TIED-STATE BASED DISCRIMINATIVE TRAINING OF CONTEXT-EXPANDED REGION-DEPENDENT FEATURE TRANSFORMS FOR LVCSR

Zhi-Jie Yan *Qiang Huo* *Jian Xu** *Yu Zhang**

Microsoft Research Asia, Beijing, China
{zhijiey, qianghuo}@microsoft.com

ABSTRACT

We present a new discriminative feature transform approach to large vocabulary continuous speech recognition (LVCSR) using Gaussian mixture density hidden Markov models (GMM-HMMs) for acoustic modeling. The feature transform is formulated with a set of context-expanded region-dependent linear transforms (RDLTs) utilizing both long-span features and contextual weight expansion. The RDLTs are estimated by lattice-free, tied-state based discriminative training using maximum mutual information (MMI) criterion, while the GMM-HMMs are trained by conventional lattice-based, boosted MMI training. Compared with two baseline systems, which use RDLTs with either long-span features or weight expansion only and are trained using the conventional lattice-based discriminative training for both RDLTs and HMMs, the proposed approach achieves a relative word error rate reduction of 10% and 6% respectively on Switchboard-1 conversational telephone speech transcription task.

Index Terms— region-dependent linear transform, maximum mutual information, discriminative training, tied-state, HMM

1. INTRODUCTION

In the past decade, several discriminatively trained feature transform approaches have been successfully used in large vocabulary continuous speech recognition (LVCSR) systems using Gaussian mixture density hidden Markov models (GMM-HMMs) for acoustic modeling. Among them, feature-space minimum phone error (fMPE) [1,2] and region-dependent linear transform (RDLT) [3, 4] are two most popular methods. The fMPE method transforms each input feature vector by using a bias term projected from a high-dimensional space and uses weight expansion (WE) to incorporate contextual information (a.k.a. acoustic context expansion in [1, 2], which calculates several averaged posterior probability vectors from neighboring frames using a Gaussian codebook to derive the weights in fMPE). The RDLT method applies, on a long-span input feature vector, a weighted-sum of a set of linear transforms, with each transform corresponding to a "region" in the original feature space. Hereinafter, we call this RDLT approach the feature expansion based RDLT (FE-RDLT). Recently, another variant of RDLT was proposed [5, 6], which uses the same contextual weight expansion method as in fMPE and a full linear transform on a single frame of input feature vector. Hereinafter, we call this RDLT approach the weight expansion based RDLT (WE-RDLT). Our comparative study of the above three discriminatively trained feature transform approaches in [7] suggests that WE-RDLT and fMPE achieve similar, but better recognition accuracies than that of FE-RDLT approach, yet WE-RDLT is computationally more efficient than fMPE at run-time because less number of regions can be used in WE-RDLT than in fMPE to achieve similar recognition accuracy. In this paper, we generalize the above methods by using a set of context-expanded RDLTs utilizing both long-span features and contextual weight expansion, therefore we refer to this new feature transform as CE-RDLT hereinafter. The first motivation of this study is to investigate the effectiveness of CE-RDLT.

Recently, acoustic modeling with deep neural network based HMMs (DNN-HMMs) has been demonstrated to achieve significant word error rate (WER) reduction against the state-of-the-art GMM-HMM based systems for different LVCSR tasks and from several research groups (e.g., [8–12]). In addition to using long-span features as the input of the DNN and nonlinear feature mapping by several layers of feed-forward neural networks, an important difference between most of the DNN-HMM based systems and the conventional GMM-HMM based systems is that the former are trained by using lattice-free discriminative training (DT) for isolated tied-state classification, while the later are typically trained by using lattice-based DT for sequence classification. The only exception is the work done by IBM researchers, where a lattice-based DT for tied-state sequence classification is used to train shallow neural network based HMMs [13] and then DNN-HMMs [11], with additional WER reduction achieved against the DNN-HMMs trained using lattice-free, tied-state based DT. To the best of our knowledge, there is no study reported in the literature yet to compare the effectiveness of the above two different ways of training feature transforms and HMMs in GMM-HMM based acoustic modeling framework. So the second motivation of this paper is to share our research findings regarding this aspect. Interestingly, a lattice-free training approach was used recently to train a log-linear model for LVCSR [14], but their work is completely different from ours presented in this paper.

