KL-DIVERGENCE REGULARIZED DEEP NEURAL NETWORK ADAPTATION FOR IMPROVED LARGE VOCABULARY SPEECH RECOGNITION

Dong Yu[1], Kaisheng Yao[2], Hang Su[3,4], Gang Li[3], Frank Seide[3]

[1]Microsoft Research, Redmond, 98052, WA, USA
[2]Online Services Division, Microsoft Corporation, Redmond, 98052, WA, USA
[3]Microsoft Research Asia, Beijing, China
[4]Tsinghua University, Beijing, China

ABSTRACT

We propose a novel regularized adaptation technique for context dependent deep neural network hidden Markov models (CD-DNN-HMMs). The CD-DNN-HMM has a large output layer and many large hidden layers, each with thousands of neurons. The huge number of parameters in the CD-DNN-HMM makes adaptation a challenging task, esp. when the adaptation set is small. The technique developed in this paper adapts the model conservatively by forcing the senone distribution estimated from the adapted model to be close to that from the unadapted model. This constraint is realized by adding Kullback–Leibler divergence (KLD) regularization to the adaptation criterion. We show that applying this regularization is equivalent to changing the target distribution in the conventional backpropagation algorithm. Experiments on Xbox voice search, short message dictation, and Switchboard and lecture speech transcription tasks demonstrate that the proposed adaptation technique can provide 2%-30% relative error reduction against the already very strong speaker independent CD-DNN-HMM systems using different adaptation sets under both supervised and unsupervised adaptation setups.

Index Terms— deep neural network, CD-DNN-HMM, speaker adaptation, Kullback–Leibler divergence regularization

1. INTRODUCTION

Recently the context dependent deep neural network hidden Markov models (CD-DNN-HMMs) outperformed the discriminatively trained Gaussian mixture model (GMM) HMMs with 16% [1][2] and 33% [3] relative error reduction, respectively, on the Bing voice search [4] and the Switchboard (SWB) tasks. Similar improvement has been observed on other tasks such as broadcast news, Google voice search, Youtube and Aurora4 [5]-[11]. Most recently Kingsbury et al. [8] showed that the speaker independent (SI) CD-DNN-HMM can outperform the well-tuned state-of-the-art speaker adaptive GMM system with more than 10% relative error reduction on the SWB task by applying the sequence-level discriminative training on the CD-DNN-HMM. The potential of CD-DNN-HMMs, however, are yet to be explored. In this paper we propose a regularized adaptation technique for DNNs to further improve the recognition accuracy of CD-DNN-HMMs.

The CD-DNN-HMM is a special case of the artificial neural network (ANN) HMM hybrid system developed in 1990s, for which several adaptation techniques have been developed. These techniques can be classified into categories of linear transformation [12]-[20], conservative training [21]-[23], and subspace method [24]. However, compared to the earlier ANN/HMM hybrid systems, CD-DNN-HMMs have significantly more parameters due to wider and deeper hidden layers used and the much larger output layer designed to model senones (tied-triphone states) directly. This difference casts additional challenges to adapting CD-DNN-HMMs, esp. when the adaptation set is small.

In this paper we propose a novel regularized adaptation technique for DNNs. Our proposed technique adapts the model conservatively by forcing the senone distribution estimated from the adapted model to be close to that estimated from the unadapted model. This constraint is realized by adding Kullback–Leibler divergence (KLD) regularization to the adaptation criterion. We show that applying this regularization is equivalent to changing the target distribution in the conventional backpropagation (BP) algorithm. Experiments on Xbox voice search, short message dictation, and SWB and lecture transcription tasks demonstrate that the proposed adaptation technique can provide 2%-30% relative error reduction against the already very strong speaker independent CD-DNN-HMM systems using different adaptation sets under both supervised and unsupervised adaptation setups. This new technique also outperforms the feature discriminative linear regression (fDLR) technique proposed in [25] and the output-feature discriminative linear regression (oDLR) proposed in [20].

The rest of the paper is organized as follows. In Section 2 we briefly review CD-DNN-HMMs. In Section 3 we describe our proposed regularized adaptation algorithm. In Section 4 related work is briefly introduced. We report the experimental results in Section 5 and conclude the paper in Section 6.

