Regularized Sequence-Level Deep Neural Network Model Adaptation

Yan Huang and Yifan Gong

Microsoft Corporation, One Microsoft Way, Redmond, WA 98052

{yanhuang; ygong}@microsoft.com

Abstract

We propose a regularized sequence-level (SEQ) deep neural network (DNN) model adaptation methodology as an extension of the previous KL-divergence regularized cross-entropy (CE) adaptation [1]. In this approach, the negative KL-divergence between the baseline and the adapted model is added to the maximum mutual information (MMI) as regularization in the sequence-level adaptation.

We compared eight different adaptation setups specified by the baseline training criterion, the adaptation criterion, and the regularization methodology. We found that the proposed sequence-level adaptation consistently outperforms the cross-entropy adaptation. For both of them, regularization is critical. We further introduced a unified formulation in which the regularized CE and SEQ adaptation are the special cases.

We applied the proposed approach to speaker adaptation and accent adaptation in a mobile short message dictation task. For the speaker adaptation, with 25 or 100 utterances, the proposed approach yields 13.72% or 23.18% WER reduction when adapting from the CE baseline, comparing to 11.87% or 20.18% for the CE adaptation. For the accent adaptation, with 1K utterances, the proposed approach yields 18.74% or 19.50% WER reduction when adapting from the CE-DNN or the SEQ-DNN. The WER reduction using the regularized CE adaptation is 15.98% and 15.69%, respectively.

Index Terms: deep neural network model adaptation, regularization, sequence training

1. Introduction

While recent advances in acoustic modeling using deep neural networks (DNN) have led to significant accuracy improvement [2, 3, 4, 5, 6], diverse acoustic environments, distinct channels, and various speaking styles remain as the main challenges [7]. Model adaptation refers to a class of techniques that can "move" the model towards a specified target using moderate amount of adaptation data and achieve improved accuracy.

One key challenge in the deep neural network model adaptation is the robust parameter estimation given the large number of parameters in the DNN and the usually limited amount of adaptation data. *Catastrophic forgetting* described in [18] is a typical form of overfitting in the neural network adaptation. Consequently, a certain form of regularization is typically applied in the deep neural network model adaptation.

The transform-based adaptation [9, 10, 11, 12] only adapt the partial network while keeping the large body of the neural network unchanged. This can be viewed as applying the regularization at the topological-level. The regularization-based adaptation [1, 16, 17, 18] operates in the full model parameter space with regularization. The combination of these two approaches can lead to more effective adaptation with small amount of adaptation data [19, 20]. In the third category, the adaptation context is represented in a certain form and input to the neural network for adaptation [13, 14, 15].

In this paper, we propose a regularized sequence-level (SEQ) deep neural network model adaptation methodology as an extension of the previous KL-divergence regularized cross-entropy (CE) model adaptation [1]. In this approach, a frame-level regularization, defined as the negative KL-divergence between the baseline and the adapted model, is added to the sequence-level maximum mutual information (MMI) objective to avoid overfitting during the sequence-level model adaptation.

We compared the convergence pattern and the adaptation performance of different adaptation setups specified by the baseline training criterion, the model adaptation criterion, and the regularization methodology. We found that the sequence-level adaptation outperforms the cross-entropy adaptation when adapting from either a cross-entropy DNN or a sequence DNN. In both cases, regularization is critical to the best adaptation performance. Under certain circumstances, such as conducting the cross-entropy or the sequence-level adaptation from a baseline DNN trained using the sequence-level criterion, without applying the regularization, the adaptation exhibits severe overfitting with large performance degradation. We further introduced a unified formulation in which the regularized CE and SEQ adaptation are the special cases.

We applied the proposed regularized sequence-level DNN adaptation methodology to speaker adaptation and accent adaptation in a mobile short message dictation task (SMD). For the speaker adaptation, with 25 or 100 adaptation utterances, the proposed regularized sequence-level adaptation yields 13.72% or 23.18% WER reduction when adapting from the CE-DNN. Correspondingly, the cross-entropy adaptation yields 11.87% or 20.18% WER reduction. For the accent adaptation, with 1K adaptation utterances, the proposed approach yields 18.74% or 19.50% WER reduction when adapting from the CE-DNN or the SEQ-DNN. In comparison, the WER reduction using the regularized cross-entropy adaptation is 15.98% and 15.69%, respectively.

