MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Discriminative Method for Recurrent Neural Network Language Models

Tachioka, Y.; Watanabe, S.

TR2015-033 April 2015

Abstract

A recurrent neural network language model (RNN-LM) can use a long word context more than can an n-gram language model, and its effective has recently been shown in its accomplishment of automatic speech recognition (ASR) tasks. However, the training criteria of RNN-LM are based on cross entropy (CE) between predicted and reference words. In addition, unlike the discriminative training of acoustic models and discriminative language models (DLM), these criteria do not explicitly consider discriminative criteria calculated from ASR hypotheses and references. This paper proposes a discriminative training method for RNN-LM by additionally considering a discriminative criterion to CE. We use the log-likelihood ratio of the ASR hypotheses and references as an discriminative criterion.

The proposed training criterion emphasizes the effect of misrecognized words relatively compared to the effect of correct words, which are discounted in training. Experiments on a large vocabulary continuous speech recognition task show that our proposed method improves the RNN-LM baseline. In addition, combining the proposed discriminative RNN-LM and DLM further shows its effectiveness.

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

DISCRIMINATIVE METHOD FOR RECURRENT NEURAL NETWORK LANGUAGE MODELS

Yuuki Tachioka

Information Technology R&D Center
Mitsubishi Electric Corporation
5-1-1, Ofuna, Kamakura, Kanagawa, Japan

Shinji Watanabe

Mitsubishi Electric Research Laboratories
201, Broadway, Cambridge, MA, US

ABSTRACT

A recurrent neural network language model (RNN-LM) can use a long word context more than can an n-gram language model, and its effective has recently been shown in its accomplishment of automatic speech recognition (ASR) tasks. However, the training criteria of RNN-LM are based on cross entropy (CE) between predicted and reference words. In addition, unlike the discriminative training of acoustic models and discriminative language models (DLM), these criteria do not explicitly consider discriminative criteria calculated from ASR hypotheses and references. This paper proposes a discriminative training method for RNN-LM by additionally considering a discriminative criterion to CE. We use the log-likelihood ratio of the ASR hypotheses and references as an discriminative criterion. The proposed training criterion emphasizes the effect of misrecognized words relatively compared to the effect of correct words, which are discounted in training. Experiments on a large vocabulary continuous speech recognition task show that our proposed method improves the RNN-LM baseline. In addition, combining the proposed discriminative RNN-LM and DLM further shows its effectiveness.

Index Terms— Speech recognition, recurrent neural network, language model, discriminative criterion, log-likelihood ratio

1. INTRODUCTION

Neural network methods have garnered much attention in the field of automatic speech recognition (ASR). One of the most successful examples is the deep neural network (DNN) used in acoustic modeling, and neural networks have been recently introduced and used for language processing. Among them, the recurrent neural network based language model (RNN-LM) has become popular due to its high performance [1, 2] as well as the availability of open source software [3, 4]. RNN is a neural network (NN) that contains one or more hidden layers with recursive inputs. Although their computational costs are high, RNN-LM greatly improves ASR performance. The greatest difference between RNN-LM and conventional n-gram models is the available word context length. The role of a language model is to estimate posterior probabilities of target words based on previous words context. A long context provides much information. However, the simple use of a long context (i.e., 4-gram or 5-gram) by a conventional n-gram language model encounters data sparsity problems. To address these problems, RNN-LM first maps a high-dimensional 1-of-N representation of a target word to a low-dimensional continuous space in a hidden layer and directly estimates the posterior probability of the target word. The hidden-layer units from the previous frame are then connected to the input vector in the next frame. These recursive inputs collect the history of words in the low-dimensional hidden-layer units. RNN-LM implicitly considers the entire history of words, whereas widely used n-gram models consider only previous $(n-1)$ words. Although there are several trials [5, 6, 7], using RNN-LM directly for decoding is essentially difficult because feed-forward propagation of RNN is much more expensive than using a table lookup method with an n-gram model. Therefore, RNN-LM is typically used for post-processing such as N-best or lattice rescoring.