2. CD-DNN-HMM

The CD-DNN-HMM [1][3] is a special ANN/HMM hybrid system in which DNNs are in place of the shallow ANNs and are used to model senones directly. The DNN accepts an input observation x, which typically consists of 9-13 frames of acoustic features, and process it through many layers of nonlinear transformation

$$h_i^\ell = \sigma\left(z_i^\ell\left(v^\ell\right)\right) = \sigma\left(\left(w_i^\ell\right)^T v^\ell + a_i^\ell\right), \quad (1)$$

where w^ℓ and a^ℓ are the weight matrix and bias, respectively, at hidden layer ℓ, h_i^ℓ is the output of the i-th neuron,

$$z^\ell\left(v^\ell\right) = \left(w^\ell\right)^T v^\ell + a^\ell \quad (2)$$

is the excitation vector given input v^ℓ, $v^\ell = h^{\ell-1}$ when $\ell > 0$ and $v^0 = x$, and $\sigma(x) = 1 / (1 + \exp(-x))$ is the sigmoid function applied element-wise. At the top layer L, the softmax function

$$p(y = s|v^L) = \frac{exp((w_s^L)^T v^L + a_s^L)}{\sum_{y'} exp\left(\left(w_{y'}^L\right)^T v^L + a_{y'}^L\right)} \quad (3)$$

is used to estimate the state posterior probability $p(y = s|x)$,

which is converted to the HMM state emission probability as

$$p(x|y=s)=\frac{p(y=s|x)}{p(y=s)}\cdot p(x), \quad (4)$$

where $s\in\{1,2,\cdots,S\}$ is a senone id, S is the total number of senones, $p(y=s)$ is the prior probability of senone s, and $p(x)$ is independent of state s.

The parameters of DNNs are typically trained to maximize the negative cross entropy

$$\overline{D}=\frac{1}{N}\sum_{t=1}^{N}D(x_t)=\frac{1}{N}\sum_{t=1}^{N}\sum_{y=1}^{S}\tilde{p}(y|x_t)\log p(y|x_t), \quad (5)$$

where N is the number of samples in the training set and $\tilde{p}(y|x_t)$ is the target probability. Often times we use a hard alignment from an existing system as the training label, under which condition $\tilde{p}(y|x_t)=\delta(y=s_t)$, where δ is Kronecker delta and s_t is the label of the t-th sample, i.e. the t-th observation frame in our training corpus. The training is carried out using the BP algorithm, speeded up with GPU and minibatch updates.

3. REGULARIZED ADAPTATION

An obvious approach to adapting DNNs is adjusting all the DNN parameters with the adaptation data, starting from the SI model. However, doing so may destroy previously learned information and overfit the adaptation data, esp. if the adaptation set is small. To prevent this from happening, adaptation needs to be done conservatively. The technique proposed here does exactly this.

The intuition behind our proposed approach is that the posterior senone distribution estimated from the adapted model should not deviate too far away from that estimated using the unadapted model, esp. when the adaptation set is small.

Since the DNN outputs are probability distributions, a natural choice in measuring the deviation is the Kullback–Leibler divergence (KLD). By adding this divergence as a regularization term to eq. (5) and removing the terms unrelated to the model parameters we get the regularized optimization criterion

$$\widehat{D}=(1-\rho)\overline{D}+\rho\frac{1}{N}\sum_{t=1}^{N}\sum_{y=1}^{S}p^{SI}(y|x_t)\log p(y|x_t), \quad (6)$$

where $p^{SI}(y|x_t)$ is the posterior probability estimated from the SI model and computed with a forward pass using the SI model, and ρ is the regularization weight. Eq. (6) can be reorganized to

$$\begin{aligned}\widehat{D}&=\frac{1}{N}\sum_{t=1}^{N}\sum_{y=1}^{S}[(1-\rho)\tilde{p}(y|x_t)+\rho p^{SI}(y|x_t)]\log p(y|x_t)\\&=\frac{1}{N}\sum_{t=1}^{N}\sum_{y=1}^{S}\hat{p}(y|x_t)\log p(y|x_t), \quad (7)\end{aligned}$$

where we have defined

$$\hat{p}(y|x_t)\triangleq(1-\rho)\tilde{p}(y|x_t)+\rho p^{SI}(y|x_t). \quad (8)$$

