi, \Omega_\mathrm{X}, W_c)p(\varphi, \Omega_\mathrm{X}, W_c)]. \quad (4)$$

We can calculate parameters, $\hat{\varphi}_{MAP}$, in G_φ by the optimization of:

$$\hat{\varphi}_{MAP} = \underset{\varphi}{argmax}\, M(\mathrm{O_Y}, \varphi, \Omega_\mathrm{X}, W_c). \quad (5)$$

Many previous studies have verified that the likelihood ratio (LR) score can provide crucial discrimination information for acoustic modeling [1-3]. To calculate LR, we require competing lists to the phone/word units in the correct transcription, W_c. In this study, we use *N*-best hypotheses by decoding adaptation utterances to obtain the competing units. Then, we define an LR-based objective function by:

$$D(\mathrm{O_Y}, \varphi, \Omega_\mathrm{X}, W_c, \widetilde{W}_1 \dots \widetilde{W}_N) = \sum_{n=1}^{N} \lambda_n log\left[\frac{P(\mathrm{O_Y}|\varphi, \Omega_\mathrm{X}, W_c)}{P(\mathrm{O_Y}|\varphi, \Omega_\mathrm{X}, \widetilde{W}_n)}\right], \quad (6)$$

where λ_n is a scaling factor that determines the weights of each of the *N*-best hypotheses, with $\sum_{n=1}^{N} \lambda_n = 1$, and $\widetilde{W}_n$ is the *n*-th competing phone/word sequence of the *N*-best lists.

In this study, we derive a new objective function that combines the objective function of Eq-(4) and the objective function of Eq-(6):

$$K(\mathrm{O_Y}, \varphi, \Omega_\mathrm{X}, W_c, \widetilde{W}_1 \dots \widetilde{W}_N) = \alpha_1 M(\mathrm{O_Y}, \varphi, \Omega_\mathrm{X}, W_c) + \alpha_2 D(\mathrm{O_Y}, \varphi, \Omega_\mathrm{X}, W_c, \widetilde{W}_1 \dots \widetilde{W}_N), \quad (7)$$

where α_1 and α_2 are weighting parameters. Accordingly, we estimate the parameters, $\hat{\varphi}_{DMAP}$, in the mapping function, G_φ, by:

$$\hat{\varphi}_{DMAP} = \underset{\varphi}{argmax}\, K(\mathrm{O_Y}, \varphi, \Omega_\mathrm{X}, W_c, \widetilde{W}_1 \dots \widetilde{W}_N). \quad (8)$$

3. DISCRIMINATIVE MAPLR AND MAP-BASED ESSEM

In this section, we present DMAPLR and DMAP-based ESSEM that use the objective function in Eq-(7) for estimating mapping function.

3.1. Discriminative MAPLR

When using the affine transformation as the mapping function, the environment structure, Ω_X, in Eq-(1) only takes one set of acoustic models. For the *m*-th Gaussian in the acoustic model, we have:

$$\bar{\mu}_m = \Gamma\, \xi_m\,, \quad (9)$$

where ξ_m is the augmented vector: $[\,\mu_m^{(1)}, \dots, \mu_m^{(D)}, 1]'$, $\mu_m^{(i)}$ is the *i*-th element of the *m*-th Gaussian component. $\bar{\mu}_m$ is the adapted mean vector. MLLR calculates the affine transformation, Γ, by [4]:

$$\Gamma_{(i)} = (G_{(i)})^{-1} k_{(i)}, \quad (10)$$

along with

$$G_{(i)} = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) \left[\frac{1}{\Sigma_{s_{(i)}}} \xi_s \xi_s'\right], \quad (11)$$

$$k_{(i)} = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) \left[\frac{1}{\Sigma_{s_{(i)}}} o_{t_{(i)}} \xi_s\right], \quad (12)$$

where o_t is the *t*-th observation. $r_s(t)$ is the posterior probability at the *t*-th observation, ξ_s is the augmented mean vector, and $\Sigma_{s_{(i)}}$ is the *i*-th element of the covariance matrix of the *s*-th Gaussian, respectively. $s \in W_c$ indicates that the *s*-th Gaussian belongs to a target model in the correct transcription, W_c.

