MULTI-BASIS ADAPTIVE NEURAL NETWORK FOR RAPID ADAPTATION IN SPEECH RECOGNITION

Chunyang Wu & Mark J.F. Gales

Cambridge University Engineering Dept,
Trumpington St., Cambridge, CB2 1PZ U.K.
Email: {*cw564, mjfg*}@*eng.cam.ac.uk*

ABSTRACT

Recent progress in acoustic modeling with deep neural network has significantly improved the performance of automatic speech recognition systems. However, it remains as an open problem how to rapidly adapt these networks with limited, unsupervised, data. Most existing methods to adapt a neural network involve modifying a large number of parameters thus rapid adaptation is not possible with these schemes. In this paper, the multi-basis adaptive neural network is proposed, a new neural network configuration which only requires very few parameters for adaptation. By modifying the topology of a single multi-layer perception, a set of sub-networks with restricted connectivity are introduced to collaboratively capture different acoustic properties. The outputs of those sub-networks are combined by speaker-dependent interpolation weights. In addition, the complete system can be optimized in an adaptive training fashion when non-homogeneous training data are used. The performance of unsupervised adaptation is evaluated on two datasets. It outperforms the speaker-independent hybrid DNN-HMM baseline both on the Broadcast News English and the AURORA-4 tasks.

Index Terms— Adaptation, deep neural network, speech recognition

1. INTRODUCTION

Recently, the deep neural network (DNN) has been successfully applied to automatic speech recognition (ASR). It outperforms the conventional Gaussian mixture hidden Markov model (GMM-HMM) system in a variety of large vocabulary continuous speech recognition tasks [1, 2, 3]. Although significant improvements have been achieved, the general speaker-independent (SI) DNN-HMM system still cannot overcome the variations in different speakers or environmental conditions [4]. Thus the speaker adaptation remains a challenge in the DNN based acoustic models. For speaker adaptation, the speaker-dependent transform must be powerful enough to represent acoustic properties whilst the transform should be robustly estimated on limited adaptation data. Speaker adaptation has been studied in the GMM-HMM framework. The parameters of the GMM-HMM can naturally be interpreted as belonging to groups thus the issues above can be addressed. Popular methods include maximum a posteriori (MAP) approaches [5] and the linear transform based models such as MLLR [6] and CMLLR [7]. However, it is difficult to find meaningful structures in DNN parameters to enable similar transforms as that used in GMM-HMMs. Additionally, due to the enormous number of neural network parameters, a DNN is likely to be over-fitted, resulting in a significant performance degradation. There have been a number of attempts to adapt neural networks. Conservative training methods [8, 9, 10] introduce regularization on the adaptation training criterion. Another category of approaches appends supplementary indicators to compensate the DNN capability in different acoustic conditions, *e.g.*, using i-vector [11, 12] or underlying factors [13] as additional input features. A method in [14, 15] jointly trains a DNN with automatic speaker-specific features, referred to as speaker codes. In addition, the transformation based schemes treat the SI neural network as canonical model and add additional linear hidden layers as speaker-dependent transforms prior to the input layer [16, 17], to the hidden layer [18, 19], or to the output layer [20]. Instead of modeling additional transformations, the Hermitian based activation function in [21] is adapted while keeping the DNN weights fixed. Recent researches also investigate adaptive training schemes in DNN rather than using the SI neural network as canonical model. In [22], one hidden layer is modeled as the speaker-dependent transform and in [23], additional bottom normalization layers along with i-vector are introduced to project raw acoustic feature into a speaker-normalized space. However, except for the feature-appending schemes like speaker code or i-vector, most existing models still involve a large number of parameters to adapt, hence they would rarely handle rapid adaptation scenarios with limited data.

This paper proposes a novel configuration of neural network and its associated adaptive training scheme, named as multi-basis adaptive neural network (MBA-NN). This approach is inspired from a similar concept of the cluster adaptive training (CAT) in the GMM-HMM framework [24, 25]. The topology of multi-layer perceptron is modified and a set of sub-networks are introduced, referred to as *bases*. The hidden nodes are restricted to connect within a single basis and different bases share no connectivity. The outputs among different bases are subsequently combined via interpolation. The interpolation weights, *basis weight vector*, can be utilized to adapt the neural network into the speaker-dependent acoustic space. In this adaptation scheme, it only requires to estimate a small number of parameters for a particular speaker, hence it allows to be adapted rapidly.