By comparing eqs. (5) and (7) we can see that applying the KLD regularization to the original training criterion equals to changing the target probability distribution from $\tilde{p}(y|x_t)$ to $\hat{p}(y|x_t)$, which is a linear interpolation of the distribution estimated from the unadapted model and the ground truth alignment of the adaptation data. This interpolation prevents overtraining by keeping the adapted model from straying too far from the SI model. Note that this differs from L2 regularization [23], which constrains the model parameters themselves rather than the output probabilities. This also indicates that the normal BP algorithm can be directly used to adapt the DNN. The only thing needs to be changed is the error signal at the output layer, which is now defined based on $\hat{p}(y|x)$.

The interpolation weight, which is directly derived from the regularization weight ρ, can be adjusted, typically using a development set, based on the size of the adaptation set, the learning rate used, and whether the adaptation is supervised or unsupervised. When ρ=1, we trust completely the unadapted model and ignore all new information from the adaptation data. When ρ=0, we adapt the model solely on the adaptation set, ignoring information from the unadapted model except using it as the starting point. Intuitively we should use a large ρ for a small adaptation set and a small ρ for a large adaptation set.

4. RELATED WORK

Many ANN adaptation techniques have been developed in the past. These techniques can be classified into three categories: linear transformation, conservative training, and subspace method.

4.1 Linear Transformation

The simplest and most popular approach to adapting ANNs is applying a linear transformation, either to the input feature (as in the linear input network (LIN) [12]-[15], [17]-[19] and the very similar feature discriminative linear regression (fDLR) [25]), to the activation of a hidden layer (as in the linear hidden network (LHN) [16]), or to the softmax layer (as in the linear output network (LON) [15] and in the output-feature discriminative linear regression (oDLR) [20]). No matter where the linear transformation is applied, it is typically trained from an identity weight matrix and zero bias to optimize the criterion specified in eq. (5), keeping fixed the weights of the original ANN.

4.2 Conservative Training

Another popular category of adaptation techniques is conservative training (CT) [22]. CT can be achieved by adding regularizations to the adaptation criterion. For example, in [23], L2 regularization was used. The KLD regularization approach proposed in this paper is another regularization technique. An alternative CT technique is adapting only selected weights, e.g., in [21] only weights connected to the hidden nodes with maximum variance (computed on the adaptation data) are adapted. Adaptation with very small learning rate and early stopping can also be considered as CT.

4.3 Subspace Method

Subspace method aims to find a speaker subspace and then construct adapted ANN weights or transformations as a point in the subspace. For example, in [24] subspace is used to estimate an affine transformation matrix by considering it as a random variable. Principal components analysis (PCA) was performed on a set of adaptation matrices to obtain the principal directions (i.e. eigenvectors) in the speaker space. Each new speaker adaptation model is then approximated by a linear combination of the retained eigenvectors. The linear combination weights can be estimated using BP. In [26] a speaker subspace and a speech subspace are estimated and combined using tensors. The deep tensor neural

network developed in [27] is also a subspace method.

5. EMPIRICAL EVALUATION

To evaluate the proposed KLD-Reg approach, we have conducted a series of experiments on four different datasets. To obtain the results reported in this section we adapted all parameters in the DNNs, which always outperforms the setup where only the softmax layer is adapted on these datasets.

5.1 Xbox Voice Search

Our initial set of experiments were conducted on a small internal Xbox voice search (VS) task which features distant talking voice search of music catalog, games, movies, and so on. We selected this dataset since it has been used in our previous study so that we can compare the KLD-Reg approach with fDLR [25], which performs slightly better than oDLR [20] on the same dataset.