However, the training criteria of RNN-LM are based on cross entropy (CE) between predicted and reference words. That is, the CE criterion does not explicitly consider discriminative criteria calculated from ASR hypotheses and references. On the other hand, discriminative criteria show the effectiveness in GMM-based acoustic model and feature transformation training at accomplishing various ASR tasks [8, 9, 10, 11]. Moreover, those for DNN acoustic modeling can also reduce ASR errors, while maintaining a fundamental high frame-level discriminability [12, 13, 14, 15, 16]. RNN-LM CE criterion is discriminative in the sense of considering the posterior distribution of a target word given history, but a discriminative criterion of RNN-LM that considers ASR hypotheses can further correct ASR errors. In recent years, [17] and [18] have applied sequence discriminative training to RNN acoustic modeling and natural language understanding, respectively. In this study, we propose a new discriminative training method for RNN-LM.

Another discriminative model within N-best rescoring framework is known as a discriminative language modeling (DLM) [19, 20, 21]. DLM is a corrective training method based on n-gram counts obtained from reference and ASR hypothesis examples of training data. It can correct errors that are inherent to a decoder in an efficient manner especially for words of short context. However, the context of DLM is limited to an n-gram (usually a tri-gram) that is identical to that in a typical n-gram language model. In addition, a long context cannot be used for error correction because of data sparsity. Our proposed method is based on a RNN-LM framework, and can consider a long context with the consideration of ASR hypotheses. Moreover, combining DLM and our discriminative RNN-LM can improve the performance from the DLM itself, which realizes short and long context discriminative language modeling.

The remainder of this paper is organized as follows. Section 2 describes the conventional RNN-LM [1]. Our proposed discriminative approach is described in Section 3. Section 4 describes our experiments involving a large vocabulary continuous speech recognition (LVCSR) task and reveals that the proposed method improves speech recognition performance.

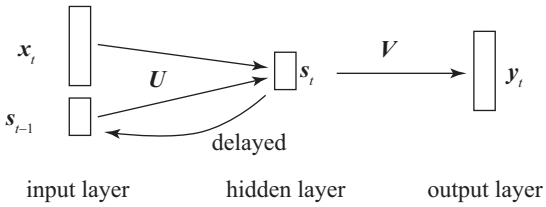

Fig. 1. Recurrent neural network language model (RNN-LM) topology. $|\mathcal{V}|$-dimensional input vector $\boldsymbol{x}_t$ is a 1-of-$|\mathcal{V}|$ representation of the t-th word of the utterance. Output vector $\boldsymbol{y}_t$ is an $|\mathcal{V}|$-dimensional posterior probability vector corresponding to input words conditioned on the previous context. The hidden layer has a low-dimensional vector $\boldsymbol{s}_t$. Hidden-layer units in the previous frame $\boldsymbol{s}_{t-1}$ are recursively concatenated to the input vector $\boldsymbol{x}_t$.

2. RNN-LM

Fig. 1 shows the topology of RNN-LM having one hidden layer, which we used for the following experiments. Hidden-layer units in the previous frame are recursively connected to the input vector. Weight matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ ($\triangleq \Theta$) are model parameters to be estimated in a training phase.

2.1. Cross-entropy training

We train the RNN-LM according to the cross-entropy (CE) criterion that minimizes the objective function $\mathcal{F}^{\mathrm{CE}}$. CE is calculated from a posterior of the predicted word $\boldsymbol{y}_t = [y_t(1), \cdots, y_t(n), \cdots, y_t(|\mathcal{V}|)]^\top$ with vocabulary $\mathcal{V}$, and a reference label sequence $C = \{c_t | t = 1, \cdots, T\}$, as follows:

$$\mathcal{F}^{\mathrm{CE}}(C) = -\sum_{n=1}^{|\mathcal{V}|}\sum_{t=1}^{T} \delta(n, c_t) \log y_t(n), \tag{1}$$

where c_t is an index of the reference label at the t-th word. $\delta(\cdot,\cdot)$ is a Kronecker delta function. The output layer has a softmax function y_t:

$$y_t(n) = \frac{\exp(a_t(n))}{\sum_{n'} \exp(a_t(n'))}, \tag{2}$$

where n is an index of elements in the output (softmax) layer and a_t is an activation of the n-th word.