The rest of this paper is organized as follows. In Section 2, we describe our CE-RDLT transform. In Sections 3 and 4, a lattice-free, tied-state based MMI criterion and the corresponding optimization method for training transform parameters are presented, respectively. Experimental results and analysis are reported in Section 5. Finally, we conclude the paper in Section 6.

2. CONTEXT-EXPANDED REGION-DEPENDENT FEATURE TRANSFORM

Our CE-RDLT method transforms the feature vector $\mathbf{o}_t$ in the *original feature space* to $\hat{\mathbf{o}}_t$ in a *new feature space* as follows:

$$\hat{\mathbf{o}}_t = \sum_{m=1}^{M} \kappa_{m,t} \cdot \mathbf{W}_m \xi_t \ , \tag{1}$$

*Jian Xu and Yu Zhang contributed to this work as interns of the Speech Group at Microsoft Research Asia.

where $\kappa_{m,t}$ is the weight of the m^{th} transform $\mathbf{W}_m$ at time t, and ξ_t is a long-span feature vector [3,4] obtained by concatenating several neighboring frames of feature vectors around $\mathbf{o}_t$, i.e.,

$$\xi_t = \left[1 \;\; \mathbf{o}_{t-L}^\top \ldots \mathbf{o}_t^\top \ldots \mathbf{o}_{t+L}^\top\right]^\top . \tag{2}$$

In the above equation, L determines the length of the context window, and the first element "1" is used to incorporate a bias term into the feature transform.

The transform weight $\kappa_{m,t}$'s are calculated by following the so-called "acoustic context expansion" method in fMPE [1, 2] as follows:

1. A codebook consisting of N_g Gaussian components is generated by clustering all the Gaussian mixture components in a GMM-HMM acoustic model trained typically using maximum likelihood (ML) criterion;
2. For each frame $\mathbf{o}_t$, an N_g-dimensional vector of posterior probabilities is computed using the above Gaussian codebook;
3. Suppose the current frame is at position 0, get 3 N_g-dimensional vectors by averaging the posterior probability vectors at position 1-2, 3-5 and 6-9 on the right and likewise another 3 averaged vectors on the left;
4. Concatenate the 7 N_g-dimensional averaged posterior probability vectors (3 on the left, 1 from the current frame, and 3 on the right) to form the final high-dimensional weight vector κ_t with $\kappa_{m,t}$ being its m^{th} element.

Consequently, the total number of linear transforms in our CE-RDLT transform is $M = 7 \times N_g$.

As discussed in the introduction section, conventional feature transforms such as fMPE, FE-RDLT and WE-RDLT can be treated as the special cases of CE-RDLT in Eq. (1). These methods differ from each other in following three aspects: 1) whether use bias only or full linear transform; 2) whether use long-span features as transform input or not, and 3) whether use contextual weight expansion or not.

3. TIED-STATE BASED MAXIMUM MUTUAL INFORMATION CRITERION

fMPE and conventional RDLT methods use the minimum phone error (MPE) criterion to train the transform parameters discriminatively. The MPE criterion approximates the expected training set phone accuracy, therefore needs to use word lattices decoded for each of the training utterances. To achieve the best performance, usually a weakened (typically uni-gram) language model should be used. So the decoded lattices are in general dense, which cause significant storage and loading overhead, especially when large-scale training data is used. The use of a language model during training also introduces task-dependent tuning factors and heuristics such as language model scaling, insertion penalty, and acoustic scaling. These parameters need be tuned for each application task.

In this study, the MMI criterion [15] is adopted to train transform parameters. The decision-tree based tied states of tri-phone HMMs are used as basic units in discriminative training. The MMI criterion can then be formulated as

$$\mathcal{F}(\mathbf{W}) = \sum_t \log p(s_t^r|\hat{\mathbf{o}}_t) = \sum_t \log \frac{p(\hat{\mathbf{o}}_t|s_t^r)p(s_t^r)}{\sum_s p(\hat{\mathbf{o}}_t|s)p(s)} , \tag{3}$$

where $\mathbf{W} = \{\mathbf{W}_m|m = 1, \ldots, M\}$ is the set of CE-RDLT transforms to be trained. In Eq. (3), s_t^r is the reference state at time t, which is determined by performing forced-alignment using the ground truth transcription of the training utterances; $\hat{\mathbf{o}}_t$ is the transformed feature vector at time t, which is a function of $\mathbf{W}$; $p(s)$ is the prior probability of each tied-state, which can be estimated conveniently from the forced-alignment results on the whole training set. In principle, the summation term in the denominator should take all tied-states into account. In practice, however, a short n-best state list can be computed beforehand for each time t to reduce the training cost.

