# DISCRIMINATIVE FEATURE TRANSFORMS USING DIFFERENCED MAXIMUM MUTUAL INFORMATION

*Marc Delcroix, Atsunori Ogawa, Shinji Watanabe*, 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,nakatani.tomohiro,nakamura.atsushi}@lab.ntt.co.jp

## ABSTRACT

Recently feature compensation techniques that train feature transforms using a discriminative criterion have attracted much interest in the speech recognition community. Typically, the acoustic feature space is modeled by a Gaussian mixture model (GMM), and a feature transform is assigned to each Gaussian of the GMM. Feature compensation is then performed by transforming features using the transformation associated with each Gaussian, then summing up the transformed features weighted by the posterior probability of each Gaussian. Several discriminative criteria have been investigated for estimating the feature transformation parameters including maximum mutual information (MMI) and minimum phone error (MPE). Recently, the differenced MMI (dMMI) criterion that generalizes MMI and MPE, has been shown to provide competitive performance for acoustic model training. In this paper, we investigate the use of the dMMI criterion for discriminative feature transforms and demonstrate in a noisy speech recognition experiment that dMMI achieves recognition performance superior to that of MMI or MPE.

***Index Terms***— Speech recognition, discriminative training, discriminative feature transforms, differenced MMI

## 1. INTRODUCTION

The use of discriminative criteria for training automatic speech recognition (ASR) systems has become a standard technique. Indeed, the optimization of such criteria is better correlated to recognition error reduction than standard maximum likelihood (ML) leading to a consistent improvement in speech recognition accuracy. Work on discriminative training approaches started with acoustic model training [1, 2] and was then extended to language model training [3] and more recently to feature extraction [4, 5, 6]. In particular, the use of discriminative training for feature transforms has recently attracted much attention, because of the significant recognition performance improvement achieved for many speech recognition tasks [4, 5, 6, 7]. These approaches share the same concept of using a Gaussian mixture model (GMM) to model the feature space and associate feature transformation parameters with each Gaussian of the GMM. A compensated feature vector is obtained by transforming an input feature vector with the transform associated with each Gaussian of the GMM, then summing up the transformed features weighted by the posterior probability of each Gaussian. This makes it possible to employ different transforms for each region of the feature space [6, 8]. The parameters of the transform associated with each Gaussian is trained using a discriminative criterion.

Many different discriminative criteria have been proposed for training acoustic models and feature transform parameters such as maximum mutual information (MMI) [1, 5], minimum classification error (MCE) [9], minimum phone error (MPE) [2, 4] or boosted MMI (BMMI) [7]. BMMI modifies the MMI criterion by incorporating margins into the denominator (corresponding to the competitor contribution) of the MMI objective function and a boosting factor, further called margin parameter. Recently, a new discriminative criterion called differenced MMI (dMMI) was proposed to generalize MPE and BMMI [10]. The objective function of dMMI is defined as the difference between two BMMI objective functions with two different margin parameters therefore combining the regularization benefits of BMMI with a loose definition of references [9, 11]. It was shown in [10] that the dMMI objective function can be derived from the integration of an MPE objective function over a margin interval. Consequently depending on the values of the margin parameters, dMMI becomes equivalent to MMI/BMMI or MPE.

The dMMI criterion has been shown to achieve competitive performance in various tasks when used for training acoustic models [10]. In this paper we investigate the use of the dMMI discriminative criterion for training the feature transform parameters. We demonstrate experimentally that dMMI is more robust to mismatches between training and testing conditions, and can provide superior recognition performance compared with conventional approaches such as MMI (MMI-SPLICE [5]), MPE (fMPE [4]) and BMMI (fBMMI [7]). In a similar way to that employed with MMI-SPLICE, we used a noisy speech recognition task to evaluate our proposal [5]. In this paper, we use the PASCAL-CHiME challenge task, which consists of speech command recognition in the presence of highly non-stationary noise [12].

The organization of the paper is as follows. In section 2 we review the principle of discriminative feature transforms and derive the transform parameters estimation using the dMMI criterion. We then present some experimental results comparing dMMI and MMI in section 3.