There is a similar structured neural network proposed in [26], which directly combines the outputs of multiple denoising autoencoders by interpolation. There are two major differences from the model proposed in this paper: one is that the MBA-NN is a general framework which can be introduced to any layer of DNN; the other is that the MBA-NN interpolation weights are estimated in an unsupervised fashion rather than predicted by a discriminative classifier.

The research leading to these results was supported by EPSRC grant EP/I031022/1 (Natural Speech Technology) and DARPA under the RATS program. The paper does not necessarily reflect the position or the policy of US Government and no official endorsement should be inferred.

The rest of this paper is organized as follows. A brief introduction to the neural network for ASR is given in Section 2. In Section 3, the general form of multi-basis adaptive neural network and the overall training strategy are proposed. Experimental results are reported in Section 4. We finally conclude this study in Section 5.

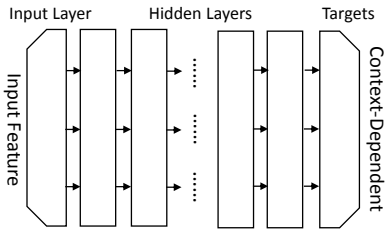
2. DEEP NEURAL NETWORK FOR ASR

In speech recognition, the hybrid DNN-HMM system uses a multi-layer perceptron (MLP) to predict the state emitting probability via computing a pseudo likelihood

$$p(\mathbf{x}|y) = \frac{p(y|\mathbf{x})p(\mathbf{x})}{p(y)} \propto \frac{p(y|\mathbf{x})}{p(y)}, \tag{1}$$

where $\mathbf{x}$ and y stand for an acoustic observation and the index of a context-dependent state respectively; $p(\mathbf{x})$ is independent from the state thus can be ignored. The multi-layer perceptron is a simple version of feed-forward neural network model that maps the input vector $\mathbf{x}$ onto a set of output targets y. It can be viewed as a directed graph, which consists of multiple hidden layers as illustrated in Fig. 1. The hidden nodes (or neurons) between two consecutive layers are fully connected while neurons within one layer are not.

Fig. 1. Multi-Layer Perceptron.



Denote the output of the l-th layer as $\mathbf{z}^l(\mathbf{x})$, the input of the successive, $(l+1)$-th, layer is then given by

$$\mathbf{h}^{l+1}(\mathbf{x}) = \boldsymbol{\sigma}\left(\mathbf{z}^l(\mathbf{x})\right). \tag{2}$$

where $\sigma_i(\mathbf{z}) = \frac{1}{1+\exp(-z_i)}$ is the sigmoid function and

$$\mathbf{z}^l(\mathbf{x}) = \mathbf{W}^l\mathbf{h}^l(\mathbf{x}) + \mathbf{b}^l \tag{3}$$

represents a transformation given on the l-th layer and the parameters of the transformation are defined as $\mathbf{W}^l$ and $\mathbf{b}^l$. The softmax activation function is usually used as the output for MLP models on multi-label classification tasks, it can be viewed as the target posterior given the input observation

$$p(y=i|\mathbf{x}) = \frac{\exp(z_i^L)}{\sum_j \exp(z_j^L)}, \tag{4}$$

where L denotes the index of the last hidden layer. The notations specified in this section will be used in the following discussions.

3. MULTI-BASIS ADAPTIVE NEURAL NETWORK

The structure of multi-basis adaptive neural network is illustrated in Fig. 2. A set of distinct sub-networks are introduced to the multi-layer perceptron, referred to as the *bases*. There is a common input layer and a common output layer. Optionally, common hidden layers can be introduced before propagating to the bases, or after their output combination. A basis is composed of multiple hidden layers. The hidden units between two successive layers within one basis are fully connected, while there is no connection between neurons coming from different bases. The outputs of bases are then combined subsequently. In this paper, we investigate the interpolation scheme: they are linearly combined using a set of adaptive weights, the *basis weight vector*,

