UNSUPERVISED DISCRIMINATIVE ADAPTATION USING DIFFERENCED MAXIMUM MUTUAL INFORMATION BASED LINEAR REGRESSION

Marc Delcroix, Atsunori Ogawa, Seong-Jun Hahm, Tomohiro Nakatani, Atsushi Nakamura

NTT Communication Science Laboratories, NTT corporation,
2-4, Hikaridai, Seika-cho (Keihanna Science City), Soraku-gun, Kyoto 619-0237 Japan
{marc.delcroix,ogawa.atsunori,seongjun.hahm,nakatani.tomohiro,nakamura.atsushi}@lab.ntt.co.jp

ABSTRACT

This paper proposes a new approach for unsupervised model adaptation using a discriminative criterion. Discriminative criteria for acoustic model training have been widely used and have provided significantly improved performance compared with models trained using maximum likelihood. However, discriminative criteria are sensitive to errors in reference transcriptions, which limits their applicability to unsupervised adaptation. In this paper, we apply the recently proposed differenced maximum mutual information (dMMI) criteria to unsupervised linear regression based adaptation because dMMI has an intrinsic mechanism that mitigates the influence of transcription errors. We report unsupervised adaptation results for a large vocabulary continuous speech recognition task showing a significant improvement over maximum likelihood based linear regression.

Index Terms— Speech recognition, acoustic model adaptation, discriminative learning, unsupervised adaptation, differenced MMI

1. INTRODUCTION

Recent recognition systems usually use acoustic models [1] and language models [2] trained with discriminative criteria. The aim of using discriminative criteria is to achieve the direct optimization of the classification accuracy, by considering both references and competing recognition hypotheses. Therefore, such criteria are better related to the word error rate than the maximum likelihood (ML) criterion and have led to the consistent improvement of recognition accuracy. Many discriminative criteria have been proposed, including maximum mutual information (MMI) [3], minimum phone error (MPE) [4, 5], boosted MMI (BMMI) [6] and more recently *differenced MMI* (dMMI) [7, 8].

There have been several attempts to use discriminative criteria for acoustic model adaptation [9, 10, 11, 12, 13, 14, 15, 16]. Adapting an acoustic model to a desired speaker or environment is important if we are to compensate for the mismatch that occurs between training and test conditions. Using a discriminative criterion for adaptation is desirable because it may provide better recognition performance, and because it can preserve the discriminative capability of the acoustic models. However, it is challenging to apply discriminative criteria to *unsupervised adaptation*. Indeed, in this case, no reference transcriptions are available beforehand. The transcriptions must thus be estimated by a first recognition pass and therefore *the transcriptions inevitably contain errors*. Such errors may be challenging when using discriminative criteria that directly attempt to optimize classification accuracy, assuming correct labels. Some approaches have been proposed to mitigate this issue by focusing for the references on the adaptation data that are expected to be correctly recognized [14, 16]. This was achieved by weighting the MPE objective function by a word/phoneme correctness estimation obtained from a confusion network [14] or estimated using support vector machine [16].

In this paper, we propose using the recently reported dMMI criterion for unsupervised acoustic model adaptation based on linear regression (LR). We refer to the proposed method as *dMMI-LR*. dMMI generalizes the MPE and BMMI criteria [7]. It is defined as an integration of margin-based MPE [5] over a margin interval, and can be simply obtained as the difference between two BMMI objective functions [8]. dMMI appears well suited for unsupervised adaptation because it *defines references in a soft manner*, i.e. as a summation of recognition candidates weighted by a margin term, and therefore has an intrinsic mechanism that *mitigates the influence of transcription errors*. In contrast to [14, 16] it does not require an explicit estimation of word correctness. We have recently investigated the use of dMMI for training discriminative feature transforms [17] and for unsupervised dynamic variance adaptation [18], both applied to a small noisy command recognition task. Here we discuss its use for adaptation based on linear regression, which is both more general and widely used than [18]. Moreover, we present experimental results for a large vocabulary task.