Using the tied-state based MMI criterion, it is unnecessary to decode, store, and load word lattices before discriminative training can be performed, therefore the training tool can be made very scalable. Furthermore, there is no trouble to set those tuning parameters (e.g. acoustic and language model scaling) which are essential in conventional lattice based methods. Because the trained transform parameters are less dependent on (or sensitive to) the language model, they could work better than the ones trained by conventional lattice based discriminative training when deployed in application tasks different from the training task.

4. PARAMETER OPTIMIZATION

The transform parameters are optimized by a batch-mode resilient backpropagation (Rprop) algorithm [16] (specifically the iRprop$^-$ algorithm described in [17]) because of its capability to adjust per-parameter learning step sizes adaptively through iterations.

The derivatives needed for optimizing $\mathbf{W}$ are similar to the ones in fMPE, except they are calculated with respect to full transform parameters instead of bias only. Using Eq. (1), it is straightforward to get

$$\frac{\partial \mathcal{F}}{\partial w_m^{ij}} = \sum_t \frac{\partial \mathcal{F}_t}{\partial \hat{o}_t^i} \cdot \kappa_{m,t} \xi_t^j , \tag{4}$$

where i and j are row and column indices, respectively. The notion of direct and indirect derivatives defined in fMPE [1] is also borrowed in this study, i.e.,

$$\frac{\partial \mathcal{F}_t}{\partial \hat{o}_t^i} = \frac{\partial \mathcal{F}_t^{\text{direct}}}{\partial \hat{o}_t^i} + \frac{\partial \mathcal{F}_t^{\text{indirect}}}{\partial \hat{o}_t^i}. \tag{5}$$

Therefore, the derivative calculated in Eq. (4) can then be fed into the Rprop optimizer to update the learning step sizes, which are used to update the corresponding transform parameters accordingly.

Although a fixed global initial learning step size for all transform parameters can be used in Rprop, it is usually more efficient to set more informative, per-parameter initial learning step sizes. This is done in the first iteration as follows. Firstly, the derivatives are accumulated separately into positive and negative parts:

$$\begin{aligned} p_m^{ij} &= \textstyle\sum_t \max\left(\frac{\partial \mathcal{F}_t}{\partial \hat{o}_t^i} \cdot \kappa_{m,t} \xi_t^j, 0\right), \\ n_m^{ij} &= \textstyle\sum_t \min\left(\frac{\partial \mathcal{F}_t}{\partial \hat{o}_t^i} \cdot \kappa_{m,t} \xi_t^j, 0\right). \end{aligned} \tag{6}$$

Secondly, the per-parameter initial learning step sizes are defined as

$$\delta_m^{ij} = \frac{\rho}{p_m^{ij} - n_m^{ij}} \cdot (p_m^{ij} + n_m^{ij}), \tag{7}$$

where ρ is determined by using a first-order approximation of the criterion improvement: Suppose there are T frames of training data and the expected per-frame criterion improvement after the first

Table 1. Comparison of WERs (relative WER reductions) (both in %) of several feature transform methods by using conventional lattice-based discriminative training while GMM-HMMs are ML trained. The WER of the ML-trained baseline system is 26.5%.

fMPE	FE-RDLT	WE-RDLT	CE-RDLT
23.4 (11.7)	23.8 (10.2)	23.2 (12.5)	23.0 (13.2)

iteration is Δ, ρ is calculated by solving

$$\frac{1}{T}\sum_m \sum_i \sum_j \frac{\rho}{p_m^{ij} - n_m^{ij}} \cdot (p_m^{ij} + n_m^{ij})^2 = \Delta. \quad (8)$$

Following this method, Δ is the only control parameter that needs be set manually. A semi-automatic process is introduced to help determine it as follows:

- The transform parameters are updated in the first iteration using an initial guess of the expected criterion improvement Δ, and the actual criterion improvement is then evaluated in the second iteration;
- If the actual improvement is much larger than expected, Δ can be increased to enlarge the learning step sizes; Otherwise, if the actual improvement is significantly less than Δ or even being negative, Δ should be decreased accordingly. In both cases, training should start over from the first iteration until a good Δ is found.