2.2. Update rule

We discuss gradient-descent-based update rules for training parameter Θ. Based on the chain rule property of neural network (i.e., $\partial/\partial\Theta = \partial/\partial a_t(n) \cdot \partial a_t(n)/\partial\Theta$), we focus on the differentiation of the objective function $\mathcal{F}^{\mathrm{CE}}$ w.r.t of the activation $a_t(n)$ as

$$\frac{\partial \mathcal{F}^{\mathrm{CE}}}{\partial a_t(n)} = -[\delta(n, c_t) - y_t(n)] \triangleq \varepsilon_t(n), \tag{3}$$

because $\partial/\partial a_t(n) \log y_t(n') = \delta(n, n') - y_t(n)$. This equation means that the difference of the reference word and posterior $\varepsilon_t(n)$, which is an error of word n at position t, is propagated to the estimation of the model parameters Θ. Since there is a recurrent connection, it will be solved by the back propagation through time [1].

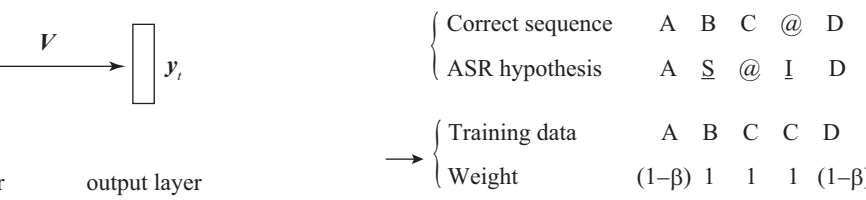

Fig. 2. Weight discount procedure of the proposed method. The weight of training data is discounted (i.e., $1-\beta$) for the correct data. A, B, C, and D are words, @ is a NULL token that follows the alignments of a correct word sequence and ASR hypothesis are fixed. S denotes a substitution and I denotes an insertion error. For insertion, repeated entry of the previous frame is used.

3. DISCRIMINATIVE TRAINING OF RNN-LM

3.1. Discriminative criterion of RNN-LM

To introduce the discriminative training into RNN-LM, we start from the word-level likelihood ratio objective function $\mathcal{F}^{\mathrm{LR}}$[1]:

$$\mathcal{F}^{\mathrm{LR}}(C, H) = -\sum_{t} \log \frac{y_t(c_t)}{y_t(h_t)^\beta}, \tag{4}$$

where h_t is an index of the t-th word of the 1-best ASR hypothesis aligned with the reference sequence C, and $H = \{h_t | t = 1, \cdots, T\}$ denotes the 1-best ASR sequence. β is a scaling factor, and the meaning of this factor will be discussed later. Note that this log likelihood ratio has a property of a discriminative criterion (used in Minimum classification error (MCE) training [22][2] and DLM [19]) so that minimizing $\mathcal{F}^{\mathrm{LR}}(C, H)$ corresponds to correct misrecognized h_t approaches to reference c_t.

Equation (4) can also be rewritten as

$$\begin{aligned}\mathcal{F}^{\mathrm{LR}}(C, H) &= -\sum_{n}\sum_{t} \delta(n, c_t) \log y_t(n) - \beta\delta(n, h_t) \log y_t(n) \\ &= \mathcal{F}^{\mathrm{CE}}(C) - \beta\mathcal{F}^{\mathrm{CE}}(H).\end{aligned} \tag{5}$$

Therefore, Equation (4) can be interpreted as a weighted difference of CE for the correct label and ASR hypothesis.