By using the combined score of Eq-(7) for the objective function, along with the *N*-best information, DMAPLR calculates the affine transformation, Γ, also using Eq-(10), but with:

$$G_{(i)} = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) \left[\frac{h_1}{\Sigma_{s_{(i)}}} \xi_s \xi_s' + \frac{h_2}{\mathrm{V}_{s_{(i)}}} \xi_s \xi_s'\right] - \sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{l \in \widetilde{W}_n} \lambda_n r_l(t) \left(\frac{h_3}{\Sigma_{l_{(i)}}} \xi_l \xi_l'\right), \quad (13)$$

$$k_{(i)} = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) \left[\frac{h_1}{\Sigma_{s_{(i)}}} o_{t_{(i)}} \xi_s + \frac{h_2}{\mathrm{V}_{s_{(i)}}} \eta_{s_{(i)}} \xi_s\right] - \sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{l \in \widetilde{W}_n} \lambda_n r_l(t) \left(\frac{h_3}{\Sigma_{l_{(i)}}} o_{t_{(i)}} \xi_l\right), \quad (14)$$

where $l \in \widetilde{W}_n$ represents the *l*-th Gaussian belonging to a competing model in the *n*-th *N*-best list, $\widetilde{W}_n$. From Eq-(7) to Eqs-(13) and (14), we set $h_1 = (\alpha_1 + \alpha_2)$, $h_2 = (\alpha_1/\pi)$, and $h_3 = \alpha_2$, where π is a scaling factor that controls the weight of priors. ξ_l is the augmented mean vector, and $\Sigma_{l_{(i)}}$ is the *i*-th element of the covariance matrix, for the *l*-th Gaussian. For the *s*-th Gaussian (from the correct transcription), we prepare hyper-parameters, η_s and V_s, in the offline. With the calculated Γ, we can adapt mean parameters by Eq-(9).

3.2. Discriminative MAP-based ESSEM

In our previous ESSEM study, we introduced several different types of mapping functions, G_φ, in Eq-(1). Here, we present a linear combination function as an example. Other types of mapping function can be derived in a similar manner. For the ESSEM approach, an environment structure consisting of multiple acoustic model sets is taken for Ω_X in Eq-(1). Accordingly, for the *m*-th Gaussian, we have an environment structure, $\mathrm{H}_m = \{\mu_m^1, \mu_m^2 \dots \mu_m^P\}$, where μ_m^p is the *m*-th mean vector for the *p*-th speaker and speaking environment. ESSEM calculates the adapted mean vector $\bar{\mu}_m$ by:

$$\bar{\mu}_m = \mathrm{H}_m\, \omega. \quad (15)$$

We estimate the parameters of the linear combination function, ω, by:

$$\omega = G^{-1} k\,. \quad (16)$$

If the ML criterion is used to calculate ω, we have:

$$G = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) \left[\mathrm{H}_s' \Sigma_s^{-1} \mathrm{H}_s\right], \quad (17)$$

$$k = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) \left[\,\mathrm{H}_s' \Sigma_s^{-1} o_t\right]. \quad (18)$$

For DMAP-based ESSEM, we also use Eq-(16) to calculate ω, while:

$$G = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) \left[h_1\, \mathrm{H}_s' \Sigma_s^{-1} \mathrm{H}_s + h_2\, \mathrm{H}_s' \mathrm{V}_s^{-1} \mathrm{H}_s\right] - \sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{l \in \widetilde{W}_n} \lambda_n r_l(t) \left(\,h_3\, \mathrm{H}_l' \Sigma_l^{-1} \mathrm{H}_l\right), \quad (19)$$

$$k = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) \left[h_1\, \mathrm{H}_s' \Sigma_s^{-1} o_t + h_2\, \mathrm{H}_s' \mathrm{V}_s^{-1} \eta_s\right] - \sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{l \in \widetilde{W}_n} \lambda_n r_l(t) \left(\,h_3\, \mathrm{H}_l' \Sigma_l^{-1} o_t\right). \quad (20)$$

Similarly, $h_1 = (\alpha_1 + \alpha_2)$, $h_2 = (\alpha_1/\pi)$, and $h_3 = \alpha_2$, π is a scaling factor. H_s and H_l are the environment structures, and Σ_s and Σ_l are the covariance matrices, respectively, for the *s*-th Gaussian and the competing *l*-th Gaussian. For the *s*-th Gaussian, we also prepare hyper-parameters, η_s and V_s, for its prior density.

4. EXPERIMENTS

In this section, we first introduce the experimental setup. Then, we present experimental results of DMAPLR and DMAP-based ESSEM and their ML- and MAP-based counterparts. Finally, we compare the results of these approaches and discuss our findings.