## 2. DISCRIMINATIVE FEATURE TRANSFORMS

There have been several proposals regarding the implementation of discriminative feature transforms [4, 5, 6]. These approaches share the common idea of transforming input feature vectors $\mathbf{o}_t$ given

*Shinji Watanabe is now with Mitsubishi Electric Research Laboratories (MERL), watanabe@merl.com.

some affine transforms and some posterior probabilities. The parameters of the affine transforms are estimated using discriminative training. The posterior probabilities are usually derived from a GMM that is trained on the training data. A general formulation of the discriminative feature transform can be expressed as [8],

$$\mathbf{x}_t = \sum_k p(k|\mathbf{o}_t)(\mathbf{A}_k\mathbf{o}_t + \mathbf{m}_k), \quad (1)$$

where $\mathbf{x}_t$ is the transformed feature for time frame $t$, $p(k|\mathbf{o}_t)$ is the posterior probability of a GMM component $k$, given the input feature vector $\mathbf{o}_t$, $\mathbf{A}_k$ is a transformation matrix and $\mathbf{m}_k$ is a bias vector. Note that eq. (1) is performed for each time frame, which enables frame level feature compensation. It is possible to introduce context information by incorporating adjacent features [4].

In this work, we follow the implementation of MMI-SPLICE [5] where only the bias term $\mathbf{m}_k$ is considered and thus assume a unity matrix for $\mathbf{A}_k$. Moreover, no context information is included and $\mathbf{o}_t$ consists of a conventional MFCC feature vector. The main difference between this paper and [5] is that we use the dMMI criterion for estimating the bias parameters $\mathbf{m}_k$ rather than MMI.

### 2.1. Differenced MMI

dMMI is a recently proposed discriminative criterion that generalizes the MPE and MMI/BMMI criteria. The dMMI objective function can be derived from the objective function of BMMI, which is defined as [7],

$$\mathcal{F}^{BMMI}_{\mathbf{\Lambda},\sigma} = \frac{1}{\psi}\log\frac{P(S_r)^{\psi\eta}p_{\mathbf{\Lambda}}(X_r|S_r)^{\psi}}{\sum_j P(S_j)^{\psi\eta}p_{\mathbf{\Lambda}}(X_r|S_j)^{\psi}e^{\psi\sigma\mathcal{E}_{j,r}}}, \quad (2)$$

where $X_r$ is the sequence of feature vectors for the training data, $S_r$ is the corresponding reference transcription and $S_j$ is a recognition candidate for $X_r$. $\mathcal{E}_{j,r}$ represents the error between the recognition candidate $S_j$ and the reference $S_r$. $p_{\mathbf{\Lambda}}(X_r|S_r)$ corresponds to the acoustic model, which is represented here by hidden Markov models (HMMs) with HMM state posterior probability modeled by GMMs. $\mathbf{\Lambda}$ represents the acoustic model parameters. The parameter $\eta$ is the language model scaling and $\psi$ is the acoustic scaling [4]. Note that to simplify the expressions in eq. (2) we omitted the summation over the training utterances. The numerator of the BMMI objective function corresponds to the contribution to the correct reference transcription, and the denominator accounts for the contribution of the competing recognition candidates. BMMI includes a margin term with parameter $\sigma$ in the denominator. The error term, $\mathcal{E}_{j,r}$, can be defined as the phone error, word error or phone frame error. In the following, we use the phone frame error as defined in [13].

Eq. (2) can be simplified if we introduce a function $\Psi_{\mathbf{\Lambda},\sigma}$ as,

$$\Psi_{\mathbf{\Lambda},\sigma} \triangleq \sum_j P(S_j)^{\psi\eta}p_{\mathbf{\Lambda}}(X_r|S_j)^{\psi}e^{\psi\sigma\mathcal{E}_{j,r}}. \quad (3)$$

Given $\Psi_{\mathbf{\Lambda},\sigma}$ the BMMI objective function can be expressed as [11] [1],

$$\mathcal{F}^{BMMI}_{\mathbf{\Lambda},\sigma} = \frac{1}{\psi}\log\frac{\Psi_{\mathbf{\Lambda},-\infty}}{\Psi_{\mathbf{\Lambda},\sigma}}. \quad (4)$$

In a similar way, the objective function of conventional MMI (without a margin) is given by,

$$\mathcal{F}^{MMI}_{\mathbf{\Lambda}} = \frac{1}{\psi}\log\frac{\Psi_{\mathbf{\Lambda},-\infty}}{\Psi_{\mathbf{\Lambda},0}}. \quad (5)$$

[1] Note that 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).