The remainder of this paper is organized as follows: Section 2 reviews the cross-entropy DNN and the KLD-regularized cross-entropy DNN adaptation; Section 3 introduces the regularized sequence-level DNN adaptation methodology; Section 4 presents the experimental results on speaker adaptation and accent adaptation tasks; Section 5 concludes this study.

2. Review of CE-DNN and KLD-regularized CE-DNN

A deep neural network is a stack of log-linear models parameterized by the layer-wise weight matrix, bias, and partition function. For a given objective function, the gradient of the top-level error signal can be back propagated for the full network optimization through the error back-propagation (BP).

2.1. Cross-Entropy DNN

The cross-entropy objective ($\mathcal{F}_{CE}$) is defined as the total negative log-posterior of the senone state (s) given the acoustic observation (o) accumulated on all frames (t) of all utterances (u):

$$\mathcal{F}_{CE} = -\sum_{u,t,s} p^T(s|o_{ut}) \log p(s|o_{ut}), \quad (1)$$

where $p^T(s|o_{ut})$ is the state-level target, defined as the Kronecker delta function ($\delta_{s;s_{ut}}$) based on the senone-state alignment; $p(s|o_{ut})$ is the posterior probability obtained by passing the output layer neuron activation ($a_{ut}(s)$) through a softmax:

$$p(s|o_{ut}) = \frac{exp^{a_{ut}(s)}}{\sum_{s'} exp^{a_{ut}(s')}}. \quad (2)$$

The corresponding gradient at the output layer is:

$$\frac{\partial \mathcal{F}_{CE}}{\partial a_{ut}(s)} = p(s|o_{ut}) - \delta_{s;s_{ut}}. \quad (3)$$

Minimizing $\mathcal{F}_{CE}$ is equivalent to maximize the mutual information between $p(s|o_{ut})$ and $\delta_{s;s_{ut}}$.

2.2. KLD-regularized Cross-Entropy Adaptation

The KLD-regularized cross-entropy model adaptation was proposed in [1]. In this methodology, the KL-divergence between the baseline and the adapted model is added to the standard cross-entropy objective. The new regularized cross-entropy objective ($\hat{\mathcal{F}}_{CE}$) can be written as:

$$\begin{aligned}\hat{\mathcal{F}}_{CE} &= (1-\rho)\mathcal{F}_{CE} + [-\rho \sum_{u,t,s} p^{SI}(s|o_{ut}) \log p(s|o_{ut})] \\ &= -\sum_{u,t,s} \hat{p}(s|o_{ut}) \log p(s|o_{ut}),\end{aligned} \quad (4)$$

where $\hat{p}(s|o_{ut}) \triangleq (1-\rho)\delta_{s;s_{ut}} + \rho p^{SI}(s|o_{ut})$, ρ is the regularization weight, $p^{SI}(s|o_{ut})$ is the adaptation baseline model.

The corresponding gradient at the output layer is:

$$\frac{\partial \hat{\mathcal{F}}_{CE}}{\partial a_{ut}(s)} = p(s|o_{ut}) - \hat{p}(s|o_{ut}). \quad (5)$$

Comparing to Eq. (3), $\delta_{s;s_{ut}}$ is replaced by $\hat{p}(s|o_{ut})$, a soft target defined as the linear combination of the true label and the posterior estimated from the adaptation baseline model.

3. Regularized Sequence-Level Adaptation

The sequence-level objective seeks to maximize the posterior of the correct utterance given the model. It takes into account of the language model, lexical, and HMM constraints. Significant accuracy improvement has been reported in applying the sequence-level deep neural network acoustic model for large vocabulary speech recognition tasks [22, 23, 24].

In this paper, we focus on the Maximum Mutual Information (MMI) criterion [21] for the DNN model adaptation.

3.1. Sequence-Level MMI Objective

The MMI objective ($\mathcal{F}_{MMI}$) can be written as [21]:

$$\mathcal{F}_{MMI} = \sum_u \log \frac{p^k(O_u|S_u)p(W_u)}{\sum_W p^k(O_u|S)p(W)}, \quad (6)$$

where O_u is the acoustic observation sequence for the utterance u; S_u and W_u are the corresponding senone-state sequence and word sequence; k is the acoustic model scaling factor.