4.1. Experimental Setup

We evaluated the proposed DMAPLR and DMAP-based ESSEM approaches on the Aurora-2 task [11]. The multi-condition training set in Aurora-2 was used to estimate acoustic models. This training set included 17 different speaking environments from the same four types of noise as in test SetA, at different SNRs: 5dB, 10dB, 15dB, 20dB, and clean condition. By further dividing the training set by genders, we obtained training data for 34 different speaker and speaking environments. In this paper, we report performance on 50 different conditions (ten noise types at five SNR levels: 0dB, 5dB, 10dB, 15dB, and 20dB). Each speech frame was characterized by 39 coefficients consisting of 13 MFCC with their first- and second-order derivatives. A cepstral mean subtraction (CMS) was performed for normalization. All digits were modeled by 16-state whole word hidden Markov models (HMMs) with each state characterized by three Gaussian mixture components. The silence and the short pause were modeled by three states and one state, respectively, with each state characterized by six Gaussian mixture components.

In the experiments, we constructed an environment structure that consists of environment clustering (EC) and environment partitioning (EP) hierarchical tree structures [5] by using the training data. We first clustered the training data into C groups based on the EC tree. The root node comprised the entire set of training data. The second layer consisted of two nodes, each including training data for one gender. Then, each gender-specific cluster in the second layer was further divided into two clusters based on high/low SNR conditions. Therefore, the 34 training environments were classified into seven clusters (C=7). Then, we trained a representative HMM set for each EC cluster using the training data belonging to that cluster. Each representative HMM set was then used to build an EP tree structure to cluster mean parameters in the representative HMM set. Each EP tree had one root, three intermediate, and six leaf nodes. The weighted Euclidean distance was used as the distance measure between each pair of mean vectors. During adaptation, we estimated a mapping function for every EP node. Each mean vector searched for the EP node containing sufficient adaptation statistics and used its mapping function for adaptation. Along with the EC and EP environment structures, we also prepared a hyper-parameter set for each mean vector: hyper-parameters, $\{\eta_m, \mathrm{V}_m\}$, for the m-th Gaussian. The hyper-parameters were used to calculate mapping functions in Eqs-(13) and (14) and in Eqs-(19) and (20). Detailed information about the construction of the EC and EP structures and hyper-parameter estimation can be found in our previous study [7, 8].

Each of the 50 testing conditions in Aurora-2 has 1001 speech utterances recorded from 104 speakers (52 male and 52 female). Each speaker pronounced nine to ten utterances, and these testing speakers did not participate in the training phase. We used the first three utterances as the adaptation set and the remaining six or seven utterances for the testing set. Accordingly, each testing condition included 312 (104×3) adaptation utterances and 689 (1001-312) testing utterances. For all the model adaptation experiments in this paper, we performed four steps on each testing speaker. First, we implemented a cluster selection (CS) to locate the best suitable EC cluster whose representative HMM set gave the highest likelihood for the speaker's adaptation data. Second, with the located EC cluster, we conducted another searching process through its EP tree to choose an EP node with sufficient adaptation statistics. Third, with the chosen EP node, each mean vector was adapted using the mapping function for that EP node. Fourth, we used the adapted HMMs to decode the testing utterances from that same speaker. For each testing condition, we performed the above four steps 104 times and calculated an average WER of the overall results (from the 689 testing utterances pronounced by the 104 testing speakers).

4.2. Experimental Results

This section presents our experimental results. A baseline result is provided for comparison and listed as Baseline in the following discussion. To obtain this Baseline result, a CS process is first performed to find a representative HMM set. The selected HMM set is directly used for testing recognition without performing adaptation.

4.2.1. Discriminative MAPLR

Figure 1 reports average WERs of MLLR, DMLLR (discriminative MLLR), MAPLR, and DMAPLR (discriminative MAPLR). In a preliminary experiment, we tested performance using different combinations of h_1, h_2, and h_3, in Eqs-(13) and (14). Here, we only present the setup that gives the best performance. Our setup in this set of experiments is: for MLLR, $\{h_1$=1.0, h_2=0.0, h_3=0.0$\}$; for DMLLR, $\{h_1$=1.0, h_2=0.0, h_3=0.6$\}$; for MAPLR, $\{h_1$=1.0, h_2= 0.2, h_3=0.0$\}$; for DMAPLR, $\{h_1$=1.0, h_2=0.2, h_3=0.6$\}$. In this study, we used a diagonal regression matrix for Γ in Eq-(9). Meanwhile, we set $\lambda_n = 1/N, \forall n$, and N=8 in Eqs-(13) and (14). Each result in Figure 1 shows an average WER over 50 testing conditions.