The features were 13-dimensional Mel filter-bank cepstral coefficients (MFCC) with up to third derivatives, further transformed to 36-dimension with heterogeneous linear discriminate analysis (HLDA). Per-device cepstral mean subtraction was applied. The baseline SI models were trained using 40 hours of voice search data. The GMM-HMM model had 70k Gaussian components and 1509 senones optimized with the standard maximum likelihood estimation (MLE) procedure. The SI CD-DNN-HMM system used an 9-frame input layer, three 2048-neuron hidden layers, and a 1509-neuron output layer. The DNN system was trained using the frame-level cross entropy criterion and the senone alignment generated from the MLE system. The trigram language model (LM) used in evaluating both SI and adapted models was trained on the transcriptions of all the data — including the test set. This is obviously suboptimal, but needed due to the unfortunate nature of the data set which has a significant out-of-vocabulary rate; we basically trade skewing the result due to OOVs for skewing the result due to test-set-inside-LM (an overly strong LM should make the results tend to be a lower bound for the effectiveness of acoustic adaptation). Thus, when interpreting the result in this section, please be aware of this experimental shortcoming, and note that the same limitation also applies to previously reported fDLR/oDLR results on this data set. The other data sets reported in this paper do not have this problem.

The adaptation experiments were conducted on 6 speakers (excluded from the training set), for whom we have over 300 utterances, the newest of which were used as test utterances. The total number of tested words is 5185. Without adaptation, the GMM system achieved 43.6% word error rate (WER) and the DNN system obtained 34.1% WER. The goal of the experiments is to improve the recognition accuracy using each speaker's past data. We pretended we have 200, 100, 50, 25, 10, and 5 past utterances for adaptation for each speaker by sampling from his/her past utterances. 200 utterances and 5 utterances roughly equal to 12 minutes and 18 seconds of audio data, respectively, in this dataset. In all these experiments we used the adaptation strategy optimized for fDLR, as demonstrated in [20], with 10 passes of data and $1x10^{-5}$ learning rate per sample.

Figure 1 compares the WER using the supervised fDLR and KLD-Reg adaptation techniques. From this figure we can make three observations. First, KLD-Reg helps, esp. when the adaptation set is small. In fact, when only 5 or 10 utterances were used for adaptation, the KLD-Reg adapted model degraded the accuracy. However, when the regularization weight ρ was increased to greater than 0.0625 the adapted model consistently outperformed the SI model. Second, the WER reduction seems to be robust to the choice of ρ once it's in the right range (e.g., [0.0625 0.5]). In addition, when the adaptation set is large (e.g., 200 utterances), good WER reduction can be obtained even with a very small ρ. Third, the KLD-Reg outperformed fDLR across all adaptation set sizes. The 5-1 cross validation indicates the average relative WER reductions using supervised KLD-Reg are 20.7%, 17.7%, 17.5%, 11.1%, 7.0%, and 5.3%, with 200, 100, 50, 25, 10, and 5 utterances of adaptation data, respectively.

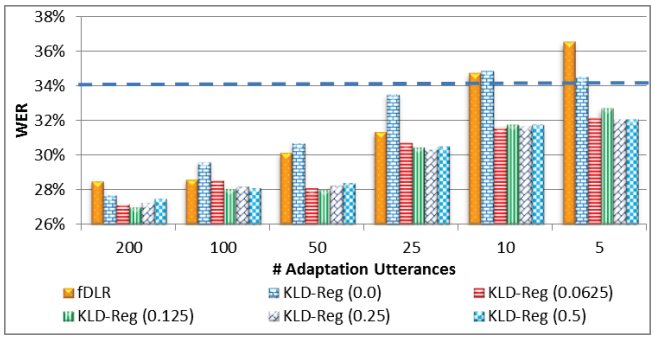


Figure 1: WERs on the Xbox voice search dataset for supervised fDLR and KLD-Reg adaptation. Numbers in parentheses are regularization weights ρ. The dashed line is the baseline WER using the SI DNN model. Note the remark in the text on LM used.

5.2 Short Message Dictation

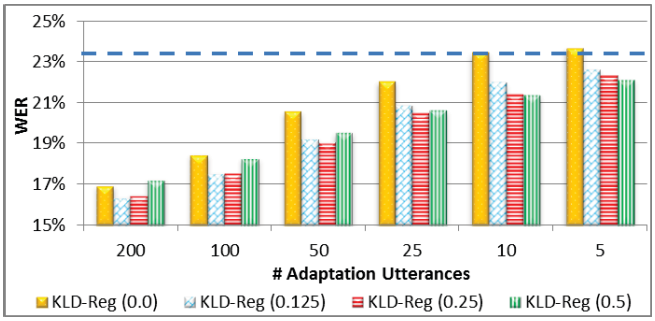
The second experiment was conducted on a short message dictation (SMD) task, which has longer utterances than VS. The goal is to see whether KLD-Reg is effective on other tasks. The baseline SI models were trained using 300hr voice search and SMD data. The evaluation was conducted on data from 9 speakers, out of which 2 were used as the development set and 7 were used as the test set. The total number of test set words is 20668. There is no overlap among training, adaptation and test sets.