The paper is organized as follows. In Section 2 we review the principles of LR based adaptation. We discuss the dMMI criterion in Section 3, and its application to LR adaptation in Section 4. Then we compare dMMI-LR with previous studies in Section 5. Finally, before concluding, we discuss experimental results for the MIT lecture recognition task.

2. ADAPTATION USING LINEAR REGRESSION

Adaptation using linear regression (LR) such as MLLR, has been widely used. It is very flexible and can be employed for speaker or environment adaptation [19]. LR adaptation consists of transforming the parameters of an acoustic model according to the following equation [20],

$$\hat{\boldsymbol{\mu}}_l = \mathbf{A}\boldsymbol{\mu}_l + \mathbf{b} = \mathbf{L}\boldsymbol{\xi}_l, \quad (1)$$

where $\hat{\boldsymbol{\mu}}_l$ is the compensated mean vector of the l^{th} Gaussian of the acoustic model, and $\mathbf{A}$ and $\mathbf{b}$ are a transformation matrix and a bias vector, respectively. The second part of Eq. (1) shows the simplified expression obtained by defining $\mathbf{L} \triangleq [\mathbf{A}\ \mathbf{b}]$ and $\boldsymbol{\xi}_l \triangleq [\boldsymbol{\mu}_l^\intercal 1]^\intercal$. In this paper we consider only the adaptation of the mean parameters of the Gaussians of the acoustic model although the proposed method could be extended to variance adaptation [19, 18].

The adaptation parameters $\mathbf{L}$ are usually shared among a cluster of Gaussians of the acoustic model, which is generated using a binary tree clustering of the Gaussians. C_k is the set of Gaussians belonging to the k^{th} cluster, and $\mathbf{L}_k$ represents the corresponding adaptation parameters. The set of all the adaptation parameters $\boldsymbol{\Theta} \triangleq [\mathbf{L}_1, \ldots, \mathbf{L}_K]$ is optimized as,

$$\hat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{\Theta}} \mathcal{F}_{\boldsymbol{\Theta}}(X, S_r), \tag{2}$$

where $\mathcal{F}_{\boldsymbol{\Theta}}(X, S_r)$ is an objective function, X is the set of feature vectors $\mathbf{x}_t$, i.e. $X \triangleq [\mathbf{x}_0, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T]$ and S_r is the corresponding reference transcription. To simplify the notations, we assume that X includes all the adaptation data. For unsupervised adaptation, S_r is obtained from a first recognition pass.

Conventional MLLR [20] uses the likelihood as an objective function. In this paper, we discuss the use of the dMMI criterion for the optimization.

3. DIFFERENCED MAXIMUM MUTUAL INFORMATION

dMMI was recently proposed for generalizing existing discriminative criteria such as MPE [4, 5] and BMMI [6]. It can be easily explained by defining a pseudo-probability function $\Psi_{\boldsymbol{\Theta},\sigma}$ as [21],

$$\Psi_{\boldsymbol{\Theta},\sigma}(X, S_r) \triangleq \sum_j P(S_j)^{\psi\eta} p_{\boldsymbol{\Theta}}(X|S_j)^{\psi} e^{\psi\sigma\mathcal{E}_{j,r}}, \tag{3}$$

where $\sum_j$ is a summation over recognition hypotheses S_j, $P(S_j)$ represents the language model, and $p_{\boldsymbol{\Theta}}(X|S_j)$ represents the acoustic model. ψ and η are the acoustic scaling [4] and language weight, respectively. The term $e^{\psi\sigma\mathcal{E}_{j,r}}$ represents a margin or boosting term [5, 6], where σ is a margin parameter and $\mathcal{E}_{j,r}$ represents the error between the recognition candidate S_j and the reference S_r, that is expressed in this paper by the phone frame error as defined in [22]. In the following, to simplify the notations, we drop the arguments (X, S_r) of $\Psi_{\boldsymbol{\Theta},\sigma}(X, S_r)$ and $\mathcal{F}_{\boldsymbol{\Theta}}(X, S_r)$.