The gradient at the output layer is:

$$\frac{\partial \mathcal{F}_{MMI}}{\partial a_{ut}(s)} = k(\delta_{s;s_{u,t}} - \gamma_{ut}^{DEN}(s)), \quad (7)$$

where $\gamma_{ut}^{DEN}(s))$ is the posterior probability of being in state s at time t computed over the lattices.

Maximizing $\mathcal{F}_{MMI}$ is equivalent to maximizing the mutual information between $\delta_{s;s_{u,t}}$ and $\gamma_{ut}^{DEN}(s)$. Comparing to Eq. (3), $p(s|o_{ut})$ is replaced by $\gamma_{ut}^{DEN}(s)$, a state occupancy stats calculated from the denominator lattices.

3.2. Regularized MMI Adaptation

In the regularized MMI adaptation, we propose to add a frame-level regularization term to the standard MMI.

The frame-level regularization is defined as the negative KL-divergence between the base model and the adapted model with respect to the frame-level senone posterior estimation. The regularized MMI objective ($\hat{\mathcal{F}}_{MMI}$) can be written as:

$$\hat{\mathcal{F}}_{MMI} = (1-\rho)\mathcal{F}_{MMI} + \rho \sum_{u,t,s} p^{SI}(s|o_{ut}) \log p(s|o_{ut}), \quad (8)$$

where $\mathcal{F}_{MMI}$ is the standard MMI defined as Eq. (6), ρ is the regularization weight.

The corresponding gradient at the output layer is:

$$\begin{aligned}\frac{\partial \hat{\mathcal{F}}_{MMI}}{\partial a_{ut}(s)} &= (1-\rho)k[\delta_{s;s_{u,t}} - \gamma_{ut}^{DEN}(s)] + \rho p_{det}^{SI}(s) \\ &= (1-\rho)k\{\delta_{s;s_{u,t}} - [\gamma_{ut}^{DEN}(s) - \frac{\rho}{(1-\rho)k} p_{det}^{SI}(s)]\},\end{aligned} \quad (9)$$

where $p_{det}^{SI}(s) \triangleq p(s^{SI}|o_{ut}) - p(s|o_{ut})$.

Comparing to Eq. (7), $\gamma_{ut}^{DEN}(s)$ is replaced by $\gamma_{ut}^{DEN}(s) - \frac{\rho}{(1-\rho)k} p_{det}^{SI}(s)$.

3.3. Regularized MMI Adaptation with F-smoothing

The frame-level negative cross-entropy was used as a regularization to prevent the model from "running away" in the MMI sequence DNN implementation in [22]. It was referred to as F-smoothing.

The MMI objective with F-smoothing ($\mathcal{F}_{MMI}^{(f)}$) is:

$$\mathcal{F}_{MMI}^{(f)} = (1-\rho_F)\mathcal{F}_{MMI} + (-\rho_F \mathcal{F}_{CE}), \quad (10)$$

where ρ_F is the weight of F-smoothing.

The objective of the regularized MMI adaptation with F-smoothing ($\hat{\mathcal{F}}_{MMI}^{(f)}$) can be written as:

$$\begin{aligned}\hat{\mathcal{F}}_{MMI}^{(f)} &= (1-\rho)\mathcal{F}_{MMI}^{(f)} + \rho \sum_{u,t,s} p^{SI}(s|o_{ut}) \log p(s|o_{ut}) \\ &= (1-\rho)(1-\rho_F)F_{MMI} - (1-\rho)\rho_F \mathcal{F}_{CE} \\ &\quad +\rho \sum_{u,t,s} p^{SI}(s|o_{ut}) \log p(s|o_{ut}).\end{aligned} \quad (11)$$

The regularized MMI with F-smoothing is a generalized formulation. When $\rho_F = 1$, it becomes the KL-regularized cross-entropy objective ($\hat{\mathcal{F}}_{CE}$); when $\rho_F = 0$, it becomes the KL-regularized sequence-level MMI objective ($\hat{\mathcal{F}}_{MMI}$).