3.2. Update rule

For our proposed model, the update rule corresponds to (3) is also derived from the differentiation of (5) such that

$$\frac{\partial \mathcal{F}^{\mathrm{LR}}(C, H)}{\partial a_t(n)} = -[\delta(n, c_t) - \beta\delta(n, h_t) - (1-\beta) y_t(n)]. \tag{6}$$

In our implementation, we assume $(1-\beta)y_t(n)$ as $y_t(n)$ for simplicity, thus we obtain

$$\frac{\partial \mathcal{F}^{\mathrm{LR}}(C, H)}{\partial a_t(n)} \approx -[\delta(n, c_t) - \beta\delta(n, h_t) - y_t(c_t)]. \tag{7}$$

[1]This is not a sequence discriminative training but a word-level discriminative training based on an alignment between reference and 1-best ASR hypothesis.

[2]We can also consider an MMI-type discriminative criterion by summing up all possible hypotheses in the denominator.

Fig. 2 shows a weight discount of the proposed method. First, alignments of correct word sequences and ASR hypotheses are fixed using dynamic programming. Second, the weight for the correct label is discounted (i.e., $1-\beta$) and the model is re-trained with these discounted weights. Note that we assume that $\delta(n,c_t)-\beta\delta(n,h_t)=0$ when $\delta(n,c_t)-\beta\delta(n,h_t)<0$ to avoid that the value of target reference word becomes negative.

3.3. Use of word-level confidence measure

Word-level confidence measure ν_t ($0 \leq \nu_t \leq 1$), which is calculated from a confusion network, can be used to adjust the discount factor β. Errors with high confidence are more problematic and should be weighted more than errors with low confidence. Equation (7) is modified as follow.

$$\frac{\partial \mathcal{F}^{\mathrm{LR}}(C,H)}{\partial a_t(n)} = -[\delta(n,c_t)-\beta(1-\nu_t(h_t))\delta(n,h_t)-y_t(c_t)]. \tag{8}$$

Thus, we can control the discount value according to the confidence in the update rule.

3.4. Smoothing with original cross-entropy model

Finally, RNN-LM models are obtained by smoothing parameters obtained by the proposed discriminative method $\boldsymbol{U}^{\mathrm{LR}}, \boldsymbol{V}^{\mathrm{LR}}$ with the original CE model $\boldsymbol{U}^{\mathrm{CE}}, \boldsymbol{V}^{\mathrm{CE}}$ such that

$$\{\boldsymbol{U},\boldsymbol{V}\} \leftarrow \tau\{\boldsymbol{U}^{\mathrm{CE}},\boldsymbol{V}^{\mathrm{CE}}\}+(1-\tau)\{\boldsymbol{U}^{\mathrm{LR}},\boldsymbol{V}^{\mathrm{LR}}\}, \tag{9}$$

where τ is a smoothing factor. This avoids over-training.

4. EXPERIMENTS

4.1. Experimental setup

We evaluated the observed performance improvement on the Corpus of Spontaneous Japanese (CSJ) [23], which is one of the most widely used LVCSR tasks to build Japanese ASR systems. Vocabulary size is about 70k. We used three types of test sets wherein each set consists of lecture-style examples from 10 speakers. Test sets E1, E2, and E3 contain 22,682, 23,226, and 14,896 words, respectively.

We trained the DNN-HMM with CE training using 23 dimensional mel-filter bank coefficients + Δ+$\Delta\Delta$. The number of context-dependent HMM states was 3,500 and the DNN contained seven hidden layers and 2,048 nodes per layer in accordance with settings used in a previous study [24]. The initial learning rate was 0.01 and decreased to 0.001 at the end of training. After a CE DNN acoustic model was obtained, boosted MMI discriminative training for DNN [15] was conducted. We used Povey's implementation of DNN training tools in a Kaldi toolkit [25].