We can see from Eq. (3) that by setting σ at a positive value the recognition candidates with large numbers of errors (i.e. $\mathcal{E}_{j,r} \gg$) will be emphasized in the summation. In contrast, by setting σ at a negative value, recognition candidates that are "close" to the reference (i.e. small $\mathcal{E}_{j,r}$) will be emphasized. For $\sigma \to -\infty$, only the term with $\mathcal{E}_{j,r} = 0$ (i.e. corresponding to the reference) remains in the summation of Eq. (3).

With the above definition, the objective function of BMMI can be expressed as [6, 21],

$$\begin{aligned}\mathcal{F}^{BMMI}_{\boldsymbol{\Theta},\sigma} &= \frac{1}{\psi} \log \frac{P(S_r)^{\psi\eta} p_{\boldsymbol{\Theta}}(X|S_r)^{\psi}}{\sum_j P(S_j)^{\psi\eta} p_{\boldsymbol{\Theta}}(X|S_j)^{\psi} e^{\psi\sigma\mathcal{E}_{j,r}}}, \\ &= \frac{1}{\psi} \log \frac{\Psi_{\boldsymbol{\Theta},-\infty}}{\Psi_{\boldsymbol{\Theta},\sigma}}. \end{aligned} \tag{4}$$

In a similar way, it is possible to express the MPE objective function as [21],

$$\mathcal{F}^{MPE}_{\boldsymbol{\Theta},\sigma} = \frac{1}{\psi} \frac{\frac{d}{d\sigma}(\Psi_{\boldsymbol{\Theta},\sigma})}{\Psi_{\boldsymbol{\Theta},\sigma}} \tag{5}$$

The objective function of dMMI is defined as the integration of the MPE loss (i.e. $-\mathcal{F}^{MPE}_{\sigma}(\boldsymbol{\Theta}, X, S_r)$) over a margin interval [8],

$$\begin{aligned}\mathcal{F}^{dMMI}_{\boldsymbol{\Theta},\sigma_1,\sigma_2} &= \frac{1}{(\sigma_2 - \sigma_1)} \int_{\sigma_1}^{\sigma_2} -\mathcal{F}^{MPE}_{\sigma} d\sigma \\ &= \frac{1}{(\sigma_2 - \sigma_1)} (\mathcal{F}^{BMMI}_{\sigma_2} - \mathcal{F}^{BMMI}_{\sigma_1}) \\ &= \frac{1}{\psi(\sigma_2 - \sigma_1)} \log \frac{\Psi_{\boldsymbol{\Theta},\sigma_1}}{\Psi_{\boldsymbol{\Theta},\sigma_2}}. \end{aligned} \tag{6}$$

Comparing Eqs. (4) and (6), we observe that the dMMI and BMMI objective functions can both be expressed as a ratio of two $\Psi_{\boldsymbol{\Theta},\sigma}$ functions with different margins. dMMI generalizes BMMI in the sense that we can choose any margin parameter value σ_1 for the numerator. By setting σ_1 at a negative value, the numerator of Eq. (6) becomes equivalent to the contribution of references defined in a soft manner, i.e. by considering the recognition candidates close to the reference. This soft definition of the references may mitigate the influence of transcription errors that inevitably occur when performing unsupervised adaptation.

Note that by setting σ_1 at a large negative value, dMMI becomes equivalent to BMMI [8]. In contrast, by setting $\sigma_1 = -\epsilon$ and $\sigma_2 = \epsilon$, where ϵ is a small positive number, dMMI becomes equivalent to MPE [8].