The gradient of the regularized MMI objective with F-smoothing is:

$$\begin{aligned}\frac{\partial \hat{\mathcal{F}}_{MMI}^{(f)}}{\partial a_{ut}(s)} &= \{(1-\rho)[(1-\rho_F)k + \rho_F]\delta_{s;s_{u,t}} + \rho p(s^{SI}|o_{ut})\} \\ &-\{(1-\rho)(1-\rho_F)k\gamma_{ut}^{DEN}(s) + [(1-\rho)\rho_F + \rho]p(s|o_{ut})\}.\end{aligned} \quad (12)$$

It degenerates to the gradient of $\hat{\mathcal{F}}_{CE}$ or $\hat{\mathcal{F}}_{MMI}$ as in Eq. (5) and Eq. (9) respectively, when $\rho_F = 1$ or $\rho_F = 0$.

3.4. Special Treatment

In MMI, the language model and the HMM constraints are usually estimated from the training set. The estimation of these parameters from limited amount of adaptation data is usually unreliable. As a special treatment, we simply use the baseline training set to estimate these parameters. For a complete regularized approach, we could use the regularized estimation from the baseline training and the adaptation data. It was not adopted since empirically we found no performance difference.

To conclude Section 3, we point out that the optimization of the regularized sequence-level MMI adaptation proposed in this paper can proceed using the standard BP with no need to change the learning procedure. The proposed methodology can be applied to the full network as well as the partial network. It has larger computation cost due to the lattice generation and the lattice-based gradient calculation.

4. Experiments and Results

In this section, we present our experimental results in applying the regularized sequence-level DNN model adaptation on the mobile short message dictation task.

4.1. Baseline

We use a pair of mobile short message dictation DNN models trained using the cross-entropy or the sequence-level MMI criterion as the baseline models throughout this paper.

The baseline DNNs have 5-hidden layers. Each hidden layer has 2048 hidden units. The input consists of a 726-dim feature vector formed by a 66-dim log filter bank feature (LFB) with a context window of 11 frames. The output layer has 5980 senone states. The training data consists of 400 hr mobile speech data which are aligned at the senone-state level. For this task, each utterance consists of 1.5 second speech in average.

4.2. Convergence Pattern and Adaptation performance

We studied the convergence pattern and the model adaptation performance of different adaptation methodologies using a channel adaptation task. In this task, we adapt the mobile speech DNN to a lecture room close-talk speech task with distinct channel mismatch and other scenario differences. We compared eight adaptation setups, specified by the baseline training criterion, the adaptation criterion, and the regularization methodology.

The baseline models are the mobile speech models as described in Section 4.1. The adaptation data consists of 1K lecture room close-talk speech utterances. The resulting models are evaluated using a lecture room close-talk test set consisting of 3K utterances. For the sequence-level adaptation, we adopted the regularized sequence-level adaptation with F-smoothing as formulated in Section 3.3. ρ_F and ρ were set to 0.095 and 0.5 respectively throughout this paper without further tuning.

Figure 1 illustrates the adaptation convergence pattern and the accuracy results. "CE-DNN" and "SE-DNN" refer to adapting from the cross-entropy or the sequence-level baseline; "CE-Adapt" and "SE-Adapt" refer to applying the cross-entropy or the sequence-level adaptation criterion; "-Reg" refers to adaptation with the KL-divergence based regularization.

- Starting from the CE-DNN, the cross-entropy adaptation yields accuracy gains with/without regularization. The regularized CE adaptation achieves better performance improvement. This observation is consistent with [1].

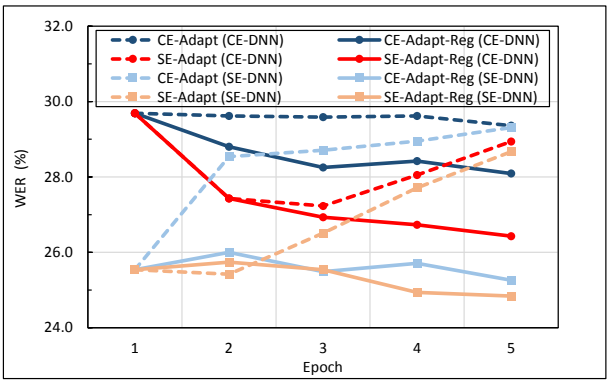

Figure 1: *Convergence pattern and model adaptation performance comparison of eight different adaptation setups. "CE-DNN" and "SE-DNN" refer to the cross-entropy or the sequence-level baseline models; "CE-Adapt" and "SE-Adapt" refer to the cross-entropy or the sequence-level adaptation; "-Reg" refers to adaptation with regularization.*