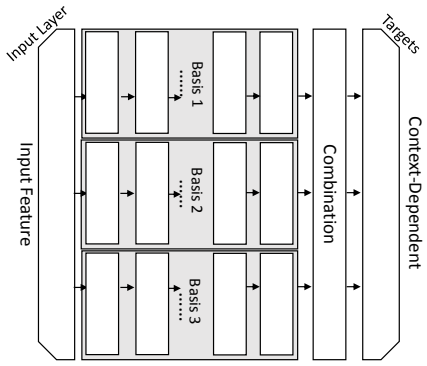
$$\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_K]^T. \tag{5}$$

where K denotes the number of basis and the result

$$\bar{\mathbf{h}}^l(\mathbf{x}) = \sum_{k=1}^{K} \lambda_k \mathbf{h}_k^l(\mathbf{x}) \tag{6}$$

is propagated to the successive common layers, where $\mathbf{h}_k^l(\mathbf{x})$ represents the input to the l-th layer of the k-th basis.

Fig. 2. Multi-Basis Adaptive Neural Network.



3.1. MBA-NN Training

The MBA-NN inherently falls into an adaptive training framework. In this training scheme, the canonical model $\mathcal{M}$ is defined as

$$\mathcal{M} = \{\Psi_1, \ldots, \Psi_K, \Omega_{\text{shared}}\}, \tag{7}$$

where Ψ_k stands for the parameters of the k-th basis and Ω_{shared} denotes the shared layers in MBA-NN. The set of speaker-dependent transforms Λ for S training speakers is given by

$$\Lambda = \left\{\boldsymbol{\lambda}^{(1)}, \ldots, \boldsymbol{\lambda}^{(S)}\right\}, \tag{8}$$

where S denotes the total of training speakers. Both $\mathcal{M}$ and Λ are jointly optimized in this training scheme. The aim of the training procedure is to minimize the cross entropy over the training set with associated state alignments and speaker information:

$$\mathcal{L}(\mathcal{M}, \Lambda) = -\sum_{s=1}^{S}\sum_{t\in\mathbb{I}_s} \log p(y_t|\mathbf{x}_t; \mathcal{M}, \boldsymbol{\lambda}^{(s)}). \tag{9}$$

where $\mathbb{I}_s$ stands for the index set of training frames belonging to speaker s. In order to break the symmetry among the bases, the parameters of $\mathcal{M}$ and Λ are updated iteratively. This interleaving training mode of parameter optimization is listed in Algorithm 1.

The canonical model $\mathcal{M}$ of MBA-NN is initialized by the speaker-independent hybrid DNN-HMM. The hidden layers given

Algorithm 1 Interleaving Training Mode of MBA-NN.

initialize the MBA-NN from the hybrid SI model
initialize $\boldsymbol{\lambda}^{(s)}$ for all the training speakers
while not convergence **do**
  update $\mathcal{M}$ via back-propagation for one iteration
  for $s := 1$ **to** S **do**
    update $\boldsymbol{\lambda}^{(s)}$ for one iteration
  end for
end while

on this SI system are duplicated to build the multiple bases. The speaker-dependent transform Λ could be initialized according to various techniques, *e.g.*, by random values, from prior knowledge like gender information or via automatic approaches such as i-vector. However, to ensure that the initial performance is the same as the hybrid SI system, the initial sum of $\lambda_1^{(s)}, \ldots, \lambda_K^{(s)}$ should be equal to one,

$$\sum_{k=1}^{K} \lambda_k^{(s)} = 1. \tag{10}$$

The interleaving update of $\mathcal{M}$ and Λ will not terminate until convergence or the maximum iteration time is reached. In each round, the parameters in the canonical model $\mathcal{M}$ can be directly updated using the standard error back propagation algorithm. For the basis weight vector for each speaker, it can be updated via an analogous gradient descent scheme as well.