4. DMMI LINEAR REGRESSION

To solve the problem of Eq. (2) for the dMMI objective function, we use a gradient optimization method. The gradient is obtained with a lattice-based Forward-Backward algorithm as,

$$\frac{\partial \mathcal{F}^{dMMI}_{\boldsymbol{\Theta},\sigma_1,\sigma_2}}{\partial \mathbf{L}_k} = \sum_{q \in Q_t} \sum_{l \in C_{k,q}} \gamma^{dMMI}_{q,l,t} \Sigma_l^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_l) \boldsymbol{\xi}_l^\intercal, \tag{7}$$

where Q_t is the set of all lattice arcs that contain the feature vector $\mathbf{x}_t$, and $C_{k,q}$ is the set of Gaussians within arc q belonging to cluster C_k. $\gamma^{dMMI}_{q,l,t}$ is the product of the posterior probability of Gaussian l and the dMMI arc posterior probability or occupancy, calculated by running the Forward-Backward algorithm twice on the same lattice once with σ_1 and once with σ_2 [8]. Note that $\sum_{q \in Q_t} \gamma^{dMMI}_{q,l,t} \Sigma_l^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_l)$ corresponds to the statistics accumulated for the discriminative training of the acoustic models. Consequently, the gradient for dMMI-LR can be obtained by the simple modification of an existing gradient based discriminative training system. We use the RPROP algorithm for the gradient optimization [23] as it has been widely used for discriminative training [1].

For adaptation using binary cluster trees, the clusters are usually chosen by comparing the cluster occupancy count with an occupancy count threshold. When performing dMMI-LR, we use the occupancy count obtained with ML, since the occupancy count obtained with dMMI may take unrealistic values (i.e. negative values). This is equivalent to [10].

5. RELATION TO PREVIOUS STUDIES

dMMI-LR shares similarities with other approaches to unsupervised adaptation that attempt to mitigate the influence of transcription errors. For example, in [24, 25] lattice representation of the references was proposed. This achieves a similar soft definition of the references as dMMI-LR without considering the margin term. However, [24, 25] applied the method to MLLR and did not employed it with a discriminative criterion.

In [14, 16] errors in the reference transcriptions were mitigated by weighting the numerator MPE occupancies by using an estimate of the word correctness. This is therefore similar in principle to the smoothing effect of the margin term of the numerator of the dMMI objective function. However, for dMMI-LR, the smoothing is theoretically motivated and intrinsic to the objective function, whereas [14, 16] require a separate module to estimate the word correctness and are more heuristic. Moreover, [14, 16] employ different lattices generated with strong and weak language models for the numerator and denominator of the objective function. The proposed dMMI-LR uses the same lattices for the numerator and denominator, which may simplify the implementation.

6. EXPERIMENTS

We conducted experiments using the MIT lecture speech corpus [26, 27].

6.1. Experimental settings

The training data consists of 104 lectures corresponding to 110 hours of speech. The test data consist of development and evaluation sets that contain 2 and 8 lectures, respectively.

We used 12-dimension MFCCs with energy, delta and delta delta (39 dimensions in total). The features were processed with cepstral mean subtraction. The acoustic model consists of left-to-right phone HMMs with HMM state probability densities modeled by GMMs with 32 components. There were a total of 2546 context dependent states, which was automatically determined by variational Bayes [28]. The acoustic model was trained with dMMI with margin parameters $\sigma_1 = -2$ and $\sigma_2 = 3$. The language model consists of word trigram models trained using 6.2 M words of manually transcribed lecture speech. The vocabulary size of the lexicon is 16.5 K.

We performed unsupervised batch adaptation on a per lecture basis. This corresponds to speaker and environment adaptation. We used full transformation matrices and set the occupancy count threshold at 5000. For BMMI-LR and dMMI-LR we first adapted the acoustic model with MLLR to obtain good initial conditions. This is equivalent to the common practice in discriminative training of using an acoustic model trained with ML as the initial value. Similar approaches have also been previously used for discriminative adaptation [10, 13, 14, 16]. For a fair comparison of the results, we used the same dMMI baseline system with the same language model described above to generate recognition lattices and 1-best recognition results used for all the adaptation experiments. All the results are evaluated with respect to the word error rate (WER).