The SI GMM-HMM acoustic model has approximately 288k Gaussian components and 5976 senones trained with the MLE procedure, followed by fMPE and BMMI. The baseline SI CD-DNN-HMM system used 24 log-filter bank features with up to second derivatives and a context window size of 11, forming a vector of 792-dimension (72x11) input. On top of the input layer, there are 5 hidden layers with 2048-neurons each. The output layer has a dimension of 5976. The DNN system was trained using the senone alignments from the GMM-HMM system. The baseline SI GMM-HMM system and CD-DNN-HMM system achieved 30.4% and 23.4% WER, respectively, on the 7-speaker test set.

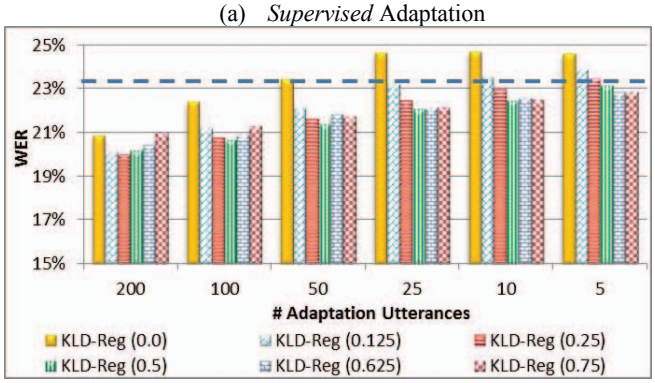
Same as in the VS experiment, we varied the number of adaptation utterances from 5 (32 seconds) to 200 (22 minutes). Different from the previous experiments, though, we used the 2-speaker development set to determine the learning rate ($4x10^{-5}$ per sample) and applied it to the 7-speaker test set.

Figure 2 (a) and (b) summarize the WER on the SMD dataset using supervised and unsupervised KLD-Reg adaptation, respectively. From these figures we can observe that KLD-Reg is very effective on this task as well. With the optimal regularization weights determined by the development set we get 30.3%, 25.2%, 18.6%, 12.6%, 8.8%, and 5.6% relative WER reduction using supervised adaptation, and 14.6%, 11.7%, 8.6%, 5.8%, 4.1%, 2.5%, using unsupervised adaptation, respectively, with 200, 100,

50, 25, 10, and 5 utterances of adaptation data. These figures also show clearer (since the learning rate is optimized using a dev set) that larger regularization weights should be used for smaller adaptation set and smaller regularization weights should be used for larger adaptation set, although the error reductions are quite robust as long as the regularization weight is within [0.125 0.5] on this task. Compared to the supervised adaptation, the unsupervised setup can benefit from larger regularization weights. We believe this is because the labels in the unsupervised adaptation setup are less reliable and thus we should trust the output from the SI model even more during the adaptation.



(a) *Supervised* Adaptation



(b) *Unsupervised* Adaptation

Figure 2: WERs on the SMD dataset using KLD-Reg adaptation for different regularization weights ρ (numbers in parentheses). The dashed line is the SI DNN baseline.

5.3 Switchboard and Lecture Transcription

In the SWB task we want to use the unsupervised adaptation to improve the recognition accuracy of that utterance itself (45 segments or 3 mins). The SI DNN model was trained using 309hr SWB-I training set with the exact same configuration (11 frames of input features, seven 2048-neuron hidden layers, and a 9.3K-neuron output layer) as that described in [3][25]. The evaluation was conducted on the 1831-segment SWB part of the NIST 2000 Hub5 eval set. On this task we achieved on average 2.7% relative WER reduction which is slightly better than the 2.0% obtained with fDLR [25]. The 2.7% reduction, however, is only half of that achieved on the SMD task using similar amount of adaptation data (3 mins or 25 utterances in SMD). This might because the SI DNN in the SMD task was trained with a mixture of VS and SMD data.