This paper investigates a special version of MBA-NN that the multiple bases are combined on the last hidden layer, in which the shared parameters are given by

$$\Omega_{\text{shared}} = \left\{ \mathbf{W}^L, \mathbf{b}^L \right\}. \tag{11}$$

The output vector $\mathbf{z}^L(\mathbf{x})$ right before the softmax activation can be written as[1]

$$\mathbf{z}^L(\mathbf{x}) = \mathbf{W}^L \left\{ \sum_{k=1}^{K} \lambda_k \mathbf{h}_k^L(\mathbf{x}) \right\} + \mathbf{b}^L = \mathbf{W}^L \mathbf{H}(\mathbf{x}) \boldsymbol{\lambda} + \mathbf{b}^L, \tag{12}$$

where

$$\mathbf{H}(\mathbf{x}) = \left[\mathbf{h}_1^L(\mathbf{x}), \ldots, \mathbf{h}_K^L(\mathbf{x}) \right] \tag{13}$$

is a matrix consisting of the outputs from different bases. An important merit of the interpolation on the last hidden layer is that, by fixing $\mathcal{M}$ and deriving the loss function with respect to $\boldsymbol{\lambda}$,

$$\mathcal{L}(\boldsymbol{\lambda}) = \sum_{t \in \mathbb{I}_s} \left\{ \log \sum_j \exp\left(\boldsymbol{\lambda}^T \mathbf{f}(\mathbf{x}_t, j) + b_j^L \right) - \boldsymbol{\lambda}^T \mathbf{f}(\mathbf{x}_t, y_t) - b_{y_t}^L \right\} \tag{14}$$

where $\mathbf{f}(\mathbf{x}_t, j) = \mathbf{H}^T(\mathbf{x}_t) \left(\mathbf{w}_j^L \right)^T$ and $\mathbf{w}_j^L$ denotes the j-th row of $\mathbf{W}^L$, it gives the same form as the log-linear model. Thus the optimization problem of $\boldsymbol{\lambda}$ degrades to a convex one. It avoids getting stuck into local minima, which is a common issue in the parameter estimation of neural networks. In this paper, stochastic gradient descent is used both in the training and adaptation phases to update $\boldsymbol{\lambda}^{(s)}$, the gradient with respect to $\boldsymbol{\lambda}^{(s)}$ is given by

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}^{(s)}} = \sum_{t \in \mathbb{I}_s} \left\{ \frac{\sum_j \exp\left(\mathbf{f}^T(\mathbf{x}_t, j) \boldsymbol{\lambda}^{(s)} + b_j \right) \mathbf{f}(\mathbf{x}_t, j)}{Z(\mathbf{x}_t)} - \mathbf{f}(\mathbf{x}_t, y_t) \right\} \tag{15}$$

[1]In Eq. 12 and 14, $\boldsymbol{\lambda}^{(s)}$ is abbreviated as $\boldsymbol{\lambda}$ in which the superscript is omitted.

where

$$Z(\mathbf{x}_t) = \sum_{\tilde{j}} \exp\left(\mathbf{f}^T(\mathbf{x}_t, \tilde{j}) \boldsymbol{\lambda}^{(s)} + b_{\tilde{j}} \right). \tag{16}$$

3.2. Adaptation

After the training phase, the transforms Λ belonging to the training speakers are wiped out and only the canonical model $\mathcal{M}$ is used for adaptation. By keeping $\mathcal{M}$ to be fixed, the speaker-dependent basis weight vector $\boldsymbol{\lambda}^{(s)}$ is updated according to Eq. 15 until convergence. The estimated $\hat{\boldsymbol{\lambda}}^{(s)}$ is then combined with $\mathcal{M}$ to decode testing utterances.

4. EXPERIMENTS

The effectiveness of the proposed multi-basis adaptive neural network was evaluated on two tasks: Broadcast News English (BNE) and AURORA-4. On both tasks, the performance of rapid, utterance-level, unsupervised adaptation was evaluated: the SI hybrid system was initially used to generate decoding hypothesis and the associated state alignments. The basis weight vector $\boldsymbol{\lambda}$ was then estimated on these alignments for each utterance. Besides, the MBA-NN training phase terminated when the criterion on the cross validation set began to increase.