Table 1. WER for unsupervised adaptation for the development and evaluation sets of the MIT lecture speech corpus. The boldface fonts indicate the best performance.

	Dev.	Eval.
Baseline (dMMI AM)	36.7 %	30.6 %
Baseline + MLLR	34.5 %	26.9 %
Lattice-MLLR	34.7 %	26.8 %
BMMI-LR	34.6 %	27.1 %
dMMI-LR (proposed)	**33.6 %**	**25.8 %**

6.2. Results

Table 1 shows the WER of the development and evaluation sets for the baseline system trained with dMMI, and with adaptation using MLLR, Lattice-MLLR [24, 25], BMMI-LR and dMMI-LR. For MLLR, the EM algorithm is usually used [20] for the optimization. Here, to obtain consistent results, we used the same gradient optimizations in all cases. Note that in this task we observed that the gradient-based MLLR performed as well as or slightly better than EM-based MLLR.

For BMMI-LR and dMMI-LR we choose the margin parameters and the number of iterations to realize optimal performance on the development set (i.e. $\sigma = 0.1$ for BMMI, and $\sigma_1 = -10$, $\sigma_2 = 0.1$ for dMMI). For dMMI-LR we did not perform I-smoothing. MLLR achieved an absolute WER reduction of 2 to 4 points. Lattice-based MLLR slightly improved the performance for the evaluation set only. BMMI-LR provided no improvement in performance compared with MLLR. Here we used only the 1-best recognition results as reference transcriptions, and therefore transcription errors may prevent any improvement over MLLR. In contrast, dMMI provided an additional absolute WER reduction of 1 point for both the development and evaluation sets. We confirmed that dMMI-LR improved the performance compared with MLLR for all lectures with relative improvements ranging from 1 to 7.4 % depending on the lectures. We also confirmed that the improvement brought about by dMMI-LR over MLLR was significant according to the matched pair sentence segment test (significance level below 0.001) calculated using the NIST scoring toolkit [29].

6.3. Discussion

6.3.1. Influence of margin parameters

We set the margin parameters for dMMI to achieve optimal performance on the development set. We focused on the tuning of σ_1, which is related to the definition of the soft references (σ_2 was fixed at 0.1, according to some preliminary experiments).

Figure 1 shows the WER as a function of σ_1 for the development set. Note that a σ_1 value close to 0 is equivalent to MPE while $\sigma_1 \to -\infty$ (here $\sigma_1 = -50$) is equivalent to BMMI. We observed that improved performance could be achieved for a wide range of margin parameters, and that the best performance was obtained for an intermediate value of $\sigma_1 = -10$. This is consistent with our intuition that the soft references can improve performance for unsupervised adaptation. Note that the difference between the values of the margin parameters of dMMI used for acoustic model training and

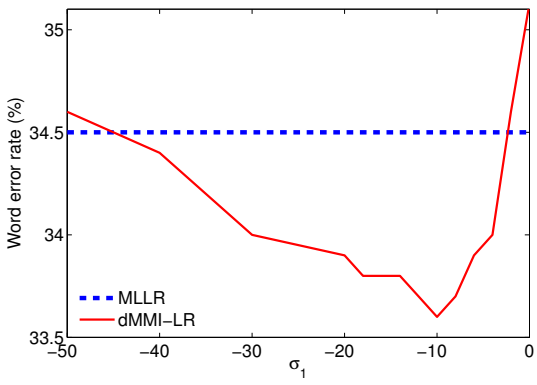

Fig. 1. WER as a function of σ_1 for $\sigma_2 = 0.1$ for the development set. For reference, the dashed line shows the performance obtained by MLLR.