- Starting from the SEQ-DNN, without applying the regularization, the cross-entropy adaptation quickly exhibits overfitting. The learned sequence-level pattern in the baseline may experience "catastrophic forgetting". After applying the regularization, the convergence is significantly improved and finally results in a better accuracy.
- When conducting the sequence-level adaptation from the CE-DNN, we can observe accuracy gains with/without regularization. These gains are primarily due to the new sequence-level pattern learned through the model adaptation. With the regularization, the adaptation converged to a better model.
- When conducting the sequence-level adaptation from the SEQ-DNN, without applying the regularization, the model quickly "runs away". After applying the regularization, the adaptation exhibits a well-behaved convergence pattern and results in a better performed model.

In summary, the sequence-level MMI adaptation consistently outperforms the frame-level cross-entropy adaptation when adapting from a cross-entropy DNN or a sequence DNN. In both cases, regularization is critical. Under certain circumstances, such as applying the cross-entropy or the sequence-level adaptation to a sequence baseline, without applying the regularization, the adaptation exhibits severe overfitting.

We note that the convergence pattern and the adaptation performance also depend on the amount of adaptation data and the specific adaptation task. Typically heavier regularization is needed for adaptation with smaller amount of adaptation data.

4.3. Regularized MMI adaptation for Speaker Adaptation

We applied the proposed regularized sequence adaptation methodology to the speaker adaptation for the SMD task.

The baseline models are as described in Section 4.1. The speaker adaptation data is collected from actual life service, collected over long period of time, representing real mobile speech application scenario. The speaker adaptation data set consists of 4 speakers, each speaker with 100 utterances for the model adaptation and 400$\sim$500 utterances for the model evaluation.

We conducted the speaker adaptation experiments using 25 or 100 adaptation utterances. For the SEQ adaptation,

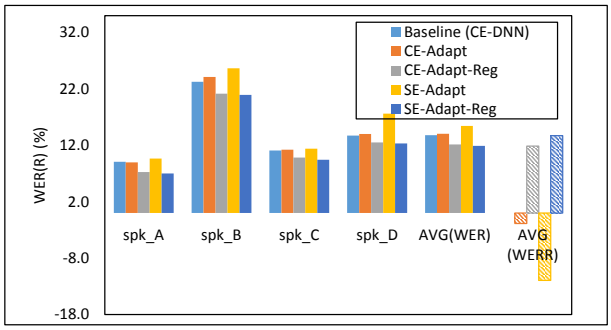

Figure 2: *Speaker adaptation performance of the regularized cross-entropy and the sequence-level adaptation with 25 or 100 adaptation utterances. The baseline is the CE-DNN.*

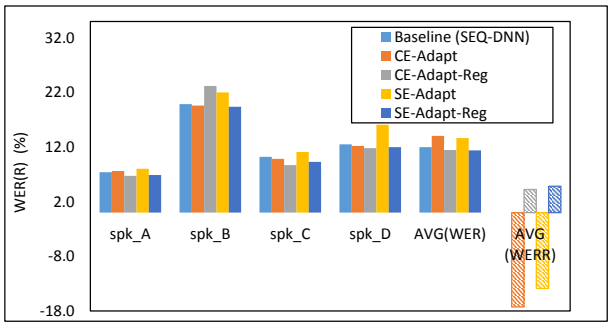

Figure 3: *Speaker adaptation performance of the regularized cross-entropy and the sequence-level adaptation with 25 or 100 adaptation utterances. The baseline is the SEQ-DNN.*

we adopted the regularized sequence-level adaptation with F-smoothing with the same parameter setup as before. The experimental results are depicted in Figure 2 and Figure 3:

- With 25 speaker adaptation utterances, when adapting from the cross-entropy DNN, the regularized sequence adaptation yields 13.72% WER reduction as opposed to 11.87% for the regularized CE adaptation. When adapting from the sequence DNN, the regularized sequence adaptation yields 4.84% WER reduction comparing to 4.24% for the regularized CE adaptation.
- With 100 speaker adaptation utterances, the regularized sequence adaptation yields 23.18% WER reduction as opposed to 20.18% for the regularized CE adaptation when adapting from the CE baseline model. When adapting from the sequence DNN, the regularized SEQ adaptation yields 19.92% WER reduction comparing to 17.37% when using the regularized CE adaptation.