Our last experiment was conducted on a lecture transcription task where enough adaptation data (6 lectures 3.8hrs total) were available. The SI DNN model was trained with the 2000hr SWB data and has seven 2048-neuron hidden layers and an 18K output layer. The development and test sets are each a lecture and have 5612 and 8481 words, respectively. Table 1 summarizes the experimental results and indicates that good improvement can be achieved with the proposed adaptation technique. For comparison, we also optimized a speaker-dependent (SD) CD-DNN-HMM with the 6 lectures (with transcription) used for adaptation. In this task, the SD model outperformed the DNN baseline trained using 2000hr of (mismatched) SWB data on both the development and test sets. However, it underperformed the adapted models. This confirmed the effectiveness of the proposed adaptation technique.

Table 1. *WER and Relative WER Reduction (in parentheses) on Lecture Transcription Task.*

	SI DNN	Supervised	Unsupervised	SD DNN
Dev Set	16.0%	14.3% (10.6%)	14.9% (6.9%)	15.0%
Test Set	20.9%	19.1% (8.6%)	19.4% (7.2%)	20.2%

6. CONCLUSION AND DISCUSSION

In this paper we have proposed a novel conservative adaptation technique for DNNs. The basic idea of our approach is to force the senone posterior probability estimated from the adapted model to not deviate too much from that estimated using the unadapted model. To achieve this goal we have applied a KLD regularization term to the adaptation criterion. We showed that this is equivalent to adjusting the target distribution of the training samples and thus can be easily incorporated into the existing BP training procedure. Experiments on four datasets demonstrated the superiority of our approach with 2-30% relative WER reduction against already very strong CD-DNN-HMM systems under various adaptation setups.

We believe the research on the DNN adaptation just started. The simple extension of this work would be to come up with a regression model, trained on many different tasks and adaptation set sizes, for determining the regularization weight. This would make it easier to apply the KLD-Reg method to new tasks. Alternatively, Bayesian optimization techniques might be used to search for the best hyper-parameters (which we did not optimize aggressively in our experiments) for different adaptation tasks [28]. Another direction of research is to build variable parameter DNNs similar to the variable parameter HMMs [29], in which the model parameters are functions of some control parameters such as signal-to-noise ratio. The strength of such a model is its ability to adjust the model parameters automatically based on even continuous control parameters. The fourth direction is to factorize the effect of the speaker, environment and channel to the model parameters so that we can find subspaces of each and combine them to estimate a new adapted model even for a speaker-environment combination that has never been observed before. This is motivated by the fact that the adapted models in the lecture transcription task, although outperformed the unadapted DNN on both the development and test sets, increased WER from 14.2% (baseline) to 16.0% (supervised) and 14.6% (unsupervised), when applied to the same speaker under different channel and environment conditions (mismatched to the adaptation data). The last possible direction is to use speaker adaptive training technique [30] to separate the canonical model and the adaptation model.

7. AKNOWLEDGEMENT

We would like to thank Drs. Yifan Gong, Li Deng, Alex Acero, Mike Seltzer, Qiang Huo and Jasha Droppo at Microsoft Corporation for valuable discussions and suggestions.

8. REFERENCES

[1] 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.

[2] D. Yu, L. Deng, and G. Dahl, "Roles of pretraining and fine-tuning in context-dependent DNN-HMMs for real-world speech recognition," in *Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, Dec. 2010.

[3] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. Interspeech'11*, pp. 437-440, 2011.

[4] D. Yu, Y.-C. Ju, Y.-Y. Wang, G. Zweig, and A. Acero, "Automated Directory Assistance System - from Theory to Practice", in *Proc. Interspeech'07*, pp. 2709-2712, 2007.

[5] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[6] N. Jaitly, P. Nguyen, and V. Vanhoucke, "application of pretrained deep neural networks to large vocabulary speech recognition", in *Proc. Interspeech'12*, 2012.

[7] T. N. Sainath, B. Kingsbury, B. Ramabhadran, P. Fousek, P. Novak, and A.-r. Mohamed, "*Making deep belief networks effective for large vocabulary continuous speech recognition*", in *Proc. ASRU'11*, pp. 30-35, 2011.