Note that this semi-automatic method is only necessary when initializing the Rprop optimizer for the first iteration. The learning step sizes are adjusted adaptively and fully automatically in succeeding iterations.

5. EXPERIMENTS AND RESULTS

Switchboard-1 conversational telephone speech transcription task [18] was used in our experiments. We used 4,870 sides of conversations (about 300 hours of speech) from 520 speakers in training, and 40 sides of conversations (about 2 hours of speech) from the 2000 Hub5 evaluation (Eval2000) for testing.

For front-end feature extraction, the baseline system used 13-dimensional PLP features with windowed mean and variance normalization, and up to third-order delta features were used to form the raw 52-dimensional input feature vectors. A (39×52) HLDA [19] transform was estimated to reduce the feature dimension. For acoustic modeling, we used phonetic decision-tree based tied-state triphone GMM-HMMs with 9,304 states and 40 Gaussian components per state. A speaker-independent baseline GMM-HMM set was trained by ML criterion and used to perform the forced-alignment of each training utterance to generate the set of training feature vectors for each tied state. For each frame of training data, an n-best state list containing 1,000 most probable states is generated to speed up tied-state based RDLT training. Our recognition vocabulary contained 22,641 unique words. The pronunciation lexicon contained multiple pronunciations per word with a total of 28,649 unique pronunciations. A trigram language model trained on the transcription of the Switchboard-1 training data and broadcast news data was used in decoding. Recognition experiments were performed with a Microsoft in-house decoder and the results were evaluated by using the NIST Scoring Toolkit SCTK [20].

All the RDLT methods used a 1,000-component GMM got by clustering the ML-trained GMM-HMM set to calculate the high-dimensional posterior probability vectors. The raw 52-dimensional feature vector was used as the transform input. If long-span features were used, 11 frames including the current frame plus 5 preceding and 5 succeeding frames were concatenated, therefore 1,000 (39×573) RDLTs were trained. If contextual weight expansion was used, the posterior probability vectors were averaged and augmented exactly as in fMPE, so 7,000 (39×53) RDLTs would be trained. Similarly, if both long-span features and weight expansion were used, there would be 7,000 (39×573) RDLTs. In all cases, the HLDA transform used in the baseline GMM-HMM set was borrowed to initialize the RDLTs.

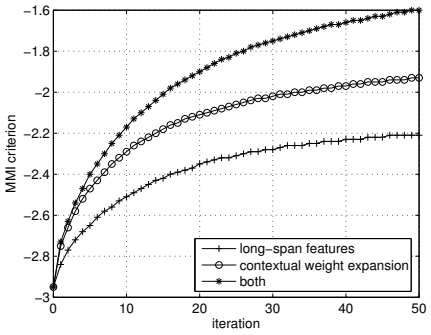

Fig. 1. MMI criterion improvement on the training set.

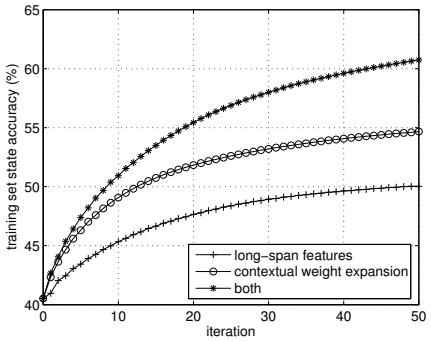

Fig. 2. State accuracy improvement on the training set.

5.1. Comparison of feature transform methods using lattice-based discriminative training

The WER using the ML-trained baseline GMM-HMM set was 26.5% on the Eval2000 set. Table 1 compares the recognition performances of several feature transform methods using the conventional word-lattice based discriminative training, while the GMM-HMMs are ML-trained. The number of "regions" used for fMPE had been increased to 50,000 for better performance. This also made fMPE having similar number of transform parameters (13.7 million) as that of WE-RDLT method (14.5 million), and slightly less than that of FE-RDLT method (22.3 million). For our CE-RDLT method, there were 156.4 million free parameters.

It is observed that FE-RDLT using only long-span features performs slightly worse than fMPE and WE-RDLT using contextual weight expansion only. Our proposed CE-RDLT using both long-span features and contextual weight expansion performs the best with increased number of transform parameters. Compared with fMPE, RDLT-based methods are more efficient in run-time because much less Gaussian components need be evaluated for calculating

Table 3. Comparison of WERs (relative WER reductions) (both in %) of several feature transform methods by lattice-based or tied-state based discriminative training while GMM-HMMs are trained with lattice-based BMMI training. The ML-trained baseline system has a WER of 26.5%.