The regularized sequence adaptation consistently outperforms the regularized cross-entropy adaptation. Nevertheless, when only a small number of adaptation utterances is available, the benefit of the sequence adaptation is small. Regularization is critical for both the cross-entropy and the sequence adaptation. Without regularization, they both exhibit sever overfitting and result in degraded performance.

4.4. Regularized MMI adaptation for Accent Adaptation

We also conducted experiments on the accent adaptation for the mobile short message dictation task.

The baseline models are as described in Section 4.1. The training data consists of 1K accent adaptation utterances for reasonable accent phonetic coverage. The resulting models were evaluated using an accent test set consisting of 5K utterances.

As before, we adopted the regularized sequence adaptation with F-smoothing with similar parameter setup. The accent adaptation experimental results are summarized in Table 1:

- Starting from the CE-DNN, the regularized sequence adaptation yields 18.74% WER reduction comparing to 11.98% for the regularized cross-entropy adaptation. Without the regularization, the WER reduction drops to 5.03% and 3.72% for the sequence adaptation and the cross-entropy adaptation, respectively.
- Starting from the SEQ-DNN, the regularized sequence adaptation yields 18.23% WER reduction as opposing to 15.69% for the regularized cross-entropy adaptation. Without the regularization, large performance degradation was observed due to overfitting.

With more adaptation data, we observe larger adaptation performance gain from the regularized sequence adaptation. This is due to the fact that more sequence-level patterns can be learned when more adaptation data is available. When the adaptation data increases to certain amount, the regularization may no longer be necessary and a simple model update with the sequence-level objective would be sufficient.

Table 1: *Accent adaptation performance of the regularized cross-entropy and the sequence.*

Model	WER	WERR
Baseline (CE-DNN)	27.95	NA
CE-Adapt	27.95	3.72
CE-Adapt-Reg	24.39	15.98
SE-Adapt	27.57	5.03
SE-Adapt-Reg	23.59	18.74
Baseline (SEQ-DNN)	23.64	NA
CE-Adapt	27.01	-14.26
CE-Adapt-Reg	19.93	15.69
SE-Adapt	24.16	-2.20
SE-Adapt-Reg	19.03	19.50

5. Conclusion

In this paper, we proposed a regularized sequence-level deep neural network model adaptation methodology. In this approach, a frame-level regularization is added to the sequence-level maximum mutual information objective to avoid overfitting in the sequence-level model adaptation.

We studied eight different adaptation setups specified by the baseline training criterion, the adaptation criterion, the regularization methodology. We found that the sequence-level adaptation outperforms the cross-entropy adaptation. In both cases, regularization is critical for the best adaptation performance.

We applied the proposed regularized sequence-level DNN adaptation methodology to speaker adaptation and accent adaptation in a mobile short message dictation task. In all cases, the proposed regularized sequence-level adaptation yields better adaptation performance than the cross-entropy adaptation.

6. Acknowledgements

The authors would like to thank Dr. Chaojun Liu and Dr. Dong Yu for the helpful discussions on this work.

7. References

[1] Yu, D., Yao, K., Su, H., Li, G., and Seide, F., "KL-Divergence Regularized Deep Neural Network Adaptation for Improved Large Vocabulary Speech Recognition," in the Proceedings of ICASSP 2013.

[2] Sainath, T. N., Kingsbury, B., Ramabhadran, B., Fousek, P., Novak, P., and Mohamed, A. -R., "Making Deep Belief Networks Effective for Large Vocabulary Continuous Speech Recognition," in the Proceedings of 2011 IEEE ASRU, 2011.

[3] Dahl, G.E., Yu, D., Deng, L., and Acero, A., "Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition," IEEE Transactions on Audio, Speech, and Language Processing (TASLP) - Special Issue on Deep Learning for Spech and Language Processing, Volume: 1, No. 1, Page(s): 33-42, Jan 2012.

[4] Seide, F., Li, G., and Yu, D., "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks," in the Proceedings of Interspeech 2012.

[5] Jaitly, N., Nguyen, P., Senior, A., and Vanhoucke, V., "Application of Pretrained Deep Neural Networks to Large Vocabulary Speech Recognition," in the Proceedings of Interspeech 2012.