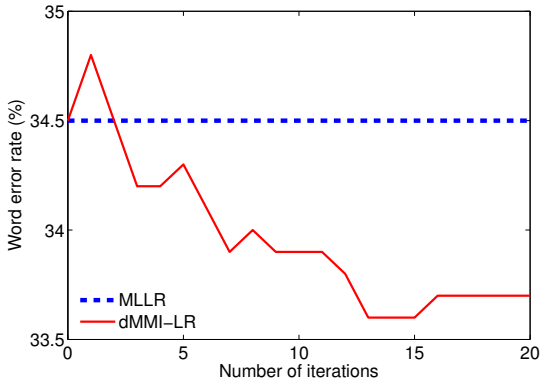

Fig. 2. WER as a function of the number of iterations for $\sigma_1 = -10$ and $\sigma_2 = 0.1$ for the development set. For reference, the dashed line shows the performance obtained by MLLR.

Table 2. WER as a function of the occupancy count threshold for MLLR and dMMI-LR with $\sigma_1 = -10$ and $\sigma_2 = 0.1$.

Occupancy Threshold		5000	1000	500
Nb transforms		≈80	≈400	≈800
Dev.	MLLR	34.5 %	34.6 %	34.6 %
	dMMI-LR	33.6 %	34.0 %	34.5 %
Eval.	MLLR	26.9 %	26.8 %	26.9 %
	dMMI-LR	25.8 %	26.3 %	26.7 %

adaptation can be explained by the fact that the training is supervised and therefore is less subject to transcription errors. Consequently a larger value of σ_1 is used for training (i.e. $\sigma_1 = -2$).

6.3.2. Influence of occupancy count threshold

Table 2 shows the WER for different occupancy count thresholds, for MLLR and dMMI-LR. The second line in Table 2 shows the approximate number of clusters used in the regression tree, which is equivalent to the estimated number of transforms. We found that the best performance was obtained for an occupancy threshold of 5000. With MLLR, the performance remained relatively stable as the number of transforms increased. This may be due to the gradient-based implementation of MLLR that does not require the matrix inversion of conventional EM-based MLLR [20], making it possible to use fewer data with each transform [13]. This intuition was confirmed by observing that with EM-based MLLR, the performance degradation increased as the number of transforms increased. The performance improvement realized by dMMI-LR decreased with increased complexity, i.e. an increased number of transforms. This is consistent with the widely known characteristics of discriminative training [1].

6.3.3. Convergence

Finally, Figure 2 shows the WER as a function of the number of iterations for dMMI-LR. We observed a relatively smooth convergence, and obtained optimal performance for about 15 iterations. Although we did not perform I-smoothing for dMMI-LR, no significant overfitting was observed at least for up to 20 iterations. We may expect the performance to degrade if we further increase the number of iterations.

7. CONCLUSION

In this paper, we discussed dMMI-based LR adaptation. We showed that dMMI has an intrinsic mechanism that mitigates the influence of transcription errors, making it particularly adequate for unsupervised adaptation. Results for a large vocabulary recognition task showed a significant improvement compared with MLLR. Future work may include combining dMMI-LR with approaches that incorporate estimated phoneme/word accuracy more directly as in [16, 30].

8. REFERENCES

[1] G. Heigold, H. Ney, R. Schluter, and S. Wiesler, "Discriminative training for automatic speech recognition: Modeling, criteria, optimization, implementation, and performance," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 58–69, 2012.

[2] H.-K. J. Kuo, E. Fosler-Lussier, H. Jiang, and C.-H. Lee, "Discriminative training of language models for speech recognition," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2002, vol. 1, pp. 325–328.

[3] A. Nádas, D. Nahamoo, and M.A. Picheny, "On a model-robust training method for speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 9, pp. 1432–1436, 1988.

[4] D. Povey and P.C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2002, vol. 1, pp. 105–108.

[5] G. Heigold, T. Deselaers, R. Schlüter, and H. Ney, "Modified MMI/MPE: a direct evaluation of the margin in speech recognition," in *Proc. International Conference on Machine Learning*, 2008, pp. 384–391.

[6] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2008, pp. 4057–4060.