Discriminative	fMPE	FE-RDLT		WE-RDLT		CE-RDLT	
Feature Transform	Lattice	Lattice	Tied-State	Lattice	Tied-State	Lattice	Tied-State
+BMMI HMM Training	22.6 (14.7)	22.8 (14.0)	21.5 (18.9)	21.9 (17.4)	21.3 (19.6)	21.8 (17.7)	20.6 (22.3)

Table 2. Comparison of WERs (relative WER reductions) (both in %) of three RDLT-based methods by tied-state based discriminative training while GMM-HMMs are ML trained. The WER of the ML-trained baseline system is 26.5%.

FE-RDLT	WE-RDLT	CE-RDLT
23.7 (10.6)	23.2 (12.5)	22.0 (17.0)

the posterior probability vectors.

5.2. Comparison of RDLT-based methods using tied-state based discriminative training

Three RDLT-based methods were compared by using tied-state based MMI training. 50 Rprop iterations were performed for each setup to train the RDLT transforms. The expected criterion improvement was set to 0.3 for initializing the Rprop learning step sizes. The value of the MMI objective function and tied-state classification accuracy on the training set were monitored and shown in Figs. 1 and 2. It is observed that the FE-RDLT method using long-span features performs the worst. It increases the training set state accuracy from 40.5% to 50.0%. With slightly less number of transform parameters, the WE-RDLT method using contextual weight expansion performs much better than FE-RDLT. Again, the proposed CE-RDLT method achieves the best performance, which achieves a training set state accuracy of 60.7%.

Table 2 compares the testing set WERs of the three RDLT-based methods by tied-state based discriminative training, while GMM-HMMs are ML trained. The FE-RDLT method performs the worst, the WE-RDLT method performs slightly better, and the proposed CE-RDLT method performs the best with a significant margin. By comparing these numbers with the corresponding ones in Table 1, RDLT-based methods by tied-state discriminative training outperform the ones by conventional lattice-based discriminative training. This is especially true for the CE-RDLT method. Apparently, it is helpful to use both long-span features and contextual weight expansion to increase the degree of freedom of RDLTs, which can be leveraged effectively by using tied-states as the training target. This observation is consistent with what was observed in DNN-HMM based acoustic modeling (e.g., [8–12]).

It is also interesting to compare the tied-state based RDLT methods with a DNN trained by our colleagues on the same task and same setup [9]. The DNN achieves a similar state classification accuracy on the training set ($\sim$60%), but much better WER ($\sim$17.1%) on the testing set. Further study is needed to investigate why DNN-HMM seems to generalize much better than GMM-HMM in this case.

5.3. Feature transform combined with discriminative training of GMM-HMMs

In addition to RDLTs, GMM-HMM parameters can also be trained by using the same tied-state based MMI criterion. Starting from the above CE-RDLTs trained, 10 iterations of extended Baum-Welch (E-B) optimization [21, 22] were performed to train the GMM-HMM parameters. The training set MMI criterion and tied-state classification accuracy were further increased to -1.5 and 63.1%, respectively. However, such an improvement failed to translate into WER reduction on the Eval2000 set: the WER remains 22.0%. This seems consistent with the results reported in [23], where a so-called "frame discrimination" (FD) method was used, and no performance gain against their ML-trained baseline system was observed either, although the FD criterion had been improved.

Then we evaluated the performances of all the above RDLT variants by combining them with the conventional word-lattice based discriminative GMM-HMM training. The boosted MMI (BMMI) criterion [24] was used in this case. Given the superior performance of WE-RDLT and CE-RDLT against fMPE, we did not conduct experiments on tied-state based DT for fMPE. Therefore only fMPE using lattice-based DT training was combined with BMMI GMM-HMM training, which is actually the method adopted in many state-of-the-art LVCSR systems. The results of this set of experiments are summarized and compared in Table 3. Several observations can be made. Firstly, by combining with the lattice-based BMMI training of GMM-HMMs, system performance can be improved further against the ML-trained GMM-HMMs using all the discriminative feature transform methods we studied. Secondly, for all the RDLT-based methods, transforms trained with tied-state based DT outperforms the lattice-based DT. Thirdly, the proposed CE-RDLT method with tied-state based DT plus word-lattice based BMMI training of GMM-HMMs achieves the best WER, which represents a relative WER reduction of about 10%, 6%, and 9% against three state-of-the-art methods, namely FE-RDLT(lattice), WE-RDLT(lattice), and fMPE(lattice), respectively.