4.1. Broadcast News English

The training set for this task included the 144-hour 1996 & 1997 Hub-4 English Broadcast News Speech datasets (`LDC97S44`, `LDC98S71`), containing 288 shows with approximately 8k speakers. Both the testsets `dev03` and `eval03` of DARPA RT03 were used for evaluation. The utterances of both testing sets were given by automatic segmentation. Decoding was performed with the RT04 tri-gram language model.

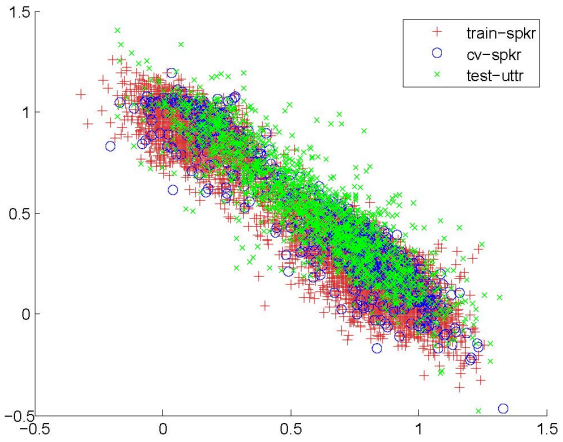
The 39-dimensional PLP features with their first- and second-order derivatives processed by both show-level cepstral mean normalisation (CMN) and cepstral variance normalisation (CVN) were firstly used to train a GMM-HMM model consisting of 6k tied triphone states on the maximum likelihood estimation. It was further extended to include the triple feature using HLDA [27] and discriminatively trained on the MPE [28] criterion. This MPE model was then used to give the state alignment for the SI Hybrid system. The 468-dimensional input feature to the neural network was formed by 52-dimensional PLP+Δ+$\Delta\Delta$+$\Delta\Delta\Delta$ with a context window of 9 frames. The neural network consisted of five hidden layers with 1k neurons on each layer. The parameters of this DNN were initialized using layer-by-layer discriminative pre-training and further fine-tuned by back-propagation. During the fine-tuning phase, 28 shows with about 600 speakers were used as the cross validation set.

Table 1. Word Error Rate Comparison on BNE.

System	dev03	eval03
SI Hybrid	13.1	11.4
MBA-NN Gender Slct	12.6	11.3
MBA-NN	12.2	11.1

For MBA-NN, the training phase optimized speaker-level basis weight vectors while the adaptation in evaluation was performed in the utterance-level scheme. The 2-dimensional $\boldsymbol{\lambda}^{(s)}$ for training speaker s was initialized with its gender information ($[1; 0]$ for male

Fig. 3. The basis-weight-vector space given by the BNE training & CV speakers and the DEV03 test utterances.



and $[0; 1]$ for female). Table 1 reported the word error rate (WER) achieved by the SI Hybrid system and the MBA-NN both on `dev03` and `eval03`. The MBA-NN model reduced the WER from 13.1% to 12.2%, 7% relative error rate reduction on `dev03` while on `eval03`, it reduced the WER from 11.4% to 11.1%. The *MBA-NN Gender Slct* represents the MBA-NN system after the first neural network iteration when the speaker-dependent weights are fixed as the gender initialization, which can be treated as the gender-dependent hybrid baseline. The gender-basis hypothesis with a higher alignment likelihood was selected to form the decoding results of this baseline. This gender-dependent model both gave a better performance than the SI system on the two testing sets.

Figure 3 illustrated the basis-weight-vector space given by the training / CV speakers and the `dev03` test utterances. They laid in a consistent location and formed a line close to $\lambda_1 + \lambda_2 - 1 = 0$.

4.2. AURORA-4

The AURORA-4 corpus is a medium vocabulary task derived from Wall Street Journal (WSJ0). The 16kHz multi-style training set is used in this series of experiments. It consists of 7138 utterances from 83 speakers, in which half are recorded using the primary Sennheiser microphone whilst the other half are recorded by a number of secondary microphones. 6 different types of noises are added to this training set with the SNR ranging from 10dB to 20dB. The evaluation dataset of AURORA-4 is based on the 330-utterance WSJ0 5K-word closed vocabulary test set from 8 speakers. It consists of 14 subsets, the clean set 01 (`Set A`), the noise set from 02 to 07 (`Set B`), the clean set with channel distortion 08 (`Set C`) and the noise set with channel distortion from 09 to 14 (`Set D`). `Set B` and `D` are corrupted by the same 6 types of noises as those in the training data with randomly selected SNRs at 5dB $\sim$ 15dB. The decoding of evaluation was performed with the standard WSJ0 bi-gram language model.