The dMMI objective function further generalizes the MMI objective function. It is defined as the difference between two BMMI objective functions with different margin parameters and can be expressed as [10],

$$\begin{aligned}\mathcal{F}^{dMMI}_{\mathbf{\Lambda},\sigma_1,\sigma_2} &= \frac{1}{\psi(\sigma_2-\sigma_1)}\log\frac{\sum_j P(S_j)^{\psi\eta}p_{\mathbf{\Lambda}}(X_r|S_j)^{\psi}e^{\psi\sigma_1\mathcal{E}_{j,r}}}{\sum_j P(S_j)^{\psi\eta}p_{\mathbf{\Lambda}}(X_r|S_j)^{\psi}e^{\psi\sigma_2\mathcal{E}_{j,r}}},\\ &= \frac{1}{\psi(\sigma_2-\sigma_1)}\log\frac{\Psi_{\mathbf{\Lambda},\sigma_1}}{\Psi_{\mathbf{\Lambda},\sigma_2}}. \quad (6)\end{aligned}$$

dMMI includes margins in both numerator and denominator terms. For negative $\sigma_1$, the numerator emphasizes low error recognition candidates (close to the reference) and therefore plays a similar role to the numerator term of the BMMI objective function of eq. (4). By setting $\sigma_2$ at a positive value, the denominator accentuates the contribution of recognition candidates with a high error rate.

It is possible to show that dMMI becomes equivalent to MPE when the margin interval between $\sigma_1$ and $\sigma_2$ becomes narrow around 0 i.e. [10],

$$\lim_{\sigma_1=\sigma_2\to 0}\mathcal{F}^{dMMI}_{\mathbf{\Lambda},\sigma_1,\sigma_2} = \mathcal{F}^{MPE}_{\mathbf{\Lambda}}. \quad (7)$$

Moreover, comparing eq. (6) and eq. (4) we can see that dMMI becomes equivalent to BMMI when $\sigma_1$ tends to have large negative values. Therefore by setting appropriate values for $\sigma_1$ and $\sigma_2$ we can approach the objective functions of MPE, MMI or BMMI.

### 2.2. dMMI for training feature transforms

We propose estimating the set of bias vector parameters $\theta = \{\mathbf{m}_k\}$ of eq. (1) using the dMMI criterion. $\theta$ can be obtained by maximizing the dMMI objective function as,

$$\hat{\theta} = \arg\max_{\theta}\mathcal{F}^{dMMI}_{\mathbf{\Lambda},\sigma_1,\sigma_2}(X_r(\theta)). \quad (8)$$

Eq. (8) can be optimized using the gradient ascent method. The gradient can be expressed as,

$$\frac{\partial\mathcal{F}^{dMMI}_{\mathbf{\Lambda},\sigma_1,\sigma_2}(X_r(\theta))}{\partial\mathbf{m}_k} = \sum_t\frac{\partial\mathcal{F}^{dMMI}_{\mathbf{\Lambda},\sigma_1,\sigma_2}(X_r(\theta))}{\partial\mathbf{x}_t}\frac{\partial\mathbf{x}_t}{\partial\mathbf{m}_k}, \quad (9)$$

where $\mathbf{x}_t$ is a feature vector for time frame $t$ as defined in eq. (1).

By calculating the gradient over lattices, we can express it as,

$$\begin{aligned}\frac{\partial\mathcal{F}^{dMMI}_{\mathbf{\Lambda},\sigma_1,\sigma_2}}{\partial\mathbf{m}_k} &= -\sum_{t,q_t,n_t,m}p(k|\mathbf{o}_t)\gamma^{dMMI}_{q_t}\\ &\quad \gamma_{n_t,m}(t)\mathbf{\Sigma}^{-1}_{n_t,m}(\mathbf{x}_t-\boldsymbol{\mu}_{n_t,m}), \quad (10)\end{aligned}$$

where $\sum_{t,q_t,n_t,m}$ is a summation over the time index, the corresponding lattice arcs, the associated HMM states and the GMM mixture components. $\gamma^{dMMI}_q$ is equivalent to the arc posterior probability or occupancy, calculated by running the Forward-Backward algorithm twice on the same lattice once with $\sigma_1$ and once with $\sigma_2$ [10]. $\mathbf{\Sigma}_{n,m}$, $\boldsymbol{\mu}_{n,m}$ and $\gamma_{n,m}(t)$ are the covariance matrix, mean vector and the posterior probability, respectively, of the GMM mixture component $m$ of the HMM state $n$.