6. CONCLUSION AND DISCUSSION

Based on the above results, we draw the following conclusions:

- Both the long-span features and the contextual weight expansion are helpful in the proposed context-expanded RDLT (CE-RDLT) feature transform;
- The best practice is to train the feature-space CE-RDLTs by using lattice-free, tied-state based discriminative training, while model-space GMM-HMMs are trained by using a conventional word-lattice based discriminative training method.

Interestingly, as demonstrated in [11], DNN-HMMs trained by a lattice-based discriminative training method can achieve better recognition accuracy than the DNN-HMMs trained by tied-state based discriminative training. Based on our research findings in this study, we are just wondering whether an additional gain could be achieved if the output (softmax) layer of the DNN was trained by lattice-based discriminative training, while other layers of the DNN was trained by lattice-free tied-state based discriminative training. This is our future work and we will report the results elsewhere once they become available.

7. REFERENCES

[1] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," in *Proc. ICASSP-2005*, pp. 961–964.

[2] D. Povey, "Improvements to fMPE for discriminative training of features," in *Proc. InterSpeech-2005*, pp. 2977–2980.

[3] B. Zhang, S. Matsoukas, and R. Schwartz, "Discriminatively trained region dependent feature transforms for speech recognition," in *Proc. ICASSP-2006*, pp. 313–316.

[4] B. Zhang, S. Matsoukas, and R. Schwartz, "Recent progress on the discriminative region-dependent transform for speech feature extraction," in *Proc. InterSpeech-2006*, pp. 1495–1498.

[5] M. Karafiát, L. Burget, P. Matějka, O. Glembek, and J. H. Černocký, "iVector-based discriminative adaptation for automatic speech recognition," in *Proc. ASRU-2011*, pp. 152–157.

[6] M. Karafiát, M. Janda, J. Černocký, and L. Burget, "Region dependent linear transforms in multilingual speech recognition," in *Proc. ICASSP-2012*, pp. 4885–4888.

[7] J. Xu, Z.-J. Yan, and Q. Huo, "A comparative study of fMPE and RDLT approaches to LVCSR," in *Proc. ISCSLP-2012*, 4 pages.

[8] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[9] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-depedent deep neural networks," in *Proc. InterSpeech-2011*, pp. 437–440.

[10] N. Jaitly, P. Nguyen, A. Senior, and V. Vanhoucke, "Application of pretrained deep neural networks to large vocabulary speech recognition," in *Proc. InterSpeech-2012*, 4 pages.

[11] B. Kingsbury, T. N. Sainath, and H. Soltau, "Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization," in *Proc. InterSpeech-2012*, 4 pages.

[12] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[13] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *Proc. ICASSP-2009*, pp. 3761–3764.

[14] S. Wiesler, R. Schlüter, and H. Ney, "Accelerated batch learning of convex log-linear models for LVCSR," in *Proc. InterSpeech-2012*, 4 pages.

[15] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. ICASSP-1986*, pp. 49–52.

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

[17] C. Igel and M. Hüsken, "Improving the Rprop learning algorithm," in *Second International Symposium on Neural Computation*, 2000, pp. 115–121.

[18] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. ICASSP-1992*, pp. 517–520.

[19] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, vol. 26, no. 4, pp. 283–297, 1998.

[20] "The NIST scoring toolkit SCTK," see the following site for details: `http://itl.nist.gov/iad/mig/tests/rt/2002/software.htm`.

[21] Y. Normandin, *Hidden Markov Models, Maximum Mutual Information Estimation, and the Speech Recognition Problem*, Ph.D. thesis, McGill University, 1991.

[22] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech and Language*, vol. 16, no. 1, pp. 25–47, 2002.

[23] P. C. Woodland, T. Hain, G. L. Moore, T. R. Niesler, D. Povey, A. Tuerk, and E. W. D. Whittaker, "The 1998 HTK broadcast news transcription system: development and results," in *Proc. DARPA Broadcast News Workshop*, 1999, pp. 265–270.

[24] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. ICASSP-2008*, pp. 4057–4060.