From Figure 1, we can first see an obvious improvement by using the prior density (Ⓐ to Ⓒ), and an additional clear gain is achieved by incorporating the discriminative term (Ⓒ to Ⓓ) when using one, two, and three adaptation utterances. Next, we observe that though the improvement from MLLR to DMLLR is marginal (Ⓐ to Ⓑ), DMAPLR gives clear improvements over MAPLR (Ⓒ to Ⓓ). This result indicates that by increasing discrimination on an ML-based transformation estimation, the improvement may not be clear, especially when there is limited adaptation data (one adaptation utterance). However, by incorporating the discriminative term into the objective function, MAP-based estimation can gain clear benefits.

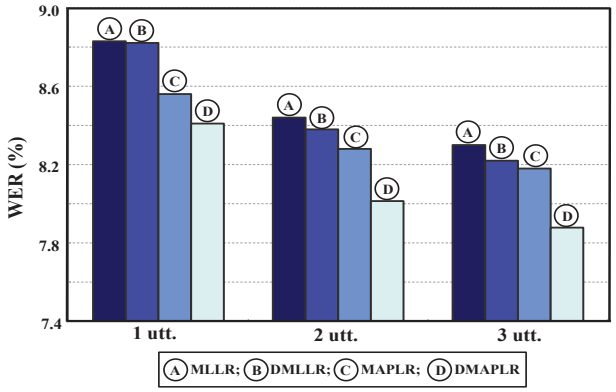

Figure 1. WERs (%) with different numbers of adaptation utterances

In real world applications, we are more interested in rapid adaptation with a very small amount of adaptation data. Therefore, we list the results of using one adaptation utterance in Table I. Each block in Table I shows an average WER over ten conditions, and the best result for each SNR condition (each column) is marked with bold font. In addition to the four tests in Figure 1, we design another setup $\{h_1$=1.0, h_2=0.2, h_3=1.0$\}$ and name it the maximum likelihood ratio linear regression (MLRLR) in the following discussion.

From Table I, we observe that DMAPLR consistently provides lower WERs than not only Baseline but also the other four approaches over 0dB to 20dB conditions (only except SNR=10dB). Special note is made that DMAPLR uses a combined score of likelihood plus LR in the objective function, MLRLR uses LR alone, and MAPLR uses likelihood alone, wherein all three methods adopt prior density. Therefore, the improvements achieved by DMAPLR over MAPLR and MLRLR suggest that the combined score

incorporating multiple knowledge sources can further enhance affine transformation-based model adaptation methods.

TABLE I. AVERAGE WER (%) FOR USING ONE ADAPTATION UTTERANCE

SNR(dB)	20	15	10	5	0	Ave.
Baseline	1.56	2.10	**3.13**	8.21	31.04	9.21
MLLR	1.55	2.09	3.59	8.77	28.15	8.83
DMLLR	1.55	2.09	3.60	8.75	28.12	8.82
MAPLR	1.53	2.05	3.45	8.37	27.40	8.56
DMAPLR	**1.51**	**2.03**	3.41	**8.16**	**26.94**	**8.41**
MLRLR	1.55	2.05	3.21	8.30	27.99	8.62

4.2.2. Discriminative MAP-based ESSEM

Figure 2 reports results of ML-based ESSEM, DML-based ESSEM, MAP-based ESSEM, and DMAP-based ESSEM. We use the same setups as that used in the previous experiment: for ML-based ESSEM, $\{h_1$=1.0, h_2=0.0, h_3=0.0$\}$; for DML-based ESSEM, $\{h_1$ =1.0, h_2=0.0, h_3=0.6$\}$; for MAPLR, $\{h_1$=1.0, h_2=0.2, h_3=0.0$\}$; for DMAPLR, $\{h_1$ =1.0, h_2 =0.2, h_3 =0.6$\}$. Similarly in this set of experiments, we set $\lambda_n = 1/N, \forall n$, N=8 in Eqs-(19) and (20).

From Figure 2, we observe very similar phenomena to Figure 1 for one, two, or three adaptation utterances: First, an obvious gain is obtained by incorporating the prior probability, (Ⓐ to Ⓒ), and an additional improvement is achieved by enhancing discrimination, (Ⓒ to Ⓓ). Second, by increasing discrimination, MAP-based estimation can achieve more improvements (Ⓒ to Ⓓ) than the ML-based estimation, (Ⓐ to Ⓑ). We also reported the results of ESSEM using one adaptation utterance in Table II. In addition to the four results in Figure 2, we tested the maximum likelihood ratio-based ESSEM (indicated as MLR-based ESSEM) with $\{h_1$=1.0, h_2=0.2, h_3=1.0$\}$.