Eq. (10) is similar to MMI-SPLICE; the main difference is the calculation of $\gamma^{dMMI}_q$. Note that here we do not carry out I-smoothing on the gradient. In this paper, we use the RPROP [14] algorithm for gradient optimization.

## 3. EXPERIMENTS

Here we describe some preliminary experiments for a noisy command recognition task.

### 3.1. Settings

We tested dMMI based feature transforms on the CHiME noisy keyword recognition task. The CHiME task consists of 6-word commands spoken by 34 English speakers. The commands are corrupted by background noise that was collected in a real living room. The noise is highly non-stationary and includes noise sources such as TV, children's voices or music. The recognition target consists of two keywords consisting of a letter followed by a digit, which are included in the command. The training data consist of 17,000 utterances and 6 hours of background noise data. The training utterances are corrupted by reverberation but do not include noise. The test data consist of a development set and an evaluation set that both include 600 reverberant utterances at 6 different SNRs ranging from -6 to 9 dB. Note that the training data set and the test data sets all consist of reverberant speech for the same room (reverberation time of 300 msec.) but with different speaker positions and room configurations (doors open/close ...) and therefore with different reverberant characteristics. A detailed description of the CHiME task can be found in [12].

We used the DOLPHIN enhancement algorithm to extract the target speech from the noisy signals [15]. DOLPHIN is a recently proposed algorithm that performs speech-noise separation using spatial and spectral information about speech and noise. We created multi-condition training data by adding background noise samples to the training data set. The amount of training data was 42 times the amount of clean training data (seven noise environments obtained from the background noise data provided by the CHiME challenge by six SNR levels). The multi-condition noisy speech data were then processed with the DOLPHIN enhancement algorithm. The obtained multi-condition training data were used to train acoustic models. We used the speech recognizer platform SOLON [16], which was developed at NTT Communication Science Laboratories, to train the acoustic model and perform decoding. The acoustic models consisted of conventional left-to-right HMMs with a total of 253 states each modeled by a GMM consisting of 20 Gaussian components. We trained speaker dependent acoustic models according to the CHiME regulation using the ML criterion. The GMM trained for feature transforms was a speaker independent GMM. The number of Gaussian components of the GMM was set at 512.

The results presented in this paper are expressed in terms of keyword error rate averaged over the 34 speakers and 6 SNR conditions. The keyword error rate of the noisy speech was 16.8 % and 15.3 % for the development and evaluation sets, respectively, using the multi-condition acoustic models. Using DOLPHIN, the keyword error rate was 12.8 % for the development set and 11.4 % for the evaluation set. Note that in the experiments below we did not use I-smoothing.

### 3.2. Experimental results

Figure 1 plots the average keyword error rate as a function of the number of iterations for the *development* set. The results are given for the ML, MMI (as in eq. (5)) and the dMMI criterion of eq. (6) with different values for the margin parameters $\sigma_1$ and $\sigma_2$. We fixed $\sigma_2 = 0.1$[2] and evaluated the performance for different $\sigma_1$ values ranging from -20 to -0.1. Setting $\sigma_1$ at a large negative value (here -20), enables us to approach the BMMI criterion of eq. (4). When we set $\sigma_1 = -0.1$, dMMI approaches the MPE criterion [10]. Finally, with the intermediate cases of $\sigma_1 = -1$ and $\sigma_1 = -3$, we observe the effect of both margins for the numerators and denominator.

[2]In this experiment, although we did not tune $\sigma_2$ for this task, we followed recommendation in [7] mentioning that for training feature transforms a small margin parameter between 0 and 0.5 tends to give good performance.