A GMM-HMM system was initially trained using the maximum likelihood criterion, which consisted of approximately 3k tied triphone states with 8 Guassians per state. The feature for the GMM-HMM system was the 39-dimensional PLP features with their Δ and $\Delta\Delta$, processed by utterance-level CMN. This system was further extended to include $\Delta\Delta\Delta$ using HLDA and discriminatively trained on the MPE criterion. Instead of using the PLP feature[2], the 72-dimensional FBANK with the first- and second- order dynamic features processed by utterance-level CMN was used to train the SI hybrid DNN-HMM system with the context dependent state alignments given by the MPE model. The neural network configuration of the SI system is $648 \times 1000 \times 500 \times 500 \times 500 \times 3$k, with a context window of 9 frames as input feature. The parameters of this DNN are initialized using layer-by-layer discriminative pre-training and further fine-tuned by back-propagation. During the fine-tuning phase, 650 utterances belonging to 8 speakers are used as the cross validation set.

Table 2. Word Error Rate Comparison on AURORA 4.

System	#Bases	A	B	C	D	Avg
SI Hybrid	–	4.7	9.8	11.0	22.8	15.1
MBA-NN	2	4.3	8.9	9.6	21.3	14.0
	4	4.3	8.8	9.6	21.5	14.0
	6	4.3	8.9	9.6	21.5	14.0
MBA-NN (REF)	2	4.2	8.6	9.3	20.7	13.5
	4	4.1	8.4	9.0	20.2	13.2
	6	3.9	8.3	8.9	20.1	13.1

The MBA-NN is then initialized with this SI hybrid model. In this series of experiments, both the $\boldsymbol{\lambda}^{(s)}$ estimation in the training and testing phases are conducted in the utterance level. For training, we first clustered the utterance i-vectors by k-means and divided the utterances into 2,4 and 6 clusters. A *1-of-K* vector (a vector with one element containing a 1 and all other elements as 0) was specified to each utterance as its initial basis weight vector, representing its cluster index.

Table 2 summarized the decoding performance of different models. All three configurations outperformed the SI Hybrid baseline with around 7% relative error reduction, dropping the word error rate (WER) from 15.1% to 14.0%. However, as the dimension of basis weight vector increased, the system started to be more sensitive to the quality of hypothesis: the systems with 4 and 6 bases gave slightly worse performance on `Set D`. However, better performance was obtained with more number of bases in oracle experiments which used the reference transcription to estimate $\boldsymbol{\lambda}^{(s)}$. As shown in the `MBA-NN(REF)` part of Table 2, the system with 2, 4 and 6 bases yielded 13.5%, 13.2% and 13.1% in WER, respectively.

5. CONCLUSION

This paper has introduced the multi-basis adaptive neural network, a novel topology of neural network for rapid adaptation in speech recognition. The adaptation scheme on MBA-NN only requires a simple but compact representation of a speaker, referred to as the basis weight vector $\boldsymbol{\lambda}$. Thus rapid adaptation scenarios with limited data could be resolved within this framework. The performance of utterance-level unsupervised adaptation is evaluated on the Broadcast News English and the AURORA-4 datasets. On both tasks, the improvement of recognition performance is obtained over the SI hybrid DNN-HMM baseline. Future work will look at the performance of different output combination schemes. Rather than interpolation, combining the bases with multiple linear transforms will be a proper extension.

[2]On this task, the DNN with the filter bank feature yielded a better performance comparing to that with PLP or MFCC.

6. REFERENCES

[1] George E Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.

[2] Geoffrey Hinton et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[3] Xie Chen, Adam Eversole, Gang Li, Dong Yu, and Frank Seide, "Pipelined back-propagation for context-dependent deep neural networks.," in *INTERSPEECH*, 2012.