Although the size of the original language model was 70k, the vocabulary size of RNN-LM was limited to 10k, which corresponds to the number of input layer dimensions (i.e., $|\mathcal{V}|$). The number of hidden-layer units was 30. RNN-LM was constructed using the RNN-LM toolkit [3]. The language model score was obtained by linear interpolation of the RNN-LM score and the original n-gram model score. The weight of interpolation was 0.5 and 100-best hypotheses for each utterance were used for rescoring. We combined the RNN-LM and the proposed discriminative RNN-LM with DLM.

Table 1. WER [%] on CSJ using a DNN acoustic model with a conventional n-gram and discriminative language model (DLM).

	E1	E2	E3	Avg.
baseline	12.81	10.64	11.13	11.53
+DLM	12.60	10.52	10.82	11.31

Table 2. WER [%] on CSJ using a DNN acoustic model with RNN-LM-based and DLM-based rescoring.

	E1	E2	E3	Avg.
+RNN-LM	11.97	10.18	10.51	10.89
+RNN-LM+DLM	11.74	9.98	10.03	10.58

Table 3. WER [%] on CSJ with the proposed discriminative RNN-LM (d-RNN-LM).

β	τ	η	E1	E2	E3	Avg.
0.05	0.85	0.1	11.99	10.19	10.50	10.89
		0.05	**11.84**	10.07	10.61	10.84
	0.9	0.1	11.91	**10.02**	10.51	10.81
		0.05	**11.84**	10.03	10.49	10.79
0.10	0.85	0.1	12.20	10.45	10.69	11.11
		0.05	11.86	10.09	10.47	10.81
	0.9	0.1	11.93	10.19	10.41	10.84
		0.05	11.90	10.04	**10.39**	**10.78**
0.15	0.85	0.1	12.06	10.38	10.49	10.98
		0.05	11.93	10.09	10.40	10.81
	0.9	0.1	11.98	10.17	**10.39**	10.85
		0.05	11.98	10.03	**10.39**	10.80

Table 4. WER [%] on CSJ with the proposed discriminative RNN-LM (d-RNN-LM) and DLM rescoring.

β	τ	η	E1	E2	E3	Avg.
0.05	0.85	0.1	12.00	10.20	10.51	10.90
		0.05	11.68	9.98	10.04	10.57
	0.9	0.1	11.72	10.01	10.04	10.59
		0.05	11.63	**9.90**	10.05	10.53
0.10	0.85	0.1	12.07	10.19	10.70	10.99
		0.05	11.75	10.03	10.28	10.69
	0.9	0.1	11.77	10.03	10.12	10.64
		0.05	11.64	9.94	10.08	10.55
0.15	0.85	0.1	11.81	10.07	10.26	10.71
		0.05	11.63	10.00	10.14	10.59
	0.9	0.1	11.61	9.95	10.01	10.52
		0.05	**11.60**	9.95	**9.99**	**10.51**

4.2. Baseline results

Table 1 shows the baseline results when using the discriminatively trained DNN acoustic model, which was state-of-the-art performance for this CSJ corpus [20, 24]. Using DLM rescoring, the word error rate (WER) was improved by 0.22% on average.

For this high baseline, RNN-LM rescoring significantly improved the WER, as shown in Table 2 by 0.64% on average. In addition to RNN-LM, the DLM was also effective for this result, which shows the effectiveness of the discriminative model.

Table 5. WER [%] on CSJ with the proposed discriminative RNN-LM (d-RNN-LM) using word-level confidence measures.