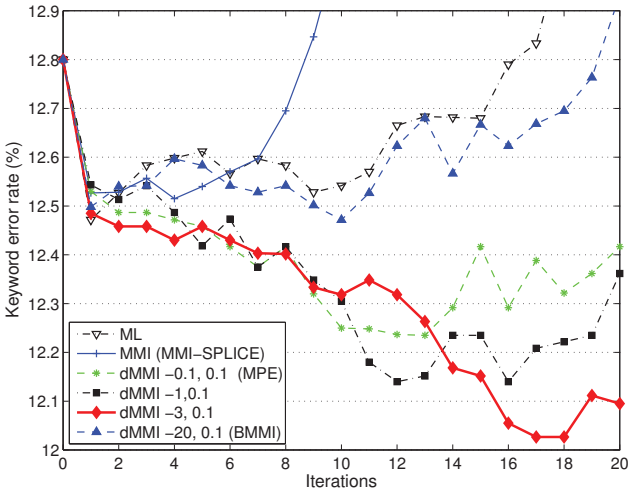


**Fig. 1**. Results of the CHiME *development* test set when using ML, MMI and dMMI criteria to train feature transforms. For dMMI, the legend is as "dMMI $\sigma_1, \sigma_2$".

Using the ML criterion to calculate the feature transforms already provides some small improvement between 0.2 and 0.3 %. With MMI, the keyword recognition error starts by decreasing by up to 0.3 %, and then it rapidly increases because of overtraining. These results show the same tendency as MMI-SPLICE presented in [5]. Note that the overtraining could be mitigated to some extent if we used I-smoothing. However, even with I-smoothing, strong overtraining has also been reported for other tasks [5]. For MMI, we limited the number of iterations to 10 because by that number the performance had already degraded significantly compared with the initial baseline value and we cannot therefore expect any gains with more iterations. For BMMI (i.e. dMMI $-20, 0.1$), performance close to MMI was observed but the overtraining problem was reduced. This may be because of the regularization effect role of the margins in the BMMI objective function.

dMMI with a larger $\sigma_1$ ($\sigma_1 = -0.1, -1, -3$) achieves a larger keyword recognition error reduction and shows less overtraining. The dMMI criterion considers a set of correct recognition hypotheses instead of a single hypotheses [11]. This may mitigate the uncertainty in the reference transcription and reduce the influence of "don't care" variations around target keywords [9]. $\sigma_1$ controls the number of hypotheses considered in the numerator. Note that dMMI appears to perform better than MPE (i.e. dMMI $-0.1, 0.1$) especially in the case of $\sigma_1 = -3$, which leads to an absolute keyword recognition error reduction of about $0.8\%$.

Figure 2 plots the average keyword error rate as a function of the number of iterations for the *evaluation* set for ML, MMI and dMMI with the same range of margin parameters as for the development set. Note that the differences between the development and evaluation sets of the CHiME task originate mainly from the room impulse responses and the background noise. For this test set, we observe that discriminative approaches (MMI and MPE (i.e. dMMI $-0.1, 0.1$)) provide less improvement than that observed for the development set. This seems to indicate a larger difference between

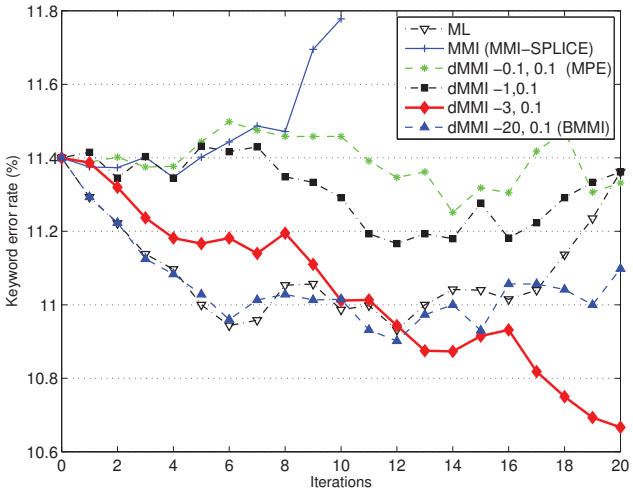


**Fig. 2**. Results of the CHiME *evaluation* test set when using ML, MMI and dMMI criteria to train feature transforms. For dMMI, the legend is as "dMMI $\sigma_1, \sigma_2$".

noise and reverberation conditions of the training data and the test data for the evaluation set than for the development set[3]. In this case, discriminative training approaches may be less effective than ML because they tend to overfit the model to the training data. In contrast ML tends to have a better generalization capability, which explains its superior performance in this case.

When margins are incorporated into the objective function, better performance can be achieved. In particular, as with the development set, we observe that including margins for both the numerator and denominator terms in the dMMI criterion can improve performance with limited overtraining. Here again dMMI with margin parameter $\sigma_1 = -3$ provided optimal performance with an absolute keyword recognition reduction of close to 0.8 %. For this task, dMMI with margin parameter $\sigma_1 = -3$ achieves a significant performance improvement for both the development and evaluation sets, and is more robust than the MPE and BMMI criteria.