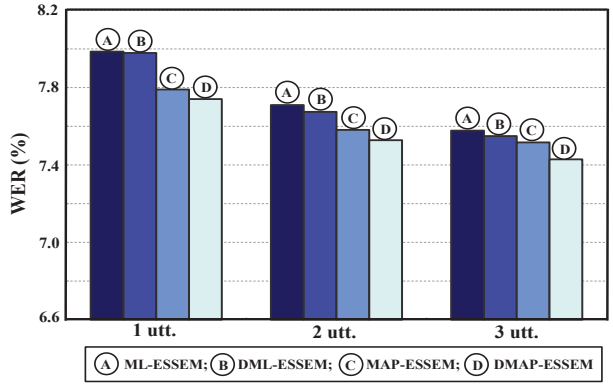

Figure 2. WERs (%) with different numbers of adaptation utterances

Similar observations are obtained to that from Table I: DMAP-ESSEM gives better performance than MAP-ESSEM, MLE-ESSEM, or MLR-based ESSEM consistently over almost all SNR conditions. The results again confirm that the objective function using a combined score performs better than those using individual scores. Comparing to Baseline, DMAP-ESSEM provides a performance improvement of 15.96% (9.21% to 7.74%) average WER reduction.

TABLE II. AVERAGE WER (%) FOR USING ONE ADAPTATION UTTERANCE

SNR(dB)	20	15	10	5	0	Ave.
Baseline	1.56	2.10	3.13	8.21	31.04	9.21
ML-ESSEM	1.31	**1.73**	2.96	7.87	26.06	7.99
DML-ESSEM	**1.30**	1.74	2.94	7.91	26.02	7.98
MAP-ESSEM	1.31	1.77	2.94	7.60	25.35	7.79
DMAP-ESSEM	1.31	1.77	**2.92**	**7.53**	**25.18**	**7.74**
MLR-ESSEM	1.32	1.88	2.95	**7.53**	25.26	7.79

5. CONCLUSIONS

We designed an objective function that uses a combined score of posterior probability and likelihood ratio. We used the proposed objective function to estimate mapping functions and developed DMAPLR and DMAP-based ESSEM. Compared to their ML-based, MAP-based, and MLR-based counterparts, DMAPLR and DMAP-based ESSEM provide notable improvements in a supervised adaptation mode on Aurora-2. The results indicate that increasing discriminative capability can effectively enhance MAP-based model adaptation performance. Moreover, from the improvements, it is confirmed that by using multiple knowledge sources in the objective function, the adaptation performance can be improved. Comparing to our baseline, DMAP-ESSEM achieves a 15.96% (9.21% to 7.74%) average WER reduction using only one adaptation utterance.

In the SME algorithm [3], a frame selection procedure is taken to improve the separations between parameters in acoustic models. In the future, we will incorporate that procedure into DMAPLR and DMAP-based ESSEM approaches. Meanwhile, some previous studies pointed out that by setting different weights for each λ_n (n=1…N) in Eqs-(13) and (14), (19), and (20), the discrimination can be further increased. We will also explore that research direction in the future.

6. REFERENCES

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

[2] D. Povey and P. C. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Proc. ICASSP*, pp. I105-I108, 2002.

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

[4] 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.

[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] 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.

[7] Y. Tsao, R. Isotani, H. Kawai, and S. Nakamura, "An environment structuring framework to facilitating suitable prior density estimation for MAPLR on robust speech recognition," to appear in *Proc. ISCSLP*, 2010.

[8] Y. Tsao, S. Matsuda, S. Nakamura, and C.-H. Lee, "MAP estimation of online mapping parameters in ensemble speaker and speaking environment modeling," in *Proc. ASRU,* pp. 271-275, 2009.

[9] J. Wu and Q. Huo, "A study of minimum classification error (MCE) linear regression for supervised adaptation of MCE-trained continuous-density hidden Markov models," *IEEE Trans. Speech Audio Processing*, vol. 15, pp. 478-488, 2007.

[10] S. Matsuda, Y. Tsao, J. Li, S. Nakamura, and C.-H. Lee, "A study on soft margin estimation of linear regression parameters for speaker adaptation," in *Proc. Interspeech*, *pp.* 1603-1606, 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] M.-W. Koo, C.-H. Lee, and B.-H. Juang, "Speech recognition and utterance verification based on a generalized confidence score," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 821-832, 2001.