β	τ	η	E1	E2	E3	Avg.
0.05	0.85	0.1	12.15	10.34	10.49	10.99
		0.05	11.88	10.08	10.54	10.83
	0.9	0.1	11.93	10.17	10.44	10.85
		0.05	**11.84**	**10.03**	10.46	**10.78**
0.10	0.85	0.1	12.39	10.43	10.92	11.25
		0.05	11.89	10.13	10.52	10.85
	0.9	0.1	12.02	10.17	10.51	10.90
		0.05	11.93	10.09	10.35	10.79
0.15	0.85	0.1	12.18	10.41	10.60	11.06
		0.05	11.95	10.11	**10.31**	10.79
	0.9	0.1	12.01	10.21	10.45	10.89
		0.05	11.95	10.04	10.35	**10.78**

Table 6. WER [%] on CSJ with the proposed discriminative RNN-LM (d-RNN-LM) using word-level confidence measures and DLM rescoring.

β	τ	η	E1	E2	E3	Avg.
0.05	0.85	0.1	11.84	10.28	10.41	10.84
		0.05	**11.56**	10.03	10.12	10.57
	0.9	0.1	11.66	9.99	10.11	10.59
		0.05	11.63	9.94	10.06	10.54
0.10	0.85	0.1	11.71	10.00	10.22	10.64
		0.05	11.65	10.02	10.27	10.65
	0.9	0.1	11.65	9.95	10.19	10.60
		0.05	11.66	9.94	10.14	10.58
0.15	0.85	0.1	11.76	10.04	10.18	10.66
		0.05	12.01	10.20	10.57	10.93
	0.9	0.1	11.63	**9.84**	**10.01**	**10.49**
		0.05	11.69	9.99	10.14	10.61

4.3. Proposed method

Table 3 shows the proposed discriminative RNN-LM (d-RNN-LM). Three parameters exist in the prosed method and parametric studies were conducted. In nearly all cases, average WER was better than that of the RNN-LM result in Table 2. This result suggest that the parameter tuning was not so difficult. Table 4 shows that DLM was effective when used with the proposed method because the explicit use of short context by the n-gram model was powerful whereas the proposed method implicitly used short context.

Table 5 shows the proposed method using word-level confidence measures. Unfortunately, little performance gain was observed, but similar tendencies were noticeable. DLM was also effective as shown in Table 6.

Although the performance gain of the proposed method was small in our experiments overall, this is simply due to very high baseline of this setting. We believe that this modeling increases model estimation robustness for a task that contains many errors.

5. CONCLUSION AND FUTURE WORK

We proposed a discriminative training method for RNN-LM. In addition to the CE training of correct examples, discriminative training against ASR hypotheses was proposed. The proposed discriminative training yielded a difference of CE that was similar to the difference statistics revealed in the discriminative training of acoustic modeling. Experimental results showed that our proposed method improved the performance of an LVCSR task. Combining the proposed discriminative RNN-LM, which uses short and long context implicitly, and the DLM, which uses short context explicitly, was also effective because the two complement one another. Future research will examine sequential discriminative training and the use of N-best hypotheses in the training.

Acknowledgment

We would thank Dr. Jonathan Le Roux and Dr. John R. Hershey at Mitsubishi Electric Research Laboratories for their valuable suggestions.

6. REFERENCES

[1] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proceedings of INTERSPEECH*, 2010, pp. 1045–1048.

[2] M. Sundermeyer, I. Oparin, J. Gauvain, B. Freiberg, R. Schlüter, and H. Ney, "Comparison of feedforward and recurrent neural network language models," in *Proceedings of ICASSP*, 2013, pp. 8430–8434.

[3] T. Mikolov, S. Kombrink, A. Deoras, L. Burget, J. Černocký, and S. Khudanpur, "RNNLM–recurrent neural network language modeling toolkit," in *Proceedings of ASRU*, 2011, pp. 1–4.

[4] M. Sundermeyer, R. Schlüter, and H Ney, "rwthlm - The RWTH Aachen University neural network language modeling toolkit," in *Proceedings of INTERSPEECH*, 2014, pp. 2093–2097.