## 4. CONCLUSION

In this paper, we investigated the use of the newly proposed dMMI criterion for training discriminative feature transforms. The dMMI criterion includes margins on both numerator and denominator terms of the objective function that enable us to combine the benefit provided by the BMMI margins and the advantage of considering multiple reference candidates for the numerator term. We demonstrated in a preliminary experiment that both margins play important roles, and that dMMI could achieve performance superior to MPE and BMMI when the margin parameters were appropriately chosen. These preliminary results seem to demonstrate the potential of the dMMI criterion for training feature transforms. Future work will include further testing with large vocabulary continuous speech recognition tasks as well as including the context in the feature transformation definition.

[3]The same phenomenon was observed with MMI-SPLICE on the Aurora 2 database for test set B, which contains unseen noise types [5].

## 5. REFERENCES


[1] Nadas, A., Nahamoo, D. and Picheny, M. A., "On a model robust training method for speech recognition," IEEE Trans. ASSP, vol. 39, no. 9, pp. 1432 - 1435, 1988.

[2] Povey, D. and Woodland, P., "Minimum phone error and I-smoothing for improved discriminative training," In Proc. ICASSP'02, vol. 1, pp. 105-108 , 2002.

[3] Kuo, H.-K. J., Fosler-Lussier, E., Jiang, H. and Lee, C.-H., "Discriminative training of language models for speech recognition," In proc. ICASSP'02, vol.1, pp. 325-328, 2002.

[4] Povey, D., "fMPE: discriminatively trained features for speech recognition," In Proc. ICASSP'05, pp. 961-964, 2005.

[5] Droppo, J. and Acero, A., "Maximum mutual information SPLICE transform for seen and unseen conditions," In Proc. Interspeech'05, pp. 989-992, 2005.

[6] Zhang, B., Matsoukas, S. and Schwartz, R., "Discriminatively trained region dependent transforms for speech recognition," In Proc. ICASSP'06, pp. 313-316, 2006.

[7] Povey, D. , Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G. and Visweswariah, K., "Boosted MMI for model and feature-space discriminative training," In Proc. ICASSP'08, pp. 4057-4060, 2008.

[8] Deng, L., Wu, J., Droppo, J. and Acero, A., "Analysis and comparison of two speech feature extraction/compensation algorithms," IEEE Signal Processing Letters, vol. 12, no. 6, pp. 477-480, 2005.

[9] McDermott, E., Hazen, T.J., Le Roux, J., Nakamura, A. and Katagiri, S., "Discriminative training for large vocabulary speech recognition using Minimum Classification Error," IEEE Trans. ASLP, vol. 15, no. 1, pp. 203-223, 2007.

[10] McDermott, E., Watanabe, S. and Nakamura, A., "Discriminative training based on an integrated view of MPE and MMI in margin and error space," In Proc. ICASSP'10, pp. 4894 - 4897, 2010.

[11] Nakamura, A., McDermott, E., Watanabe, S. and Katagiri, S., "A unified view for discriminative objective functions based on negative exponential of difference measure between strings," In Proc. ICASSP'09, pp. 1633-1636, 2009.

[12] Christensen, H., Barker, J., Ma, N., and Green, P., "The CHiME corpus: a resource and a challenge for computational hearing in multisource environments," In Proc. Interspeech'10, pp. 1918-1921, 2010.

[13] Zheng, J. and Stolcke, A., "Improved discriminative training using phone lattices," In Proc. Interspeech'05, pp. 2125-2128, 2005.

[14] Riedmiller, M. and Braun, H., "A direct adaptive method for faster backpropagation learning: The RPROP algorithm," In Proc. ICNN'93, pp. 586-591, 1993.

[15] Nakatani, T., Araki, S., Delcroix, M., Yoshioka, T. and Fujimoto, M., "Reduction of highly nonstationary ambient noise based on spectral and locational characteristics of speech and noise for robust ASR," In Proc. Interspeech'11, 2011.

[16] Hori, T., Hori, C., Minami, Y. and Nakamura, A., " Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," IEEE Trans. ASLP, vol. 15, no. 4, pp. 1352-1365, 2007.