[6] Hinton G., Deng, L., Yu, D., Dahl G., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath,T., and Kingsbury, B., "Deep Neural Networks for Acoustic Modeling in Speech Recognition," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97, November 2012.

[7] Huang, Y., Yu, D., Liu, C., and Gong, Y., "A Comparative Analytic Study on the Gaussian Mixture and Context Dependent Deep Neural Network Hidden Markov Models," in the Proceedings of Interspeech 2014.

[8] Albesano, D., Gemello, R., Laface, P., Mana, F., and Scanzio, S., "Adaptation of Artificial Neural Networks Avoiding Catastrophic Forgetting," in the Proceedings of the 2006 International Joint Conference on Neural Networks, 2006.

[9] Gemelloa, R., Manaa, F., Scanziob, S. Lafaceb, and P., De Mori, R., "Linear Hidden Transformations for Adaptation of Hybrid ANN/HMM Models," Speech Communication 49, no. 10, pp. 827-83, 2007.

[10] Li, B. and Sim, K.C., "Comparison of Discriminative Input and Output Transformations for Speaker Adaptation in the Hybrid NN/HMM Systems," in the Proceedings of Interspeech 2010.

[11] Yu, D., Seltzer, M., Li, J., Huang, J., and Seide, F., "Feature Learning in Deep Neural Networks - Studies on Speech Recognition Tasks," in the Proceedings of 2013 International Conference on Learning Representation, 2013.

[12] Yao, K., Yu, D., Seide, F., Su, H., Deng, L., and Gong, Y., "Adaptation of Context-Dependent Deep Neural Networks for Automatic Speech Recognition," in the Proceedings of 2012 Spoken Language Technology Workshop (SLT), 2012.

[13] Abdel-hamid, O. and Jiang, H., "Fast Speaker Adaptation of Hybrid NN/HMM Model for Speech Recognition Based on Discriminative Learning of Speaker Code," in the Proceedings of ICASSP 2013.

[14] Saon, G., Soltau, H., Nahamoo, D., Picheny, M., "Speaker Adaptation of Neural Network Acoustic Models Using I-vectors," in the Proceedings of 2013 IEEE Automatic Speech Recognition and Understanding (ASRU) workshop, 2013.

[15] Senior, A. and Moreno I., "Improving DNN speaker independence with i-vector inputs", in the Proceedings of ICASSP 2014.

[16] Gemello, R., Mana, F., Scanzio, S., Laface, P., and Mori, R.D., "Adaptation of Mybrid ANN/HMM Models Using Linear Hidden Transformations and Conservative Training,"in the Proceedings of ICASSP 2006.

[17] Li, X. and Bilmes. J., "Regularized Adaptation of Discriminative Classifiers,"in the Proceedings of ICASSP 2006.

[18] Liao, H., "Speaker Adaptation of Context Dependent Deep Neural Networks,"in the Proceedings of ICASSP 2013.

[19] Xue, J., Li, J., Yu, D., Seltzer, M., and Gong, Y., "Singular Value Decomposition Based Low-footprint Speaker Adaptation and Personalization for Deep Neural Network," in the Proceedings of ICASSP 2014.

[20] Huang, Y., Yu, D., Liu, C., and Gong, Y., "Multi-Accent Deep Neural Network Acoustic Model With Accent Specific Top Layer Using The KLD-Regularized Model Adaptation," in the Proceedings of Interspeech 2014.

[21] Povey, D, Kingsbury, B., Ramabhadran, B., Saon, G., Soltau H., and Visweswariah, K., "Boosted MMI for Model and Feature-space Discriminative Training," in the Proceedings of ICASSP 2008.

[22] Su, H., Li, G., Yu, D., and Seide, F., "Error Back Propagation for Sequence Training of Context-Dependent Deep Networks for Conversational Speech Transcription," in the Proceedings of ICASSP 2013.

[23] Kingsbury, B., Sainath, N. T., and Soltau, H., "Scalable Minimum Bayes Risk Training of Deep Neural Network Acoustic Models Using Distributed Hessian-free Optimization," in the Proceedings of Interspeech 2012.

[24] Vesely, K., Ghoshal, A., Burget, L., and Povey, D., "Sequence-discriminative Training of Deep Neural Networks," in the Proceedings of Interspeech 2013.