[5] Y. Shi, W.-Q. Zhang, M. Cai, and J. Liu, "Efficient one-pass decoding with NNLM for speech recognition," *IEEE Signal Processing Letters*, vol. 21, pp. 377–381, 2014.

[6] T. Hori, Y. Kubo, and A. Nakamura, "Real-time one-pass decoding with recurrent neural network language model for speech recognition," in *Proceedings of ICASSP*, 2014, pp. 6414–6418.

[7] Z. Huang, G. Zweig, and B. Dumoulin, "Cache based recurrent neural network language model inference for first pass speech recognition," in *Proceedings of ICASSP*, 2014, pp. 6404–6407.

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

[9] M. Gales, S. Watanabe, and E. Fosler-Lussier, "Structured discriminative models for speech recognition: An overview," *IEEE Signal Processing Magazine*, vol. 29, pp. 70–81, 2012.

[10] Y. Tachioka, S. Watanabe, and J.R. Hershey, "Effectiveness of discriminative training and feature transformation for reverberated and noisy speech," in *Proceedings of ICASSP*, 2013, pp. 6935–6939.

[11] Y. Tachioka, S. Watanabe, J. Le Roux, and J.R. Hershey, "Sequential maximum mutual information linear discriminant analysis for speech recognition," in *Proceedings of INTERSPEECH*, 2014, pp. 2415–2419.

[12] G. Wang and K.C. Sim, "Sequential classification criteria for NNs in automatic speech recognition," in *Proceedings of INTERSPEECH*, 2011, pp. 441–444.

[13] B. Kingsbury, T. Sainath, and H. Soltau, "Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization," in *Proceedings of INTERSPEECH*, 2012, pp. 485–488.

[14] N. Jaitly, P. Nguyen, A. Senior, and V. Vanhoucke, "Application of pretrained deep neural networks to large vocabulary speech recognition," in *Proceedings of INTERSPEECH*, 2012.

[15] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proceedings of INTERSPEECH*, 2013.

[16] Y. Kubo, T. Hori, and A. Nakamura, "Large vocabulary continuous speech recognition based on WFST structured classifiers and deep bottleneck features," in *Proceedings of ICASSP*, 2013, pp. 7629–7633.

[17] H. Sak, O. Vinyals, G. Heigold, A. Senior, E. McDermott, R. Monga, and M. Mao, "Sequence discriminative distributed training of long short-term memory recurrent neural networks," in *Proceedings of INTERSPEECH*, 2014, pp. 1209–1213.

[18] K. Yao, B. Peng, G. Zweig, D. Yu, X. Li, and F. Gao, "Recurrent conditional random field for language understanding," in *Proceedings of ICASSP*, 2014, pp. 4105–4109.

[19] B. Roark, M. Saraçlar, M. Collins, and M. Johnson, "Discriminative language modeling with conditional random fields and the perceptron algorithm," in *Proceedings of ACL*, 2004, pp. 47–54.

[20] T. Oba, T. Hori, A. Nakamura, and A. Ito, "Round-robin duel discriminative language models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 1244–1255, 2012.

[21] E. Dikici, M. Semarci, M. Saraçlar, and E. Alpaydin, "Classification and ranking approaches to discriminative language modeling for ASR," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 291–300, 2013.

[22] E. McDermott, T.J. Hazen, J. Le Roux, A. Nakamura, and S. Katagiri, "Discriminative training for large-vocabulary speech recognition using minimum classification error," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 203–223, 2007.

[23] S. Furui, K. Maekawa, and H. Isahara, "A Japanese national project on spontaneous speech corpus and processing technology," in *Proceedings of ASR*, 2000, pp. 244–248.

[24] N. Kanda, R. Takeda, and Y. Obuchi, "Elastic spectral distortion for lowresource speech recognition with deep neural networks," in *Proceedings of ASRU*, 2013, pp. 309–314.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, M. Petr, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Proceedings of ASRU*, 2011, pp. 1–4.