[7] E. McDermott, S. Watanabe, and A. Nakamura, "Margin-space integration of MPE loss via differencing of MMI functionals for generalized error-weighted discriminative training," in *Proc. Interspeech*, 2009, pp. 224–227.

[8] E. McDermott, S. Watanabe, and A. Nakamura, "Discriminative training based on an integrated view of MPE and MMI in margin and error space," in *Proc. International Conference on Acoustics Speech and Signal Processing (ICASSP)*. IEEE, 2010, pp. 4894–4897.

[9] A. Gunawardana and W. Byrne, "Discriminative speaker adaptation with conditional maximum likelihood linear regression," in *Proc. Eurospeech*, 2001.

[10] L.F. Uebel and P.C. Woodland, "Discriminative linear transforms for speaker adaptation," in *Proc. ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*, 2001, pp. 61–64.

[11] D. Povey, P.C. Woodland, and M.J.F. Gales, "Discriminative MAP for acoustic model adaptation," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2003, vol. 1, pp. 312–315.

[12] J.-T. Chien and C.-H. Huang, "Aggregate a posteriori linear regression adaptation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 797–807, 2006.

[13] 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 Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 478–488, 2007.

[14] L. Wang and P. C. Woodland, "MPE-based discriminative linear transforms for speaker adaptation," *Computer Speech & Language*, vol. 22, no. 3, pp. 256–272, 2008.

[15] 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*, 2009, pp. 1603–1606.

[16] M. Gibson and T. Hain, "Correctness-adjusted unsupervised discriminative acoustic model adaptation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2648–2656, 2012.

[17] M. Delcroix, A. Ogawa, S. Watanabe, T. Nakatani, and A. Nakamura, "Discriminative feature transforms using differenced maximum mutual information," in *Proc. Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4753–4756.

[18] Delcroix M., Ogawa A., Nakatani T., and Nakamura A., "Dynamic variance adaptation using differenced maximum mutual information," in *Proc. Symposium on Machine Learning in Speech and Language Processing (MLSLP)*, 2012.

[19] M.J.F. Gales and P.C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech & Language*, vol. 10, no. 4, pp. 249–264, 1996.

[20] Leggetter C.J. and Woodland P.C., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[21] A. Nakamura, E. McDermott, S. Watanabe, and S. Katagiri, "A unified view for discriminative objective functions based on negative exponential of difference measure between strings," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2009, pp. 1633–1636.

[22] J. Zheng and A. Stolcke, "Improved discriminative training using phone lattices," in *Proc. Interspeech*, 2005, pp. 2125–2128.

[23] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: the RPROP algorithm," in *Proc. International Conference on Neural Networks*. IEEE, 1993, pp. 586–591.

[24] M. Padmanabhan, G. Saon, and G. Zweig, "Lattice-based unsupervised MLLR for speaker adaptation," in *Proc. ISCA Tutorial and Research Workshop (ITRW) on Automatic Speech Recognition: Challenges for the New Millenium*, 2000, pp. 128–131.

[25] L.F. Uebel and P.C. Woodland, "Improvements in linear transform based speaker adaptation," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2001, vol. 1, pp. 49–52.

[26] H.-A. Chang and J.R. Glass, "Discriminative training of hierarchical acoustic models for large vocabulary continuous speech recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2009, pp. 4481–4484.

[27] J. Glass, T. J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent progress in the MIT spoken lecture processing project," in *Proc. Interspeech*, 2007, pp. 2553–2556.

[28] S. Watanabe, A. Sako, and A. Nakamura, "Automatic determination of acoustic model topology using variational Bayesian estimation and clustering for large vocabulary continuous speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 855 – 872, 2006.

[29] J. Fiscus, "sc_stats - SCLITE's statistical system comparison program," in `ftp://jaguar.ncsl.nist.gov/current_docs/sctk/doc/sc_stats.htm`, Cited Nov. 11 2012.

[30] A. Ogawa, T. Hori, and A. Nakamura, "Error type classification and word accuracy estimation using alignment features from word confusion network," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4925–4928.