[8] 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'12*, 2012.

[9] Hang Su, Gang Li, Dong Yu, Frank Seide, "Error back propagation for sequence training of context-dependent deep networks for conversational speech transcription", in *Proc. ICASSP* 2013.

[10] Michael Seltzer, Dong Yu, Yongqiang Wang, "An investigation of deep neural networks for noise robust speech recognition", in *Proc. ICASSP* 2013.

[11] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, 2012.

[12] V. Abrash, H. Franco, A. Sankar, and M. Cohen, "Connectionist speaker normalization and adaptation," in *Proc. EUROSPEECH'95*, pp. 2183–2186, 1995.

[13] J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, and S. Renals, T. Robinson, "Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system," in *Proc. EUROSPEECH'95*, pp. 2171–2174, 1995.

[14] D. Albesano, R. Gemello, and F. Mana, "Hybrid HMM-NN modeling of stationary-transitional units for continuous speech recognition", in *Proc. NIPS'97*, pp. 1112–1115, 1997.

[15] B. Li and K. C. Sim, "Comparison of discriminative input and output transformations for speaker adaptation in the Hybrid NN/HMM systems", in *Proc. Interspeech'10*, pp. 526-529, 2010.

[16] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. De Mori, "Linear hidden transformations for adaptation of hybrid ANN/HMM models", *Speech Communication* 49, no. 10, pp. 827–83, 2007.

[17] X. Liu, M. J. F. Gales, and P. C. Woodland. "Improving LVCSR system combination using neural network language model cross adaptation," in Proc. *Interspeech'11,* Pp. 2857-2860, 2011.

[18] Y. Xiao, Z. Zhang, S. Cai, J. Pan, and Y. Yan, "A initial attempt on task-specific adaptation for deep neural network-based large vocabulary continuous speech recognition", in *Proc. Interspeech'12,* 2012.

[19] J. Trmal, J. Zelinka, and L. Müller. "Adaptation of a feedforward artificial neural network using a linear transform," *Text, Speech and Dialogue. Springer Berlin/Heidelberg*, pp. 423-430, 2010.

[20] K. Yao, D. Yu, F. Seide, H. Su, L.i Deng, and Y. Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition", in *Proc. SLT'12*, 2012.

[21] J. Stadermann and G. Rigoll, "Two-stage speaker adaptation of hybrid tied-posterior acoustic models," in *Proc. ICASSP'05*, vol. I, pp. 997-1000, 2005.

[22] D. Albesano, R. Gemello, P. Laface, F. Mana, and S. Scanzio, "Adaptation of artificial neural networks avoiding catastrophic forgetting," in *Proc. Int. Jnt. Conference on Neural Networks 2006*, pp. 2863-2870, 2006.

[23] X. Li and J. Bilmes, "Regularized adaptation of discriminative classifiers," in *Proc. ICASSP'06*, 2006.

[24] S. Dupont and L. Cheboub, "Fast speaker adaptation of artificial neural networks for automatic speech recognition", in *Proc. ICASSP'00*, vol.3, pp. 1795-1798, 2000.

[25] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. ASRU'11*, 2011.

[26] D. Yu, X. Chen, and L. Deng, "Factorized deep neural networks for adaptive speech recognition", *International workshop on statistical machine learning for speech processing*, 2012.

[27] D. Yu, L. Deng, and F. Seide, "The deep tensor neural network with applications to large vocabulary speech recognition", *IEEE Trans. on Audio, Speech, and Language Processing*, 2013.

[28] J. Snoek, H. Larochelle, and RP. Adams, "Practical bayesian optimization of machine learning algorithms," *arXiv*:0912.4896, 2012.

[29] D. Yu, L. Deng, Y. Gong, and A. Acero, "A novel framework and training algorithm for variable-parameter hidden markov models", *IEEE Trans. on Audio, Speech, and Language Processing*, vol 17, no. 7, pp. 1348-1360, 2009.

[30] J. Trmal, J. Zelinka, and L. Müller, "On speaker adaptive training of artificial neural networks", in *Proc. Interspeech'10*, pp. 554-557, 2010.