[4] Yan Huang, Dong Yu, Chaojun Liu, and Yifan Gong, "A comparative analytic study on the gaussian mixture and context dependent deep neural network hidden markov models," *Interspeech*, 2014.

[5] Jean-Luc Gauvain and Chin-Hui Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *Speech and audio processing, ieee transactions on*, vol. 2, no. 2, pp. 291–298, 1994.

[6] Christopher J Leggetter and Philip C Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[7] Mark JF Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[8] Jan Stadermann and Gerhard Rigoll, "Two-stage speaker adaptation of hybrid tied-posterior acoustic models.," in *ICASSP (1)*, 2005, pp. 977–980.

[9] Xiao Li and Jeff Bilmes, "Regularized adaptation of discriminative classifiers," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. IEEE, 2006, vol. 1, pp. I–I.

[10] Dong Yu, Kaisheng Yao, Hang Su, Gang Li, and Frank Seide, "Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7893–7897.

[11] Andrew Senior and Ignacio Lopez-Moreno, "Improving dnn speaker independence with i-vector inputs," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 225–229.

[12] George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 55–59.

[13] Jinyu Li, Jui-Ting Huang, and Yifan Gong, "Factorized adaptation for deep neural network," in *Proc. ICASSP*, 2014.

[14] Ossama Abdel-Hamid and Hui Jiang, "Fast speaker adaptation of hybrid nn/hmm model for speech recognition based on discriminative learning of speaker code," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7942–7946.

[15] Shaofei Xue, Ossama Abdel-Hamid, Hui Jiang, and Lirong Dai, "Direct adaptation of hybrid dnn/hmm model for fast speaker adaptation in lvcsr based on speaker code," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6339–6343.

[16] Frank Seide, Gang Li, Xie Chen, and Dong Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 24–29.

[17] Joao Neto, Luís Almeida, Mike Hochberg, Ciro Martins, Luís Nunes, Steve Renals, and Tony Robinson, "Speaker-adaptation for hybrid hmm-ann continuous speech recognition system," 1995.

[18] Roberto Gemello, Franco Mana, Stefano Scanzio, Pietro Laface, and Renato De Mori, "Linear hidden transformations for adaptation of hybrid ann/hmm models," *Speech Communication*, vol. 49, no. 10, pp. 827–835, 2007.

[19] Kaisheng Yao, Dong Yu, Frank Seide, Hang Su, Li Deng, and Yifan Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition.," in *SLT*, 2012, pp. 366–369.

[20] Bo Li and Khe Chai Sim, "Comparison of discriminative input and output transformations for speaker adaptation in the hybrid nn/hmm systems.," in *INTERSPEECH*, 2010, pp. 526–529.

[21] Sabato Marco Siniscalchi, Jinyu Li, and Chin-Hui Lee, "Hermitian based hidden activation functions for adaptation of hybrid hmm/ann models.," in *INTERSPEECH*, 2012.

[22] Tsubasa Ochiai, Shigeki Matsuda, Xugang Lu, Chiori Hori, and Shigeru Katagiri, "Speaker adaptive training using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6349–6353.

[23] Yajie Miao, Hao Zhang, and Florian Metze, "Towards speaker adaptive training of deep neural network acoustic models," in *Proc. Interspeech*, 2014.

[24] Mark JF Gales, "Cluster adaptive training of hidden markov models," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 4, pp. 417–428, 2000.

[25] Roland Kuhn, Patrick Nguyen, Jean-Claude Junqua, Lloyd Goldwasser, Nancy Niedzielski, Steven Fincke, Ken Field, and Matteo Contolini, "Eigenvoices for speaker adaptation.," in *ICSLP*, 1998, vol. 98, pp. 1774–1777.

[26] Forest Agostinelli, Michael R Anderson, and Honglak Lee, "Adaptive multi-column deep neural networks with application to robust image denoising," in *Advances in Neural Information Processing Systems*, 2013, pp. 1493–1501.

[27] Nagendra Kumar and Andreas G Andreou, *Investigation of silicon auditory models and generalization of linear discriminant analysis for improved speech recognition*, Ph.D. thesis, Johns Hopkins University, 1997.

[28] Daniel Povey and Philip C Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. IEEE, 2002, vol. 1, pp. I–105.